Léxicos Computacionais: Desafios na Construção de um Léxico de Português Brasileiro

Marcelo Caetano Martins Muniz

Léxicos Computacionais: Desafios na Construção de um Léxico de Português Brasileiro

Marcelo Caetano Martins Muniz

Profª Drª Maria das Graças Volpe Nunes

Monografia apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, para o Exame de Qualificação, como parte dos requisitos para obtenção do título de Mestre em Ciências de Computação e Matemática Computacional.

USP - São Carlos
Fevereiro/2003

Resumo

A escassez de recursos lingüístico-computacionais é um dos maiores entraves para o avanço das pesquisas, e conseqüente desenvolvimento de sistemas, na área de PLN no Brasil. Este trabalho propõe a construção de um léxico computacional para português brasileiro no formato DELA (formato de dicionários da plataforma INTEX), além de uma biblioteca de acesso e manipulação a este léxico, para que qualquer aplicação de PLN possa utilizá-lo. Pretende-se também a extensão de novas funcionalidades para a ferramenta de processamento de corpus Unitex, que é uma versão de código aberto do INTEX. Os recursos construídos deverão ser úteis tanto a usuários leigos quanto a pesquisadores das áreas de Lingüística e PLN.

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

Um dos maiores entraves para o avanço das pesquisas, e conseqüente desenvolvimento de sistemas, na área de Processamento de Língua Natural (PLN) no Brasil é a escassez de recursos lingüístico-computacionais que, em última análise, fornecem todo o conhecimento do domínio necessário nessa área. Por serem muito especializados, volumosos e complexos, sua construção exige equipes interdisciplinares treinadas, cujo custo de manutenção tem impedido que as pesquisas em PLN/português cheguem a patamares compatíveis com os da língua inglesa. A importância desses recursos para o desenvolvimento dessa área é visível nos inúmeros trabalhos para a língua inglesa que compartilham estudos, corpora e sistemas dependentes de língua.

Os recursos necessários para o desenvolvimento de aplicações robustas e abrangentes podem ser divididos em dois grupos: o daqueles que oferecem conhecimento lingüístico, mas não o processam automaticamente, como os Dicionários Eletrônicos (*Machine-Tractable Dictionaries- MTDs*), Corpora (bancos de textos autênticos) e *Thesaurus*; e aqueles que se caracterizam por processar a língua para efeito de algum resultado pré-definido, como os *Taggers* (etiquetadores morfossintáticos) ou os analisadores morfológicos, sintáticos (Parsers) e semânticos. Pode-se afirmar que a maior parte do tempo e esforço necessários para o desenvolvimento de uma aplicação de PLN é dedicada à construção dos recursos lingüísticos que dão suporte ao funcionamento da mesma.

Nos últimos anos, têm-se notado um grande esforço dos pesquisadores da área de PLN para a padronização na construção desses recursos, visando principalmente a reusabilidade. Alguns padrões e ferramentas têm-se destacado no cenário internacional e vêm sendo utilizados por vários grupos de pesqui-

sas em muitos países. Um exemplo de padrão foi o desenvolvido no LADL (*Laboratoire d'Automatique Documentaire et Linguistique*) na França, o DELA (*Dictionnarie Electronique du LADL*), juntamente com a ferramenta de análise de corpora INTEX[1](Silberztein 2000a).

O DELA tornou-se um padrão de dicionários eletrônicos (também conhecidos como léxicos computacionais) que é utilizado pela rede de pesquisa européia RELEX[2]. Esses dicionários foram primeiramente desenvolvidos para serem utilizados pela ferramenta de análise de corpora, o INTEX, e mais atualmente para sua versão de código fonte aberto, o Unitex[3].

O Unitex é um ambiente de desenvolvimento lingüístico que inclui dicionários e gramáticas de grande cobertura de várias línguas (Inglês, Francês, Grego, Português de Portugal, Tailandês), e processa textos com milhões de palavras em tempo real. Ele inclui ferramentas para criar e manter recursos lexicais e esses dicionários e gramáticas podem ser aplicados aos textos para localizar padrões morfológicos, lexicais e sintáticos, remover ambigüidades, e etiquetar palavras simples e compostas.

O *Núcleo Interinstitucional de Lingüística Computacional de São Carlos* (NILC) tem desenvolvido, desde sua criação em 1991, vários aplicativos e recursos lingüísticos para o português brasileiro (PB). Entre os aplicativos desenvolvidos, destaca-se o Revisor Gramatical ReGra (Martins et al. 1998), desenvolvido com o apoio da Itautec-Philco, da Fapesp, do CNPq e da Finep, que está comercialmente disponível no produto Microsoft Office versão português, desde 2000.

Este revisor conta com um léxico de aproximadamente 500 mil entradas de palavras simpes (incluindo derivações), cada uma podendo pertencer a uma ou mais categorias sintáticas, com atributos específicos e distintos.

Os dados do léxico, mais as unidades de verbos flexionados de ênclise e mesóclise, totalizam mais de 1.500.000 entradas (Nunes et al. 1996) e fazem parte ainda de uma base de dados lexicais, Diadorim, disponível na Web para consulta[4]. Essa base de dados centraliza todas as informações lexicais do NILC, resultado de cerca de 10 anos de pesquisas. Porém, o acesso de aplicações diretamente a essa base de dados é extremamente lento, uma vez que esses dados estão em um banco de dados relacional e suas tabelas possuem muitos milhares de entradas (Greghi 2002).

Uma saída para esse problema é a utilização de métodos de compactação e manipulação de léxicos baseados em autômatos finitos. O Revisor Gramatical

[1]Veja `http://www.nyu.edu/pages/linguistics/intex/`

[2]Veja `http://ladl.univ-mlv.fr/Relex/introduction.html`

[3]Veja `http://www-igm.univ-mlv.fr/~unitex/`

[4]Disponível para consulta em `http://www.nilc.icmc.usp.br/`, Tool & Resources, DIADORIM

ReGra utiliza, no seu léxico, essa tecnologia desenvolvida por Kowaltowski, Lucchesi e Stolfi (Kowaltowski & Lucchesi 1993; Kowaltowski et al. 1995b; Kowaltowski et al. 1995a), mas não se tem acesso ao seu código.

Como o Unitex é uma ferramenta de código fonte aberto e possui métodos de compactação e manipulação de léxicos igualmente baseados em autômatos finitos, é nossa intenção utilizá-lo para compactar os dados lexicais atualmente disponíveis na Diadorim, para que outras aplicações de PLN possam acessar essas informações de forma eficiente. Além disso pretende-se estender o conjunto de funcionalidades do Unitex ao incorporar funções atualmente inexistentes, como o acesso a um analisador morfológico, um parser e um analisador morfossintático de português brasileiro.

Para isso, deve ser construído um filtro no intuito de converter os dados da Diadorim para o formato DELA, objetivando manter todas informações gramaticais já disponíveis na base de dados. Com o dicionário convertido, ele pode ser utilizado tanto pela ferramenta Unitex, quanto por qualquer aplicação que necessite de um acesso eficiente a um léxico, o que seria de grande utilidade para qualquer grupo de pesquisa em PLN que necessite de um léxico de grande cobertura e de acesso rápido.

Dessa forma, os objetivos desse projeto podem ser resumidos em: a) construir um léxico computacional para o português brasileiro baseado no formalismo DELA, incorporando todas informações disponíveis na Diadorim, daqui em diante referenciado como Unitex-PB; b) criar e implementar um projeto de expansão do Unitex-PB, e conseqüentemente da Diadorim, para incluir entradas de palavras compostas; c) construir uma biblioteca de acesso e manipulação a esse léxico computacional, para que outras aplicações, além do Unitex, possam utilizá-lo; d) investigar, projetar e implementar novas funções para o Unitex;

Esta monografia está organizada da seguinte forma: no Capítulo 2 será apresentada uma revisão bibliográfica sobre léxicos computacionais, seus modos de representação e alguns trabalhos relacionados; no Capítulo 3 serão apresentadas as ferramentas INTEX e UNITEX, além de seus formalismos; nos Capítulos 4 e 5 é apresentada a proposta deste projeto juntamente com o cronograma. Por fim, é apresentado um Apêndice com os campos e códigos utilizados na primeira versão do léxico Unitex-PB.

CAPÍTULO 2

Léxicos Computacionais

Nos últimos anos, a pesquisa lexical vem se tornando cada vez mais o ponto chave em engenharia de língua. As razões para que isto ocorra são ambas práticas e teóricas. Com os lingüistas, o papel do léxico tem se tornado central: o léxico não é mais considerado uma simples lista de palavras (como nas décadas de 1960 e 1970); supõe-se que devam estar contidas num léxico quase todas as informações morfológicas, sintáticas, semânticas e fonológicas de uma língua (Tiberius 1999). Atualmente, a crescente necessidade de aplicações de PLN fez ressaltar a carência de dados linguísticos de dimensões reais, e, em particular, de léxicos e gramáticas de grande cobertura (Ranchhod 2001).

Neste capítulo estaremos abordando o que é um léxico computacional, suas formas de representação e alguns trabalhos relacionados.

2.1 Léxico computacional

O léxico computacional ou dicionário é uma estrutura fundamental para a maioria dos sistemas e aplicações de *processamento de língua natural* (PLN), sendo que ele é uma estrutura de dados contendo os itens lexicais e as informações correspondentes a estes itens. Na realidade, os itens que constituem as entradas de um léxico podem ser palavras isoladas (como *lua*, *mel*, *casa*, *modo*) ou composições de palavras as quais, reunidas, apresentam um significado específico (por exemplo, *lua de mel* ou *Casa de Cultura* ou *a grosso modo*). Entre as informações associadas aos itens lexicais, no léxico, encontra-se a categoria gramatical (*part-of-speech*) do item, além de valores para variáveis morfo-sintático-semânticas como gênero, número, grau, pessoa, tempo,

modo, regência verbal ou nominal etc. Também são associadas ao item lexical, no léxico, uma ou mais representações ou descrições semânticas. No entanto, associações a representações contextuais são raramente encontradas. Esses léxicos ou dicionários são especialmente elaborados para serem utilizados em operações automáticas de processamento de textos.

Todas as aplicações que têm como objetivo o tratamento computacional de língua natural consideram o léxico como componente central, o que tem provocado uma demanda constante de informações léxicas detalhadas sobre áreas amplas de vocabulários. A finalidade fundamental do processamento de língua natural é a automatização dos processos lingüísticos, tais como a compreensão, produção ou aquisição de uma língua, tarefas que os usuários de uma língua realizam de forma fluente e natural. Tanto para os humanos como para as máquinas, todas essa tarefas implicam num conhecimento profundo do vocabulário de uma língua (Ortiz 2000).

O trabalho de construção de um léxico para uma língua é enorme. O dicionário *Oxford English Dictionary*, por exemplo, contém 250.000 entradas de palavras independentes e, apesar do número aparentemente elevado, ele não inclui muitas palavras pertencentes a vocabulário técnico. O processo de construir um léxico manualmente é muito custoso, tanto em recursos humanos como em tempo e dinheiro. Isso tem levado muitos pesquisadores a considerar como fonte potencial de informações léxicas as versões eletrônicas de dicionários impressos, que podem ser convertidos de forma automática ou semi-automática em sistemas de PLN.

Em termos gerais, pode-se identificar ao menos cinco tipos de conhecimento que são relevantes para qualquer sistema de PLN (Nunes et al. 1999). São eles:

1. fonético-fonológico: quando se trata de depreender a identidade sonora dos elementos que constituem a palavra.

2. morfológico: quando as unidades mínimas dotadas de significado são isoladas para a compreensão do processo de formação e flexão das palavras.

3. sintático: quando a distribuição das palavras resulta em determinadas funções que elas desempenham na sentença.

4. semântico: quando o conteúdo significativo da palavra implica relações de natureza ontológica e referencial para a identificação dos objetos no mundo.

5. pragmático-discursivo: quando a força expressiva das palavras remete à

identificação dos objetos do mundo em termos do seu contexto de enunciação e condições de produção discursiva.

