

# PAPEL

---

Palavras Associadas Porto Editora Linguateca

Apresentação das relações extraídas

Hugo Gonçalo Oliveira, Paulo Gomes  
Linguateca, pólo de Coimbra, DEI - FCTUC, CISUC

---

Dezembro 2008

# Índice

1	Introdução . . . . .	2
2	Metodologia . . . . .	3
	2.1 Procura dos nós com as palavras relacionadas . . . . .	3
	2.2 Processamento por categoria gramatical . . . . .	4
	2.3 Construção das gramáticas . . . . .	4
3	Apresentação das relações . . . . .	5
	3.1 Características comuns a todas as gramáticas . . . . .	5
	3.2 As relações . . . . .	6
	3.3 Visão quantitativa . . . . .	14
4	Limitações e pistas para o futuro . . . . .	16
	4.1 Hiperonímia entre verbos . . . . .	16
	4.2 Cruzamento com hiperónimos . . . . .	17
	4.3 Melhor extracção de locais . . . . .	17
	4.4 Tratamento da SINONÍMIA . . . . .	18
	4.5 Substituições . . . . .	19
	4.6 Correção de relações . . . . .	19
5	Agradecimentos . . . . .	21

# 1 Introdução

Depois de fazer uma revisão do estado da arte [3, 4] e de mostrar de que forma é possível extrair relações a partir das definições de um dicionário [2] este relatório apresenta as relações extraídas do Dicionário da Língua Portuguesa (DLP) [1] no âmbito do PAPEL, assim como padrões indicadores importantes, alguns exemplos dessas mesmas relações e ainda uma visão quantitativa das relações extraídas. São ainda referidas algumas limitações que existem actualmente e sugestões para que no futuro este trabalho possa ser melhorado.

## 2 Metodologia

A metodologia utilizada para a extracção de relações foi aquela que se encontra descrita no relatório anterior [2], ou seja, foram construídas gramáticas com o objectivo de encontrar determinadas relações entre palavras definidas no DLP e palavras que ocorrem na sua definição. As gramáticas são processadas pelo analisador sintáctico PEN e de todas as derivações obtidas aquela que tiver pelo menos um nó com o nome da relação procurada e melhor se adequar<sup>1</sup> é automaticamente seleccionada.

### 2.1 Procura dos nós com as palavras relacionadas

As gramáticas construídas visam a extracção de um leque de relações que se podem inserir dentro dos tipos HIPERONIMIA, CAUSA, PARTE, FINALIDADE, LOCAL E SINONIMIA. Por sua vez estes tipos dividem-se em especificações das relações, de acordo com as categorias gramaticais esperadas para os seus argumentos.

Cada gramática pode ter uma grande quantidade de regras para identificar vários padrões utilizados na decomposição das definições, mas apenas uma pequena parte dessas regras incluem palavras relacionadas.

Para que o programa responsável pela extracção das relações consiga extrair automaticamente o conteúdo dos nós relativos a regras com palavras relacionadas, é utilizado um ficheiro que descreve as relações que se pretendem procurar (`descricao_relacoes.dat`).

A Figura 1 mostra a descrição de uma relação neste ficheiro. O nome da relação é escrito em maiúsculas e antecede um bloco que tem em cada linha a categoria esperada para os argumentos da relação directa, a identificação da relação directa e a identificação da relação inversa. As relações directa e inversa são identificadas por um nome que coincide com o nome das regras que nas gramáticas contém a palavra relacionada na definição.

As relações convencionadas como inversas que forem extraídas podem utilizar esta informação para serem transformadas no tipo directo.

```
PARTE{
nome:nome * PARTE_DE:INCLUI;
nome:adj * PARTE_DÊ_ALGO_COM_PROPRIEDADE:PROPRIEDADE_DE_ALGO_QUE_INCLUI;
}
```

Figura 1: Descrição da relação PARTE

---

<sup>1</sup>Remetemos para o relatório anterior, onde a forma de seleccionar a derivação mais adequada é seleccionada.

As descrições das relações são processadas com o auxílio do PEN e de uma gramática muito simples, construída para o efeito.

## 2.2 Processamento por categoria gramatical

Para detectar as categorias gramaticais das palavras definidas foram utilizadas as categorias atribuídas no DLP. Além disso, a maior parte dos padrões utilizados conseguem prever a categoria gramatical das palavras que os seguem. Foram no entanto tidas em conta apenas palavras de categoria aberta, ou seja, os nomes, verbos, adjectivos e advérbios.

