



Pontifícia Universidade Católica do Rio Grande do Sul  
Faculdade de Informática  
Programa de Pós-Graduação em Ciência da Computação



## **$E\chi$ ATO<sub>LP</sub> – Extrator Automático de Termos para Ontologias em Língua Portuguesa**

Lucelene Lopes, Renata Vieira

**Relatório Técnico N<sup>o</sup> 054**

Porto Alegre, Agosto de 2009

## Resumo

Este trabalho apresenta a ferramenta  $E\chi ATOLP$  desenvolvida para extrair sintagmas nominais de *corpus* anotados linguisticamente. Os sintagmas nominais extraídos são os candidatos a conceito de uma ontologia. Este relatório técnico apresenta aspectos gerais, de implementação, interface com usuário e aplicação desta ferramenta. Finalmente, apresenta-se também uma breve comparação da aplicação do  $E\chi ATOLP$  com outras ferramentas de função análoga.

# Lista de Figuras

2.1	Exemplo de Anotação feita pelo <i>parser</i> PALAVRAS . . . . .	11
3.1	Representação Gráfica do Exemplo da Figura 3.2 . . . . .	28
3.2	Exemplo do Formato TIGER-XML . . . . .	29
3.3	Exemplo de Lista de Termos nas Formas Originais e Canônica	30
3.4	Exemplo de Tabela de Sintagmas Nominais . . . . .	32
4.1	Exemplo de Sintagmas Nominais . . . . .	44
4.2	Exemplo de Texto Considerado como <i>corpus</i> . . . . .	45
4.3	Sintagmas Nominais Extraídos do Texto da Figura 4.2 . . . . .	46
4.4	Exemplo de Lista de Termos a Comparar . . . . .	49
4.5	Exemplo de Resultados de Comparação . . . . .	49
4.6	Exemplo de Relatório de Localização . . . . .	50
5.1	Tela de Seleção de Sintagmas pelo Número de Palavras . . . . .	52
5.2	Tela de Escolha de Ponto de Corte e Limpeza . . . . .	53
5.3	Tela de Comparação e Cálculo de Métricas . . . . .	54
5.4	Exemplo de Saída com Cálculo de Métricas . . . . .	55
5.5	Tela de Localização de Sintagma . . . . .	56
5.6	Tela de Escolha de Ponto de Corte de Lista Individual . . . . .	57
5.7	Tela de Comparação de Listas Individuais . . . . .	57

# Lista de Tabelas

2.1	Adjetivos com variação morfológica de grau sintético . . . . .	13
2.2	Formas verbais e etiquetas atribuídas pelo PALAVRAS . . . . .	17
3.1	Caracteres Aceitos em Listas de Entrada . . . . .	31
6.1	Sintagmas Extraídos pela Ferramenta $E\chi ATOLP$ . . . . .	59
6.2	Comparação de Métricas do $E\chi ATOLP$ com Outras Ferramentas	60

# Lista de Pseudo-Códigos

4.1.1 Extração de Sintagma . . . . .	35
4.1.2 Armazena Terminais . . . . .	35
4.1.3 Armazena Não-Terminais . . . . .	36
4.1.4 Busca Sintagmas Nominais . . . . .	37
4.1.5 Função Recursiva Monta( <i>nt</i> , <i>sintag</i> ) . . . . .	38
4.1.6 Função Válido ( <i>sintag</i> ) . . . . .	39
4.1.7 Função Aprimora ( <i>sintag</i> ) . . . . .	41
4.1.8 Função Armazena ( <i>sintag</i> ) . . . . .	43
4.1.9 Cálculo das Frequências Absoluta e Relativa . . . . .	43

# Sumário

<b>1</b>	<b>Introdução</b>	<b>6</b>
<b>2</b>	<b>Conceitos Linguísticos Básicos</b>	<b>10</b>
2.1	Categorias Gramaticais dos Terminais . . . . .	10
2.1.1	Substantivos . . . . .	12
2.1.2	Adjetivos . . . . .	13
2.1.3	Pronomes . . . . .	14
2.1.4	Numerais . . . . .	14
2.1.5	Artigos . . . . .	15
2.1.6	Verbos . . . . .	16
2.1.7	Advérbios . . . . .	18
2.1.8	Preposições . . . . .	19
2.1.9	Conjunções . . . . .	19
2.1.10	Interjeições . . . . .	20
2.2	Classificação dos Não-Terminais . . . . .	20
2.2.1	Classificação de Sintagmas . . . . .	21
<b>3</b>	<b>Visão Geral da Ferramenta E<sub>χ</sub>ATOLP</b>	<b>23</b>
3.1	Funcionalidades . . . . .	23
3.1.1	Extração de Sintagmas Nominais . . . . .	24
3.1.2	Aplicação de Pontos de Corte . . . . .	25
3.1.3	Comparação de Listas de Termos . . . . .	26
3.1.4	Cálculo de Métricas . . . . .	27
3.2	Formatos de Entrada . . . . .	27
3.2.1	<i>Corpus</i> – Texto Anotado . . . . .	27
3.2.2	Listas de Termos . . . . .	30
3.3	Formatos de Saída . . . . .	31

<b>4</b>	<b>Implementação</b>	<b>34</b>
4.1	Módulo Básico da Ferramenta – Módulo Extrator . . . . .	34
4.1.1	Procedimento Geral . . . . .	34
4.1.2	Primeiro Passo – Armazena Terminais . . . . .	35
4.1.3	Segundo Passo – Armazena Não-terminais . . . . .	36
4.1.4	Terceiro Passo – Busca Sintagmas Nominais . . . . .	37
4.1.5	Montagem de Sintagmas – Função Monta . . . . .	38
4.1.6	Descarte de Sintagmas – Função Válido . . . . .	39
4.1.7	Alteração de Sintagmas – Função Aprimora . . . . .	41
4.1.8	Finalização de Sintagma – Função Armazena . . . . .	42
4.1.9	Cálculo das Frequências dos Sintagmas . . . . .	43
4.1.10	Exemplo de Aplicação . . . . .	44
4.2	Módulos Acessórios da Ferramenta . . . . .	47
4.2.1	Módulo Cortador . . . . .	47
4.2.2	Módulo Comparador . . . . .	48
4.2.3	Módulo Localizador . . . . .	50
<b>5</b>	<b>Interface com o Usuário</b>	<b>51</b>
5.1	Extração de Termos . . . . .	51
5.1.1	Seleção do Número de Palavras . . . . .	52
5.1.2	Aplicação de Pontos de Corte e Limpeza . . . . .	52
5.1.3	Comparação com Referência e Métricas . . . . .	54
5.2	Operações Acessórias . . . . .	55
5.2.1	Localização de Termos . . . . .	56
5.2.2	Aplicação de Pontos de Corte . . . . .	56
5.2.3	Comparação de Listas . . . . .	57
<b>6</b>	<b>Exemplos de Utilização</b>	<b>58</b>
6.1	Resultados Numéricos . . . . .	59
<b>7</b>	<b>Conclusão</b>	<b>61</b>

# Capítulo 1

## Introdução

É clara a importância e a dificuldade da construção de ontologias para a estruturação, organização e disseminação de um conhecimento específico. Dentre as formas de construir ontologias, a construção a partir de textos é aquela que mais se presta a uma automatização e a tarefa de extração de termos é o ponto de partida para este processo [16]. Além disso, trata-se de uma etapa fundamental, pois, dela depende o sucesso de todas as demais etapas, uma vez que os termos extraídos devem ser a representação conceitual do domínio alvo.

Via de regra, os processos de extração automática de termos baseiam-se na análise de um conjunto de textos (*corpus*) do domínio de interesse [9]. A abordagem de extração automática utilizada neste relatório técnico situa-se neste campo de pesquisa.

É um consenso da área de processamento de linguagem natural que os métodos de extração de termos podem ser agrupados, segundo a abordagem utilizada em:

- Abordagens estatísticas - os documentos contidos no *corpus* são vistos como um conjunto de termos e são medidas suas frequências de ocorrência, medidas de proximidade e outras informações puramente numéricas;
- Abordagens linguísticas - os textos são anotados com informações linguísticas: morfológicas, sintáticas e/ou semânticas;
- Abordagens Híbridas - nestes casos são utilizadas técnicas comuns às duas abordagens citadas (estatística e linguística).



No entanto, esta divisão raramente é estanque, pois praticamente todos os métodos sempre têm ao menos algum componente de cada uma das abordagens. Métodos baseados em informações linguísticas sempre levam em consideração algum critério de frequência, assim como métodos baseados em informações estatísticas usualmente consideram algumas listas de palavras que seguem critérios linguísticos (*stoplist*). Desta forma, a quase totalidade dos métodos poderiam ser vistos como híbridos, porém para fins de classificação a área denomina de métodos linguísticos aqueles que tem a maior parte das decisões baseadas neste tipo de informação e, analogamente, denomina-se métodos estatísticos aqueles em que não se considera explicitamente informações linguísticas.

Um exemplo de abordagem puramente estatística é o trabalho de Baroni e Bernadini [2], o qual utiliza de forma randômica algoritmos para extrair termos automaticamente da Web. Uma particularidade deste trabalho é que, ao contrário dos demais citados a seguir, a própria construção do *corpus* faz parte do processo. Especificamente, um conjunto inicial de termos é usado para fazer busca no Google, construindo o *corpus* que será utilizado para extração de termos. O processo de extração de termos é relativamente simples e consiste em buscar termos que frequentemente aparecem seguidos ou precedidos de conectores (por exemplo: *de, do, da, etc*). Em seguida, uma lista de *stop words* é construída, ou seja, uma lista de palavras irrelevantes que aparecem com muita frequência no texto, mas não são conectores. A busca de termos compostos é baseada em heurísticas, como por exemplo, considerar somente termos que estão acima de um limiar de frequência e não considerar termos que começam ou terminam por conectores.

Bourigault *et al.* [8] utiliza a abordagem linguística apresentando uma ferramenta denominada analisador de *corpus* SYNTEX para extração de termos em *corpus* de língua francesa. A extração de termos é feita através dos sintagmas nominais, levando em consideração as categorias morfossintáticas e as principais relações sintáticas como por exemplo, sujeito, objeto direto e complemento proposicional (de nome, de verbo e de adjetivo). Para o desenvolvimento do SYNTEX não se utiliza um léxico rico, mas sim textos anotados por um *parser* e de acordo com o *corpus* a ser tratado, um léxico específico do domínio é construído ao mesmo tempo que é feita a análise sintática. Segundo Bourigault *et al.*, esta técnica permite uma melhor adequação aos *corpora*, pois estes tem particularidades do domínio que são específicas e imprevisíveis [8].

Como sequência da sua abordagem linguística em SYNTEX, Bourigault

parte em outro trabalho para uma abordagem híbrida com o módulo UPERY [7], uma ferramenta que faz análise distribucional dos termos extraídos linguisticamente pelo SYNTAX. Partindo de um contexto sintático, os termos são analisados e se constrói uma rede de palavras a partir de cada frase do *corpus*. Após a construção da rede, parte-se para a análise distribucional que é feita através de medidas de dependência estatística entre cada termo, ou seja, é calculada a distância (proximidade) entre as frases e termos da rede. Esse cálculo é baseado sempre no contexto sintático, mas todas as definições para considerar ou não termos é baseada em medidas numéricas calculadas a partir de informações estatísticas.

Outros exemplos mais recentes de extração híbrida de termos são os trabalhos de Aubin e Hamon [1] e Fortuna, Lavrac e Velardi [11]. Ambos trabalhos descrevem experiências com extração de termos a partir de textos (*corpus*), com o propósito de construir ontologias (hierarquia de conceitos). Enquanto Aubin e Hamon [1] utilizam-se de uma ferramenta específica (YATEA) para extrair os termos, Fortuna, Lavrac e Velardi [11] utilizam-se de um ambiente integrado, denominado, OntoGen para o processo de extração de termos e determinação de hierarquia de conceitos.

O trabalho desenvolvido neste relatório técnico é a proposta para extração de termos com relevância conceitual a partir de um corpus anotado. A ferramenta proposta chama-se  $E\chi ATOLP$  – Extrator Automático de Termos para Ontologias em Língua Portuguesa. O método utilizado para a ferramenta  $E\chi ATOLP$  é baseado em informações linguísticas e propõe uma ferramenta automática que parte de um *corpus* anotado sintaticamente e extrai os termos utilizando uma análise baseada na busca dos sintagmas nominais mais frequentes.

Neste sentido, este método é semelhante ao trabalho de Bourigault *et al.* [8] que também extrai sintagmas nominais levando em consideração as categorias morfossintáticas e as principais relações sintáticas como por exemplo, sujeito, objeto direto e complemento proposicional (de nome, de verbo e de adjetivo). No entanto, o trabalho de Bourigault e seus colaboradores está baseado em uma ferramenta desenvolvida para a extração de termos sobre um *corpus* composto de textos em língua francesa.

Uma ferramenta semelhante ao  $E\chi ATOLP$  é a ferramenta OntoLP [21] que implementa entre outras possibilidades de extração uma abordagem igualmente baseada em informações linguísticas. OntoLP na verdade é um plug-in para o editor de ontologias Protégé [12], um editor bastante utilizado na comunidade científica dando suporte à construção de ontologias, seguindo

as tecnologias da Web Semântica, como por exemplo, a construção de ontologias OWL Web Ontology Language, conforme o padrão definido pelo World Wide Web Consortium (W3C) [19].

Outra ferramenta que executa extração de termos chama-se NSP – *Ngrams Statistic Package* [3]. Esta ferramenta é um conjunto de programas escritos na linguagem Perl desenvolvido para identificar e extrair  $n$ -gramas, uma sequência contínua de palavras (tokens). Atualmente na versão 1.09, o NSP ([www.d.umn.edu/~tpederse/nsp.html](http://www.d.umn.edu/~tpederse/nsp.html)) é utilizado principalmente para a extração e análise de  $n$ -gramas a partir de textos ou *corpus* textuais. No entanto, ao contrário do OntoLP e  $E\chi$ ATO $LP$ , o NSP está baseado em uma abordagem puramente estatística que não utiliza informações linguísticas.

O texto deste relatório inicia pela definição de uma série de conceitos linguísticos básicos (capítulo 2) importantes para a compreensão dos demais capítulos. A contribuição central deste trabalho inicia no capítulo 3 que descreve uma visão geral da ferramenta  $E\chi$ ATO $LP$ . A seguir, detalhes de implementação da ferramenta são vistos no capítulo 4. Noções genéricas da interface com o usuário atual são vistas no capítulo 5. O capítulo 6 apresenta um experimento de extração de termos utilizando a ferramenta  $E\chi$ ATO $LP$  para um *corpus* específico da área de Pediatria e faz uma breve comparação da ferramenta  $E\chi$ ATO $LP$  com ferramentas que tem o mesmo propósito. Finalmente, a conclusão sumariza a contribuição do trabalho desenvolvido e sugere trabalhos futuros.

## Capítulo 2

# Conceitos Linguísticos Básicos

Este capítulo descreve brevemente alguns conceitos linguísticos básicos úteis para a compreensão de algumas decisões tomadas na construção da ferramenta  $E\chi ATOLP$ . Junto com estes conceitos é apresentada a forma como estes são reconhecidos no *parser* PALAVRAS [5] que faz a anotação linguística dos *corpora* para a ferramenta  $E\chi ATOLP$ .

