1
Introdução

Este trabalho tem por objetivo prover um critério operacional para caracterizar substantivos em combinações *S-Adj*, em que o substantivo se apresenta em situação análoga à dos chamados verbos leves ou verbos-suporte, largamente estudados em Lingüística e Processamento de Linguagem Natural nos últimos anos. O trabalho se situa na confluência entre estudos lingüísticos, lexicográficos e computacionais e pretende explorar a potencialidade da análise automática de corpora e instrumentos quantitativos em busca de uma maior objetividade na utilização e evidenciação de conceitos que norteiam a atividade de análise lingüística.

A utilização de análises lingüísticas baseadas em corpus vem ampliando as possibilidades de detecção de padrões construcionais das línguas. O apoio computacional disponível viabiliza o trabalho de pesquisa em corpora gigantescos, de grande cobertura e variabilidade textual. As principais características metodológicas da pesquisa baseada em corpus são (Biber, Conrad, & Reppen 1998):

- é uma pesquisa empírica, com base em padrões de uso, efetivamente produzidos, em textos reais;
- utiliza uma coleção de textos, o **corpus**, caracterizável por uma ou mais dimensões textuais (língua, gênero, registro, datas, etc);
- é extensivamente apoiada por computadores, de forma totalmente automática ou interativamente;
- utiliza mecanismos quantitativos e qualitativos de análise dos dados.

Do ponto de vista operacional, construções complexas, restritas por condições gramaticais de diversas naturezas, podem ser codificadas para que sejam identificadas em contexto, liberando o lingüista da tarefa enfadonha da busca manual de exemplos. O resultado da busca automática é mais consistente, pois se dá de maneira homogênea ao longo do tempo, o que é extremamente difícil para o lingüista individualmente e, ainda mais, em trabalhos cooperativos.

O léxico é um sistema dinâmico, em constante rearranjo, que não só armazena formas significantes, mas também fornece processos de produção de palavras e de expressões. O delineamento da classe de substantivos-suporte proposto no trabalho pressupõe um modelo de léxico que dê conta de construções regulares e semi-regulares, tais como as construções verbais com verbo-suporte, dentro de um quadro teórico que rejeita a separação entre a sintaxe e o léxico. Os modelos lexicais computacionais inspirados em teorias lingüísticas de orientação funcionalista, tais como (Fillmore 1976), oferecem um quadro adequado para a utilização dos resultados desta pesquisa em sistemas computacionais.

1.1 Caracterização do problema

Na lingüística, o termo **substantivo vazio** é utilizado para fazer referência a substantivos que não denotam conceitos, ou substantivos com um conteúdo semântico mínimo, identificados, na análise sintática, com elementos sem expressão fonológica que atuam como núcleos de sintagmas nominais. De acordo com Panagiotidis (2003), no léxico de qualquer língua existe um número limitado de substantivos vazios, que formam uma classe fechada. Eles são considerados palavras gramaticais, não lexicais, com forte característica pronominal. Tendo em vista sua reduzida capacidade de denotação, substantivos vazios distinguem-se entre si por meio de seus traços fonológicos e morfo-semânticos, tais como o gênero. O substantivo vazio típico do inglês seria ONE, como em *a big* ONE.

Na teoria lexical, tem havido muito pouco interesse no fenômeno, com algumas exceções tais como Schmid (2000), que apresenta um estudo sobre substantivos abstratos. De acordo com Schmid, há um subconjunto dos substantivos abstratos em inglês que atuam como conchas conceituais, que realizam seu potencial semântico em associação com outros itens lexicais. Palavras como THING, FACT, CASE, POINT, IDEA, REASON, PROBLEM e QUESTION, entre outros substantivos de alta freqüência no inglês, são exemplos dessa classe.

Estudos lexicais aplicados ao processamento automático de textos ampliaram essa noção, incluindo palavras que podem ser apagadas ou ignoradas para o propósito de sumarização, classificação, agrupamento e outras operações computacionais sobre textos. Substantivos ocorrendo como quantificadores em expressões multi-vocabulares, tais como GROUP, BUNCH e LOTS, em GROUP *of students*, BUNCH *of students*, LOTS *of students*, são exemplos especificamente mencionados em (Muresan, Tzoukermann, & Klavans 2001) para o inglês. O caso geral, em que esse tipo de substantivo ocorre legitimamente como núcleo

de um sintagma nominal, não foi explorado.

No âmbito da Recuperação de Informações, a indexação automática do texto completo exige operações de compressão de texto tais como apagamento de palavras vazias, palavras que não contribuam para a precisão nem para a abrangência nas operações de busca. Tradicionalmente, palavras vazias são listas de palavras funcionais e verbos auxiliares. Substantivos-suporte não são palavras vazias, a julgar por esse critério usual.