Assim sendo, cada gramática foi construída com o objectivo de processar apenas definições de uma das quatro categorias. A directoria com as gramáticas foi organizada em quatro subdirectorias de forma a que cada uma dessas subdirectorias corresponda a uma categoria gramatical. O sistema responsável pela extracção faz depois a correspondência entre gramáticas e categorias baseando-se na estrutura da directoria.

## 2.3 Construção das gramáticas

A construção das gramáticas foi feita inicialmente com base na observação das frequências dos ngramas presentes nas definições do DLP e também num conjunto de 5000 definições utilizadas para teste, constituídas por definições dos domínios de medicina e informática. Pontualmente eram também consultadas outras definições, através do sítio Infopédia<sup>2</sup>.

Sempre que foi considerado oportuno, as gramáticas foram enviadas para a Porto Editora, onde todas as definições do DLP eram processado pelo PEN e pelo sistema de extracção. Os resultados eram-nos depois enviados e podiam ser comparados com versões anteriores, através de um programa que lista todas as relações novas e removidas e que serviu de auxílio à melhoria das regras das gramáticas. Depois de extraídas as primeiras relações, as definições que originavam relações eram também observadas e podiam passar a ser utilizadas para testes.

---

<sup>2</sup><http://www.infopedia.pt/>

## 3 Apresentação das relações

Nesta secção são apresentadas as relações extraídas no âmbito do PAPEL. Para cada relação são apresentadas as diferentes especificações de acordo com os seus argumentos; os padrões ou palavras chave utilizados na sua extracção; e ainda alguns exemplos de relações extraídas. Antes das relações serem apresentadas são referidas algumas características comuns às gramáticas e que podem ajudar a perceber alguns dos exemplos.

### 3.1 Características comuns a todas as gramáticas

#### Partilha de regras

Regras utilizadas por mais de um gramática são, por norma, declaradas numa terceira gramática que é depois incluída pelas primeiras. Desta forma é possível que, por exemplo, uma gramática para processar definições de nomes possa utilizar exactamente os mesmos padrões que uma gramática utilizada para processar adjectivos. Esses padrões podem ser depois tratados de forma diferente por cada gramática e podem também ser inseridos num padrão maior.

Esta funcionalidade é também utilizada para manter símbolos terminais (ficheiro `terminais.txt`), como palavras funcionais (necessárias para quase todas as gramáticas), de onde destacamos preposições, determinantes e pronomes. Além das palavras funcionais, o conjunto de símbolos terminais tem ainda sinais de pontuação, adjectivos genéricos e advérbios frequentes, entre outros. Os verbos foram colocados num outro ficheiro (ficheiro `verbos.txt`), onde fomos adicionando verbos consoante eram necessários para construir padrões.

#### Enumerações

A palavra relacionada que procuramos com cada gramática não costuma ter restrições e pode ser qualquer coisa. O que nos indica se é a palavra que pretendemos é o padrão que a antecede. Nas gramáticas que construímos consideramos que, após o padrão, pode ocorrer uma enumeração de palavras relacionadas. A enumeração consiste em palavras separadas por vírgulas ou conjunções. As palavras podem ser introduzidas por determinantes, ser modificadas por um adjectivo genérico ou outros, dependendo da gramática.

#### Entidades complexas

Em algumas gramáticas (essencialmente aquelas para as quais palavras relacionadas na definição são nomes) considera-se que a palavra relacionada pode ser

aquilo que chamamos de entidade complexa. Uma entidade complexa pode ser apenas uma palavra ou então uma palavra modificada. Actualmente apenas são consideradas palavras modificadas através de preposições, pois esta é o único tipo de modificação facilmente detectável a nível sintáctico.

## 3.2 As relações

### HIPERONIMIA

A relação de HIPERONIMIA é provavelmente a relação mais estudada e que deu origem a mais trabalhos na comunidade NLP. Pode ser utilizada para a construção de taxonomias porque ocorre quando uma entidade é uma especificação de outra, ou seja, um tipo ou uma subclasse da primeira entidade.

Nesta relação existe assim a restrição de ambas a entidades terem de pertencer à mesma categoria gramatical. Apesar de ser possível estabelecer relações de HIPERONIMIA tanto entre nomes, como entre verbos, por enquanto apenas procuramos extrair HIPERONIMIA entre nomes.

A extração de HIPERONIMIA foi feita seguindo três estratégias diferentes:

- Utilização de padrões indicadores de hiperonímia: **tipo/forma/gênero de**;
- Palavras (eventualmente modificadas por adjetivos genéricos) que ocorrem no início das definições (e que constituem o *genus*), antes de padrões indicadores de outras relações;
- Definições iniciadas por hiperónimos frequentes<sup>3</sup>, como por exemplo: **pessoa, planta, instrumento, propriedade**;

Alguns exemplos de relações de HIPERONIMIA extraídos encontram-se na Tabela 1.