Dentro do escopo deste relatório é importante saber que o reconhecimento dos conceitos é feito sobre um conjunto de palavras que compõem uma frase. Cada frase reconhecida é armazenada pelo *parser* como uma estrutura em árvore composta por nós terminais (as folhas da árvore) que representam as palavras e nós não-terminais que representam estruturas gramaticais. No contexto deste capítulo, vamos nos referenciar a um exemplo anotado pelo PALAVRAS apresentado na figura 2.1, onde está representada a anotação linguística realizada pelo *parser* para a frase “*Estas duas cidades são os maiores e mais importantes centros de pesquisa no Brasil.*”.

As demais seções deste capítulo descrevem as categorias gramaticais dos termos identificados como terminais (seção 2.1) e as estruturas mais complexas identificadas como não-terminais (seção 2.2).

### 2.1 Categorias Gramaticais dos Terminais

Um fato importante a observar na figura 2.1 é que cada um dos termos reconhecidos pelo PALAVRAS são identificados como um único terminal. Caso o termo seja composto de mais do que uma palavra, o *parser* armazena este termo colocando um caracter sublinhado (“\_”) entre as palavras.

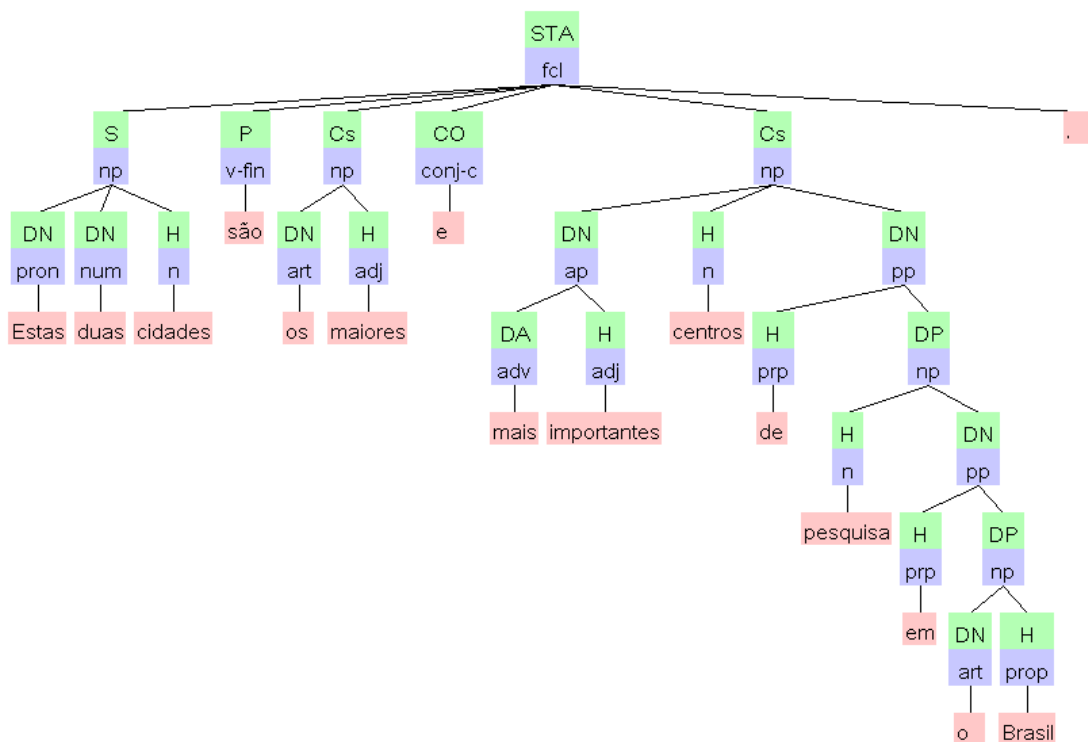


Figura 2.1: Exemplo de Anotação feita pelo *parser* PALAVRAS

Por exemplo, o termo “Porto Alegre” é armazenado pelo PALAVRAS como “Porto\_Alegre”, pois se trata de um nome próprio. Igualmente, o termo “duzentos e oitenta e quatro” que representa um único número é armazenado como “duzentos\_e\_oitenta\_e\_quatro”. Neste sentido, esta seção trata da definição das categorias gramaticais dos termos identificados como terminais pelo PALAVRAS, independente do número de palavras que o compõem.

De um ponto de vista linguístico [4], os termos em língua portuguesa são classificadas em: substantivos (seção 2.1.1); adjetivos (seção 2.1.2); pronomes (seção 2.1.3); numerais (seção 2.1.4); artigos (seção 2.1.5); verbos (seção 2.1.6); advérbios (seção 2.1.7); preposições (seção 2.1.8); conjunções (seção 2.1.9); e interjeições (seção 2.1.10).

### 2.1.1 Substantivos

Substantivo é o nome com que designamos entidades em geral, sejam seres ou coisas. Os substantivos podem ser classificados em substantivos comuns e substantivos próprios.

O substantivo comum designa a entidade como pertencente a uma classe com o mesmo conjunto de qualidades. São exemplos de substantivos comuns as palavras “cidades”, “centros” e “pesquisa”. O substantivo próprio é o que designa individualmente as entidades, sem referência a suas qualidades, ou seja, é o nome próprio de uma entidade específica. É um exemplo de substantivo próprio a palavra “Brasil” anotada na figura 2.1.

No *parser* PALAVRAS os substantivos são anotados com as etiquetas sintáticas “n” e “prop” caso sejam substantivos comuns ou próprios, respectivamente.

#### Variações Morfológicas dos Substantivos

Os substantivos comuns podem ter variações morfológicas de gênero, número e grau.

As variações de gênero se aplicam a substantivos biformes que possuem formas originárias do mesmo radical para masculino e feminino, por exemplo “menino” e “menina”, ou para substantivos heterônimos que possuem formas distintas, por exemplo “bode” e “cabra”. Os substantivos comuns na frase exemplo não se enquadram nesta categorias, pois eles são uniformes, ou seja, possuem sempre a mesma forma em relação ao gênero.

As variações de número, plural e singular, aparecem neste exemplo nos termos “cidades” e “centros” que estão no plural, enquanto “pesquisa” está no singular.

As variações de grau também não aparecem na frase exemplo, pois todos os substantivos comuns estão em sua forma normal de grau. As demais formas de grau são aumentativo, por exemplo “centrões”, e diminutivo, por exemplo, “cidadezinha”.

No PALAVRAS os substantivos são salvos tanto em sua forma normal, quanto em uma forma canônica que ignora as variações morfológicas e salva os substantivos comuns de gênero no masculino (quando não for uniforme), de número no singular e de grau em normal. Adicionalmente, o *parser* também salva para cada termo uma etiqueta morfológica com as variações morfológicas do termo na forma original.

## 2.1.2 Adjetivos

Adjetivo é a expressão modificadora que denota qualidade, condição ou estado de ser. Usualmente os adjetivos são utilizados junto com substantivos, mas eventualmente podem ser encontrados também com pronomes que substituem um substantivo. São exemplos de adjetivos as palavras “maiores” e “importantes” encontradas no exemplo da figura 2.1.

A etiqueta sintática que o *parser* atribui para os adjetivos é “adj”.

### Variações Morfológicas dos Adjetivos

A exemplo dos substantivos, os adjetivos também tem variações morfológicas de gênero, número e grau. As variações de gênero e número concordam com as variações do substantivo e seguem o mesmo padrão. A variação de grau, no entanto, é mais sutil, pois implica em mudanças de superlativos que nem sempre são representadas por flexões.

A variação de superlativo mais simples é aquela feita por flexão que é usualmente aplicada a grande maioria dos adjetivos. Este é o caso, por exemplo, do adjetivo “importantes” pode ser flexionado para “importantíssimos”. No entanto, alguns adjetivos não admitem superlativos, como é o caso de “maiores” na frase exemplo, que representa a mesma idéia do adjetivo “grande”, porém faz a comparação entre diversas entidades. Especificamente, existem na língua portuguesa apenas quatro adjetivos básicos que possuem esta característica de ter uma forma canônica e duas outras formas para fazer comparação e superlativo sintético. A tabela 2.1 apresenta estes adjetivos.

Tabela 2.1: Adjetivos com variação morfológica de grau sintético

Forma canônica	Comparativo	Superlativo
pequeno	menor	mínimo
grande	maior	máximo
mau	pior	péssimo
bom	melhor	ótimo

No PALAVRAS os adjetivos são salvos na forma original e também na forma canônica. Logo, o adjetivo “maiores” é salvo tanto na sua forma original, quanto na sua forma canônica: “grande”.

### 2.1.3 Pronomes

Pronome é uma expressão que designa um entidade (um substantivo) sem dar-lhes nome nem qualidade, indicando-os apenas como pessoa do discurso. Na frase exemplo da figura 2.1 a palavra “Estas” que inicia a frase é um exemplo de pronome.

Os pronomes podem ser classificados quanto a sua utilização em pronomes substantivos quando fazem referência a um substantivo ausente na frase, ou em pronomes adjetivos quando fazem referência a um substantivo presente.

Quanto a sua característica, os pronomes podem ser classificados em pronomes pessoais (“eu”, “contigo”, “se”, *etc*), pronomes possessivos (“meu”, “nosso”, “seus”, *etc*), pronomes indefinidos (“algo”, “nenhuma”, “quem”, *etc*), pronomes demonstrativos (“esta”, “isso”, “aquele”, *etc*) e pronomes relativos (“que”, “cujo”, “onde”, *etc*).

De acordo com a classe referente à característica dos pronomes, o *parser* atribui a etiqueta sintática “pron-pers” caso sejam pessoais, “pron-poss” caso sejam possessivos, “pron-indef” caso sejam indefinidos, “pron-dem” caso sejam demonstrativos e “pron-rel” caso sejam relativos, respectivamente.

### Variações Morfológicas dos Pronomes

Os pronomes podem ter variações morfológicas de gênero, número e pessoa, mas não aceitam variações de grau. No que diz respeito ao reconhecimento feito pelo PALAVRAS, apenas as variações de gênero e número são anotadas. Por exemplo, os pronomes “este”, “esse” e “aquele” que são todos pronomes demonstrativos no masculino e singular são vistos como pronomes distintos.

O pronome “Estas” presente na frase exemplo é um pronome adjetivo demonstrativo, pois ele se refere ao substantivo “cidades” que está presente na frase. Este pronome aparece na primeira pessoa do discurso, plural e feminino, logo na forma canônica, o *parser* o armazenará como “este” que está no singular e masculino, não identificando na sua etiqueta morfológica que se trata de um pronome na primeira pessoa.

### 2.1.4 Numerais

Numeral é a palavra que denota uma quantidade ou posição de uma entidade. Na frase exemplo da figura 2.1 a palavra “duas” é um numeral que denota a



quantidade de cidades.

Os numerais são classificados em cardinais (“dois”, “oitenta e quatro”, *etc*), coletivos (“dúzia”, “década”, *etc*), fracionários (“metade”, “um terço”, *etc*), multiplicativos (“dobro”, “quíntuplo”, *etc*) e ordinais (“primeiro”, “último”, *etc*).

Apesar desta classificação gramatical, o *parser* reconhece apenas como numerais os cardinais. Para os cardinais é atribuída a etiqueta sintática “num”.

Para os numerais coletivos, fracionários e multiplicativos o PALAVRAS atribui etiqueta “n”, ou seja, considera-os como substantivos comuns. Os numerais ordinais são considerados adjetivos, e, portanto, recebem a etiqueta “adj”.

### Variações Morfológicas dos Numerais

Os numerais coletivos, fracionários, multiplicativos e ordinais podem ter variação de gênero e número, e são frequentemente tratados pelo PALAVRAS como adjetivos e não como numerais, sempre que acompanham um substantivo especificando-o e caracterizando-o. Os numerais cardinais apresentam apenas variação de gênero e ainda assim, somente os cardinais terminados em “um/uma” e “dois/duas” apresentam esta flexão.

Este é justamente o caso da palavra “duas” da frase exemplo que é salva na forma canônica como “dois” e tem sua etiqueta morfológica indicando que a palavra está originalmente no gênero feminino e no plural.

Note-se que está informação de plural é irrelevante para a anotação, posto que não faz sentido imaginar variações de número para os numerais cardinais. A palavra “duas”, por exemplo sempre será plural, enquanto que a palavra “um” sempre será singular.

### 2.1.5 Artigos

Artigo é a palavra que se antepõe aos substantivos designando seres determinados (“o”, “a”, “os” e “as”) ou indeterminados (“um”, “uma”, “uns” e “umas”). Os artigos não necessariamente adicionam informação aos substantivos aos quais eles precedem, no entanto, conforme será visto na próxima seção (2.2), a presença de artigos antes de substantivos pode ser significativa para sua identificação pelo *parser* PALAVRAS.

Os artigos são anotados sintaticamente pelo PALAVRAS com a etiqueta “art”, porém com certa frequência artigos podem ser anotados erroneamente como pronomes. Os artigos indefinidos podem ser identificados erroneamente como pronomes indefinidos (“pron-indef”), enquanto os artigos definidos podem ser identificados como pronomes pessoais (“pron-pers”) ou demonstrativos (“pron-dem”).

Cabe salientar, que algumas vezes as palavras “o”, “a”, “os”, “as”, “um”, “uma”, “uns” e “umas” podem ser corretamente identificados como pronomes. Como, por exemplo, na frase “A medicação prescrita o curou conforme era esperado.”, onde a palavra “o” foi identificada como um pronome pessoal na terceira pessoa do singular.

Outro fato importante referente a maneira como o *parser* trata dos artigos é o fato do PALAVRAS desfazer as combinações e contrações feitas de preposições com artigos. Este é o caso, bastante comum, de contração com preposições como “dos” ou “à”, que são expandidas em “de os” e “a a”, respectivamente. Na frase exemplo da figura 2.1, encontra-se um exemplo onde a preposição “em” estava contraída com o artigo “o” na palavra “no”. A combinação de preposição com artigo também é desfeita como no caso da palavra “ao” que é expandida em “a o”.

### Variações Morfológicas dos Artigos

Os artigos podem ter variação de gênero e número. Desta forma, apenas duas formas canônicas de artigo são salvas, uma para artigos definidos (“o”) e outra para artigos indefinidos (“um”). Por exemplo, na frase exemplo da figura 2.1, os artigos “os” e “o” são salvos na sua forma canônica como “o”, porém o primeiro é salvo com etiqueta morfológica masculino e plural, enquanto que o segundo é salvo com a etiqueta masculino e singular.

#### 2.1.6 Verbos

“Verbo é a palavra que exprime ação ou apresenta estado ou mudança de um estado a outro” [4]. Ainda que bastante complexo, o reconhecimento dos verbos tem pouca importância para o tema deste relatório que tem como objetivo extrair termos candidatos a conceitos de uma ontologia.

A etiqueta sintática atribuída pelo *parser* para os verbos faz distinções de acordo com a forma do verbo atribuindo as etiquetas distintas para os verbos em todas conjugações usuais (modos indicativo, subjuntivo e imperativo) e

as formas nominais de infinitivo, particípio passado e gerúndio. A tabela 2.2 apresenta esta atribuição de etiquetas e alguns exemplos.