Segundo (Pustejovsky 2001), entre esses tipos de conhecimento, três devem fazer parte da estrutura de uma entrada lexical: fonológico, sintático e semântico. Existem dois tipos de conhecimento sintático associados a um item lexical: sua *categoria* e sua *subcategoria gramatical*. Em léxicos são incluídas classificações tradicionais tanto para as maiores categorias gramaticais, como os nomes (substantivos), verbos, adjetivos, advérbios e preposições como para as categorias menores, como os determinantes e conjunções.

O conhecimento de subcategoria de um item lexical é geralmente a informação que diferencia categorias em classes distribucionais e distintas. Esta forma de informação pode ser separada em: *característica contextual* e *característica inerente*. Estas características podem ser definidas em termo do contexto no qual uma dada entrada lexical pode ocorrer. A informação de subcategorização pode marcar tanto o contexto sintático local para uma palavra, como o contexto semântico. Ela é a informação que garante, por exemplo, que o verbo *devour* é sempre transitivo no inglês, requerendo um objeto direto. A entrada lexical codifica este requerimento como uma característica contextual de subcategorização (contexto sintático).

Já as *características inerentes* são propriedades de entradas lexicais que não são facilmente deduzidas a partir de uma definição contextual, mas se referem a um tipo de entidade ontológica (contexto semântico). Isso engloba características como contável, abstrato, animado, humano, físico, etc.

Informações semânticas também podem ser classificadas em duas categorias. A primeira identifica a qual classe semântica um item lexical pertence (como entidade, evento, propriedade) e a segunda especifica as características semânticas de argumentos de itens lexicais.

2.2 Modos de representação

Até o início dos anos 80, o processo de desenvolvimento de léxicos e bases de informação lexical era realizado sem grandes preocupações com a padronização na elaboração e organização dos dados utilizados ou mesmo na construção do recurso propriamente dito, o que tornava a modificação e a reutilização dos dados duas tarefas praticamente impossíveis de serem executadas. A partir de então, vários pesquisadores passaram a se preocupar com a reutilização dos dados e, conseqüentemente, com a diminuição do esforço inicial para o desenvolvimento de novas aplicações (Evans & Kilgarriff 1995).

As principais formas de representação que então surgiram foram: dicionário legível por máquina (*machine-readable dictionary*), dicionário tratável por

máquina (*machine tractable dictionary*) e, posteriormente, base de dados lexicais (*lexical database*) e base de conhecimento lexical (*lexical knowledge base*). A tipologia utilizada nesse trabalho é baseada nos trabalhos de (Correia 1994; Correia 1996).

2.2.1 Dicionários legíveis por máquinas

A oposição entre *dicionário legível por máquina* (do inglês, MRD) e *dicionário tratável por máquina* (do inglês, MTD) é proposta por (Wilks et al. 1988).

Os MRDs são dicionários feitos por lexicógrafos e concebidos para uso humano. São geralmente dicionários que, ou foram inicialmente construídos em formato digital, ou foram criados no formato papel e posteriormente transferidos para formato digital. Desses dicionários são publicadas versões impressas e versões digitais. A denominação MRD pode, portanto, corresponder a produtos diferentes em termos de concepção e metodologia de trabalho. No entanto, todos estes produtos apresentam como denominador comum as características de serem concebidos para uso humano e de se encontrarem disponíveis em formato digital.

A estrutura interna desses dicionários é semelhante à dos dicionários impressos, isto é, basicamente as unidades lexicais são descritas em artigos distintos, apresentando a estrutura tripartida clássica, como visto na Figura 2.1.

Entrada - categoria - definição (eventualmente, exemplificação)

Figura 2.1: Exemplo da estrutura de um dicionário legível por máquina.

Segundo (Correia 1994), os MRDs, embora se beneficiando das virtudes do formato digital, que se traduzem em grande diversificação e aumento de possibilidades de consulta, não são susceptíveis de serem utilizados diretamente em sistemas de PLN, devido fundamentalmente ao fato de serem concebidos para uso humano, isto é, a informação é dada em língua natural, pouco formalizada, não reconhecível pelos programas de PLN, que pressupõem grande formalização da informação.

2.2.2 Dicionários tratáveis por máquina

Segundo (Wilks et al. 1988) MTD é um MRD transformado, apresentando um formato que o torne apto a ser usado em sistemas de PLN. Esta aptidão resulta basicamente na descrição do conhecimento lexical num formalismo no qual o sistema possa facilmente o reconhecer, traduzindo a informação

que nos dicionários humanos é apresentada em língua natural, bem como na explicitação de todo o conhecimento que nos dicionários para uso humano permanece implícito na sua descrição. Os MTDs são, em primeira instância, apenas utilizáveis em sistemas de PLN (Correia 1994).

O presente projeto utilizará este formato de dicionário.

2.2.3 Base de dados lexicais

Uma *base de dados lexicais* (BDL) é uma estrutura computacional criada para ser capaz de suportar os mais variados tipos de conhecimento sobre cada unidade lexical, permitindo estabelecer conexões, tanto entre unidades lexicais distintas, quanto entre características pertencentes a unidades lexicais distintas. Isto permite observar e acessar as unidades lexicais sob as mais variadas formas.

Uma das principais características das BDLs, do ponto de vista teórico, é o fato de corresponderem a uma concepção de léxico bastante diferente da dos dicionários: numa BDL, o léxico é entendido como uma complexa rede de relações (morfológicas, sintagmáticas, semânticas, paradigmáticas), onde o conhecimento sobre uma unidade lexical é composto de vários níveis ou camadas. Por outro lado, nos dicionários em geral (dicionários impressos, MRDs ou MTDs), o léxico é encarado como uma listagem de unidades a descrever de forma atomística, não sendo potenciadas (ou, pelo menos, não de modo sistemático e/ou exaustivo) as relações interlexicais (Calzolari 1990).

Dados lexicais são muito mais complexos do que os tipos de dados usados para a maioria das pesquisas na área de banco de dados (Ide & Véronis 1992). Dessa forma, é necessário que seja feito um levantamento sobre os possíveis modelos de representação dos dados e também sobre *sistemas de gerenciamento de bancos de dados* (SGBDs) disponíveis, para que se encontre um modelo adequado à implementação.

2.2.4 Bases de conhecimento lexical

Uma base de conhecimento lexical (BCL) representa explicitamente uma teoria do léxico, sendo, por isso, um corpo de informação representada num tipo de notação especial - a LRL (*lexical representation language*), que contém uma sintaxe e uma semântica explícitas e que suporta operações lexicais capazes de realizar transformações válidas dessa informação (Briscoe 1991).

Em outras palavras, enquanto uma BDL é concebida como uma representação estática das propriedades das unidades lexicais extraíveis de MRDs, uma BCL é concebida como uma representação dinâmica, à medida que, além de conter informação lexical estruturada, pressupõe a construção de uma LRL

capaz de analisar essa informação e de gerar produções lingüísticas. A definição dessa LRL é feita explicitamente de acordo com uma teoria semântica determinada. No interior da BCL, é possível *navegar* pelo léxico, caminhando nele através dos conceitos ou relações semânticas, o que o faz se assemelhar conceitualmente a um *thesaurus* (Correia 1996).

2.3 Trabalhos relacionados

A seguir, serão apresentados alguns exemplos de projetos que tratam de construção de léxicos.

2.3.1 GENELEX

O projeto GENELEX (*GENEric LEXicon*) é um projeto da União Européia que iniciou-se em 1990 na França, sendo uma abstração do modelo de dicionário monolingüe francês, que foi progressivamente expandido para outros países. Foi um dos primeiros projetos com o intuito de reusar os recursos léxicais.

O objetivo do projeto é a criação de BDLs monolingües de várias línguas européias, com uma especificação comum. A informação a ser inserida nessas bases diz respeito principalmente às unidades do léxico comum, não havendo, portanto, nenhum enfoque particular nas terminologias científicas ou técnicas.

Como o próprio nome sugere, a principal característica no desenvolvimento do GENELEX é a generalidade do seu formato, o que possibilita a recuperação do maior número de informações lingüísticas de uma dada entrada, a maior portabilidade do sistema e o esforço para evitar condições que possam impossibilitar a completeza do léxico em conformidade com a especificação estabelecida (EAGLES 1993; EAGLES 1996).

O resultado dessa iniciativa foi uma gramática formal em SGML (*Standard Generalised Markup Language*), que descreve os elementos permitidos e as relações permitidas entre eles.

2.3.2 WordNet

WordNet é um sistema de referência lexical do inglês, *on-line*[1], desenvolvido por um grupo de pesquisadores no *Cognitive Science Laboratory*, na Universidade de Princeton, nos EUA.

Esse sistema baseou-se em teorias psicolingüísticas concernentes à organização do léxico na memória humana, ou seja, o léxico mental (Miller et al. 1990). Ele tenta organizar as informações lexicais em termos do significado

[1]Veja `http://www.cogsci.princeton.edu/~wn/`

das palavras, mais do que de suas formas, o que torna o sistema mais semelhante a um *thesaurus* do que a um dicionário propriamente dito.

A WordNet foi desenvolvida para o tratamento da língua inglesa e ela divide o léxico em cinco categorias: substantivos, verbos, adjetivos, advérbios e palavras funcionais. Porém, esse sistema atualmente possui substantivos, verbos, adjetivos e advérbios. Esse tipo de categorização o diferencia de um dicionário tradicional e, apesar de causar certa redundância nas informações armazenadas (algumas palavras podem ser classificadas em mais de uma categoria), traz a vantagem de que diferenças fundamentais na organização semântica dessas categorias sintáticas podem ser claramente observadas e facilmente exploradas.

A idéia básica utilizada na WordNet é a representação das palavras e de seus significados em uma matriz lexical. O mapeamento entre as formas e seus significados é N:N, ou seja, algumas formas podem ter diferentes significados, e alguns significados podem ser expressos por várias formas diferentes. A WordNet distingue relações semânticas de relações lexicais. O significado de uma palavra P1 pode ser representado por uma lista de palavras que podem ser usadas para expressar P1 {S1, S2,...}. Uma matriz lexical consiste, assim, em um mapeamento entre palavras e conjuntos de sinônimos (*synsets*). Os sinônimos são relações lexicais entre palavras.

A Tabela 2.1 é um exemplo de ilustração de uma matriz lexical. As formas das palavras (*word forms*) estão nos cabeçalhos das colunas e os significados (*word meanings*) nos cabeçalhos das linhas. Uma entrada em uma célula da matriz implica que a forma naquela coluna pode ser usada (em um contexto apropriado) para expressar o significado daquela linha. Deste modo, a entrada $E_{1,1}$ significa que a forma F_1 pode ser usada para expressar o significado M_1. Se existem duas entradas na mesma coluna, a forma é polissêmica. Se existem duas entradas na mesma linha, as duas formas são sinônimos (relativos a um contexto).

Significado das Palavras	Forma das Palavras				
	F_1	F_2	F_3	$\ldots$	F_n
M_1	$E_{1,1}$	$E_{1,2}$			
M_2		$E_{2,2}$			
M_3			$E_{3,3}$		
$\vdots$				$\ddots$	
M_m					$E_{m,n}$

Tabela 2.1: Ilustrando o Conceito de Matriz Lexical.

O modelo da WordNet tem sido amplamente adotado em outros projetos de mesma natureza para outras línguas. Temos, como exemplo, o projeto

WordNet-BR (da Silva et al. 2002), para o português brasileiro, em desenvolvimento no NILC[2].

2.3.3 Relex

Relex é uma rede informal de laboratórios de grupos de pesquisas europeus[3] (França, Alemanha, Itália, Portugal) que trabalham no domínio de lingüística computacional para a construção de léxicos eletrônicos e gramáticas. Cada grupo trabalha em sua língua nacional e todas as equipes estão usando métodos idênticos. Pelo menos uma vez ao ano eles se encontram para confrontar seus problemas, apresentar seus resultados e adotar futuras padronizações.

Dicionários de tamanhos significativos foram construídos para cada língua e programas que incorporam esses dicionários foram construídos para processar *corpora*. Uma característica muito importante deste trabalho é que, em todos os níveis, os grupos de pesquisa trabalharam nos mesmos itens (dicionários e gramáticas) e que seus resultados parciais têm sido unidos sem grandes dificuldades. A metodologia comum garante a acumulação de dados.

Este projeto utiliza, como formato padrão para os dicionários, o padrão DELA (Silberztein 1990; Courtois 1990) desenvolvido na França (esse padrão será visto com mais detalhes no capítulo 3). Nestes grupos de pesquisa estão sendo desenvolvidos dicionários de palavras simples, de palavras compostas e dicionários fonológicos.