### CAUSA

Definimos que ocorre uma relação de CAUSA quando uma entidade (causador) pode causar, provocar ou originar uma outra entidade (resultado). Essa entidade pode ser por exemplo um agente, uma acção, um sintoma, um evento, um fenómeno, ...

No DLP a relação de CAUSA pode encontrar-se através da utilização de padrões que contenham palavras e expressões chave, como as que se seguem:

---

<sup>3</sup>A lista de hiperónimos frequentes é constituída por todos os nomes que iniciam mais de 80 definições.

Entrada	Definição	Relações de HIPERONIMIA
fardamento, s. m.	tipo de farda	tipo HIPERONIMO _DE fardamento
fotojornalismo, s. m.	gênero de jornalismo em que as fotografias constituem o principal material informativo	jornalismo HIPERONIMO _DE fotojornalismo
detonação, s. f.	ruído causado por explosão	ruído HIPERONIMO _DE detonação
bioacústica, s. f.	ciência que tem por objectivo o estudo dos sons produzidos por animais	ciência HIPERONIMO _DE bioacústica
esfera armilar, s. f.	dispositivo formado por armilas que representam círculos da esfera celeste	dispositivo HIPERONIMO _DE esfera _armilar
curvígrafo, s. m.	instrumento que traça curvas	instrumento HIPERONIMO _DE curvígrafo

Tabela 1: Exemplos de relações do tipo HIPERONIMIA.

- Verbos que indicam causa:
  - No presente (causa, provoca, origina, suscita)
  - No particípio passado (causado, provocado, originado, suscitado)
  - No infinitivo (causar, provocar, originar, suscitar)
- Indicadores de causador (causador, provocador, causa, origem)
- Indicadores de resultado (resultado, consequência)
- Outras expressões (devido a, efeito de)

As gramáticas que construímos procuram extrair relações de CAUSA não só entre nomes, mas também entre outras categorias gramaticais. A especificação da relação do tipo CAUSA é feita através do padrão utilizado para a representar e também com a categoria gramatical da palavra definida onde o padrão ocorre.

Na Tabela 2 é possível verificar as várias especificações da relação CAUSA que definimos e a categoria gramatical que os atributos de cada especificação devem respeitar.

Alguns exemplos das relações que se podem extrair com estes padrões encontram-se na Tabela 3.

Directa	Arg 1	Arg 2
CAUSADOR_DE	nome	nome
ACCAO_QUE_CAUSA	verbo	nome
CAUSADOR_DA_ACCAO	nome	verbo
CAUSADOR_DE_ALGO_COM_PROPRIEDADE	nome	adj
PROPRIEDADE_DE_ALGO_CAUSADOR_DE	adj	nome

Tabela 2: Relações do tipo CAUSA.

Entrada	Definição	Relações de CAUSA
detonação, s. f.	ruído causado por explosão	explosão CAUSADOR_DE detonação
dardada, s. f.	ferimento provocado por golpe de dardo	golpe_de_dardo CAUSADOR_DE dardada
ópio, s. m.	o que causa adormecimento, entorpecimento	ópio CAUSADOR_DE entorpecimento ópio CAUSADOR_DE adormecimento
friagem, s. f.	tempo frio, em geral por causa do vento	vento CAUSADOR_DE friagem
prova, s. f.	resultado de um ensaio ou teste	ensaio CAUSADOR_DE prova teste CAUSADOR_DE prova
assadura, s. f.	irritação da pele devido a calor ou fricção	fricção CAUSADOR_DE assadura calor CAUSADOR_DE assadura
reactivo, adj.	que suscita reacção	reactivo PROPRIEDADE_DE_ALGO_CAUSADOR_DE reacção
renzilhar, v. intr.	provocar quezílias	renzilhar ACCAO_QUE_CAUSA quezílias
purgação, s. f.	acto ou efeito de purgar, limpar ou purificar	purgar ACCAO_QUE_CAUSA purgação purificar ACCAO_QUE_CAUSA purgação limpar ACCAO_QUE_CAUSA purgação
fumigar, v. tr.	desinfectar (local) ou exterminar (parasitas) por acção de fumo ou gases	fumo CAUSADOR_DA_ACCAO fumigar gases CAUSADOR_DA_ACCAO fumigar

Tabela 3: Exemplos de relações do tipo CAUSA.