Tabela 2.2: Formas verbais e etiquetas atribuídas pelo PALAVRAS

Formas	Etiqueta	Exemplos
indicativo subjuntivo imperativo	v-fin	“é”, “medicava”
infinitivo	v-inf	“ser”, “medicar”
particípio passado	v-pcp	“sido”, “medicados”
gerúndio	v-ger	“sendo”, “medicando”

### Variações Morfológicas dos Verbos

As variações morfológicas dos verbos correspondem a suas conjugações que podem fazer indicação de pessoa, número, modo, tempo e voz. As variações de pessoa indicam se o sujeito é o emissor da mensagem (primeira pessoa), o receptor da mensagem (segunda pessoa) ou a própria mensagem (terceira pessoa). As variações de número indicam singular ou plural. As variações de modo indicam se o verbo está: no indicativo quando apresenta o fato de uma maneira real, certa, positiva; no subjuntivo quando apresenta o fato de forma duvidosa ou incerta; ou no imperativo quando exprime uma ordem ou solicitação. As variações de tempo podem indicar presente, pretérito perfeito, pretérito imperfeito, pretérito mais-que-perfeito, futuro do presente, futuro do pretérito. As variações de voz podem ser: voz ativa quando o sujeito é o autor da ação descrita pelo verbo; voz passiva quando o sujeito sofre a ação; e voz reflexiva quando o sujeito faz e sofre a ação.

Para verbos nas formas nominais de infinitivo (“v-inf”) e gerúndio (“v-ger”) não existe nenhum tipo de flexão.

Para os verbos nas formas nominais de particípio passado (“v-pcp”) podem haver flexões de número e gênero, logo o salvamento da forma canônica é feito sempre no infinitivo, sendo anotado na etiqueta morfológica as flexões da forma original. Por exemplo, a palavra “nascidas” será anotada como um verbo no particípio passado (“v-pcp”) e a forma canônica salvará o forma infinitiva “nascer” e será indicado na etiqueta morfológica que este verbo apareceu no plural e feminino.

O tratamento de verbos no particípio passado torna-se muito semelhante àquele feito para adjetivos. Inclusive, podendo o verbo no particípio passado assumir funções de substantivos.

Para os verbos nas formas não nominais, ou seja, formatos usuais com conjugação (“v-fin”) são possíveis flexões de pessoa, número, modo e tempo, logo o seu salvamento é bastante distinto dos demais verbos. O verbo é salvo na forma canônica no infinitivo e as etiquetas morfológicas indicam todas as flexões possíveis. Utilizando a palavra “são” presente na frase exemplo da figura 2.1, o *parser* anotará na etiqueta morfológica o tempo verbal (presente), a pessoa (terceira), o número (plural) e o modo (indicativo).

### 2.1.7 Advérbios

Advérbio é a palavra que modifica um verbo, um adjetivo ou mesmo um outro advérbio e em todas estas situações indica as circunstâncias em que acontece a ação verbal. Os advérbios são anotados sintaticamente pelo PALAVRAS com a etiqueta “adv”. A exemplo dos verbos, os advérbios tem pouca relevância para o propósito da extração de termos deste relatório, pois usualmente eles não tem valor terminológico.

Na frase exemplo da figura 2.1 a palavra “mais” é um advérbio que modifica o adjetivo “importantes” que vem logo a seguir.

### Variações Morfológicas dos Advérbios

Os advérbios não flexionam em gênero ou número, apenas em grau podendo indicar usualmente superlativos ou diminutivos, como é o caso, por exemplo, do advérbio “pouco” que flexiona em “pouquíssimo” (superlativo) e “pouquinho” (diminutivo). Fora esta situação usual, os advérbios “bem” e “mal” possuem uma flexão de grau adicional, o comparativo de superioridade que flexiona para as formas “melhor” e “pior”, respectivamente.

No entanto, o *parser* não percebe estas sutilezas de linguagem e considera os advérbios flexionados erroneamente como substantivos (“n”) quando flexionados para superlativos e diminutivos, enquanto os comparativos de superioridade (“melhor” e “pior”) são considerados como adjetivos (“adj”).

A única consideração de variação morfológica de advérbios encontrada utilizando o PALAVRAS é no caso do advérbio “mais” que aparece na frase exemplo, que é salvo na sua forma canônica como uma flexão do advérbio “muito”.

### 2.1.8 Preposições

Preposição é a expressão que, posta entre duas outras, estabelece uma subordinação da segunda a primeira. Desta forma, ela liga dois substantivos, um substantivo a um verbo ou um advérbio, e um adjetivo, verbo ou advérbio a um substantivo.

Na língua portuguesa existe um grande número de palavras que são empregadas sempre como preposição. Essas são chamadas de preposições essenciais e são: “a”, “ante”, “após”, “até”, “com”, “contra”, “de”, “desde”, “em”, “entre”, “para”, “per”, “perante”, “por”, “sem”, “sob”, “sobre” e “trás”. A maioria destas preposições pode ser contraída com artigos, como é o caso da preposição “no” na frase exemplo da figura 2.1 que é o resultado da contração da preposição essencial “em” e o artigo “o”.

Uma outra possibilidade de preposições são as preposições acidentais que são palavras que podem ser utilizadas como preposição em situações específicas. Dentre as preposições acidentais, as mais frequentes são as palavras: “como”, “conforme”, “exceto”, “feito”, “mediante”, *etc.*

As preposições não admitem variação morfológica, portanto o *parser* salva sempre as preposições na forma canônica e as identifica com a etiqueta ”**prp**”. A única transformação feita pelo PALAVRAS é a separação das preposições contraídas com artigos.

### 2.1.9 Conjunções

Conjunção é a expressão que liga orações ou, dentro da mesma oração, palavras que tenham o mesmo valor ou função. Um fato importante das conjunções é que com grande frequência elas não são representadas por uma única palavra, mas sim por uma expressão que denomina-se locução conjuntiva. Assim como é feito no PALAVRAS, no contexto deste trabalho trataremos de forma igual as conjunções propriamente ditas (com uma única palavra) e as locuções conjuntivas, com o termo genérico conjunções.

As conjunções se dividem em dois grupos, conjunções coordenativas e subordinativas, segundo a relação de dependência sintática dos termos que relacionam.

As conjunções coordenativas são utilizadas para conectarem duas orações ou dois termos pertencentes a um mesmo nível sintático, ou seja, duas orações ou termos que se invertidos mantém o mesmo sentido. Elas podem ser aditivas (por exemplo: “e”, “mas também”), adversativas (por exemplo: “mas”, “no

entanto”), alternativas (por exemplo: “ou”, “ora”), explicativa (por exemplo: “pois”, “porque”) ou conclusivas (“logo”, “então”).

As conjunções subordinativas são utilizadas para conectar duas orações que possuem diferentes níveis sintáticos, ou seja, a segunda oração fica subordinada à primeira. Elas podem ser integrantes (por exemplo: “se”), causais (por exemplo: “já que”), comparativas (por exemplo: “tanto . . . quanto”), concessivas (por exemplo: “embora”), condicionais (por exemplo: “caso”), conformativas (por exemplo: “conforme”), consecutivas (por exemplo: “de forma que”), explicativas (por exemplo: “pois”), finais (por exemplo: “para que”), proporcionais (por exemplo: “à medida que”) ou temporais (por exemplo: “quando”).

As conjunções não admitem variação morfológica, portanto o *parser* salva sempre as preposições na forma canônica. Sua única distinção é que o PALAVRAS identifica com a etiqueta ”conj-c” as conjunções coordenativas e com a etiqueta ”conj-s” as conjunções subordinativas.

Na frase exemplo da figura 2.1 temos apenas uma conjunção, a palavra “e” que é anotada corretamente como uma conjunção coordenativa (”conj-c”) que liga os termos “maiores” e “mais importantes”.

### 2.1.10 Interjeições

Interjeição é a expressão com que traduzimos os nossos estados emotivos, logo, é muito raro aparecer interjeições nos textos científicos que são o alvo usual da ferramenta *E $\chi$ ATOLP*, e da extração de termos. No entanto, o *parser* detecta as interjeições e as identifica com a etiqueta sintática “intj”. Evidentemente, as interjeições não possuem variação morfológica.

## 2.2 Classificação dos Não-Terminais

Uma sentença, do ponto de vista gramatical, pode ser composta de diversas orações. Na língua portuguesa a classificação de orações é bastante complexa e sua plena compreensão foge ao escopo deste trabalho.

Neste sentido, esta seção se limita a analisar definições segundo a classificação escolhida pelo *parser* PALAVRAS. Segundo esta classificação, as orações podem ser de quatro tipos: orações finitas (etiquetada como “fcl”), orações infinitas (etiquetada como “icl”), orações averbais (etiquetada como

“acl”) e paratagmas (etiquetada como “cu”) que são orações que cumprem a função de um substantivo.

Em todas estas orações o *parser* reconhece os elementos que a compõem como termos individuais, ou como sintagmas que são um conjunto de termos que desempenha uma função na frase mantendo entre si relações de dependência e de ordem[22].

### 2.2.1 Classificação de Sintagmas

Os sintagmas são uma unidade de informação presente na frase que organizam-se em torno de um elemento fundamental, denominado núcleo, que pode, por si só, constituir o sintagma[22]. Desta forma, a natureza do sintagma depende do tipo de elemento que constitui o seu núcleo, podendo ser:

- sintagma nominal (etiquetado como “np”) cujo núcleo pode ser um substantivo comum ou próprio, um pronome, um adjetivo ou até um verbo no particípio passado desde que estes estejam substituindo um substantivo;
- sintagma verbal (etiquetado como “vp”) cujo núcleo é um verbo;
- sintagma adjetival (etiquetado como “ap”) cujo núcleo é um adjetivo; ou
- sintagma preposicional (etiquetado como “pp”) cujo núcleo é uma preposição.

Dentre estes tipos de sintagmas, para os propósitos do trabalho desenvolvido neste relatório, apenas os sintagmas nominais são relevantes, pois eles são os melhores candidatos a conceitos de uma ontologia[15].

#### Sintagmas Nominais

Sintagmas nominais (SN) podem ter como núcleo um substantivo ou pronome substantivo (pessoal, demonstrativo, indefinido, interrogativo, possessivo ou relativo). Excepcionalmente, o núcleo do SN pode ser um adjetivo que substitua um substantivo previamente citado. Análogo aos adjetivos, os verbos no particípio passado possuem como característica fundamental a capacidade de desempenhar como forma nominal a função de adjetivo [4]. Desta forma,

substantivos, pronomes, adjetivos e verbos no particípio passado podem ser núcleos de um SN.

Além do núcleo, um SN pode ser precedido por artigos, pronomes e numerais (determinantes) e precedido ou sucedido por adjetivos, locuções adjetivas ou orações subordinadas adjetivas (modificadores). De um ponto de vista prático, o modificador de um SN pode ser constituído de um sintagma adjetival ou de um sintagma preposicionado (formado de preposição + sintagma nominal) [20].

O parser PALAVRAS anota sintaticamente o núcleo do SN com uma etiqueta “H”. Este é o caso da palavra “cidades” do primeiro SN da frase exemplo da figura 2.1 (“Estas duas cidades”), mas é também o caso da palavra “maiores” do segundo SN (“os maiores”). Note-se que neste segundo caso o núcleo é um adjetivo.

Neste mesmo exemplo é interessante observar que o SN “mais importantes centros de pesquisa no Brasil” tem como núcleo a palavra “centros”, mas ele engloba dois outros SN: “pesquisa no Brasil” e “o Brasil”, cada um deles com seus próprios núcleos, respectivamente “pesquisa” e “Brasil”.

Cabe salientar que a implementação do *parser* PALAVRAS só considera como SN termos composto por pelo menos duas palavras, como é o caso do último SN da frase exemplo: “o Brasil”. Porém de um ponto de vista terminológico, o artigo que serve como determinante “o” não possui nenhuma relevância, ou seja, apenas a palavra “Brasil” é terminologicamente relevante. Desta forma, ainda que os SN detectados pelo *parser* tenham, por definição, pelo menos duas palavras, é perfeitamente possível que apenas uma das palavras do SN tenha valor terminológico. Por esta razão, no contexto deste relatório, os SN identificados pelo PALAVRAS serão considerados candidatos a extração de termos independente do número de palavras que o compõe.



# Capítulo 3

## Visão Geral da Ferramenta

### $E\chi$ ATO<sub>LP</sub>

$E\chi$ ATO<sub>LP</sub> – Extrator Automático de Termos para Ontologias em Língua Portuguesa – é uma ferramenta que recebe um *corpus* anotado e extrai automaticamente todos os sintagmas nominais (SN) deste texto classificando-os segundo o número de palavras. Os sintagmas extraídos são salvos em listas que podem conter tanto os SN na sua forma original no texto, como em sua forma canônica. A ferramenta ainda oferece algumas opções de manipulação usuais para listas de termos como a aplicação de pontos de corte, comparação de listas e cálculo de medidas usuais de precisão e abrangência.

As seções a seguir descrevem, respectivamente: as funcionalidades da ferramenta (seção 3.1); os formatos de entrada para o *corpus* sintaticamente anotado e listas de termos a manipular (seção 3.2); e os formatos de saída das listas de termos extraídos e diversas medidas calculáveis (seção 3.3).

### 3.1 Funcionalidades

As funcionalidades da ferramenta vão desde tarefas fortemente baseadas em conceitos linguísticos como a extração de sintagmas nominais, até tarefas puramente estatísticas como o cálculo de métricas de avaliação, passando por tarefas usuais de processamento de dados como corte de listas de strings, neste caso listas de termos.

As seções a seguir detalham cada uma destas funcionalidades mantendo uma visão descritiva em alto nível das tarefas executadas pela ferramenta.

Definições mais precisas da implementação destas funcionalidades só serão vistas no capítulo 4, assim como definições mais objetivas de como o usuário irá executá-las só serão vistas no capítulo 5.

### 3.1.1 Extração de Sintagmas Nominais

A função primária da ferramenta  $E\chi ATO\mathcal{LP}$  é extrair termos candidatos a conceitos de uma ontologia a ser construída. Desta forma, a principal funcionalidade da ferramenta é o processo de extração de sintagmas nominais, pois estes são, segundo especialistas da área [15], os melhores candidatos a conceitos. Ao contrário das palavras isoladas cujo significado depende fortemente do contexto, quando SN são extraídos de um texto, seus significados permanecem os mesmos[15].

De um ponto de vista objetivo, a ferramenta utiliza um conjunto de heurísticas para refinar o processo de extração. Estas heurísticas tem base linguística com o propósito de eliminar ou refinar termos identificados pelo *parser* como SN que não sirvam como possíveis conceitos de uma ontologia, seja por eventual erro de identificação do parser, seja por falta de relevância terminológica.

Especificamente, as heurísticas aplicadas aos termos identificados como SN pelo PALAVRAS são:

- são eliminados SN que terminam com preposição, *e.g.*, “*criança acrescida de*”, “*dosagem diária para*”;
- são eliminados SN que possuem números, *e.g.*, “*década de 50*”, “*dois estudos*”;
- são aceitos apenas os SN cujo o núcleo é substantivo, nome próprio, adjetivo ou verbo no particípio passado. Não foram encontrados nos experimentos SN que não possuem núcleos em categorias diferentes destas;
- são aceitos apenas sintagmas que possuem letras (acentuadas ou não) ou hífen, ou seja, SN que contém caracteres especiais são eliminados, *e.g.*, “*remédio+profilaxia*”, “*dupla mãe/neonato*”;
- SN que iniciam com pronomes, *e.g.*, “*estas condições*” “*todas as crianças*” “*seus acompanhantes*”, “*esses dados*”, são armazenados sem a primeira palavra (o pronome);

- SN que terminam com conjunções (“e” e “ou”) são armazenados sem a conjunção, *e.g.*, “*baixo peso e*” e “*leite materno ou*” são armazenados, respectivamente como “*baixo peso*” e “*leite materno*”;
- SN que contém artigos são armazenados sem estes artigos, *e.g.*, “*a cicatriz renal*” é armazenado apenas como “*cicatriz renal*”, “*os pacientes da clínica*” é armazenado apenas como “*pacientes de clínica*”.