Alguns dicionários desse projeto, como o dicionário do Português de Portugal[4], já estão sendo disponibilizados *on-line*. Atualmente, o dicionário do português de Portugal possui 1.250.000 entradas de palavras simples flexionadas e 25.000 entradas de palavras compostas flexionadas[5] (Ranchhod et al. 1999).

O projeto aqui proposto certamente trará contribuições diretas à rede RELEX.

2.3.4 Diadorim

É um projeto de BDL para o português brasileiro desenvolvido no NILC (Greghi 2002). O NILC tem desenvolvido, desde sua criação em 1991, vários aplicativos e recursos lingüísticos para o português brasileiro e a Diadorim é uma BDL que foi desenvolvida incorporando as informações presentes no lé-

[2]Veja `http://www.nilc.icmc.usp.br/nilc/projects/wordnetbr.htm`
[3]Veja `http://ladl.univ-mlv.fr/Relex/introduction.html`
[4]Veja `http://label.ist.utl.pt/pt/resources/resources.htm`
[5]Os conceitos de palavras simples e compostas são apresentados no Capítulo 3.

xico do NILC[6], no dicionário UNL-Português utilizado no projeto UNL/Brasil[7], e as informações presentes num *thesaurus* (da Silva et al. 2000) da língua portuguesa. Seu objetivo é centralizar todas essas informações em uma única base de dados.

Atualmente essa BDL possui cerca de 1,5 milhão de entradas lexicais (palavras simples), representadas em um banco de dados, usando modelo relacional.

A base Diadorim pode ser acessada *on-line*[8] e o acesso pode ser feito por meio de dois módulos: (i) Módulo de consulta aos dados morfossintáticos; (ii) Módulo de consulta aos dados do *thesaurus*.

2.4 Considerações finais

Neste capítulo foram apresentados a definição de Léxico Computacional, que é um recurso fundamental para qualquer aplicação de PLN, suas formas de representação e trabalhos importantes que resultaram na construção de léxicos.

Sobre as representações lexicais, a principal característica sobre a escolha do modelo será a finalidade de sua representação. Esta, por sua vez, terá em cada uma de suas formas um tipo de aplicação de PLN.

[6]Maiores detalhes em (Nunes et al. 1996).
[7]Veja `http://www.nilc.icmc.usp.br/nilc/projects/unl.htm`
[8]Veja `http://www.nilc.icmc.usp.br/nilc/tools/intermed.htm`

CAPÍTULO
3

INTEX/Unitex

O INTEX[1] é um ambiente de desenvolvimento lingüístico que pode ser utilizado para analisar corpora de muitos milhões de palavras em tempo real. As descrições lingüísticas são formalizadas através de dicionários eletrônicos (léxicos) e gramáticas de grandes dimensões, representados por autômatos de estados finitos (Silberztein 2000b).

Este sistema utiliza tecnologias criadas pelo Laboratoire d'Automatique Documentaire et Linguistique[2] (LADL), fundado em 1967 por Maurice Gross, na Université de Marne-la-Vallée, na França. Ele vem sendo desenvolvido desde 1992 e sua versão inicial era para NextStep, sendo que em 1996 ele foi completamente integrado com uma interface gráfica (versão 3.0) e começou a ser distribuído para centros de pesquisa como um ambiente lingüístico de pesquisa.

Hoje em dia, mais de 200 laboratórios de pesquisa em vários países utilizam o INTEX como uma ferramenta de pesquisa ou educacional. Alguns usuários estão interessados nas funcionalidades de processamento de corpus (análise literária de textos, pesquisando informações em jornais ou documentos técnicos, etc); outros estão utilizando esta plataforma para formalizar certos fenômenos lingüísticos (por exemplo, descrevendo morfologia, léxico e expressões da língua), ou ainda por seu poder computacional (análise automática de textos).

Países como Alemanha, Coréia, Eslovênia (Vitas & Krstev 2001), Espanha, França, Grécia (Anastasiadis-Symeonidis et al. 2000), Itália (Vietri & Elia 2000), Noruega, Polônia, Portugal (Ranchhod et al. 1999) e Tailândia entre

[1] Veja `http://www.nyu.edu/pages/linguistics/intex/`
[2] Veja `http://ladl.univ-mlv.fr/`

outros, estão trabalhando para a construção de seus próprios dicionários lexicais para o sistema INTEX.

Este capítulo tem como objetivo apresentar uma visão geral da ferramenta INTEX e de sua nova vertente, o Unitex, a qual será utilizada neste projeto, bem como os padrões utilizados nessas ferramentas.

3.1 Dicionários lexicais no padrão INTEX

O sistema INTEX utiliza um conjunto de dicionários eletrônicos em um formalismo concebido pelo LADL para o francês conhecido como DELA (*Dictionnarie Electronique du LADL*). Este formalismo permite declarar entradas lexicais simples e compostas de uma língua. Entradas estas que podem ser associadas a informações gramaticais e à semântica de suas flexões. Esses dicionários são instrumentos lingüísticos especificamente concebidos para serem utilizados em operações automáticas de processamento de texto.

Os dicionários utilizados pelo INTEX para identificar palavras em um texto são os dicionários de palavras flexionadas, DELAF (DELA de palavras Flexionadas) ou o DELACF (DELA de palavras Compostas Flexionadas). Esses dicionários são geralmente gerados automaticamente a partir dos dicionários DELAS (DELA de palavras Simples) (Courtois 1990) e DELAC (DELA de palavras Compostas) (Silberztein 1990).

O DELAS, como dito anteriormente, é o dicionário de *palavras simples*, entendendo-se por palavras simples seqüências de caracteres alfabéticos delimitadas por *separadores*. Um separador é um caractere não alfanumérico.

As entradas do DELAS possuem a seguinte estrutura:

<palavra>, <descrição formal>

onde *palavra* representa a forma canônica (o lema) de uma unidade lexical simples (em geral, o masculino singular para nomes e adjetivos que têm essa variação; feminino singular para os nomes e adjetivos que são exclusivamente femininos; infinitivo para verbos) e *descrição formal* corresponde a um código alfanumérico que contém as informações gramaticais das entradas: sua classe gramatical e seus comportamentos semânticos/morfológicos.

As formas flexionadas das palavras são geradas automaticamente da associação do lema ao código flexional, gerando desta forma o DELAF. Assim, a Figura 3.1 nos mostra exemplos de entrada para o DELAS, onde *N* e *A* indicam, respectivamente, que *gato* é um nome (substantivo) e *gordo* é um adjetivo; *01* corresponde à regra de flexão para masculino, feminino, singular e plural; *D1* e *S1* explicita os tipos de sufixos de diminutivo e superlativo

gato, N01D1

gordo, A01D1S1

Figura 3.1: Exemplos de entrada paro o dicionário DELAS.

que podem ser aceitos por essas entradas. Estas regras geralmente estão em forma de autômatos finitos ou expressões regulares.

Este último exemplo gera as formas flexionadas (entradas do DELAF) vistas na Figura 3.2.

gato, gato.N: ms

gata, gato.N: fs

gatos, gato.N: mp

gatas, gato.N: fp

gatinho, gato.N: Dms

gatinha, gato.N: Dfs

gatinhos, gato.N: Dmp

gatinhas, gato.N: Dfp

gordo, gordo.A:ms

gorda, gordo.A: fs

gordos, gordo.A: mp

gordas, gordo.A: fp

gordinho, gordo.A: Dms

gordinha, gordo.A: Dfs

gordinhos, gordo.A: Dmp

gordinhas, gordo.A: Dfp

gordíssimo, gordo.A: Sms

gordíssima, gordo.A: Sfs

gordíssimos, gordo.A: Smp

gordíssimas, gordo.A: Sfp

Figura 3.2: Exemplo de formas flexionadas para as entradas *gato* e *gordo*.

O DELAC é o dicionário de *palavras compostas*, isto é, das unidades lexicais que são constituídas por uma combinação fixa de palavras simples, que representam uma parte significante de um léxico de qualquer língua. As pa-

lavras compostas são seqüências de palavras que apresentam restrições às propriedades que as palavras teriam individualmente.

A formalização das entradas dos dicionários de palavras compostas é similar à de palavras simples. Desde que advérbios compostos, preposições e conjunções não sofram flexões, seus formatos são simples (por exemplo, *para com*, PREP).

Nomes compostos, entretanto, tem geralmente formas flexionadas. As regras para flexão de nomes compostos geralmente exibem restrições flexionais em gênero ou número que não podem ser consideradas pelas propriedades morfológicas de seus constituintes. No formato DELA, as propriedades flexionais dos nomes compostos são especificadas com o mesmo critério do dicionário de palavras simples. Exemplos de entradas para o DELAC estão na Figura 3.3.

ser(21) humano(01), N + NA: ms - +

guerra fria, N + NA:fs - -

Figura 3.3: Exemplos de entrada para o dicionário DELAC.

Estes dois nomes compostos correspondem a *Adjetivo Nominal* (*NA*). Cada entrada é caracterizada pela possibilidade (+) ou impossibilidade (-) de flexão de gênero e número, respectivamente. Os elementos das palavras compostas que podem ser flexionados recebem o código que eles tem no DELAS: ambos constituintes de *ser humano* são flexionados (em número) de acordo, respectivamente, com as regras 21 e 01. O dicionário que tem como entrada as flexões das palavras compostas é o DELACF.

3.2 Funcionalidades da ferramenta

O INTEX é um sistema baseado no uso de grandes dicionários lexicais. Ele pode ser usado para analisar textos de muitos milhões de palavras. Inclui vários dicionários e gramáticas embutidos representados como autômatos de estados finitos, porém, o usuário pode adicionar seus próprios dicionários e gramáticas. Estas ferramentas são aplicadas ao texto para localizar padrões léxicos e sintáticos, gerar dicionários lexicais, remover ambigüidades e etiquetar palavras simples como também expressões complexas. Ele pode ser utilizado por lingüistas para análise de corpora, mas também pode ser visto como um sistema de recuperação de informação.

3.2.1 Reconhecimento de unidades lexicais: simples e compostas

O modo como o sistema trata um dado texto é o seguinte: primeiro identifica todos os tokens (palavras, sinais de pontuação, marcadores de frase e algarismos), depois aplica os dicionários de palavras simples e compostas da língua corrente, indexando todas as unidades lexicais, ou seja, associando a cada uma delas a informação constante dos dicionários aplicados.

O sistema associa todas as formas flexionadas (dicionários DELAF, DELACF) ao seu lema (mais do que um, no caso das ambigüidades resultantes de homografia, por exemplo: *entre,entrar.V:S1s:S4s:S3:Y4s; entre,entre.PREP*) e especifica os seus atributos lingüísticos indexando o texto. Por exemplo, *V:S1s:S4s:S3:Y4s* da forma verbal *entre* significa que esta forma corresponde à primeira, segunda, (tratamento "você") e terceira pessoas do singular do presente do subjuntivo, e à segunda pessoa do singular (tratamento "você") do imperativo. A ambigüidade nas palavras compostas é muito menor. Em todo caso, os compostos ambíguos estão incluídos em um dicionário e os não ambíguos em um outro dicionário. Por exemplo, o advérbio *em combinação* é ambíguo (devido à polissemia do nome *combinação* e à possibilidade de poder ser precedido pela preposição *em*, quer quando é um nome predicativo, relacionado com combinar, quer quando é um nome concreto) (Ranchhod & Santos 1999).

O sistema também constrói o dicionário de palavras de um dado texto, isto é, extrai dos dicionários aplicados no texto todas as entradas que ocorrem no texto. Entradas podem ser palavras simples (por exemplo, *casa*), ou compostas (por exemplo, *lua de mel*). Além disso, o sistema constrói um dicionário das palavras simples desconhecidas (que não estão nos dicionários DELA), que geralmente são palavras mal formadas ou nomes próprios não cadastrados.