O conceito de **densidade lexical** (Halliday 1985) subjaz essa decisão pela eliminação de palavras vazias na prática da Recuperação de Informações. Considerando os exemplos 1.1, o enunciado 2. seria mais "denso"que o 1. por expressar o mesmo significado proposicional de modo mais compacto.

ex. 1.1

1. *É a coisa da ausência de interesse cultural.*

2. *É a ausência de interesse cultural.*

Apesar de adotar a tradicional divisão entre classes lexicais e gramaticais (cf. capítulo 3) Halliday admite o caráter fronteiriço de certos substantivos e verbos. O autor também propões que a freqüência relativa de uma palavra pode indicar a quantidade de informação que ela traz para o enunciado.

Existe um conjunto não muito extenso de substantivos do português que se caracterizam por sua generalidade semântica. Os primeiros questionamentos sobre este tipo de palavra surgiram a partir de uma pesquisa de corpus que realizei em busca de sintagmas nominais *S-Adj*, onde *Adj* é um adjetivo denominal resultado de um processo $[X]_S \rightarrow [[X]_S\ al]_{Adj}$. O resultado da pesquisa mostrou que, combinados aos mais variados adjetivos, sempre havia termos, como FATOR, PERSPECTIVA e ASPECTO, que são aparentemente intercambiáveis, apesar de terem significados literais bastante distintos.

$$\left.\begin{array}{l} fator \\ perspectiva \\ aspecto \end{array}\right\} ambiental \qquad \left.\begin{array}{l} fator \\ perspectiva \\ aspecto \end{array}\right\} racial \qquad \left.\begin{array}{l} fator \\ perspectiva \\ aspecto \end{array}\right\} fiscal$$

Os dados mostram que a contribuição desses termos para o significado geral da expressão é bastante reduzida. As noções de significado e de contribuição de um item para o significado de uma expressão são, no mínimo, problemáticas, mas um apelo preliminar ao entendimento pré-teórico dessas noções parece confirmar que em ASPECTO AMBIENTAL a idéia mais proeminente é de 'meio-ambiente'.

A identificação de tais substantivos, aqui denominados de **substantivos-suporte**, é muito importante dentro do contexto de interpretação automática

de textos e suas aplicações. Se a base do processamento é a palavra e o sintagma, então é fundamental que a computação do significado da expressão não seja prioritariamene baseada no substantivo, mas sim em seus complementos.

1.2 Objetivos

O principal objetivo deste trabalho é a delimitação do substantivo-suporte por meio de suas características lexicográficas, funcionais e textuais.

A descrição lexicográfica do substantivo-suporte foi feita com propósitos especulativos e prospectivos. O conjunto inicial de substantivos-suporte foi obtido de uma forma empírica e assistemática e o dicionário foi muito útil em uma primeira tentativa de caracterizar esses substantivos como uma classe.

A proposta de enquadrar o substantivo-suporte dentro de um fenômeno mais amplo na linguagem, a função de suporte, é uma conseqüência da observação das ocorrências dos sintagmas nominais *S-Adj*, onde *S* pertence a um conjunto de substantivos de grande generalidade semântica. As características da construção remetem à descrição dos verbos leves apontadas por (Jespersen 1940) e (Poutsma 1926). Semelhante aos verbos-suporte, que realizam a função de suporte para o substantivo em sintagmas verbais, proponho que o substantivo-suporte realize a função de suporte nos sintagmas nominais. Dessa forma, proponho um visão lexical dos sintagmas *S-Adj*, enquadrando-os como expressões multi-vocabulares (EMV).

A questão da constituição do léxico vem à tona nessa proposta na definição da unidade lexical e na constatação de que a noção de palavra, da maneira como é delimitada tradicionalmente, não é adequada para definir o conjunto das unidades simbólicas básicas da língua. A perspectiva da análise da **construção** aproxima-se do modelo lexical da Gramática Funcional, onde a unidade é a oração, com a função de construir um modelo da experiência e das relações lógicas (significado ideacional), de realizar interação social (significado interpessoal) e conferir relevância ao contexto (significado textual) (Halliday 1994; Neves 2004). Mesmo admitindo, contrariamente a algumas vertentes do funcionalismo, que haja fronteiras possíveis entre a sintaxe e a morfologia, o trabalho é desenvolvido em torno de um modelo de léxico que prevê a formação de unidades maiores que a palavra morfológica, como explicitado em (Basilio 2005):

> "[...] a cada passo adiante na investigação sobre unidades lexicais, mais os dados me forçam a concluir que o léxico, em seu papel de produzir e armazenar formas significativas a serviço dos macrosistemas de significação e comunicação que são as línguas, utiliza

> processos tanto morfológicos quanto sintáticos para a formação de suas unidades simbólicas básicas."