Os sintagmas extraídos podem ser compostos de um número qualquer de palavras, inclusive sendo apenas um unigrama, pois alguns sintagmas que seriam originalmente compostos por duas palavras podem ser transformados pelo remoção de uma delas. Este é tipicamente o caso de termos que eram compostos por um artigo e um substantivo e que tem o artigo removido. Na prática, a ferramenta agrupa os sintagmas extraídos em dez listas que contém respectivamente os sintagmas compostos por 1 a 9 palavras e a última lista contém sintagmas compostos por 10 ou mais palavras.

A ferramenta *E $\chi$ ATOLP* gera cada uma destas dez listas de termos em ordem decrescente de frequência no *corpus*. Desta forma, estas listas podem ser facilmente submetidas a pontos de corte que levam em consideração a frequência relativa ou absoluta, ou são utilizadas na sua totalidade.

### 3.1.2 Aplicação de Pontos de Corte

Em geral, a saída do processo de extração gera uma lista de termos muito extensa, a qual inclui termos relevantes, mas também um número grande de termos irrelevantes. Neste sentido, é interessante buscar uma forma de reduzir o tamanho das listas, excluindo o mínimo possível de termos relevantes.

Para que esse tipo de redução seja feita, o primeiro passo deve ser ordenar os termos segundo sua relevância. É necessário definir um critério que traduza, da melhor maneira possível, a relevância de cada termo. A ferramenta *E $\chi$ ATOLP* já fornece como saída do processo de extração os termos ordenados segundo suas frequências no *corpus*. Note-se que tanto a ordenação pela frequência absoluta, como pela frequência relativa, resultam necessariamente em uma mesma ordem.

O passo natural para a aplicação de pontos de corte é definir a partir de que ponto desprezar os termos menos frequentes. Basicamente, entre as diversas opções, pode-se imaginar um ponto de corte arbitrário absoluto. Por exemplo, desprezar todos os termos em que a frequência relativa seja inferior a

$10^{-5}$ , ou desprezar todos os termos que aparecem menos de 4 vezes no *corpus*, ou então manter apenas os 100 primeiros termos da lista ordenada. Outra opção ainda é adotar pontos de cortes relativos, por exemplo, manter os 20% primeiros termos da lista ordenada. Na verdade, a definição destes pontos de corte pode ser feita de acordo com diversos critérios como o tamanho do *corpus*, especificidade do domínio, ou qualquer outro aspecto que o usuário possa julgar relevante. Neste sentido, a ferramenta  $E\chi ATO_{LP}$  disponibiliza as seguintes opções de ponto de corte:

- ponto de corte absoluto segundo a frequência relativa, onde um limiar mínimo (um número real entre 0 e 1) deve ser informado;
- ponto de corte absoluto segundo a frequência absoluta, onde um limiar mínimo (um número inteiro superior a 1) deve ser informado;
- ponto de corte absoluto único, onde um número específico de termos (um inteiro) deve ser informado;
- ponto de corte relativo, onde um percentual do número de termos (um valor entre 0% e 100%) deve ser informado.

Note-se que a escolha de um ponto de corte, apesar de muito importante, é uma questão aberta onde existem pesquisas que indicam ser particular a cada caso qual das opções pode ser mais interessante [17].

### 3.1.3 Comparação de Listas de Termos

Em fases experimentais e de avaliação, o processo de extração de termos deve comparar a lista de termos extraída automaticamente com uma lista de referência contendo os termos do corpus considerados relevantes por especialistas do domínio. Estas duas listas são denominadas arbitrariamente de lista de referência ( $LR$ ) e lista de extraídos ( $LE$ ).

No  $E\chi ATO_{LP}$  a comparação de listas recebe como entrada duas listas,  $LR$  e  $LE$ , e pode retornar qualquer uma das seguintes listas:

- a intersecção entre elas ( $LR \cap LE$ );
- a união entre elas ( $LR \cup LE$ );
- os termos de  $LR$  ausentes em  $LE$  ( $LR - (LR \cap LE)$ );
- os termos de  $LE$  ausentes em  $LR$  ( $LE - (LR \cap LE)$ ).

### 3.1.4 Cálculo de Métricas

Com intuito de tornar objetiva a comparação de listas, a ferramenta  $E\chi ATOLP$  disponibiliza o cálculo de métricas quantitativas que expressam a precisão e a abrangência de listas comparadas, bem como o equilíbrio entre estes dois índices (*f-measure*). A precisão ( $P$ ) indica a capacidade do método de identificar os termos corretos, considerando a lista de referência. Este índice é calculado pela primeira das fórmulas abaixo que é a razão entre o número de termos encontrados na lista de referência ( $|LR|$ ) e na lista de termos extraídos ( $|LE|$ ), ou seja, a cardinalidade da intersecção dos conjuntos  $LR$  e  $LE$  pelo total de termos extraídos (cardinalidade do conjunto  $LE$ ).

$$P = \frac{|LR \cap LE|}{|LE|}$$

Analogamente, a abrangência ( $A$ ) avalia a quantidade de termos corretos extraídos pelo método em relação ao tamanho da lista de referência.

$$A = \frac{|LR \cap LE|}{|LR|}$$

Finalmente, a *f-measure* ( $F$ ) é simplesmente a média harmônica entre a precisão e abrangência.

$$F = \frac{2 \times P \times A}{P + A}$$

## 3.2 Formatos de Entrada

O formato básico de entrada da ferramenta é um *corpus* com anotações linguísticas. Porém, a ferramenta  $E\chi ATOLP$  também recebe como entrada listas de termos que podem ser manipuladas de diversas formas. Nesta seção os formatos de entrada de *corpus* (seção 3.2.1) e de listas de termos (seção 3.2.2) são detalhados.

### 3.2.1 Corpus – Texto Anotado

Na abordagem utilizada pela ferramenta  $E\chi ATOLP$ , o processo de extração de termos inicia-se com anotação linguística dos textos que compõem o *corpus*, realizada pelo *parser* PALAVRAS [5]. O *parser* faz análise sintática e semântica através da construção de uma árvore na qual os nós terminais

(folhas da árvore) são as palavras do texto e os não terminais representam as categorias da estrutura da frase. Os diversos textos entram como arquivos ASCII (txt) e o PALAVRAS tem na saída as informações representadas em um arquivo no formato TIGER-XML [13]. Este arquivo XML contém todas as frases devidamente anotadas linguisticamente, ou seja, cada uma de suas palavras é anotada conforme sua função sintática, semântica e suas características morfológicas.

Note-se que a escolha de trabalhar como entrada da ferramenta  $E\chi ATO_{LP}$  o formato TIGER-XML é uma escolha que leva em conta tanto o fato deste formato ser a saída do *parser* PALAVRAS. Maiores detalhes do formato TIGER-XML podem ser encontrados em [14], mas no que diz respeito ao uso na ferramenta  $E\chi ATO_{LP}$ , o formato TIGER-XML é composto de um conjunto de sentenças, cada uma delas representada por um árvore (*graph*) descrita pelos suas folhas (*terminals*) e nodos intermediários (*nonterminals*). A figura 3.2 apresenta um exemplo do formato TIGER-XML para uma única frase: “Gastroesquise é um defeito da parede abdominal anterior.” e a figura 3.1 é a representação gráfica desta árvore obtida pelo visualizador VISL [6].

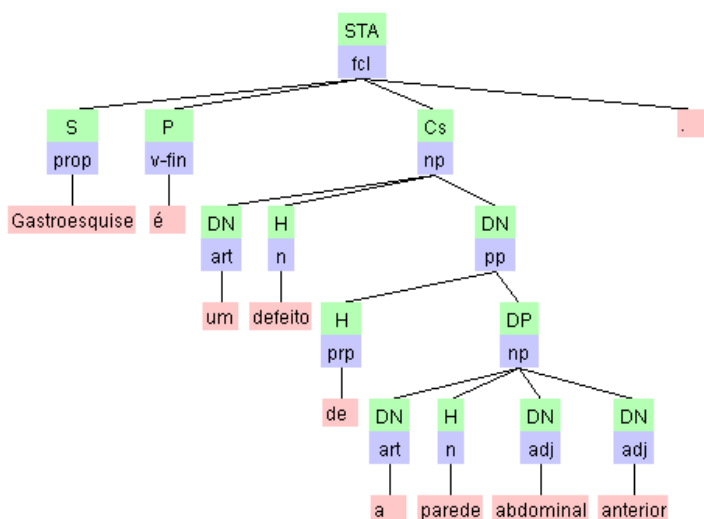


Figura 3.1: Representação Gráfica do Exemplo da Figura 3.2

```

<?xml version="1.0" encoding="iso-8859-1"?>
<body>
<corpus>
  <s id="s1" ref="1" source="Running text" forest="1" text="Gastrosquise é um defeito das
    paredes abdominais anteriores.">

    <graph root="s1_500">
      <terminals>
        <t id="s1_1" word="Gastrosquise" lemma="Gastrosquise"
          pos="prop" morph="F S" sem="--" extra="hum"/>
        <t id="s1_2" word="é" lemma="ser"
          pos="v-fin" morph="PR 3S IND VFIN" sem="--" extra="fmc mv"/>
        <t id="s1_3" word="um" lemma="um" pos="art" morph="M S" sem="--" extra="--"/>
        <t id="s1_4" word="defeito" lemma="defeito"
          pos="n" morph="M S" sem="ac" extra="--"/>
        <t id="s1_5" word="de" lemma="de" pos="prp" morph="---" sem="---" extra="sam- np-close"/>
        <t id="s1_6" word="as" lemma="o" pos="art" morph="F P" sem="---" extra=" -sam"/>
        <t id="s1_7" word="paredes" lemma="parede"
          pos="n" morph="F P" sem="part-build" extra="--"/>
        <t id="s1_8" word="abdominais" lemma="abdominal"
          pos="adj" morph="F P" sem="---" extra="np-close"/>
        <t id="s1_9" word="anteriores" lemma="anterior"
          pos="adj" morph="F P" sem="---" extra="np-long"/>
        <t id="s1_10" word="." lemma="---" pos="pu" morph="---" sem="---" extra="--"/>
      </terminals>
      <nonterminals>
        <nt id="s1_500" cat="s">
          <edge label="STA" idref="s1_501"/>
        </nt>
        <nt id="s1_501" cat="fcl">
          <edge label="S" idref="s1_1"/>
          <edge label="P" idref="s1_2"/>
          <edge label="Cs" idref="s1_502"/>
        </nt>
        <nt id="s1_502" cat="np">
          <edge label="DN" idref="s1_3"/>
          <edge label="H" idref="s1_4"/>
          <edge label="DN" idref="s1_503"/>
        </nt>
        <nt id="s1_503" cat="pp">
          <edge label="H" idref="s1_5"/>
          <edge label="DP" idref="s1_504"/>
        </nt>
        <nt id="s1_504" cat="np">
          <edge label="DN" idref="s1_6"/>
          <edge label="H" idref="s1_7"/>
          <edge label="DN" idref="s1_8"/>
          <edge label="DN" idref="s1_9"/>
        </nt>
      </nonterminals>
    </graph>
  </s>
</corpus>
</body>

```

Figura 3.2: Exemplo do Formato TIGER-XML

### 3.2.2 Listas de Termos

As listas de termos tratadas pelo  $E\chi ATOLP$  possuem características comuns que se encontram em diversas ferramentas de manipulação de dados textuais. Basicamente, as listas de termos tem o formato de um termo em cada linha, seja ele composto de quantas palavras for. Estas listas de termos podem aparecer no formato que o usuário desejar, por exemplo na figura 3.3 são apresentados dois exemplos de listas, a primeira com termos na sua forma original e a segunda com os mesmos termos na sua forma canônica, ou seja, os termos com declinações de gênero e número removidas, bem como as conjugações verbais. Neste tipo de lista as palavras são separadas entre si por um espaço em branco simples.

recém-nascidos	recém nascer
aleitamento materno	aleitamento materno
presente estudo	presente estudo
leite materno	leite materno
idade gestacional	idade gestacional
ventilação mecânica	ventilação mecânico
sexo masculino	sexo masculino
sexo feminino	sexo feminino
leite humano	leite humano
faixas etárias	faixa etário
Estados Unidos	estados unidos
vias aéreas	via aérea
hipertensão arterial	hipertensão arterial
pressão arterial	pressão arterial
período neonatal	período neonatal
baixo peso	baixo peso
perímetro cefálico	perímetro cefálico
atividade física	atividade físico
cicatriz renal	cicatriz renal

Figura 3.3: Exemplo de Lista de Termos nas Formas Originais e Canônica

Normalmente, estas listas devem conter somente letras maiúsculas ou minúsculas sejam elas acentuadas ou não, segundo as regras da língua portuguesa, tolerando o uso de trema na letra u, para aceitar textos que tenham sido escritos antes da última reforma ortográfica. Adicionalmente, aceita-se



também o uso dos caracteres espaço em branco, sublinhado e hífen. Desta forma, o conjunto total de caracteres aceito nas listas é apresentado na tabela 3.1.

Tabela 3.1: Caracteres Aceitos em Listas de Entrada

a	b	c	d	e	f	g	h	i
j	k	l	m	n	o	p	q	r
s	t	u	v	w	x	y	z	À
B	C	D	E	F	G	H	I	J
K	L	M	N	O	P	Q	R	S
T	U	V	W	X	Y	Z	á	é
í	ó	ú	â	ê	ô	à	ã	õ
ü	ç	Á	É	Í	Ó	Ú	Â	Ê
Ô	À	Ã	Õ	Ü	Ç		-	-

### 3.3 Formatos de Saída

As saídas básicas da ferramenta *E<sub>χ</sub>ATOLP* são conjuntos de termos extraídos. Esses conjuntos de termos podem variar de acordo com a opção escolhida, sendo basicamente de três tipos:

- listas de termos na forma original, onde os termos são mantidos como são encontrados no *corpus*, a exceção da remoção dos artigos.
- listas de termos na forma canônica, onde os termos são colocados na forma canônica, ou seja, são removidos artigos, as declinações de gênero, número e conjugação;
- tabelas de sintagmas na formas original e canônica, número de termos, núcleo do sintagma, etiquetas sintática e semântica do núcleo, além das frequências absoluta e relativa do sintagma no texto.

As listas de termos nas formas original e canônica apresentam o mesmo formato utilizado para a entrada de listas apresentado na figura 3.3, ou seja, um termo por linha com suas palavras separadas por espaços em branco.