Depois de indexado o mesmo, é possível construir um autômato do texto, onde cada token está associado às suas classificações morfossintáticas/semânticas. O exemplo da Figura 3.4, corresponde ao autômato do texto: "*Por outro lado, ele está vivo.*"

3.2.2 Identificação de padrões

Após ser efetuada a indexação, é possível localizar padrões morfossintáticos num corpus através de expressões regulares ou grafos. Todas expressões regulares podem ser representadas por grafos. Tais padrões podem ser:

- Uma dada palavra ou uma lista de palavras. Por exemplo, pode-se localizar em um texto todas as ocorrências da flexão do verbo *cantar* conjugado

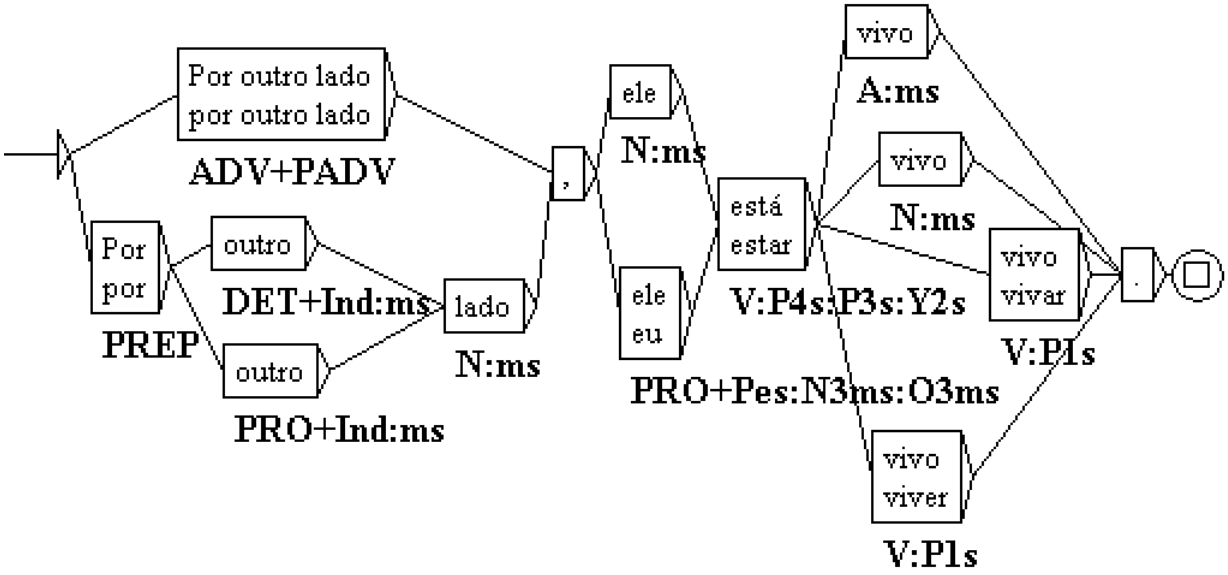


Figura 3.4: Representação do autômato de um texto.

no futuro, ou todas as ocorrências de palavras compostas.

- Uma dada categoria gramatical, como todos verbos conjugados na terceira pessoa do singular (*V:3s*) ou os nomes femininos no plural (*N:fp*). Abaixo alguns exemplos de categorias (quaisquer códigos gramaticais propostos pelo usuário criador do dicionário podem ser usados).

A:p	Adjetivo (*A*) no plural (*p*)
ADV	Advérbio
PRO	Pronome
DET:f	Determinante (*DET*) no feminno (*f*)
V+t:ms	Verbo (*V*) transitivo (*t*) no masculino singular (*ms*)

- Uma expressão regular ou um grafo. O seguinte exemplo é uma expressão regular:

(<dever>+<poder>) (<ADV> + <E>) <V:W>

Este padrão reconhece qualquer seqüência começando com o verbo *dever* ou *poder*, seguido de um advérbio opcional (<*E*> significa uma palavra vazia) e um verbo na forma infinitiva (<*V:W*>). Note que as categorias são reconhecidas tanto para palavras simples quanto para palavras compostas. Um exemplo de seqüência reconhecida por esse padrão é:

conhecido. Um fato que deverá ser alterado em Fe
ro do próximo ano, se se puderem cumprir as prev

- Conjunto de expressões de sinônimos. Grafos de diferentes línguas podem ser ligados, para que cada seqüência reconhecida numa língua fonte seja automaticamente associada a um grafo correspondente na língua alvo. Um grafo pode representar todas as expressões que designam entidades ou um processo. Indexando estes grafos (ao invés de meras palavras) pode-se recuperar informações em corpora grandes com alta precisão.

- Gramáticas locais de uma língua. O INTEX inclui um editor de grafos, o qual pode ser utilizado para edição de gramáticas locais. Operações padrões em grafos (união, intersecção, diferença) permitem ao usuário construir facilmente sistemas com centenas de grafos.

3.2.3 Resolução de ambigüidade

Como o DELAF e o DELACF são dicionários que têm, em geral, uma grande cobertura, ambos contêm palavras cuja classificação morfossintática pode ocorrer somente em domínios específicos. Como conseqüência, tais palavras podem ser analisadas de forma imprópria, ou seja, quando elas têm mais de uma classificação morfossintática. Uma saída para resolver ambigüidade é utilizar dicionários filtrados, isto é, quando o usuário sabe que em um dado corpus, uma entrada ambígua do dicionário só pode ter uma classificação morfossintática, ele remove do dicionário as outras entradas.

Muitas das palavras compostas podem ser ambíguas, pois elas podem ser analisadas como seqüências de palavras simples, entretanto, algumas palavras compostas não são ambíguas, ou porque elas contêm constituintes não autônomos ou porque são termos técnicos. Inserindo estas palavras compostas não ambíguas num dicionário filtrado, o usuário previne o INTEX de procurar em dicionários por palavras simples, uma vez que o sistema não mais processa essas palavras compostas como ambíguas.

Dicionários filtrados são usados quando é possível desambiguar uma palavra independentemente de seu contexto. Isto nem sempre é possível. Nestes casos podemos utilizar gramáticas locais.

Uma gramática local é uma regra de duas partes representada por um autômato de estado finito. Se uma dada seqüência de palavras é reconhecida, então as palavras seguintes são etiquetadas de maneira correta. Por exemplo, no Português a palavra *compra* pode ser um nome ou um verbo; a forma *o* pode ser um determinante, um pronome demonstrativo ou um pronome pessoal. Então, a combinação linear desses elementos permite seis diferentes análises (nome - determinante, nome - pronome demosntrativo, nome - pronome pessoal, verbo - determinante, verbo - pronome demosntrativo, verbo -

pronome pessoal) (Ranchhod, Mota, & Baptista 1999). Entretanto, em sentenças como:

Ela compra-o hoje.

compra é somente um verbo, e *o* é somente um pronome pessoal ligado ao verbo por um hífen. O autômato de estado finito na Figura 3.5 pode resolver esta ambigüidade.

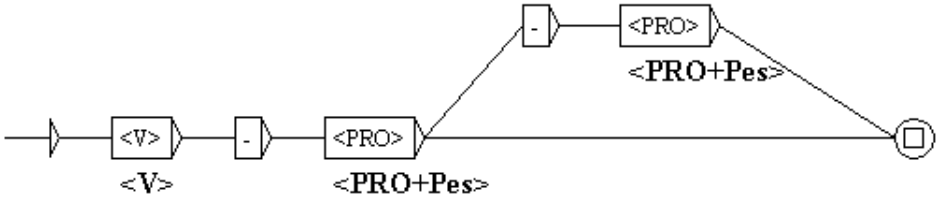


Figura 3.5: Exemplo de automato finito que resolve a ambigüidade do "*o*".

3.2.4 Etiquetação de palavras ou expressões

O INTEX, além de pesquisar um texto por padrões, pode ser utilizado para inserir informações em textos. O usuário pode adicionar informações aos grafos de busca de padrões que, ao serem reconhecidos, adicionam ao texto as informações contidas no grafo. Esses grafos especiais são chamados de *transducers* e podem tanto ser utilizado para inserir informações (etiquetar palavras ou expressões) quanto para substituir informações, isto é, quando uma expressão for reconhecida, ela pode ser substituída pela informação contida no grafo.

Um exemplo de *transducer* pode ser observado na Figura 3.6, onde o padrão reconhecido são *nomes* no *masculino singular*. Caso esse *transducer* seja utilizado para inserir informação em um texto (modo *merge*), toda vez que tal padrão for encontrado, será adicionado antes do padrão a tag *{Substantivo}*. Caso seja utilizado para substituir informação em um texto (modo *replace*), tada vez que o *transducer* encontrar esse padrão, o padrão será substituído no texto pela tag *{Substantivo}*.

3.3 Unitex

Apesar de todas essas funcionalidades, o INTEX é um sistema proprietário e não pode ser modificado. Porém, em outubro de 2002, o LADL lançou uma

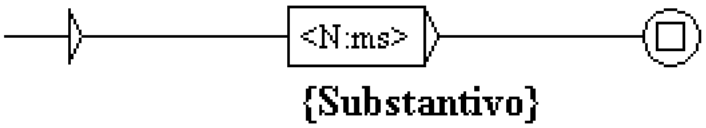


Figura 3.6: Exemplo de um *transducer*.

vertente do INTEX, mas de código fonte aberto, o Unitex[3].

O Unitex possui as mesmas funcionalidades do INTEX e, além disso, trabalha com o padrão unicode[4] 3, o qual permite a utilização de aproximadamente todos os caracteres das línguas, inclusive línguas asiáticas. A Figura 3.7 mostra a interface do sistema Unitex.

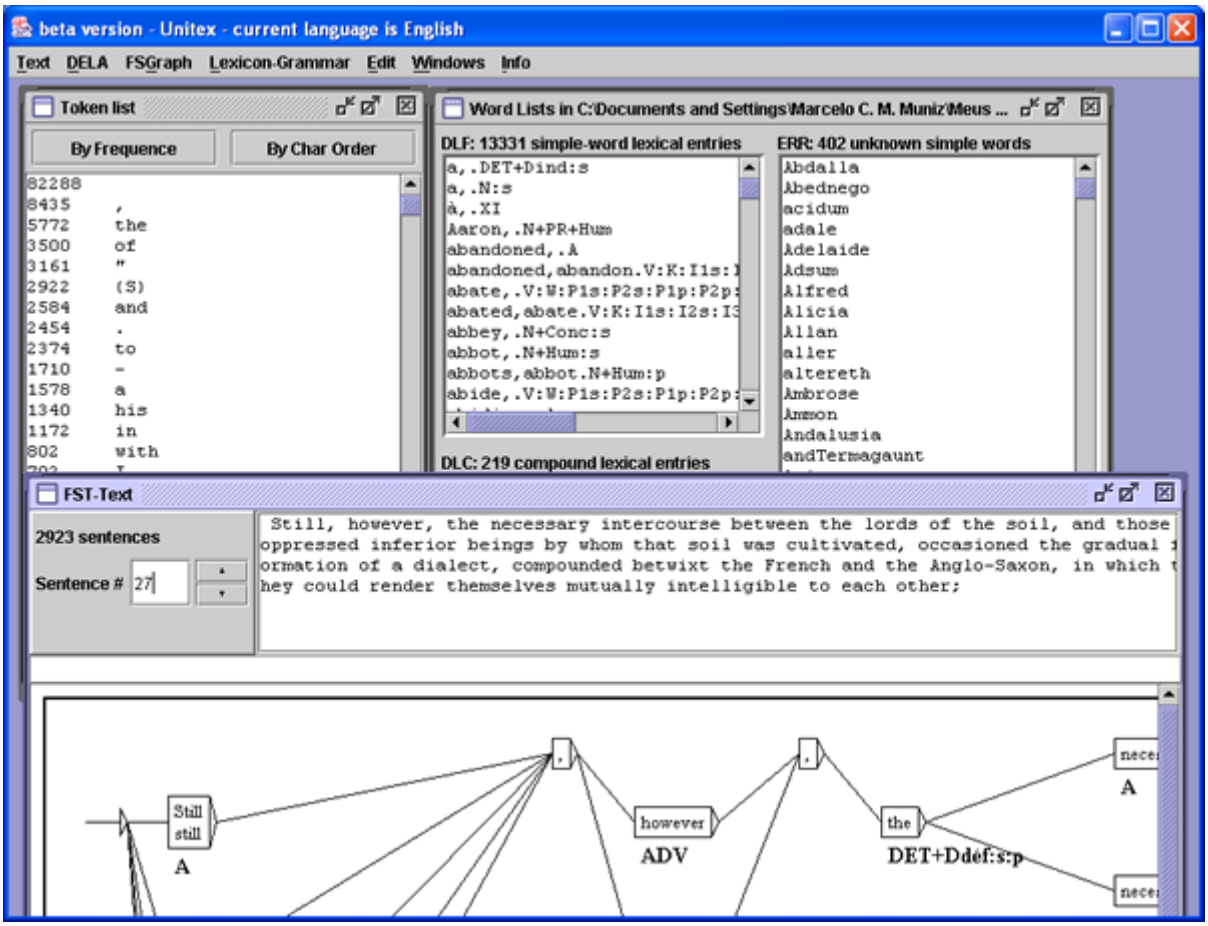


Figura 3.7: Interface do sistema Unitex.