Feito o enquadramento do substantivo-suporte como palavra de suporte, partiu-se para a sistematização da delimitação da construção **Substantivo-suporte−Adjetivo denominal**, tendo-se como base uma medição da composicionalidade semântica baseada em corpus. A valorização da ocorrência da construção no corpus marca uma posição similar à de J. R. Firth, que confere um papel primordial aos **eventos de fala** (**speech acts**) em sua Teoria Contextual do Significado, cf. (Newmeyer 1998). Em objeção à postulação do estruturalismo Saussureano de que instâncias da "parole" não passam de meras evidências para a estrutura da "langue", Firth afirma que os eventos de fala são o objeto principal da lingüística, sendo elementos concretos, em contraste com a "langue" de Saussure que seria "um sistema de valores diferenciais, não de termos concretos e positivos" (Firth 1968), apud (Joseph, Love, & Taylor 2001). Metodologicamente, o interesse de Firth pelo fenômeno da colocação ficou registrado em sua frase "You shall know a word by the company it keeps".

1.3 Posicionamento Metodológico

J. Sinclair, em uma aula sobre Lexicologia, Lexicografia e Lingüística Computacional, em Singapura, 1996, distinguiu duas posturas metodológicas frente à pesquisa em corpus (Ooi 1998), como mostra a tabela 1.1.

A distinção entre abordagens baseadas em corpus e dirigidas por corpus se assemelha ao contraste entre as abordagens **top-down** e **bottom-up** de resolução de problemas, utilizadas tradicionalmente nas disciplinas de Programação de Computadores e Engenharia de Software. No primeiro caso, o processo é analítico e os conceitos mais gerais da teoria do problema, suas abstrações de mais alto nível, são utilizadas para iniciar a análise. Os dados são utilizados em última instância, na confirmação, extensão ou rejeição da teoria. Por outro lado, a abordagem bottom-up inicia-se com os dados e, em processos de síntese, formulam a teoria que abstrai e generaliza a informação inerente aos dados.

Na prática da pesquisa lingüística, embora não na teoria, uma mistura das duas metodologias é invariavelmente necessária. No caso de uma pesquisa interdisciplinar, que busca meios lingüísticos de atingir objetivos computacionais, assim como prover meios computacionais para adicionar aos instrumentos de análise lingüística, a convergência das metodologias pode se acentuar. No entanto, minha pesquisa e seu encaminhamento tenderam mais à síntese. Em retrospectiva, foram longos períodos de observação do corpus, procurando

Lingüística baseada em corpus	**Lingüística dirigida por corpus**
o corpus é utilizado para validar, verificar e melhorar observações lingüísticas que já tenham sido realizadas	um corpus é de importância essencial no surgimento de novas idéias de como examinar os dados
o lingüista não questiona posições teóricas pre-estabelecidas ou categorias descritivas aceitas; sua posição com respeito à estrutura da língua já se estabilizou	o lingüista acredita que pode conciliar o tipo de evidências que emerge do corpus com as posições estabelecidas; ele deixa abertas as possibilidades de mudanças radicais na teoria para lidar com as evidências
o corpus é utilizado para ajudar a estender e melhorar a descrição lingüística	a evidência do corpus é soberana portanto o lingüista minimiza os pressupostos sobre a natureza das categorias teóricas e descritivas
um exemplo de questão relevante: WHOM ainda é utilizado em inglês? como?	um exemplo de questão relevante: a distinção entre gramática e léxico é necessária?

Tabela 1.1: Lingüística baseada em corpus vs. lingüística dirigida por corpus

realizar a demarcação paulatina do fenômeno e, finalmente, formular uma proposta de caracterização da construção em foco.

Entretanto, não se pode negar que esta fase foi posterior a vários cursos de Teoria Lexical e a um estudo detalhado dos adjetivos denominais, em que se menciona a relevância da sua combinação com substantivos de semântica geral. Por outro lado, tanto na Teoria Lexical quanto na Lingüística Computacional, a questão da delimitação das unidades lexicais e dos verbos-suporte vem sendo permanentemente discutida. Neste sentido, poderíamos dizer que a pesquisa tende mais à pesquisa baseada em corpus.

1.4 Organização do texto

No capítulo 2, discuto os diferentes aspectos teóricos envolvidos no tratamento lexical e lexicográfico de unidades lexicais, assim como questões relacionadas ao significado dos itens lexicais, com especial ênfase na questão polissemia / vagueza, crucial no entendimento do conceito de substantivo-suporte.

No capítulo 3, abordo a questão das classes de palavras e apresento uma visão geral das classes envolvidas no trabalho: Substantivo e Adjetivo.

Os capítulos 4 e 5 são centrais na execução dos objetivos da tese. O capítulo 4 concentra-se na conceituação do substantivo-suporte, por meio de propriedades lexicográficas e textuais. A noção de substantivo-suporte é posicionada em paralelo à do verbo-suporte.

O capítulo 5 apresenta os resultados experimentais da análise de sintagmas *S-Adj* para fundamentar a proposição de um mecanismo objetivo de identificação em corpora da entidade substantivo-suporte.

O capítulo de conclusões discute os resultados alcançados e problemas remanescentes, e aponta para algumas linhas de continuação que pretendo seguir nesta área de investigação.