As tabelas de sintagmas gerados são mais ricas em informação, mas também tem uma leitura menos intuitiva. No entanto, como será visto

na descrição da interface com o usuário (capítulo 5) a visualização destas informações na interface é facilitada. A figura 3.4 apresenta um exemplo parcial de uma tabela de sintagmas.

recém-nascidos	recém_nascer	2	recém_nascer	v-pp	?	262	0.002051
aleitamento_materno	aleitamento_materno	2	aleitamento	n	act	229	0.001793
presente_estudo	presente_estudo	2	estudo	n	sem-r	177	0.001386
leite_materno	leite_materno	2	leite	n	drink	136	0.001065
idade_gestacional	idade_gestacional	2	idade	n	f-q	130	0.001018
ventilação_mecânica	ventilação_mecânico	2	ventilação	n	act	112	0.000877
sexo_masculino	sexo_masculino	2	sexo	n	activity_f-h	103	0.000807
sexo_feminino	sexo_feminino	2	sexo	n	activity_f-h	85	0.000666
leite_humano	leite_humano	2	leite	n	drink	79	0.000619
faixas_etárias	faixa_etário	2	faixa	n	sem-l	68	0.000532
Estados_Unidos	estados_unidos	2	estados_unidos	prop	?	66	0.000517
vias_aéreas	via_aérea	2	via_aérea	n	Lpath	66	0.000517
hipertensão_arterial	hipertensão_arterial	2	hipertensão	n	sick	62	0.000485
pressão_arterial	pressão_arterial	2	pressão_arterial	n	f-q	61	0.000478
período_neonatal	período_neonatal	2	período	n	per	59	0.000462
baixo_peso	baixo_peso	2	peso	n	f-q	58	0.000454
perímetro_cefálico	perímetro_cefálico	2	perímetro	n	Labs	58	0.000454
atividade_física	atividade_físico	2	atividade	n	activity_act-d	54	0.000423
cicatriz_renal	cicatriz_renal	2	cicatriz	n	sick-c	53	0.000415

Figura 3.4: Exemplo de Tabela de Sintagmas Nominais

Nesta tabela as informações em cada coluna são separadas por espaços em branco e, respectivamente, possuem:

- sintagma na forma original que aparece no *corpus* com palavras separadas por sublinhados quando no *corpus* eram separados por espaços e separados por hífen quando a forma original possui hífen, além disto, a forma original, por efeito do PALAVRAS desmembra palavras que possuem junções entre preposições e artigos, por exemplo “da” é desmembrado na preposição “de” e no artigo “a”;
- sintagma na forma canônica onde são removidos todos os artigos e todas as letras são transformadas em minúsculas, além disto todas as palavras são sempre separadas por sublinhados;
- número de palavras no sintagma;
- núcleo do sintagma conforme informado pelo PALAVRAS, ou seja, não necessariamente o núcleo será uma única palavra caso o *parser* tenha considerado duas ou mais palavras como um único token léxico;

- a etiqueta sintática do núcleo de acordo com a codificação do PALAVRAS, usualmente esta etiqueta indica um substantivo (n) ou um nome próprio (prop);
- o conjunto de etiquetas semânticas do núcleo, pois o PALAVRAS pode associar diversas etiquetas que aparecem separadas por sublinhados, por exemplo a palavra “ventilação” recebeu a etiqueta semântica “act” que significa que se trata de uma ação;
- a frequência absoluta do sintagma no *corpus*, ou seja, o número de vezes que o sintagma foi corretamente identificado nos textos tratados;
- a frequência relativa do sintagma que é simplesmente a frequência absoluta dividida pelo total de termos identificados no *corpus*.

Além das listas de termos e tabela de sintagmas, a ferramenta  $E\chi$ ATOLP também fornece como saída algumas informações numéricas na forma de relatórios de termos extraídos, erros de saídas do PALAVRAS, cardinalidade de listas e métricas resultantes das comparações de listas. A maior parte destas informações são apresentadas na interface gráfica e na sua totalidade são salvas em relatórios textuais em arquivos simples.

# Capítulo 4

## Implementação

Este capítulo descreve os módulos centrais da ferramenta  $E\chi ATOLP$  para que o leitor possa entender em detalhe as decisões de implementação da ferramenta, bem como tirar as dúvidas que possam ficar sobre o seu funcionamento. Nesta seção descreve-se inicialmente o módulo central da ferramenta que é responsável pela extração dos termos candidatos a conceitos de uma ontologia a partir de um *corpus*, chamado de módulo extrator. A segunda seção descreve módulos de operações auxiliares neste processo que são a aplicação de pontos de corte, comparação de listas, cálculo de métricas e localização de termos em arquivos.

### 4.1 Módulo Básico da Ferramenta – Módulo Extrator

Este módulo é o ponto central da ferramenta  $E\chi ATOLP$ , pois ele é responsável pela busca dos sintagmas nominais em cada um dos textos do *corpus*. Este módulo é composto por duas tarefas que são responsáveis por identificar os sintagmas nominais nos arquivos TIGER-XML.

#### 4.1.1 Procedimento Geral

Para entender o funcionamento desta tarefa é necessário observar em detalhe o formato TIGER-XML que é a saída do *parser* PALAVRAS e no qual são codificados cada um dos textos anotados sintaticamente. Para isto é

necessário observar o exemplo da figura 3.2, que mostra a codificação do formato TIGER-XML da frase “Gastrosquise é um defeito da parede abdominal anterior.”.

A primeira informação importante para a compreensão do módulo extrator é entender que cada texto é tratado individualmente, mas que todos os sintagmas extraídos são agrupados em um mesmo arquivo. Da mesma forma que os textos, cada uma de suas sentenças são tratadas individualmente.

Logo, o processo descrito a seguir é repetido para cada uma das sentenças de cada texto. Uma vez identificado o início de uma sentença (pelos caracteres “<s”) são executados três passos para a identificação dos sintagmas. O pseudo-código 4.1.1 descreve esta função principal de extração de sintagmas.

---

#### Pseudo-Código 4.1.1 Extração de Sintagma

---

<b>para todos</b> textos <i>txt</i>	// (“*.xml”)
<b>para todas</b> sentenças <i>s</i> do texto <i>txt</i>	// (“<s”)
<b>armazena</b> terminais de <i>s</i>	// (“<terminals>”)
<b>armazena</b> não-terminais de <i>s</i>	// (“<nonterminals>”)
<b>busca</b> sintagmas nominais de <i>s</i>	

---

#### 4.1.2 Primeiro Passo – Armazena Terminais

O primeiro passo do processo de extração é salvar o conjunto de terminais e suas informações relevantes em memória. O pseudo-código 4.1.2 descreve esta função de armazenamento de terminais.

---

#### Pseudo-Código 4.1.2 Armazena Terminais

---

<b>para todos</b> terminais <i>t</i> da sentença <i>s</i>	// (“<t”)
<b>le e armazena</b> palavra na forma original	// (“word=”)
<b>le e armazena</b> palavra na forma canônica	// (“lemma=”)
<b>le</b> etiqueta morfológica	// (“morph=”)
<b>le e armazena</b> etiqueta sintática	// (“pos=”)
<b>le e armazena</b> etiqueta semântica	// (“sem=”)
<b>le</b> etiqueta adicional	// (“extra=”)

---

Este módulo procura os caracteres “<terminals>”, e em seguida para cada terminal (identificados pelos caracteres “<t”) são salvas as seguintes informações:

- a palavra na forma original (identificada após “word=”);
- a palavra na forma canônica (identificada após “lemma=”);
- a etiqueta sintática da palavra (identificada após “pos=”);
- a etiqueta semântica da palavra (identificada após “sem=”).

Note-se que neste módulo são desprezadas as informações constantes nas etiquetas morfológicas (“morph”) e adicionais (“extra”).

### 4.1.3 Segundo Passo – Armazena Não-terminais

O segundo passo do processo de extração é o armazenamento dos nodos não terminais em memória. O pseudo-código 4.1.3 descreve esta função de armazenamento de não-terminais.

---

#### Pseudo-Código 4.1.3 Armazena Não-Terminais

---

```

para todos não-terminais nt da sentença s           // (“<nt””)
  le categoria                                       // (“cat=”)
  se categoria indica sintagma nominal              // (“cat=np”)
    armazena é_sintagma_nominal(nt) = verdadeiro
  senão
    armazena é_sintagma_nominal(nt) = falso
  para todos ramos r do não-terminal nt           // (“<edge””)
    le rótulo                                       // (“label=”)
    se rótulo indica núcleo do sintagma            // (“label=H”)
      armazena é_núcleo(nt,r) = verdadeiro
    senão
      armazena é_núcleo(nt,r) = falso
    le e armazena o índice do ramo r              // (“idref=”)

```

---

Este módulo procura os caracteres “<nonterminals>”, e em seguida para cada nodo (identificados pelos caracteres “<nt””) são salvas as seguintes informações:

- o fato da sua categoria ser ou não sintagma nominal (identificado pela informação de categoria “cat=np”);

- cada um dos ramos (arcos de saída) deste nodo (identificado pelo número após “<edge” e “idref=”);
- marca-se ainda como núcleo do sintagma o ramo que aponta para o terminal anotado pelo PALAVRAS como núcleo (identificado pela informação do rótulo “label=“H””).

Cabe salientar que ocasionalmente o *parser* PALAVRAS identifica algum conjunto de palavras como um único token e, conseqüentemente, um destes conjuntos de palavras pode ser identificado como o núcleo do sintagma. Este, por exemplo, foi o caso da expressão “*pressão arterial*” que foi identificado em um *corpus* de onde o exemplo da figura 3.4 foi retirado.

#### 4.1.4 Terceiro Passo – Busca Sintagmas Nominais

Finalmente, o terceiro passo consiste em analisar as estruturas de terminais e não-terminais salvos em memória para buscar efetivamente cada um dos sintagmas nominais. Este é o passo mais delicado deste módulo, pois um conjunto de tratamentos linguísticos são aplicados através de heurísticas que ora invalidam alguns sintagmas, ora aprimoram algumas escolhas do *parser*, e até chegam a detectar alguns erros de anotação do *parser* que não são possíveis de remediar. O pseudo-código 4.1.4 descreve esta tarefa de busca de sintagmas.

---

##### Pseudo-Código 4.1.4 Busca Sintagmas Nominais

---

```

para todos não-terminais nt
  se é_sintagama_nominal(nt)
    monta (nt, sintag)
    se válido(sintag)
      aprimora (sintag)
      armazena (sintag)

```

---

Neste pseudo-código existem quatro chamadas a funções que serão detalhadas logo a seguir. Estas funções são responsáveis por montar recursivamente o sintagma que está sendo buscado (função **monta**), por verificar a sua validade (função **válido**), tentar aprimorar este sintagma (função **aprimora**) e finalmente armazenar o sintagma extraído (função **armazena**).

### 4.1.5 Montagem de Sintagmas – Função Monta

O processo de montagem de sintagmas é feito a partir dos ramos dos não-terminais indicados como sintagma nominal durante o seu armazenamento. Basicamente, para cada um destes ramos, caso o ramo aponte para um terminal; esta palavra é adicionada ao sintagma; caso aponte para um não-terminal, a função de montagem é chamada recursivamente para este não-terminal apontado. A recursão termina quando os ramos apontam apenas para terminais. Por erros observados na saída do *parser* PALAVRAS, esta função indica igualmente:

- quando um não-terminal indicado como sintagma nominal não possui ramos (sintagma vazio):
  - ▶ neste caso este não-terminal é simplesmente ignorado;
- quando existe uma referência circular de não-terminais, ou seja, um ramo de um não-terminal aponta para um outro não-terminal que possui um ramo apontando para este não-terminal:
  - ▶ este problema é detectado quando um não-terminal possui um ramo apontando para um não-terminal com índice inferior ao seu próprio índice, por exemplo, o não-terminal “s1\_504” possui um ramo que aponta para o não-terminal “s1\_502”;
  - ▶ neste caso o terminal é marcado como inválido.

O pseudo-código 4.1.5 descreve genericamente esta função.

---

**Pseudo-Código 4.1.5** Função Recursiva Monta(*nt*, *sintag*)

---

```
para todos ramos r de nt
  se r aponta para um terminal t
    concatena t no final de sintag
  senão
    se r aponta para um não-terminal nt' anterior a nt
      notifica erro do PALAVRAS
      despreza sintag
    senão
      monta (nt', sintag)
```

---





- ▶ na prática, estes casos ocorrem tipicamente quando o PALAVRAS falha em identificar a estrutura de uma frase e aglutina diversas palavras em um único token;
  - ▶ este tipo de erro pode acontecer devido a estruturas complexas demais, nomes próprios muito grandes, ou, mais frequentemente, quando o texto é mal preparado pela inclusão de palavras que não pertencem a frases;
- sintagmas que possuem mais do que 256 caracteres;
  - ▶ na prática este problema acontece ou por uma acumulação de erros como o anterior (aglutinação de diversas palavras em um único token), ou ainda por causa de sintagmas que são naturalmente muito grandes;
  - ▶ cabe salientar que sintagmas muito grandes não são, naturalmente, bons candidatos a conceitos de uma ontologia;
- sintagmas que contenham numerais, tanto na sua forma numérica, por exemplo “20”, quanto na sua forma escrita, por exemplo “vinte”;
  - ▶ estes sintagmas na grande maioria das vezes também não possuem valor terminológico<sup>1</sup>;
- sintagmas que possuem dígitos ou caracteres especiais;
  - ▶ na prática, são aceitos apenas sintagmas que possuem letras maiúsculas e minúsculas, acentuadas ou não, além dos símbolos espaço em branco, hífen e sublinhado (vide tabela 3.1);
- o núcleo do sintagma é uma palavra que não é substantivo (“n”), nome próprio (“prop”), adjetivo (“adj”), nem verbo no particípio passado (“v-pcp”);
  - ▶ na prática, em nenhum dos experimentos encontrou-se esta situação, mas é possível que caso isto ocorra seja um erro do *parser*;
- sintagmas cujo o núcleo possui mais do que oito etiquetas semânticas;

---

<sup>1</sup>O *parser* anota de forma diferenciada a palavra “um” empregada como artigo indefinido e empregada como numeral cardinal, logo, esta regra de exclusão é aplicada somente quando a palavra “um” está anotada como numeral.

- ▶ na prática, em nenhum dos experimentos encontrou-se este tipo de situação (o número máximo de etiquetas encontrado foi de cinco etiquetas);
- ▶ sintagmas que por ventura possuam mais do que oito etiquetas semânticas não seriam necessariamente inválidos, mas por questões de implementação foi necessário incluir este limite.

### 4.1.7 Alteração de Sintagmas – Função Aprimora

Análogo à função Válido, esta função “corrige” os sintagmas montados segundo um conjunto de heurísticas. O pseudo-código 4.1.7 descreve genericamente esta função.

---

#### Pseudo-Código 4.1.7 Função Aprimora (*sintag*)

---

```

se sintag termina com uma conjunção
  remove a conjunção
se sintag começa com pronome
  remove este pronome
se sintag possui artigos
  remove todos artigos

```

---

Casos onde são aprimorados os sintagmas são:

- sintagmas que terminam com palavras identificadas como conjunções (“conj-c”) tem esta última palavra removida;
  - ▶ na prática, estes sintagmas são erros *benignos* do PALAVRAS, onde provavelmente dois candidatos conceitos ocupavam uma mesma unidade semântica na frase, mas o *parser* não conseguiu extrair ambos termos;
    - ▶ por exemplo, as palavras “aleitamento materno e artificial” seriam provavelmente anotadas pelo PALAVRAS como sintagma nominal apenas “aleitamento materno e”, perdendo naturalmente o termo “aleitamento artificial”, logo esta iniciativa preserva pelo menos o termo “aleitamento materno”;
- sintagmas que começam por pronomes, por exemplo, “estes”, “vários”, “algum” tem esta primeira palavra removida;

- ▶ na prática, esta iniciativa mantém apenas o substantivo (e seus complementos) ao qual o pronome se referia;
- ▶ por exemplo, o sintagma anotado pelo PALAVRAS como “vários recém-nascidos” seria aprimorado para a “recém-nascidos”;
- sintagmas que possuem artigos em suas palavras tem estes artigos removidos;
  - ▶ na prática, esta iniciativa acontece removendo artigos que precedem substantivos, ou removendo artigos que são usualmente concatenados com preposições;
  - ▶ por exemplo, o sintagma anotado pelo PALAVRAS como “o peso de o nascimento” seria aprimorado para a “peso de nascimento”.