A ferrameta também é um sistema multi-plataforma, sendo sua interface escrita em Java e todos os outros programas em ANSI C. Isto permite que o Unitex funcione em qualquer sistema que suporte Java 1.4 e que compile

[3]Veja `http://www-igm.univ-mlv.fr/~unitex/`
[4]Veja `http://www.unicode.org/`

programas em C.

Este sistema é distribuído livremente sob os termos da General Public License[5] (GPL). Portanto todos têm acesso ao código fonte da aplicação e podem modificá-lo seguindo os termos da licença GPL.

Neste projeto, estaremos construindo os dicionários lexicais do português brasileiro no formato do LADL para ser utilizado no sistema Unitex. Além disso, temos a oportunidade de estender as utilidades do sistema construindo novas ferramentas de PLN.

3.4 Considerações Finais

Neste capítulo foram apresentadas as ferramentas INTEX e Unitex, ambas de análise de corpora. O INTEX é uma ótima ferramenta para o PLN, porém é uma ferramenta proprietária. Como o Unitex possui as mesmas funcionalidades do INTEX e, além disso, é de código fonte aberto, ele possui uma grande potencialidade e pode ser estendido.

[5]Veja `http://www.gnu.org/licenses/gpl.html`

CAPÍTULO 4

Proposta

O objetivo deste trabalho é criar um léxico para português brasileiro para o ambiente Unitex, incluindo todas as informações disponíveis nos recursos desenvolvidos pelo NILC. Além disso, pretende-se estender o conjunto de funcionalidades do Unitex.

A seguir são apresentadas a metodologia pretendida para o desenvolvimento deste trabalho e as tarefas já realizadas.

4.1 Metodologia

Este trabalho será dividido em 5 etapas, a saber:

1. Levantamento de requisitos
2. Modelagem e implementação do léxico Unitex-PB
3. Projeto e implementação de expansão da Diadorim
4. Desenvolvimento de uma biblioteca de acesso e manipulação do léxico Unitex-PB
5. Extensão de funcionalidades do Unitex

As etapas descritas acima, bem como a metodologia a ser utilizada para a execução de cada uma delas e seu status atual, serão detalhados nas subseções seguintes.

4.1.1 Levantamento de requisitos

O primeiro passo a ser realizado nesse projeto é o levantamento dos requisitos para um bom léxico computacional. Devem ser pesquisadas quais informações devem estar contidas no léxico a partir das necessidades de aplicações de PLN e de experiências prévias reportadas no NILC e pela literatura. Este estudo deverá ser realizado em conjunto com especialistas em lingüística.

Como o formalismo a ser utilizado nesse projeto será o DELA, deverá ser feito um estudo de como funciona em detalhes esse formalismo, como também um estudo da ferramenta Unitex. Essa ferramenta possui um manual e seu código fonte está disponível para consulta. Este estudo visará descobrir maneiras de integrar outras ferramentas (já existentes no NILC ou não) ao Unitex.

Essa etapa já foi cumprida, sendo que o primeiro passo realizado foi o estudo da ferramenta Unitex e de seus formalismos, conforme descrito no capítulo 3. Utilizamos o léxico no formato DELA do português de Portugal[1] para observar os padrões utilizados e tentar, ao criar o léxico Unitex-PB, utilizar os mesmos padrões, facilitando o reaproveitamento mútuo.

Foi elaborado um levantamento de requisitos para o léxico Unitex-PB juntamente com lingüístas do NILC, e optou-se primeiramente pela utlilização das mesmas informações contidas no léxico do ReGra para o léxico Unitex-PB. Observou-se que, utilizando o formalismo DELA, não haverá perda de informações no léxico. As informações presentes no léxico Unitex-PB, que dependem de cada classe gramatical, podem ser visualisadas na Tabela 4.1. Além dessas informações, cada entrada está associada a sua canônica.

4.1.2 Modelagem e implementação do léxico Unitex-PB

Atualmente a Diadorim possui aproximadamente 1.500.000 entradas lexicais e ela será a maior fonte de informações para este projeto. A Diadorim é um banco de dados relacional que centraliza todas informações lexicais do NILC. As entradas estão organizadas em um modelo lingüístico como ilustrado na Figura 4.1 (Greghi et al. 2002).

É importante ressaltar que atualmente os módulos fonético-fonológico e morfológico não contêm qualquer informação.

Além dessas informações no banco de dados, existe uma versão desses dados em arquivos textos que é utilizado para a criação do léxico do ReGra. Como as duas fontes de informações possuem os mesmos dados, com exceção de informações de sinônimos e antônimos, que somente a Diadorim possui, elas poderão ser utilizadas para gerar o léxico Unitex-PB.

[1] Veja `http://label.ist.utl.pt/pt/resources/resources.htm`

Classe Gramatical	Informações morfossintáticas
Substantivo	gênero, número, grau, regência nominal
Adjetivo	gênero, número, grau, regência nominal
Artigo	gênero, número, tipo (definido ou indefinido)
Preposição	contração (indica se a preposição é simples ou combinada)
Conjunção	tipo (coordenativa, subordinativa), subtipo
Numeral	gênero, número, tipo (cardinal, ordinal, multiplicativo, fracionário)
Pronome	gênero, número, tipo, contração (caso exista)
Nomes Próprios	gênero, número
Verbo	predicação, tempo, pessoa, colocação pronomial, regência do verbo
Advérbio	tipo, grau
Prefixo	(somente informação da classe)
Sigla	(somente informação da classe)
Abreviaturas	gênero, número
Interjeição	(somente informação da classe)

Tabela 4.1: Informações morfossintáticas presentes no léxico Unitex-PB.

A estrutura do léxico do ReGra é composta de entradas constituídas por uma palavra ou, no máximo, palavras compostas hifenizadas. O exemplo abaixo ilustra duas entradas da palavra "mata", uma para cada canônica possível: [matar], verbo, e [mata], substantivo. Pode-se notar que, além das informações morfológicas, cada verbete traz informações sobre sua(s) classe(s) gramatical(is)[2].

```
mata=<V.][][PRES.ELE.]N.][][matar]0.#S.F.SI.N.[]?.3.[mata]0.>
```

Essa etapa foi parcialmente realizada. A partir do léxico do ReGra, foi proposto um modelo de conversão para o Formato DELA. Foi projetado quais campos e códigos seriam utilizados para cada classe gramátical. O Apêndice A apresenta com detalhes os campos e códigos do léxico Unitex-PB. Em seguida foi implementado um protótipo de filtro para a conversão desses dados e chegamos a resultados animadores (Tabela 4.2).

N. de Entradas	Tamanho	Tamanho Compactado	Taxa de Compactação
1.542.563	60 MB	1.59 MB	97.35 %

Tabela 4.2: Resultados da primeira versão do léxico Unitex-PB.

O filtro trata de 14 classes gramaticais (substantivo, adjetivo, artigo, preposição, conjunção, numeral, pronome, nomes próprios, verbo, advérbio, pre-

[2]Leia-se V: verbo; PRES: presente do indicativo; ELE: 3ª pessoa; N: colocação pronomial nenhuma; S: substantivo; F: feminino; SI: singular; N: grau nulo

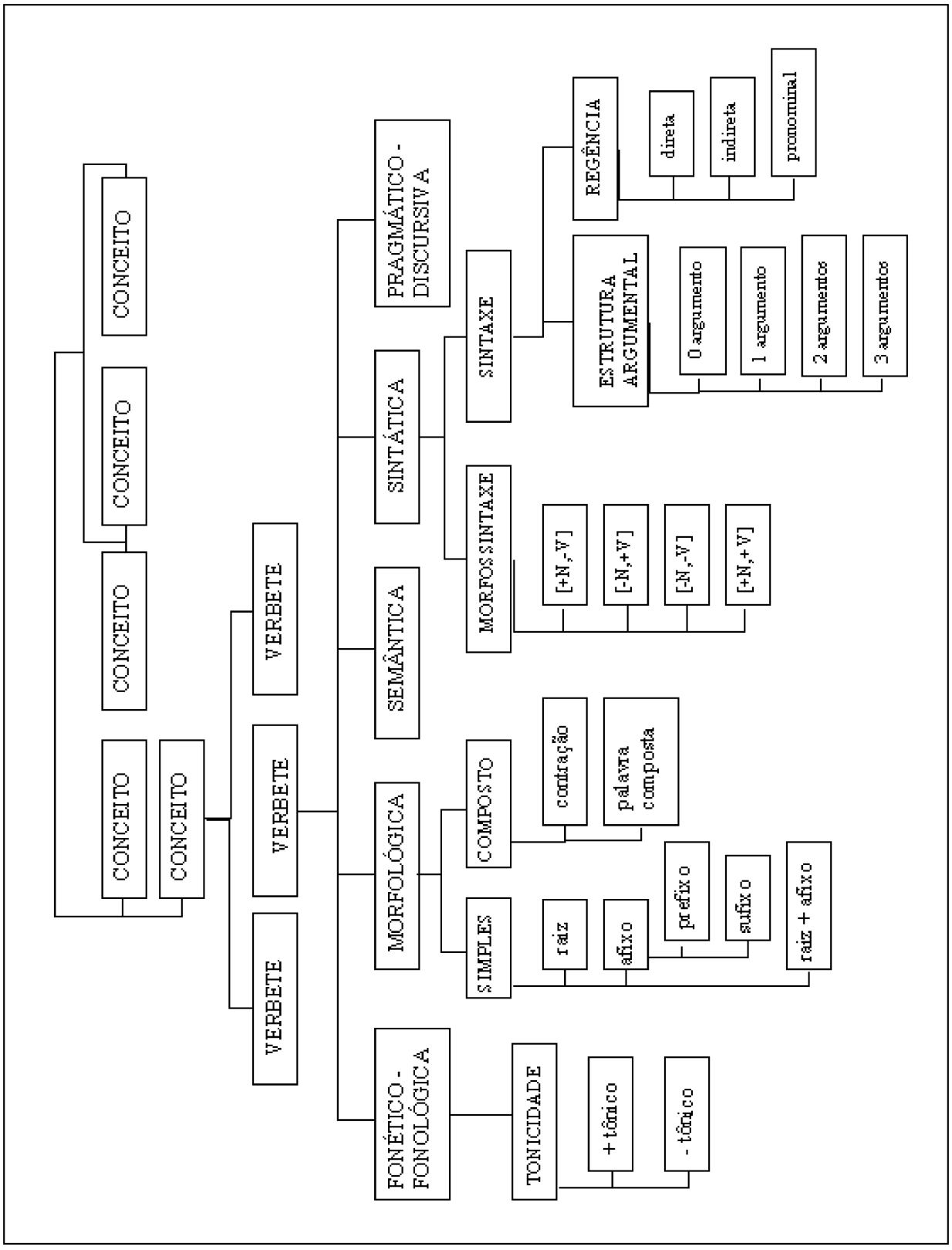


Figura 4.1: Modelo Lingüístico da Diadorim.

fixos, siglas, abreviaturas e interjeição), as mesmas que são suportadas pelo léxico do ReGra.

A primeira versão de teste do filtro, ao converter o léxico do ReGra, gerou 1.542.563 entradas em um arquivo ASCII com 60MB (120MB em formato Uni-

code). Ao compactar este dicionário, ele é convertido em dois arquivos. Um arquivo binário contém as entradas compactadas em um autômato finito, e neste teste o arquivo gerado foi de 1.04MB, e um arquivo contendo as informações gramaticais das entradas, de 560KB no teste. Desta forma, a taxa de compactação foi de aproximadamente 97.35%, considerando o arquivo no formato ASCII.