#### 4.1.8 Finalização de Sintagma – Função Armazena

Esta função é bastante simples e consiste em salvar em um arquivo de saída o sintagma nominal extraído em uma única linha que contém:

- o sintagma na forma original;
- o sintagma na forma canônica;
- o número de palavras que ele contém;
- a palavra que é o núcleo do sintagma;
- a etiqueta sintática do núcleo; e
- a etiqueta semântica do núcleo.

A única sofisticação desta função é salvar os sintagmas extraídos em dez arquivos diferentes de acordo com o número de palavras que compõem o termo extraído. Esta escolha de dez arquivos é arbitrária, ou seja, qualquer outro número de arquivos poderia ser considerado. Para isto esta função também conta o número de palavras do sintagma de acordo com a sua forma canônica. O pseudo-código 4.1.8 descreve esta função.

---

**Pseudo-Código 4.1.8** Função Armazena (*sintag*)

---

**conta palavras** (*nro*) do *sintag* na forma canônica  
**salva** no arquivo *nro.sn* o *sintag* na forma original  
**salva** no arquivo *nro.sn* o *sintag* na forma canônica  
**salva** no arquivo *nro.sn* *nro*  
**salva** no arquivo *nro.sn* o núcleo do *sintag*  
**salva** no arquivo *nro.sn* a etiqueta sintática do núcleo do *sintag*  
**salva** no arquivo *nro.sn* a etiqueta semântica do núcleo do *sintag*

---

### 4.1.9 Cálculo das Frequências dos Sintagmas

Feito o armazenamento dos sintagmas é necessário identificar o número de sintagmas iguais, ou seja, calcular a frequência absoluta dos sintagmas. Esta tarefa é feita na ferramenta  $E\chi ATO_{LP}$  através de um processo conceitualmente simples, mas cuja a implementação prática é um pouco mais complexa para aumentar a eficiência computacional da ferramenta. Este processo está descrito no pseudo-código 4.1.9.

---

**Pseudo-Código 4.1.9** Cálculo das Frequências Absoluta e Relativa

---

```
classifica lista de sintagmas
cont = 1
para i = 2 até total de sintagmas
  se sintag(i) é igual a sintag(i-1)
    cont = cont + 1
  senão
    escreve sintag(i-1)
    escreve cont // frequência absoluta de sintag(i-1)
    escreve cont/total de sintagmas // frequência relativa de sintag(i-1)
    cont = 1
escreve sintag(i)
escreve cont // frequência absoluta de sintag(i)
escreve cont/total de sintagmas // frequência relativa de sintag(i)
```

---

De um ponto de vista conceitual este processo consiste em classificar os sintagmas de maneira que os sintagmas com a mesma forma canônica fiquem adjacentes. Estes sintagmas iguais são escritos uma única vez e o número de repetições é anotado como a frequência absoluta. Após, anota-se

a frequência relativa que é igual a frequência absoluta dividida pelo total de sintagmas. Uma vez classificados, basta percorrer os sintagmas um a um contando mais um quando se encontra sintagmas iguais, ou escrevendo os sintagmas contabilizados quando se encontra um sintagma diferente.

No pseudo-código 4.1.9, o processo é relativamente simples, porém a primeira linha (classificação da lista) é um passo bastante delicado do processo, pois o número de sintagmas extraídos costuma ser bastante grande. Cabe lembrar que a lista de sintagmas é composta de todos os sintagmas extraídos com todas as suas repetições. Portanto, para evitar um problema computacional de queda de desempenho pelo grande número de sintagmas a ordenar é necessário otimizar o processo de classificação.

Isto é feito com um processo de repartição da lista segundo diversos critérios. Inicialmente, como foi dito na seção anterior, os sintagmas são separados quanto ao número de palavras que o compõem. A seguir, os sintagmas são separados em arquivos distintos segundo a sua primeira letra.

Adicionalmente, os sintagmas são ainda repartidos em até 10 arquivos com um máximo de 2048 sintagmas em cada um para que o processo de classificação seja feito sobre um conjunto consideravelmente reduzido de sintagmas. Uma vez que cada um destes arquivos com até 2048 sintagmas é classificado, é feito um (*merge*) ou seja, os arquivos são concatenados com a devida preocupação de organizar alfabeticamente o resultado com todos os sintagmas classificados.

Todo este processo torna a classificação da lista de sintagmas um processo bastante rápido capaz de tratar *corpus* de um tamanho considerável, como pode ser visto nos exemplos de aplicação do capítulo 6. Cabe salientar que não foram utilizados pacotes de classificação pré-definidos para ter certeza sobre a exatidão de todos os processos realizados.

#### 4.1.10 Exemplo de Aplicação

Retomando o exemplo apresentado na figura 3.2, que mostra a anotação da frase “Gastrosquise é um defeito das paredes abdominais anteriores.” temos como resultado da extração os sintagmas indicados na lista da figura 4.1.

defeito_de_paredes_abdominais_anteriores	defeito_de_parede_abdominal_anterior	5	defeito	n	ac
paredes_abdominais_anteriores	parede_abdominal_anterior	3	parede	n	part-build

Figura 4.1: Exemplo de Sintagmas Nominais

Esta frase possui dois sintagmas, a saber: “o defeito das paredes abdominais anteriores” e “as paredes abdominais anteriores”. Note-se que o primeiro sintagma é composto de 5 palavras e o segundo, apenas 3.

Na figura 4.1, percebe-se inicialmente aprimoramentos feitos em ambos sintagmas pela remoção de artigos, o que causa as diferenças entre os sintagmas apresentados na forma original (primeira coluna) e canônica (segunda coluna), em seguida percebe-se o número de palavras em cada termo extraído (terceira coluna), seguidos do núcleo e suas etiquetas. Para o primeiro sintagma, temos como núcleo a palavra “defeito” que foi etiquetada sintaticamente como um substantivo (“n”) e semanticamente como uma entidade abstrata e enumerável (“ac”). Para o segundo sintagma, temos como núcleo a palavra “parede” que foi etiquetada como substantivo e como parte de um prédio (“part-build”).

Antes de observarmos os cálculos de frequência absoluta e relativa vamos considerar não apenas uma única frase mas o seguinte texto da figura 4.2 como o *corpus* completo. Este texto possui quatro frases retiradas de um *corpus* consideravelmente maior que será apresentado no capítulo 6. O resultado do processo inicial de extração sobre este pequeno texto resultou em 16 sintagmas extraídos, sendo:

- cinco unigramas: *animal*, *grupos*, *laparotomia*, *nascimento* e *pacientes*;
- três bigramas: *alças intestinais* (2 vezes) e *papel fundamental*;
- dois trigramas: *paredes abdominais anteriores* e *prognósticos dos pacientes*;
- dois quadrigramas: *alças intestinais no nascimento* e *exérese das alças intestinais*;

*Gastrosquise é um defeito da parede abdominal anterior. Entretanto, o aspecto primário das alças intestinais no nascimento parece exercer um papel fundamental no prognóstico dos pacientes. Seqüencialmente foram submetidos a laparotomia exploradora através de uma incisão xifopúbica para exérese das alças intestinais. Em nenhum animal, de ambos os grupos, observou-se qualquer anormalidade macroscópica das alças intestinais.*

Figura 4.2: Exemplo de Texto Considerado como *corpus*

- dois pentagramas: *anormalidade macroscópica das alças intestinais e defeito das paredes abdominais anteriores*; e
- dois heptagramas: *aspecto primário das alças intestinais no nascimento e incisão xifopúbica para exérese das alças intestinais*.

lista de unigramas							
animal	animal	1	animal	n	Azo	1	0.0625
grupos	grupo	1	grupo	n	HH	1	0.0625
laparotomia	laparotomia	1	laparotomia	n	?	1	0.0625
nascimento	nascimento	1	nascimento	n	event	1	0.0625
pacientes	paciente	1	paciente	n	H	1	0.0625

lista de bigramas							
alças_intestinais	alça_intestinal	2	alça	n	cord	2	0.1250
papel_fundamental	papel_fundamental	2	papel	n	ac	1	0.0625

lista de trigramas							
paredes_abdominais_anteriores	parede_abdominal_anterior	3	parede	n	part-build	1	0.0625
prognóstico_de_pacientes	prognóstico_de_paciente	3	paciente	n	H	1	0.0625

lista de quadrigramas							
alças_intestinais_em_nascimento	alça_intestinal_em_nascimento	4	nascimento	n	event	1	0.0625
exérese_de_alças_intestinais	exérese_de_alça_intestinal	4	alça	n	cord	1	0.0625

lista de pentagramas							
anormalidade_macroscópica_de_alças_intestinais	anormalidade_macroscópica_de_alça_intestinal	5	alça	n	cord	1	0.0625
defeito_de_paredes_abdominais_anteriores	defeito_de_parede_abdominal_anterior	5	parede	n	part-build	1	0.0625

lista de heptagramas							
aspecto_primário_de_alças_intestinais_em_nascimento	aspecto_primário_de_alça_intestinal_em_nascimento	7	nascimento	n	event	1	0.0625
incisão_xifopúbica_para_exérese_de_alças_intestinais	incisão_xifopúbica_para_exérese_de_alça_intestinal	7	alça	n	cord	1	0.06250

Figura 4.3: Sintagmas Nominais Extraídos do Texto da Figura 4.2

Aplicando a extração de sintagmas da ferramenta  $E\chi ATOLP$  conforme descrito anteriormente, obteve-se seis listas de sintagmas como apresentado na figura 4.3. Nesta figura temos em cada uma das colunas, respectivamente:

- o sintagma na forma original segundo o *parser* PALAVRAS;
- o sintagma na forma canônica segundo o *parser* PALAVRAS;
- o número de palavras que o compõem calculado pela ferramenta  $E\chi ATOLP$ , de acordo com a forma canônica e não considerando palavras compostas como uma única palavra;
- o núcleo do sintagma na sua forma canônica;
- a etiqueta sintática atribuída pelo *parser* PALAVRAS;
- a etiqueta semântica atribuída pelo *parser*, colocando o símbolo “?” quando nada foi informado pelo PALAVRAS;



- a frequência absoluta do sintagma;
- a frequência relativa (a frequência absoluta dividida pelo número total de termos).

## 4.2 Módulos Acessórios da Ferramenta

Os módulos acessórios da ferramenta  $E\chi ATOLP$  são responsáveis por executar tarefas que estão relacionadas com a função principal de extração de termos, porém não fazem parte direta desta função. As tarefas acessórias são a aplicação de pontos de corte (seção 4.2.1 - módulo cortador), a comparação de listas e cálculo de métricas (seção 4.2.2 - módulo comparador) e a localização de termos no *corpus* (seção 4.2.3 - módulo localizador).

### 4.2.1 Módulo Cortador

Este módulo da ferramenta  $E\chi ATOLP$  é responsável por aplicar técnicas de ponto de corte as listas de sintagmas extraídos. Seja qual for das opções de ponto de corte (veja seção 3.1.2), este módulo se resume a percorrer a lista de sintagmas organizada pela frequência e selecionar:

- os termos que foram extraídos em um número mínimo de vezes, para ponto de corte absoluto, segundo a frequência absoluta;
  - ▶ por exemplo, selecionar os sintagmas com frequência relativa maior ou igual a  $10^{-4}$ ;
- os termos que tem uma frequência relativa igual ou superior a um limiar, para ponto de corte absoluto segundo a frequência relativa;
  - ▶ por exemplo, selecionar os sintagmas com frequência absoluta maior ou igual a 4;
- um número específico de termos, para ponto de corte absoluto único;
  - ▶ por exemplo, selecionar os primeiros 300 sintagmas;
- um percentual dos termos, para ponto de corte relativo.
  - ▶ por exemplo, selecionar os primeiros 10% sintagmas;

Enquanto as duas primeiras opções são inequívocas, pois fica claro quais termos devem ser mantidos, as duas últimas opções deixam margem a dúvidas. Por exemplo, caso seja escolhido manter apenas um número fixo de termos é fácil imaginar que possa existir a necessidade de um critério adicional à frequência.

Tendo como exemplo a lista de sintagmas extraídos na seção 4.1.10 onde temos 15 sintagmas distintos extraídos, se escolhêssemos um ponto de corte com apenas os 5 sintagmas mais frequentes teríamos uma decisão a tomar. Dos sintagmas extraídos neste exemplo, “alças intestinais” seria sem dúvida um dos 5 mais frequentes, porém a frequência não pode indicar qual dos demais 14 deveriam ser escolhidos. Neste sentido, a ferramenta  $E\chi ATOLP$  oferece duas opções: ou os sintagmas podem ser escolhidos segundo uma classificação alfabética, ou o número de termos pode ser estendido para incluir todos os sintagmas que possuem frequência igual ao último dos escolhidos.

Na primeira opção teríamos como resultado para a escolha dos 5 sintagmas mais frequentes do exemplo da seção 4.1.10: *alças intestinais*, *alças intestinais em nascimento*, *animal*, *anormalidade macroscópica de alças intestinais* e *aspecto primário de alças intestinais em nascimento*.

Na segunda opção, iniciariamos com os mesmos cinco termos, mas como o quinto (e último) destes termos tem frequência absoluta igual a 1, todos os demais sintagmas extraídos também seriam incluídos na lista.

Evidentemente, para o caso de ponto de corte relativo o mesmo tipo de decisão também deve ser tomada. Por exemplo, se fosse escolhido um ponto de corte de 20% dos sintagmas extraídos no exemplo da seção 4.1.10 teríamos os sintagmas *alças intestinais*, *alças intestinais em nascimento* e *animal* (com a classificação alfabética) ou todos os 15 sintagmas (com a inclusão dos sintagmas que possuem a mesma frequência).

## 4.2.2 Módulo Comparador

O módulo comparador recebe duas listas ( $\mathcal{A}$  e  $\mathcal{B}$ ) e calcula através de comparação de strings entre os termos das listas quais termos são comuns entre ambas, ou seja a intersecção entre elas ( $\mathcal{A} \cap \mathcal{B}$ ). A partir desta informação são calculadas as demais possíveis saídas deste módulo (veja a seção 3.1.3).

Sabendo a intersecção entre as listas calcula-se os termos presentes em apenas uma das listas, ou seja,  $\mathcal{A} - \mathcal{B}$  e  $\mathcal{B} - \mathcal{A}$ . Em seguida, gera-se a lista de união ( $\mathcal{A} \cup \mathcal{B}$ ) com a simples concatenação das três listas anteriormente geradas ( $\mathcal{A} \cap \mathcal{B}$ ,  $\mathcal{A} - \mathcal{B}$  e  $\mathcal{B} - \mathcal{A}$ ).

Por exemplo, as listas da figura 4.4 submetidas ao módulo comparador resultariam nas listas apresentadas na figura 4.5.

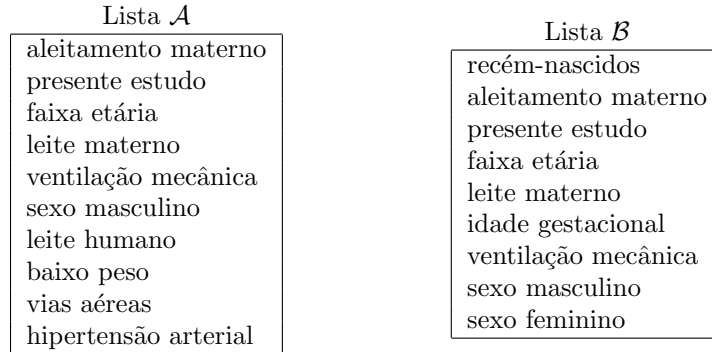


Figura 4.4: Exemplo de Lista de Termos a Comparar

Uma vez geradas as listas, o módulo comparador calcula métricas usuais de comparação apresentadas na seção 3.1.4. Para estes cálculos é preciso assumir uma das listas de entrada como lista de referência e a outra como lista de termos extraídos.

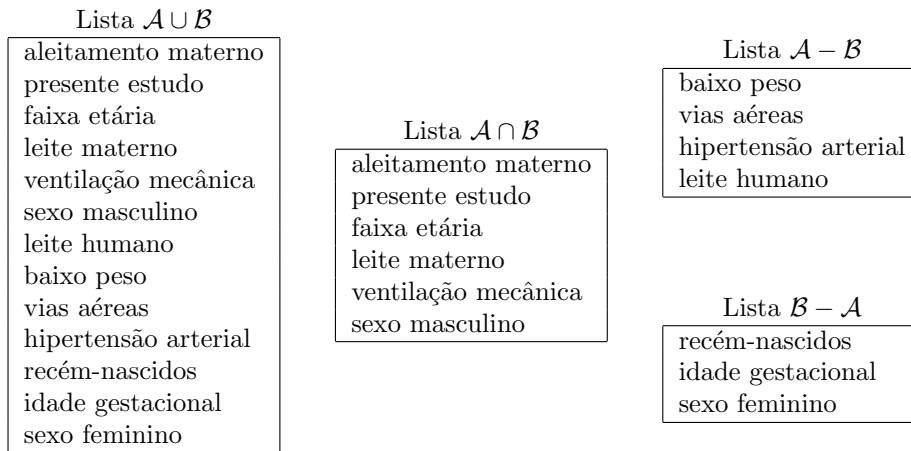


Figura 4.5: Exemplo de Resultados de Comparação

Por exemplo, para as listas apresentadas na figura 4.5, assumindo a lista  $\mathcal{A}$  como lista de referência o módulo comparador retornaria:

$$P = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{B}|} = \frac{6}{9} = 0.666667 \quad A = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A}|} = \frac{6}{10} = 0.6$$

$$F = \frac{2 \times P \times A}{P + A} = \frac{2 \times 0.66667 \times 0.6}{0.66667 + 0.6} = 0.631579$$

### 4.2.3 Módulo Localizador

O módulo localizador permite buscar um sintagma informado em sua forma normal ou canônica nos textos do *corpus*. Basicamente este módulo percorre todos textos repetindo um processo análogo ao processo de extração. No entanto, ao invés de extrair os sintagmas, este módulo apenas identifica-os e informa se este sintagma seria extraído pelo  $E\chi ATOLP$ , além de obviamente relatar em que textos ele foi encontrado.

Do ponto de vista de implementação prática, este módulo lê o sintagma em um arquivo textual de entrada e salva em um arquivo de relatório cada vez que ele foi encontrado, em qual texto, em qual frase e como ele seria extraído. A figura 4.6 apresenta um exemplo de relatório textual do módulo localiza aplicado a um *corpus*.

Sintagma 'alça_intestinal' (aceito) na forma canonica:	<i>corpus/03-79-06-525port.xml</i> (frase 2)
Sintagma 'alça_intestinal' (aceito) na forma canonica:	<i>corpus/03-79-06-525port.xml</i> (frase 5)
Sintagma 'alça_intestinal' (aceito) na forma canonica:	<i>corpus/03-79-06-525port.xml</i> (frase 6)
Sintagma 'alça_intestinal' (aceito) na forma canonica:	<i>corpus/03-79-06-525port.xml</i> (frase 26)
Sintagma 'alça_intestinal' (aceito) na forma canonica:	<i>corpus/03-79-06-525port.xml</i> (frase 43)
Sintagma 'alça_intestinal' (aceito) na forma canonica:	<i>corpus/03-79-06-525port.xml</i> (frase 67)
Sintagma 'alça_intestinal' (aceito) na forma canonica:	<i>corpus/03-79-06-525port.xml</i> (frase 70)
Sintagma 'alça_intestinal' (aceito) na forma canonica:	<i>corpus/04-80-01-77port.xml</i> (frase 17)
Sintagma 'alça_intestinal' (aceito) na forma canonica:	<i>corpus/04-80-01-77port.xml</i> (frase 19)

Figura 4.6: Exemplo de Relatório de Localização

Este exemplo da figura 4.6 mostra que o sintagma '*alça\_intestinal*' foi encontrado e aceito, ou seja, seria extraído corretamente, em 9 ocorrências no *corpus* informado. Destas ocorrências, 7 aconteceram no arquivo "*corpus/03-79-06-525port.xml*" (frases 2, 5, 6, 26, 43, 67 e 70) e outras 2 no arquivo "*corpus/04-80-01-77port.xml*" (frases 17 e 19).

# Capítulo 5

## Interface com o Usuário

A interface da ferramenta  $E\chi ATOLP$  ainda se encontra em fase experimental, porém as suas funcionalidades principais já estão disponíveis e implementadas com uma interface textual. Estas funcionalidades são: uma operação básica de extração de termos (seção 5.1) e operações acessórias (seção 5.2).

### 5.1 Extração de Termos

A extração de termos é feita conforme descrito nas seções anteriores (3.1.1 e 4.1). Nesta operação assume-se a existência de um conjunto de arquivos no formato TIGER-XML com os textos do *corpus* a ser manipulado em um diretório denominado “*corpus*”. Este diretório deve estar no diretório raiz onde a ferramenta  $E\chi ATOLP$  está sendo executada. Ao escolher executar esta operação, evidentemente, todo o processo de extração é feito com a geração de um conjunto de arquivos com o formato apresentado, por exemplo, nas figuras 3.4 e 4.3.

Esta é a execução mínima da operação extração de termos. Porém o usuário pode selecionar incrementalmente as seguintes operações adicionais a serem realizadas após este processo:

1. selecionar sintagmas segundo o número de termos;
2. aplicar pontos de corte e extrair sintagmas “limpos” na forma original ou canônica;
3. comparar os sintagmas extraídos com listas de referência e calcular métricas.

### 5.1.1 Seleção do Número de Palavras

Esta opção elementar permite ao usuário indicar quais sintagmas deverão ser efetivamente extraídos segundo o número de palavras que os compõem. Conforme dito anteriormente a operação básica de extração agrupa os termos extraídos em 10 diferentes arquivos de saída. Estes arquivos são salvos em sua forma básica no diretório “*sintag*” nos arquivos *?sn* que contém os unigramas (para *?* igual a 1), para os bigramas (para *?* igual a 2) e assim por diante até eneigramas (para *?* igual a 9) e *M.sn* que contém os n-gramas com 10 ou mais palavras.

O resultado desta seleção é escolher o tratamento de somente alguns destes arquivos. A interface da ferramenta *EχATOLP* permite selecionar cada um destes 10 conjuntos. Por exemplo, se quisermos observar apenas bigramas e trigramas extraídos seleciona-se apenas estas opções como demonstra a tela apresentada na figura 5.1.

Escolha os sintagmas a serem tratados	
<input type="radio"/> unigramas	<input type="radio"/> hexigramas
<input checked="" type="radio"/> bigramas	<input type="radio"/> heptigramas
<input checked="" type="radio"/> trigramas	<input type="radio"/> octigramas
<input type="radio"/> quadrigramas	<input type="radio"/> eneigramas
<input type="radio"/> pentigramas	<input type="radio"/> multigramas (10 ou mais)

Figura 5.1: Tela de Seleção de Sintagmas pelo Número de Palavras

### 5.1.2 Aplicação de Pontos de Corte e Limpeza

Conforme discutido na seção 4.2.1, a aplicação de pontos de corte deve informar um dos critérios possíveis, ou seja:

- informar uma frequência relativa mínima;
- informar uma frequência absoluta mínima;
- informar um número absoluto de termos a considerar e como proceder com termos tão frequentes quanto o último a ser considerado;
- informar um percentual da lista de sintagmas e como proceder com termos tão frequentes quanto o último a ser considerado.

De acordo com a opção escolhida, o ponto de corte é aplicado as listas de sintagmas selecionadas na operação de seleção de número de palavras. Por exemplo, caso seja escolhido um ponto de corte com uma frequência absoluta mínima de 4 ocorrências teríamos esta escolha feita como apresentado na tela da figura 5.2. Nesta tela, apesar de não escolhidas, temos também indicados parâmetros para as demais opções de pontos de corte. Para frequência relativa temos indicado que sintagmas com frequência superior ou igual a  $10^{-4}$  seriam mantidos. Para número absoluto de termos, seriam aceitos 300 termos e os termos seguintes (a partir do 301º), que possuísem a mesma frequência que o 300º termo, também seriam mantidos. Para percentual de termos, temos indicado que apenas 10% (exatamente) dos termos seriam mantidos.

Escolha o ponto de corte a aplicar			
<input type="radio"/>	nenhum ponto de corte		
<input type="radio"/>	frequência relativa mínima	<input type="text" value="0.0001"/>	
<input checked="" type="radio"/>	frequência absoluta mínima	<input type="text" value="4"/>	
<input type="radio"/>	número absoluto de termos	<input type="text" value="300"/>	<input checked="" type="radio"/> inclui equifrequentes
<input type="radio"/>	percentual de termos	<input type="text" value="10"/>	<input type="radio"/> inclui equifrequentes
Escolha que lista de termos gerar			
<input type="radio"/>	termos na forma original		
<input checked="" type="radio"/>	termos na forma canônica		

Figura 5.2: Tela de Escolha de Ponto de Corte e Limpeza

Esta operação adicional permite ainda extrair listas com termos nas formas original ou canônica, limpos de todas as demais informações contidas inicialmente nos arquivos *?sn* que contém diversas informações relativas ao núcleo do sintagma e frequências de ocorrência. Cabe lembrar que os arquivos *?sn* possuem os termos tanto na forma original quanto na forma canônica conforme aparece nas figuras 4.1 4.3. Esta operação visa facilitar a leitura humana dos termos, mas também possibilitar a próxima operação adicional de comparação com listas de referência que pode ser feita tanto com os termos na forma original, quanto na forma canônica.

Os arquivos gerados são salvos no diretório “*listas*” com o formato “*cnc?.lst*” e “*org?.lst*” para as listas na forma canônica e original respectivamente, onde o caracter variável é o número de palavras. Desta forma, segundo as escolhas prévias feitas na operação de seleção de número de palavras, teremos no diretório “*listas*” o salvamento dos arquivos “*cnc2.lst*” com os bigra-

mas e “*cnc3.lst*” com os trigramas que atendem a escolha de ponto de corte, ou seja, contendo somente sintagmas que foram identificados pelo menos 4 vezes no *corpus*.

### 5.1.3 Comparação com Referência e Métricas

Para esta operação deve ser informado uma lista de termos de referência para cada conjunto de sintagmas agrupado pelo número de palavras que foi tratado até o momento. Além disto deve ser dito que os resultados da comparação devem ser salvos. Estes resultados são salvos em arquivos com a extensão “.*lst*” no diretório “*listas*” onde também devem se encontrar as listas de referências.

A figura 5.3 apresenta um exemplo de tela onde escolhe-se a comparação com listas de referências “*ref2.lst*” “*ref3.lst*”. Nesta tela também é indicado quais arquivos serão salvos dentre as opções:

- positivos – termos presentes na lista de extraídos ( $LE$ ) e na lista de referência ( $LR$ ), ou seja, a intersecção das listas;
- falsos positivos – termos presentes na lista de extraídos, mas ausentes da lista de referência ( $LE - LR$ );
- falsos negativos – termos ausentes da lista de extraídos, mas presentes na lista de referência ( $LR - LE$ );
- total – termos presentes na lista de extraídos ou na lista de referência, ou seja, a união das listas.

Informe as listas de referência para	
<input checked="" type="checkbox"/> bigramas	<input type="text" value="ref2.lst"/>
<input checked="" type="checkbox"/> trigramas	<input type="text" value="ref3.lst"/>
Escolha que arquivos salvar	
<input checked="" type="checkbox"/> positivos (intersecção)	
<input type="checkbox"/> falsos positivos	
<input checked="" type="checkbox"/> falsos negativos	
<input type="checkbox"/> total (união)	

Figura 5.3: Tela de Comparação e Cálculo de Métricas



No exemplo da figura 5.3 foi escolhido o salvamento da intersecção e dos termos não encontrados na lista de extraídos, mas presentes na lista de referência. O formato do nome de cada um dos arquivos a ser salvo é o nome de cada um dos arquivos comparados com a menção “\_I\_” entre eles para intersecção, “\_D\_” para diferenças e “\_U\_” para união. Desta forma, seguindo o exemplo desta seção seriam gerados os seguintes arquivos no diretório “*listas*”:

- *cnc2\_I\_ref2.lst* com os bigramas presentes na lista de referência e na lista de extraídos;
- *ref2\_D\_cnc2.lst* com os bigramas presentes na lista de referência, mas ausentes da lista de extraídos;
- *cnc3\_I\_ref3.lst* com os trigramas presentes na lista de referência e na lista de extraídos;
- *ref3\_D\_cnc3.lst* com os trigramas presentes na lista de referência, mas ausentes da lista de extraídos.

Finalmente, as métricas (precisão, abrangência e *f-measure*) calculadas para as comparações feitas são salvas em um arquivo de relatório chamado “*metricas.txt*” que contém em cada linha uma métrica calculada para uma das comparações feitas. Para o exemplo citado uma possível saída do arquivo “*metricas.txt*” é apresentada na figura 5.4.

Precisão	cnc2.lst - ref2.lst:	53.63% =	702 / 1309
Abrangência	cnc2.lst - ref2.lst:	50.00% =	702 / 1404
F-measure	cnc2.lst - ref2.lst:	51.75%	
Precisão	cnc3.lst - ref3.lst:	44.25% =	285 / 644
Abrangência	cnc3.lst - ref3.lst:	38.99% =	285 / 731
F-measure	cnc3.lst - ref3.lst:	41.45%	

Figura 5.4: Exemplo de Saída com Cálculo de Métricas

## 5.2 Operações Acessórias

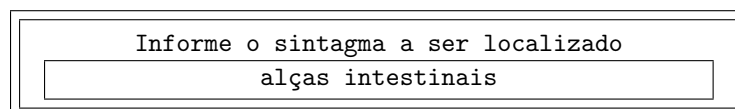
As operações acessórias são disponibilizadas na ferramenta *EχATOLP* para possibilitar algumas operações que fogem ao usual. Estas operações permitem

localizar termos em textos específicos do *corpus* (seção 5.2.1), aplicar novos pontos de corte a listas previamente extraídas (seção 5.2.2) e fazer novas comparações de listas (seção 5.2.3).

### 5.2.1 Localização de Termos

Nesta operação acessória assume-se a existência de um conjunto de arquivos no formato TIGER-XML com os textos do *corpus* em um diretório denominado “*corpus*”, que se encontre no diretório raiz no qual a ferramenta  $E\chi A T O L P$  está sendo executada. Desta forma, utiliza-se exatamente a mesma entrada da operação básica de extração descrita na seção 5.1.

Esta operação utiliza o módulo localizador descrito anteriormente na seção 4.2.3. De um ponto de vista de interface, a única diferença está no fato da informação do sintagma a ser procurado ser perguntado diretamente ao usuário. A figura 5.5 exemplifica a busca pelo sintagma “alças intestinais”.



A interface de usuário para a localização de sintagmas. Ela consiste em um formulário com duas linhas de entrada de texto. A primeira linha contém o texto "Informe o sintagma a ser localizado" e a segunda linha contém o texto "alças intestinais".

Figura 5.5: Tela de Localização de Sintagma

O relatório da localização é apresentado textualmente através de um arquivo a semelhança do exemplo apresentado anteriormente na figura 4.6.