Este protótipo do filtro de conversão ainda está em testes e pretende-se que na sua versão final possua opções de gerar dicionários personalizados (por exemplo, dicionários somente de substantivos ou contendo apenas informações de flexão).

4.1.3 Projeto e implementação de expansão da Diadorim

O sistema de dicionários DELA fornece suporte não apenas a dicionários de palavras simples, mas também a dicionários de palavras compostas. Atualmente, a Diadorim possui somente entradas formadas de palavras simples. Pretende-se neste projeto expandir a Diadorim para prover suporte para palavras compostas, expandindo também desta forma o léxico Unitex-PB.

Para expandir a Diadorim, primeiramente deveremos fazer uma nova modelagem lingüística da base de dados, com especialistas, tendo como finalidade contemplar as necessidades de diferentes aplicações de PLN para palavras compostas.

Com esse modelo pronto, deve-se passar para a fase de modelagem computacional. As tabelas no banco de dados deverão ser modificadas e depois atualizadas com os novos dados.

As fontes de informações que serão utilizadas são: listas de palavras compostas que já estão inclusas no ReGra como locuções adjetivas, adverbiais, prepositivas e conjuncionais; listas de verbos compostos (Vale 2001); entre outras.

Uma vez a Diadorim expandida, deverá ser criada uma versão do léxico Unitex-PB com palavras compostas, mantendo assim um espelho da Diadorim em formato DELA.

4.1.4 Desenvolvimento de uma biblioteca de acesso e manipulação do léxico Unitex-PB

Depois que os dicionários estiverem no formato DELA, eles devem ser compactados. O Unitex possui uma ferramenta de compactação de dicionários baseada em autômatos finitos.

O Unitex tem funções de acesso e manipulação a estes dicionários compactados, porém não existe uma biblioteca específica para isso. Um dos objetivos

desse projeto é desenvolver uma biblioteca específica para acesso e manipulação a esses dicionários compactados para que qualquer outra aplicação de PLN possa utilizá-los.

Deverá ser estudado o código fonte do Unitex e depois implementada uma biblioteca de acesso e manipulação ao léxico Unitex-PB.

4.1.5 Extensão de funcionalidades do Unitex

Pretende-se, além de construir um léxico de português brasileiro para Unitex, criar novas ferramentas estendendo as funcionalidades do Unitex. Essas ferramentas serão novos recursos de PLN, ou recursos já existentes no NILC que serão portados para o ambiente Unitex. Citamos aqui três exemplos de extensões possíveis (analisador morfológico, parser e desambiguador), sem descartar futuras extensões.

Analisador Morfológico

Uma ferramenta que deverá ser implementada será um analisador morfológico. As informações de natureza morfológica ainda não existem na Diadorim, e é de interesse do nosso grupo de pesquisa possuir essas informações.

Essas informações são muito importantes, por exemplo, o fenômeno da concordância é uma dessas relações que exige a presença de certo morfema, e não outro, no interior da palavra, já que elas não são isoladas no texto. Nesse sentido, é preciso especificar os traços morfológicos pertinentes a cada item lexical, além do mecanismo de operação que esses traços exigem na concatenação das palavras na sentença.

Informações morfológicas podem ser úteis também em aplicações como na correção ortográfica (detecção de palavra desconhecida seguida de sugestão de alternativas válidas). Estudos como o de (Pelizzoni 2002) mostram que esse tipo de informação pode ser usada para (tentar) validar ou até explicar tentativas de neologia, como ao explicitar a perfeita plausibilidade/impossibilidade de "esticabilidade"/"leiturabilidade".

Para obter essas informações, pretendemos desenvolver um módulo de análise morfológica integrado ao Unitex que, dada a grafia de uma palavra, lexicalizada ou não, levante as várias hipóteses de estrutura morfológica plausíveis para a palavra em questão, incluindo aquelas meramente potenciais, mas válidas, dentro do sistema lingüístico considerado, i.e., o português. Mais especificamente, devem-se identificar não só as várias possíveis seqüências de morfemas que possam estar subjacentes a uma dada grafia, como também as várias possíveis configurações em que cada seqüência possa estar estruturada. Naturalmente, tal ferramenta deverá embutir conhecimento sobre as

regras de formação de palavras do português, bem como dispor de um inventário de morfemas. Vale notar que os morfemas de classe aberta não precisariam ser diretamente inventariados, podendo ser, em princípio, inferidos mediante consultas a um léxico de formas livres (Unitex-PB).

De posse de tal analisador, poderemos proceder à inclusão de dados estruturais às entradas da Diadorim - de forma semi-automática, vale notar, já que será necessário um humano para selecionar a hipótese de estrutura morfológica pertinente em alguns casos. No entanto, acreditamos que, em muitos casos, os traços morfossintáticos já constantes das entradas poderão prévia e automaticamente descartar diversas hipóteses de análise morfológica, inclusive a ponto de resolver a ambigüidade e dispensar o árbitro humano.

Analisador Sintático (Parser)

Está em desenvolvimento no NILC um parser robusto de português brasileiro, denominado CURUPIRA[3] (Martins et al. 2002), atualmente em fase de testes da versão 1.0. O principal objetivo do CURUPIRA é fornecer para uma dada sentença todas as suas possibilidades de análise sintática. A ferramenta não tem compromisso com a indicação da estrutura "correta", ou "mais adequada", ou mesmo "mais provável"para a sentença de entrada. Ele não é um desambiguador de estruturas sintáticas.

Ao integrar essa ferramenta ao Unitex, poderiamos utilizá-la para gerar as possíveis análises sintáticas das sentenças de um corpus carregado no Unitex. A Figura 4.2 ilustra um exemplo de análise sintática na versão 1.0 do CURUPIRA, da sentença "O português é a lingua oficial do Brasil".

Desambiguador Morfossintático

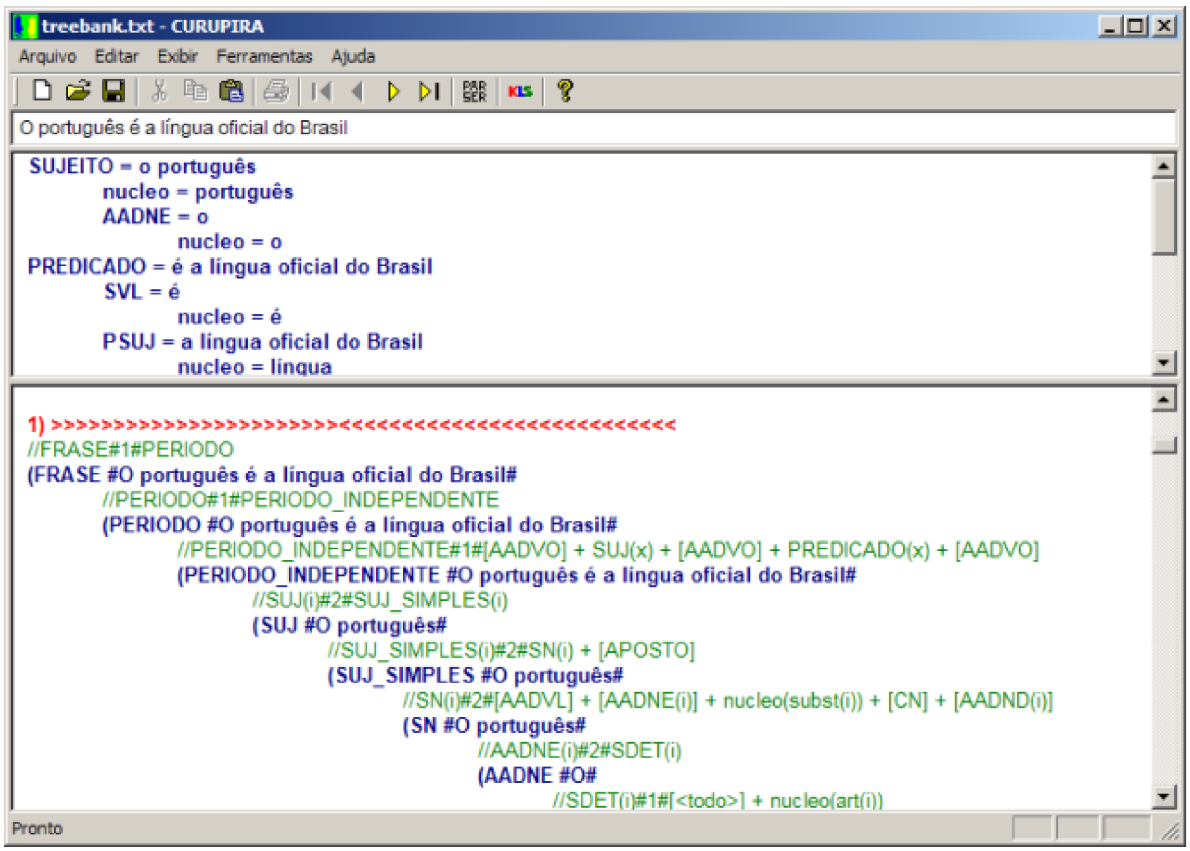
Um desambiguador morfossintático é um programa capaz de descartar etiquetas de categoria gramatical de palavras sintaticamente ambíguas, baseado no contexto em que ela aparece.

No ambiente padrão Unitex, para se remover uma ambigüidade morfossintática por homofragia, existem duas saídas: utilizar dicionários filtrados ou utilizar gramáticas locais, como visto na seção 3.2.3. Entretanto, não existem pesquisas em relação ao português brasileiro para Unitex e conseqüentemente gramáticas locais de resolução de ambigüidades .

A proposta deste trabalho é criar um desambuiguador morfossintático para o Unitex, por meio de gramáticas locais ou por meio de uma aplicação independente que seria integrada ao Unitex.

O desambiguador morfossintático será baseado no conjunto de regras de

[3]Veja `http://www.nilc.icmc.usp.br/nilc/tools/curupira.html`



```
O português é a língua oficial do Brasil
```

```
SUJEITO = o português
        nucleo = português
        AADNE = o
                nucleo = o
PREDICADO = é a lingua oficial do Brasil
        SVL = é
                nucleo = é
        PSUJ = a lingua oficial do Brasil
                nucleo = lingua
```

```
1) >>>>>>>>>>>>>>>>>>>>>>>><<<<<<<<<<<<<<<<<<<<<<<<<<<
//FRASE#1#PERIODO
(FRASE #O português é a lingua oficial do Brasil#
        //PERIODO#1#PERIODO_INDEPENDENTE
        (PERIODO #O português é a lingua oficial do Brasil#
                //PERIODO_INDEPENDENTE#1#[AADVO] + SUJ(x) + [AADVO] + PREDICADO(x) + [AADVO]
                (PERIODO_INDEPENDENTE #O português é a lingua oficial do Brasil#
                        //SUJ(i)#2#SUJ_SIMPLES(i)
                        (SUJ #O português#
                                //SUJ_SIMPLES(i)#2#SN(i) + [APOSTO]
                                (SUJ_SIMPLES #O português#
                                        //SN(i)#2#[AADVL] + [AADNE(i)] + nucleo(subst(i)) + [CN] + [AADND(i)]
                                        (SN #O português#
                                                //AADNE(i)#2#SDET(i)
                                                (AADNE #O#
                                                        //SDET(i)#1#[<todo>] + nucleo(art(i))
```

Figura 4.2: Interface gráfica do CURUPIRA.

desambigüização, presentes no corretor gramatical ReGra (Hasegawa et al. 2002).

A seguir apresenta-se um exemplo de regra para desambigüização da transitividade das ênclises e mesóclises que está presente no ReGra:

```
if Palavra = ênclise ou mesóclise and palavra = verbo transitivo
direto (independentemente de outras possibilidades) and palavra
contém {lo, la, los, las, no, na, nos, nas, me, mo, ma, mos, mas,
te, to, ta, tos, tas, se, nos, vos} then
  Retire a informação de que se trata de verbo transitivo direto
  Acrescente a informação de que se trata de verbo intransitivo
end if
```

As Figuras 4.3 e 4.4 mostram um exemplo de desambiguador ideal. Na Figura 4.3, as palavras *a*, *casa*, *bonita* estão marcadas com mais de uma etiqueta morfossintática. Um desambiguador ideal eliminaria todas as etiquetas inválidas basedadas no contexto, tendo como resultado a Figura 4.4.