### 5.2.2 Aplicação de Pontos de Corte

A operação acessória de aplicação de pontos de corte é análoga àquela apresentada anteriormente no contexto da extração de termos (seção 5.1.2). Porém neste caso é possível recortar qualquer lista já existente. Usualmente esta operação acessória é válida para aplicar diferentes pontos de corte buscando refinar uma lista de termos previamente extraídos.

A figura 5.6 apresenta a tela que executa esta operação para a lista “*cnc2.lst*” onde se aplica um ponto de corte que mantém apenas 10% dos termos presentes nesta lista. Esta nova lista reduzida de 90% dos seus termos é salva em um arquivo chamado “*cnc2\_10.lst*”.

Note-se que como esta operação se aplica a listas de termos “limpos”, ou seja, listas sem informação da frequência de cada um dos termos, apenas as opções de um número absoluto e de um percentual são possíveis.

Escolha a lista para aplicar pontos de corte	
cnc2.lst	
Escolha o ponto de corte a aplicar	
<input type="radio"/> número absoluto de termos	300
<input checked="" type="radio"/> percentual de termos	10
Escolha o nome da nova lista a ser salva	
cnc2_10.lst	

Figura 5.6: Tela de Escolha de Ponto de Corte de Lista Individual

### 5.2.3 Comparação de Listas

Igualmente à operação descrita na seção anterior, a operação utilitária de comparação de listas é uma repetição da comparação de listas feita como sequência da extração de termos. Neste contexto, porém, devem ser informadas as duas listas a serem comparadas e qual tipo de saída deve ser salva.

A figura 5.7 apresenta a tela de diálogo para executar esta operação utilitária. Neste exemplo, compara-se a lista “*org2.lst*” (utilizada como referência) com a lista “*cnc2.lst*” e apenas busca-se observar as métricas desta comparação.

Informe a lista de referência	
org2.lst	
Informe a lista ser comparada com a referência	
cnc2.lst	
Escolha que arquivos salvar	
<input checked="" type="radio"/> nenhum (apenas calcula métricas)	
<input type="radio"/> positivos (intersecção)	
<input type="radio"/> falsos positivos	
<input type="radio"/> falsos negativos	
<input type="radio"/> total (união)	

Figura 5.7: Tela de Comparação de Listas Individuais

## Capítulo 6

# Exemplos de Utilização

Um exemplo de utilização da ferramenta  $E\chi ATOLP$  na sua versão atual foi aplicado a um *corpus* da área de Pediatria. Este *corpus* é composto por 283 textos em português extraídos da revista bilíngue *Jornal de Pediatria* (<http://www.jpmed.com.br/>), num total de 785.448 palavras.

Para analisar a eficiência do processo, é necessário que haja uma lista de termos de referência. A lista de termos de referência foi construída pelo projeto TEXTQUIM-TEXTECC da Universidade Federal do Rio Grande do Sul ([www.ufrgs.br/textecc](http://www.ufrgs.br/textecc)). O objetivo deste projeto era a elaboração de dois glossários para apoio aos estudantes de tradução. Para a identificação dos itens desses glossários, o grupo realizou uma extração de  $n$ -gramas de textos puros (sem anotação lingüística) do *corpus* de Pediatria. Nesse processo foram aproveitados apenas os  $n$ -gramas com mais de 4 ocorrências no *corpus*.

A partir de uma lista de 36.741  $n$ -gramas, partiu-se para um processo de filtragem baseado em heurísticas que resultou em uma lista com 3.645  $n$ -gramas considerados possivelmente relevantes para integrar os glossários. Essas heurísticas foram desenvolvidas com o objetivo de excluir grupos de palavras que não fossem apropriados para gerar verbetes, por exemplo, termos que começavam ou terminavam por preposições, como em *para aleitamento materno*, foram transformados pela exclusão destas preposições, e termos que iniciavam com verbos foram excluídos, como em *promover o aleitamento*.

Uma etapa posterior às heurísticas foi a avaliação manual sobre a relevância dos termos realizada por 5 estudantes de tradução com algum conhecimento do domínio. Esse resultado foi novamente refinado por meio

de uma verificação manual com o objetivo de tornar a referência mais adequada ao propósito de criação de uma ontologia (definição de conceitos), visto que o objetivo inicial era a construção de glossários para estudantes de tradução. Finalmente, obteve-se uma lista com 2.135 termos, sendo 1.404 bigramas, 731 trigramas. Termos de composição maior que 3 palavras não foram considerados para as listas de referência.

## 6.1 Resultados Numéricos

Aplicando a extração da ferramenta  $E\chi ATOLP$  a este *corpus* foram extraídos 141.431 sintagmas repartidos segundo o número de palavras que o compõem de acordo com a tabela 6.1. Nesta tabela a coluna **extraídos** indica o total de sintagmas devidamente identificados nos textos. A coluna **distintos** indica quantos sintagmas distintos existem no grupo total de sintagmas extraídos. A coluna **após corte** indica quantos destes sintagmas distintos aparecem pelo menos 4 vezes (ponto de corte por frequência absoluta). A última linha contabiliza os sintagmas indicados em cada coluna.

Tabela 6.1: Sintagmas Extraídos pela Ferramenta  $E\chi ATOLP$

<b>termos</b>	<b>extraídos</b>	<b>distintos</b>	<b>após corte</b>
unigramas	38489	4523	1497
bigramas	29728	15194	1309
trigramas	21339	14858	644
quadrigramas	14186	12260	166
pentigramas	9443	8764	60
hexigramas	6686	6480	12
heptigramas	4879	4786	2
octigramas	3579	3521	1
eneigramas	2672	2634	1
multigramas	10430	10274	4
<b>total</b>	141431	83294	3696

Pode ser observado pelos números da tabela 6.1 que de fato os sintagmas compostos por um grande número de palavras são pouco frequentes. Porém, a análise comparativa feita neste relatório leva em conta a existência de listas de referência, logo foca-se exclusivamente nos bigramas e trigramas extraídos.

As primeiras linhas da tabela 6.2 apresenta os valores resultantes das métricas obtidas através da comparação das listas extraídas com este ponto de corte por frequência absoluta com as listas de referência. Nesta tabela, estão indicados o número de termos na lista de extraídos ( $|LE|$ ), o número de termos na lista de referência ( $|LR|$ ), o número de termos na intersecção destas listas ( $|LE \cap LR|$ ), e as métricas de precisão ( $P$ ), abrangência ( $A$ ) e  $f$ -measure ( $F$ ) calculados a partir destas informações.

Tabela 6.2: Comparação de Métricas do  $E\chi ATOLP$  com Outras Ferramentas

$E\chi ATOLP$						
termos	$ LE $	$ LR $	$ LE \cap LR $	$P$	$A$	$F$
bigramas	1309	1404	702	53.63%	50.00%	51.75%
trigramas	644	731	285	44.25%	38.99%	41.45%
OntoLP						
termos	$ LE $	$ LR $	$ LE \cap LR $	$P$	$A$	$F$
bigramas	964	1404	219	22.72%	15.60%	18.50%
trigramas	284	731	101	35.56%	13.82%	19.90%
NSP						
termos	$ LE $	$ LR $	$ LE \cap LR $	$P$	$A$	$F$
bigramas	3709	1404	1230	33.16%	87.61%	48.11%
trigramas	2550	731	556	21.80%	76.16%	33.90%

Os resultados preliminares do uso da ferramenta  $E\chi ATOLP$  mostram um desempenho bastante interessante quando comparado a outras ferramentas similares. A tabela 6.2 apresenta também uma comparação dos resultados obtidos com o *corpus* apresentado no capítulo 6 com a aplicação de outras duas ferramentas:

- OntoLP [21] utilizando o processo linguístico igualmente baseado em sintagmas nominais;
- NSP [3] utilizando um processo puramente estatístico.

Os resultados desta tabela mostram que a ferramenta  $E\chi ATOLP$  apresentou uma precisão superior às demais. No entanto, a abrangência foi inferior àquela obtida pela ferramenta NSP. Apesar disto, a combinação destas métricas expressa pela  $f$ -measure obtida pela ferramenta  $E\chi ATOLP$  foi superior tanto ao NSP, quanto ao OntoLP.

# Capítulo 7

## Conclusão

Este trabalho de construção de uma ferramenta para extração de termos é bastante complexo e envolve diversas decisões baseadas em heurísticas que tem forte base linguística, mas também envolve decisões de programação e manipulação de dados. Os resultados desta experiência de construção de uma ferramenta só podem ser apreciados pela sua aplicação e conseqüente comparação com outras ferramentas análogas conforme foi brevemente descrito na seção anterior e objeto de um artigo científico submetido recentemente [18].

Desta forma, a análise da validade da ferramenta proposta passa por discussões sobre esta comparação sumarizada na tabela 6.2. Ignorando as dificuldades da anotação linguística de um *corpus*, podemos concluir que a abordagem linguística utilizada pelo  $E\chi ATOLP$  fornece melhores resultados e portanto é a mais indicada para o objetivo de identificação de conceitos na construção automática de ontologias. Cabe salientar que a dificuldade de anotação sintática consiste em ter uma ferramenta (*parser*) confiável e de fácil adaptação de sua saída para a ferramenta de extração linguística. Na verdade, alguns problemas encontrados em experimentos preliminares [18] podem ser explicados, talvez, por erros de anotação herdados do *parser* PALAVRAS.

Em resumo, é possível afirmar que havendo confiança na ferramenta de anotação linguística e sua adaptação ao extrator de termos, é mais interessante usar uma abordagem linguística como a busca por sintagmas nominais utilizada no  $E\chi ATOLP$ . Obviamente, esta afirmação esta baseada no estudo de um único *corpus* comparado a uma lista de referência desenvolvida manualmente. Portanto, um trabalho futuro natural ao descrito neste relatório é

o estudo das ferramentas citadas a outros *corpus* para reforçar esta afirmação.

Outra sequência natural é continuar o processo de desenvolvimento do  $E\chi ATOLP$  incluindo novas ou aprimorando as heurísticas de extração. Por exemplo, seria possível imaginar refinamentos que permitam aumentar a abrangência obtida se aproximando da abrangência que ferramentas baseadas em abordagens estatísticas fornecem. No entanto, só após estas experiências será possível avaliar se este provável aumento de abrangência não irá implicar em variações significativas de precisão.

Cabe salientar que, no contexto do projeto de pesquisa em que o desenvolvimento desta ferramenta se insere, todas estas preocupações em otimizar o processo de extração de termos visa oferecer uma base confiável para a etapa posterior de construção de ontologias que é a determinação de hierarquia entre os conceitos. Porém, os resultados desta ferramenta não se limitam a esta aplicação.

Finalmente, cabe lembrar que a ferramenta  $E\chi ATOLP$  ainda se encontra em uma versão experimental a qual facilidades de interface mais sofisticadas deverão ser agregadas. Apesar disto, os resultados práticos da ferramenta já são significativos e esforços de pesquisa e desenvolvimento para esta ferramenta parecem claramente justificados.



# Referências Bibliográficas

- [1] AUBIN, S.; HAMON, T. Improving term extraction with terminological resources. FinTAL 2006, LNAI 4139, pp. 380-387, 2006.
- [2] BARONI, M.; BERNADINI, S. BootCaT: Bootstrapping Corpora and Terms from the Web. Proceedings of the 4th LREC, pp.1313-1316, 2004.
- [3] BANBERJEE, S.; PEDERSEN, T. The Design, Implementation, and Use of the Ngram Statistics Package. In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Feb., 2003, Mexico City.
- [4] BECHARA, Evanildo. Moderna gramática portuguesa : cursos de 1º e 2º graus. 25. ed. São Paulo : Cia. Editora Nacional, 1980. 374 p.
- [5] BICK, E. The parsing System “Palavras”: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. PhD thesis, Arhus University, 2000.
- [6] BICK, E. et alli. World of VISL – Portuguese. <http://visl.sdu.dk/visl/pt/index.php> (acessado em 30 de maio de 2009).
- [7] BOURIGAULT, D. UPERY : un outil d’analyse distributionnelle étendue pour la construction d’ontologies à partir de corpus, TALN, Nancy 2002.
- [8] BOURIGAULT, D.; FABRE, C.; FRÉROT,C.; JACQUES,M.; OZDOWSKA, S. SYNTEX, analyseur syntaxique de *corpus*, TALN, Dourdan 2005.

- [9] BUITELAAR, P.; CIMIANO, P.; MAGNINI, B. Ontology learning from text: An overview. In: P-Buitelaar, Cimiano, P.; and Magnini, B. (Ed.). *Ontology Learning from Text: Methods, Evaluation and Applications*, v. 123 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2005.
- [10] ESTOPÁ BAGOT, R. *Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada)*. Tese de Doutorado. Universidade Pompeu Fabra, 1999.
- [11] FORTUNA, B; LAVRAC, N.; VELARDI, P. Advancing topic ontology learning through term extraction. *PRICAI 2008, LNAI 5351*, pp. 626-635, 2008.
- [12] GENNARI, J. et al. The evolution of protégé: an environment for knowledge-based systems development. 2002. Technical Report SMI-2002-0943.
- [13] KONIG, E.; LEZIUS, W. The TIGER language - A Description Language for Syntax Graphs. Formal Definition. Technical report, IMS, University of Stuttgart, 2003.
- [14] KONIG, E.; LEZIUS, W. VOORMANN, H. *TIGERSearch 2.1 - User's Manual*, IMS, University of Stuttgart, 2003. disponível em: <http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/doc/html/> (acessado em 30 de maio de 2009).
- [15] KURAMOTO, H. Nominal Groups: a New Purpose to Information Retrieval. *DataGramZero - Revista de Ciência da Informação - v.3 n.1* Fev., 2002.
- [16] LOPES, L.; VIEIRA, R.; FINATTO, M. J.; ZANETTE, A.; MARTINS, D.; RIBEIRO JR, L. C. Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area. *RECIIS*, v.3, n.1, p.72-84, 2009
- [17] LOPES, L.; VIEIRA, R.; FINATTO, M. J.; MARTINS, D. Extracting Compound Terms from Domain Corpora: An experiment in health care.

28th International Conference on Conceptual Modeling - GRAMADO, BRAZIL, 2009 (submetido)

- [18] LOPES, L.; OLIVEIRA, L. H. M.; VIEIRA, R. Análise Comparativa de Métodos de Extração de Termos: Abordagens Linguística e Estatística - The 7th Brazilian Symposium in Information and Human Language Technology, São Carlos, Brazil, 2009. (submetido)
- [19] MCGUINNESS, D.L. VAN HARMELEN, F. OWL web ontology language overview. World Wide Web Consortium (W3C) recommendation. <http://www.w3.org/TR/owl-features/>. Acesso em: 01 fev. 2004.
- [20] PERINI, M.A. Princípios de linguística descritiva : introdução ao pensamento gramatical, 206 p. Parábola, São Paulo, 2007.
- [21] RIBEIRO, L.C. OntoLP: Construção semi-automática de ontologias a partir de textos da língua portuguesa. 2008. Dissertação (Mestrado em Computação Aplicada), Universidade do Vale do Rio dos Sinos - UNISINOS, São Leopoldo.
- [22] SOUZA-E-SILVA, M. C. P. ; KOCH, I. V. LINGUÍSTICA APLICADA AO PORTUGUÊS: SINTAXE. Cortez Editora, p.160, São Paulo, 1983.
- [23] TELINE, M. F. Avaliação de métodos para extração automática de terminologia de textos em português. Dissertação de Mestrado. ICMC/USP, 2004.