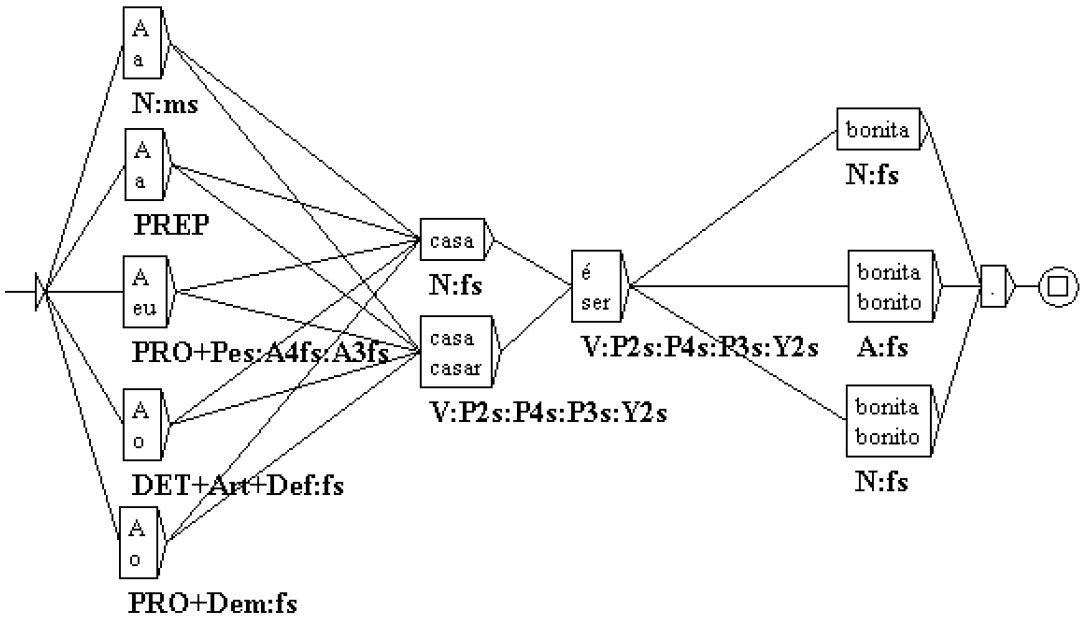


Figura 4.3: Exemplo de frase com palavras ambíguas.

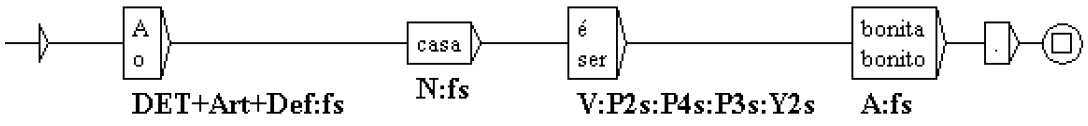


Figura 4.4: Exemplo de frase após executar um desambiguador ideal.

CAPÍTULO 5

Proposta de Tarefas e Cronograma

A seguir serão enumeradas as tarefas, já realizadas ou não, que fazem parte do desenvolvimento deste projeto. O cronograma é apresentado na Tabela 5.1.

5.1 Tarefas já realizadas

T1 - Levantamento de Requisitos

5.2 Tarefas a serem cumpridas

T2 - Modelagem e implementação do léxico unitex-PB
A modelagem e implementação do léxico unitex-PB será desenvolvida segundo a metodologia apresentada na seção 4.1.2.
Já foi desenvolvido um protótipo inicial do léxico, porém ele deve passar por um processo de aprimoramento antes de chegar a sua versão final.

T3 - Projeto e implementação de expansão da Diadorim, para que essa base de dados possa dar suporte a entradas lexicais compostas. Deverá ser criada uma versão do léxico unitex-PB para palavras compostas. Esta tarefa deverá seguir a metodologia definida na seção 4.1.3.

T4 - Desenvolvimento de uma biblioteca de acesso e manipulação do léxico Unitex-PB, utilizando a metodologia apresentada na seção 4.1.4.

T5 - Extensão de ferramentas do Unitex:
Um analisador morfológico, um parser de análise sintática e um desam-

biguador morfossintático. Esta tarefa deverá ser desenvolvida segundo a metodologia apresentada na seção 4.1.5.

A redação da dissertação deverá estar finalizada até Dezembro de 2003.

2003										
Tarefa/Mês	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
T2	•	•								
T3		•	•							
T4			•	•						
T5					•	•	•	•	•	•

Tabela 5.1: Cronograma

APÊNDICE

A

Unitex-PB

Este apêndice tem por objetivo apresentar os campos e códigos utilizados na primeira versão do léxico Unitex-PB. As entradas consideradas serão somente de palavras simples.

A.1 Estrutura das entradas:

`Palavra,canônica.Classe+traços:flexão`

Cada verbete poderá estar classificado em mais de uma classe gramatical, neste caso haverá uma entrada para cada classe.

Este dicionário é uma conversão do léxico do NILC para o formato DELA (utilizado na ferramenta Unitex).

A.2 As categorias (classes) básicas do verbete são:

A.2.1 Substantivo

Classe: `S`

Regência do Substantivo: (`p1`, ..., `pn`) uma lista (podendo ser nula) de preposições.

Gênero:

- `m`: masculino
- `f`: feminino
- `d`: dois gêneros
- `i`: invariável

Número:

s: singular
p: plural
n: dois números
j: invariável

Grau:

A: aumentativo
D: diminutivo
N: nulo

Estrutura:

`Entrada,canônica.S+Regência do Substantivo:gênero número grau`

Exemplos:

menino: `menino,menino.S+[]:msN`
meninos: `meninos,menino.S+[]:mpN`
meninão: `meninão,menino.S+[]:msA`
lápis: `lápis,lápis.S+[]:mnN`
ajuda: `ajuda,ajuda.S+[a]:fsN`

A.2.2 Adjetivo

Classe: `ADJ`
Regência do Adjetivo: (p1, ..., pn) uma lista (podendo ser nula) de preposições.
Gênero:

m: masculino
f: feminino
d: dois gêneros
i: invariável

Número:

s: singular
p: plural
n: dois números
j: invariável

Grau:

A: aumentativo
D: diminutivo
S: superlativo
N: nulo

Estrutura:

`Entrada,canônica.ADJ+Regência do Adjetivo:gênero número grau`

Exemplos:

bonito: `bonito,bonito.ADJ+[]:msN`
bonitas: `bonitas,bonito.ADJ+[]:fpN`
aprazível: `aprazível,aprazível.ADJ+[]:dsN`
simples: `simples,simples.ADJ+[]:dnN`
igual: `igual,igual.ADJ+[a]:dsN`
amabilíssimo: `amabilíssimo,amável.ADJ+[]:msS`

A.2.3 Artigo

Classe: ART
Tipo:

`Def`: Definido
`Ind`: Indefinido

Gênero:

`m`: masculino
`f`: feminino

Número:

`s`: singular
`p`: plural
`n`: dois números
`j`: invariável

Estrutura:

`Entrada,canônica.ART+Tipo:gênero número`

Exemplos:

o: `o,o.ART+Def:ms`
umas: `umas,um.ART+Ind:fp`

A.2.4 Preposição

Classe: PREP
Contração: um par da forma (preposição, palavra), ou nula.

Estrutura:

`Entrada,canônica.PREP+Contração`

Exemplos:
ante: `ante,ante.PREP`
ao: `ao,ao.PREP+C+[a.o.]`
do: `do,do.PREP+C+[de.o.]`
daqui: `daqui,daqui.PREP+C+[de.aqui.]`

A.2.5 Conjunção

Classe: `CONJ`
Tipo:
`COORD`: Coordenativa
`SUBORD`: Subordinativa
Complemento Coordenativa:
`ADIT`: Aditiva
`ADVE`: Adversativa
`ALTER`: Alternativa
`CONCL`: Conclusiva
`EXPL`: Explicativa
Complemento Subordinativa:
`INTE`: Integrante
`CAUS`: Causal
`COMP`: Comparativa
`CONC`: Concessiva
`COND`: Condicional
`CONS`: Consecutiva
`FIN`: Final
`TEMP`: Temporal
`PROPOR`: Proporcional
`CONFOR`: Conformativa

Estrutura:
`Entrada,canônica.CONJ+tipo+complemento`

Exemplos:
Mas: `mas,mas.CONJ+COORD+ADVE`
Mais: `mais,mais.CONJ+COORD+ADIT`
Mal: `mal,mal.CONJ+SUBORD+TEMP`

A.2.6 Numeral

Classe: `NUM`

Tipo:

C: cardinal
O: ordinal
M: multiplicativo
F: Fracionário
L: Coletivo

Gênero:

m: masculino
f: feminino
d: dois gêneros
i: invariável

Número:

s: singular
p: plural
n: dois números
j: invariável

Estrutura:

`Entrada,canônica.NUM:gênero número tipo`

Exemplos:

segundo: `segundo,segundo.NUM:msO`
duplo: `duplo,duplo.NUM,msM`

A.2.7 Pronome

Classe: PRON

Tipo:

TRAT: Tratamento
RET: Pessoal Reto
OBL-AT: Pessoal Oblíquo Átono
OBL-TO: Pessoal Oblíquo Tônico
POSS: Possessivo
DEM: Demonstrativo
INDE: Indefinido
INTE: Interrogativo
REL: Relativo
REFL: Reflexivo

Contração: um par da forma (preposição, palavra), ou nula

Gênero:

`m`: masculino
`f`: feminino
`d`: dois gêneros
`i`: invariável

Número:
`s`: singular
`p`: plural
`n`: dois números
`j`: invariável

Estrutura:
`Entrada,canônica.PRON+Tipo:gênero número`

Exemplos:
Senhora: `senhora,senhora.PRON+TRAT+[]:fs`
eu: `eu,eu.PRON+RET+[]:ds`
dele: `dele,dele.PRON+RET+[de.ele.]:ms`

A.2.8 Nomes Próprios

Classe: `NOM`
Gênero:
`m`: masculino
`f`: feminino
`d`: dois gêneros

Número:
`s`: singular
`p`: plural
`n`: dois números

Estrutura:
`Entrada,canônica.NOM:gênero número`

Exemplos:
Darci: `Darci,Darci.NOM:ds`
Atlântico: `Atlântico,Atlântico.NOM:ms`

A.2.9 Verbo

Classe: `V`

Predicação do Verbo: (p1, ..., pn) uma lista (podendo ser nula) de tipos de predicação.

Os tipos podem ser:

`INT`: Intransitivo

`TD`: Transitivo Direto

`TI`: Transitivo Indireto

`BI`: Bitransitivo

`LIG`: Ligação

`AUX`: Auxiliar

`PRONOM`: Pronominal

colocação pronomial:

`ENC`: ênclise

`MESOC`: mesóclise

`N`: nulo

Regência do Verbo: (p1, ..., pn) uma lista (podendo ser nula) de preposições.

Tempo:

`P`: Presente / PRES

`I`: Pretérito Imperfeito / PRET-IMPERF

`J`: Pretérito Perfeito / PRET-PERF

`M`: Pretérito Mais-que-Perfeito / PRET-M-Q-P

`F`: Futuro do Presente / FUT-PRES

`R`: Futuro do Pretérito /FUT-PRET

`S`: Presente do Subjuntivo / PRES-SUBJ

`T`: Pretérito Imperfeito do Subjuntivo / PRET-IMPERF-SUBJ

`U`: Futuro do Subjuntivo / FUT-SUBJ

`Y`: Imperativo Afirmativo / IMPER-AFIRM

`V`: Infinitivo Pessoal / INF-PESS

`G`: Gerúndio / GERUN

`K`: Particípio / PARTIC

Pessoa:

`1s`: eu

`2s`: tu

`3s`: ele

`1p`: nós

`2p`: vós

`3p`: eles

Obs: A primeira e a terceira pessoa do Infinitivo Pessoal coincidem com o Infinitvo Impessoal. Assim, não colocamos a forma nominal Infinitivo Impessoal como complemento do verbo. No caso de Particípio, são necessárias as informações de gênero - `M/F/2G` e número - `S/P/2N` e no caso de Infinitivo Pessoal,

o verbo possui Pessoa - eu, tu, ele, nós, vós, eles.

Estrutura:

`Entrada,canônica.V+Predicação do Verbo+Colocação do Pronome+Regência do Verbo:tempo pessoa`

Exemplos:
meditar: `meditar,meditar.V+[INT.TD.TI.]+N+[em.sobre.]:U3s:U1s:V3s:V1s`
meditaram-no: `meditaram-no,meditar.V+[]+ENC+[]:M3p:J3p`

A.2.10 Advérbio

Classe: `ADV`
Tipo:
`CIR-LUG`: Circunstância Lugar
`CIR-TEMP`: Circunstância Tempo
`CIR-MOD`: Circunstância Modo
`NEG`: Negação
`DUV`: Dúvida
`INT`: Intensidade
`AFIR`: Afirmação
`INT-LUG`: Interrogativo de Lugar
`INT-TEMP`: Interrogativo de Tempo
`INT-MOD`: Interrogativo de Modo
`INT-CAUS`: Interrogativo de Causa
`LAT`: Latinas
`ESTRA`: Estrangeirismos
`(podendo ser nulo)`

Grau:
`A`: Aumentativo
`D`: Diminutivo
`S`: Superlativo
`N`: Nulo
Estrutura:
`Entrada,canônica.ADV+Tipo:Grau`

Exemplos:
abaixo: `abaixo,abaixo.ADV+CIR-LUG:N`
misericordiosissimamente:

misericordiosissimamente,misericordiosissimamente.ADV+CIR-MOD:S
mesmo: mesmo,mesmo.ADV:N

A.2.11 Prefixos

Classe: PREF

Estrutura:
Entrada,canônica.PREF

Exemplos:
super: super,super.PREF
pós: pós,pós.PREF
sub: sub,sub.PREF

A.2.12 Siglas

Classe: SIGL

Estrutura:
Entrada,canônica.SIGL

Exemplos:
ONU: ONU,ONU.SIGL
PDT: PDT,PDT.SIGL
OTAN: OTAN,OTAN.SIGL
USP: USP,USP.SIGL

A.2.13 Abreviaturas

Classe: ABREV
Gênero:
m: masculino
f: feminino
d: dois gêneros

Número:
s: singular
p: plural
n: dois números

Estrutura:

`Entrada,canônica.ABREV:gênero número`

Exemplos:

ml: `ml,ml.ABREV:ms`

mm: `mm,mm.ABREV:ms`

A.2.14 Interjeição

Classe: `INTERJ`

Estrutura:

`Entrada,canônica.INTERJ`

Exemplos:

Ah: `Ah,Ah.INTERJ`

Ih: `Ih,Ih.INTERJ`

Olá: `Olá,Olá.INTERJ`

Oi: `Oi,Oi.INTERJ`

A.3 Exemplo de entrada do Léxico formato do Regra e no formato Unitex:

Formato do Regra

```
a=<ART.F.SI.DE.?.?.[o]O.#PREP.[a]O.#PRON.F.SI.[DEM.OBL-
AT.]3S.?.?.C.[][o]O.#ABREV.M.SI.[a]O.#S.M.SI.N.[]?.?.[a]O.>
```

Formato do Unitex

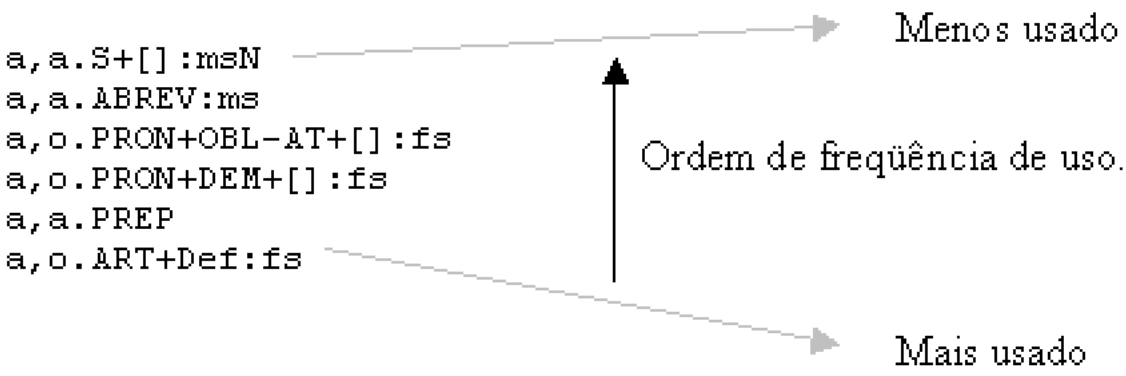


Figura A.1: Exemplo de entrada do Léxico formato do Regra e no formato Unitex.

Referências

Anastasiadis-Symeonidis, A., T. Kyriacopoulou, E. Sklavounou, I. Thilikos, & R. Voskaki (2000). A system for analysing texts in modern greek: representing and solving ambiguities. In *Proceedings of COMLEX 2000. Computational Lexicography and Multimedia Dictionaries*. Dept. of Electrical and Computer Engineering, Univ. of Patras, Greece. `http://citeseer.nj.nec.com/457067.html` (14/10/2002).

Briscoe, T. (1991). Lexical issues in natural language processing. *Natural Language and Speech*, 39–68.

Calzolari, N. (1990). Structure and access in a automated lexicon and related issues. In *Linguistica Computazionale vol. II - Computational Lexicology and Lexicography: Special Issue dedicated to Bernard Quemada*, pp. 139–161. Pisa, Giardini Editori e Stampatori.

Correia, M. (1994). Bases digitais lexicais na união europeia. In *Simpósio de Lexicologia, Lexicografia e Terminologia*. Araraquara, SP.

Correia, M. (1996). Terminologia e lexicografia computacional. In *CAMBRÉ, MT. Cicle de conferéncies*, pp. 83–91. Barcelona: Institut Universitari de Lingüística Aplicada.

Courtois, B. (1990). Un système de dictionnaires électroniques pour les mots simples du français, langue française. *Dictionnaires électroniques du français 87*, 11–22.

da Silva, B. C. D., M. F. de Oliveira, & H. R. de Moraes (2002). Groundwork for the development of the brazilian portuguese wordnet. In *PorTAL, Advances in Natural Language Processing, Third International Conference, PorTAL 2002*, Volume 2389 of *Lecture Notes in Computer Science*, pp. 189–196. Faro, Portugal: Springer.

da Silva, B. C. D., M. F. Oliveira, H. R. Moraes, R. Hasegawa, D. Amorim, C. Paschoalino, & A. C. Nascimento (2000). A construção de um thesaurus eletrônico para o português do brasil. In *V Encontro para o pro-*

cessamento computacional da Língua Portuguesa Escrita e Falada (PROPOR'2000). Atibaia, SP.

EAGLES (1993). Eagles lexicon architeture - eagles document eag-clwg-lexarch/b.

EAGLES (1996). Eagles computational lexicons working group reading guide - eagles document eag-clwg-fr-2.

Evans, R. & A. Kilgarriff (1995). Mrds, standards and how to do lexical engeneering. Technical Report ITRI-95-19, University of Brighton.

Greghi, J. G. (2002). Uma base de dados lexicais para o português do brasil. Dissertação de Mestrado, ICMC-USP, São Carlos, SP.

Greghi, J. G., R. T. Martins, & M. das Gracas Volpe Nunes (2002). Diadorim: a lexical database for brazilian portuguese. In *Proceedings of the Third International Conference on language Resources and Evaluation. LREC2002.*, Volume IV, pp. 1346–1350. Las Palmas, Ilhas Canárias.

Hasegawa, R., R. T. Martins, & M. Nunes (2002). Regra 2002: Características e desempenho. Technical Report NILC-TR-02-8, ICMC-USP.

Ide, N. & J. Véronis (1992). Modeling lexical databases. In *International Conference ALLC-ACH'92*. Oxford.

Kowaltowski, T. & C. L. Lucchesi (1993). Applications of finite automata representing large vocabularies. *Software-Pratice and Experience 23*(1), 15–20.

Kowaltowski, T., C. L. Lucchesi, & J. Stolfi (1995a). application of finite automata in debugging natural language vocabularies. *Journal of the Brazilian Computing Society 3*(1), 5–11.

Kowaltowski, T., C. L. Lucchesi, & J. Stolfi (1995b). Minimization on binary automata. *Journal of the Brazilian Computing Society 3*(1), 36–42.

Martins, R. T., R. Hasegawa, & M. Nunes (2002). Curupira: um parser functional para o português. Technical Report NILC-TR-02-26, ICMC-USP.

Martins, T. B. F., R. Hasegawa, M. G. V. Nunes, & O. N. O. Jr. (1998). Linguistic issues in the develepment of regra: a grammar checker for brazilian portuguese. *Natural Language Engineering 4*(4), 287–307.

Miller, G. A., R. Backwith, C. Fellbaum, D. Gross, & K. Miller (1990). Introduction to wordnet: An on-line lexical database. *Journal of Lexicography 3*(4), 234–244.

Nunes, M. G. V., B. C. D. da Silva, L. H. M. Rino, O. N. O. Jr., R. T. Martins, & G. Montilha (1999). Introdução ao processamento de línguas naturais. Technical Report ND-38, ICMC-USP, São Carlos, SP. 91p.

Nunes, M. G. V., F. M. C. Vieira, C. Zavaglia, C. R. C. Sossolote, & J. Hernandez (1996). A construção de um léxico de português do brasil: Lições aprendidas e perspectivas. In *Anais do II Workshop de Processamento Computacional de Português Escrito e Falado (PROPOR'96)*, pp. 61–70. CEFET-PR, Curitiba.

Ortiz, A. M. (2000). Diseño e implementación de um lexicón computacional para lexicografía y traducción automática.

Pelizzoni, J. M. (2002). Preâmbulo ao aconselhamento ortográfico para o português do brasil - uma releitura baseada em utilidade e conhecimento lingüístico. Dissertação de Mestrado, ICMC-USP, São Carlos, SP.

Pustejovsky, J. (2001). Very large lexical databases. In *ACL (Companion Volume)*. `http://citeseer.nj.nec.com/pustejovsky01very.html` (11/02/2003).

Ranchhod, E., C. Mota, & J. Baptista (1999). A computational lexicon of portuguese for automatic text parsing. In *Proceedings of SIGLEX99: Standardizing Lexical Resources, 37th Annual Meeting of the ACL*, pp. 74–80. College Park, Maryland, USA.

Ranchhod, E. & D. Santos (1999). Ambientes de processamento de corpora em português: Comparação entre dois sistemas. In *Acta IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada*, pp. 257–268. Universidade de Évora.

Ranchhod, E. M. (2001). *Tratamento das Línguas por Computador. Uma Introdução à Linguística Computacional e suas Aplicações*, Chapter O uso de dicionários e de autómatos finitos na representação lexical das línguas naturais, pp. 13–47. Lisboa: Caminho.

Silberztein, M. (1990). Le dictionnaire électronique des mots composés, langue française. *Dictionnaires électroniques du français 87*, 71–83.

Silberztein, M. (2000a). Intex: a fst toolbox. *Theoretical Computer Sciences 231*, 33–46.

Silberztein, M. (2000b). *INTEX Manuel*. LADL, Université Paris 7. `http://www.nyu.edu/pages/linguistics/intex/downloads/Manuel.pdf` (14/10/2002).

Tiberius, C. (1999). Language sampling for multilingual lexical representation. Technical Report ITRI-99-12, University of Brighton.

Vale, O. A. (2001). *Expressões Cristalizadas do Português do Brasil: uma proposta de tipologia*. Tese de Doutorado, Unesp, Araraquara, SP.

Vietri, S. & A. Elia (2000). Electronic dictionaries and linguistic analysis of italian large corpora. In *JADT 2000 : 5es Journées Internationales d'Analyse Statistique des Données Textuelles*.

Vitas, D. & C. Krstev (2001). Intex and slavonic morphology. In *Proceedings of the 4th Intex workshop*. Bordeaux, France.

Wilks, Y., D. Fass, C.-M. Guo, J. McDonald, T. Plate, & B. Slator (1988). Machine tractable dictionaries as tools and resources for natural language processing. In *Proceedings of Colling'88*, pp. 750–755. Budapest.