## PRODUTOR

A relação PRODUTOR trata-se de uma relação que ocorre entre uma entidade (produtor) que produz uma outra entidade (produto). Está muito ligada à relação do tipo CAUSA e inicialmente até a consideramos dentro desse tipo de relações. No entanto os padrões indicadores desta relação pareceram-nos suficientemente distinguíveis a nível sintáctico para que pudesse existir uma separação.

Apesar de semelhantes aos padrões indicadores da relação CAUSA, os padrões utilizados para extrair esta relação do DLP utilizam palavras chave diferentes:

- Verbos que indicam produção:
  - No presente (produz, gera)

- No particípio passado (produzido, gerado, obtido)
- No infinitivo (produzir, gerar, obter)
- Indicadores de produtor (produtor, gerador)
- Indicadores de produto (produto, fruto)

De acordo com a classe gramatical dos argumentos da relação, foram definidas especificações da relação, descritas na Tabela 4:

Relação	Arg 1	Arg 2
PRODUTOR_DE	nome	nome
PRODUTOR_DE_ALGO_COM_PROPRIEDADE	nome	adj
PROPRIEDADE_DE_ALGO_PRODUTOR_DE	adj	nome

Tabela 4: Relações do tipo PRODUTOR.

Entrada	Definição	Relações PRODUTOR
borborigmo, s. m.	ruído produzido por gases nos intestinos	gases PRODUTOR_DE borborigmo
fotógeno, adj.	que gera ou emite luz	fotógeno PROPRIEDADE_DE_ALGO_PRODUTOR_DE luz
sublimado, adj.	obtido por sublimação	sublimação PRODUTOR_DE_ALGO_COM_PROPRIEDADE sublimado

Tabela 5: Exemplos de relações do tipo HIPERONIMIA.

A relação PRODUTOR abrange também a relação entre um procedimento e um resultado da sua aplicação (podemos chamar-lhe PROCESSO\_PARA). Essa relação está muitas vezes presente através da utilização do verbo *obter*, mas não foi separada porque nem todas as utilizações deste verbo lhe dizem respeito e ainda porque a relação pode estar presente com a utilização de outros padrões, difíceis de dissociar dos padrões utilizados para a extracção da relação PRODUTOR e CAUSA.

## PARTE

Uma relação do tipo PARTE (ou meronímia) ocorre quando uma entidade maior é constituída, inclui ou se pode dividir em entidades mais pequenas. A entidade maior será o todo e as entidades mais pequenas as partes.

No DLP esta relação ocorre com a utilização das seguintes palavras e expressões chave:

- Indicadores de um todo: parte/membro

- Indicadores de um colectivo: `porção/conjunto/grupo/família`
- Indicadores de constituição: `constituído/formado/composto/provido/munido`
- Indicadores de posse: `possui/contém/inclui/tem`
- Indicadores de pertença: `pertence, pertencente`

No caso da relação de PARTE, as gramáticas construídas visam obter apenas relações entre nomes ou entre nomes e adjetivos. Na Tabela 6 é possível verificar as duas especificações que definimos de acordo com a categoria gramatical que os argumentos devem respeitar.

Relação	Arg 1	Arg 2
PARTE_DE	nome	nome
PARTE_DE_ALGO_CÔM_PROPRIEDADE	nome	adj
PROPRIEDADE_DE_ALGO_PARTE_DE	adj	nome

Tabela 6: Relações do tipo PARTE.

Alguns exemplos das relações que se podem extrair com estes padrões encontram-se na Tabela 7.

A relação PARTE acaba por ser bastante abrangente, podendo dizer respeito aos seguintes tipos de meronímia:

- Componente de um objecto (chaminé, deutolécito)
- Elementos de uma colectividade (director, celta, centáurea-maior, claquista)
- Porção de uma massa (coágulo)
- Subdivisão de uma área abstracta (citologia)
- Propriedade/bem de uma entidade (falha, armazém)
- Subdivisão de um local

Verifica-se no entanto que a parte é sempre inferior ao todo que a inclui.

A razão para não se ter feito a distinção entre estes tipos de meronímia prende-se com a difícil distinção deste tipo de relações se tivermos em conta que foram obtidas através da análise dos padrões textuais utilizados.

Os únicos padrões que apresentam muito pouca ambiguidade são `membro de` e `grupo/conjunto de` que apontam claramente para uma colectividade constituída por elementos que se caracterizam por ser todos do mesmo tipo ou terem todos algo em comum. Todos os restantes padrões podem dizer respeito a diferentes tipos de meronímia.

Entrada	Definição	Relações de PARTE
citologia, s. f.	parte da biologia que estuda as células	citologia PARTE_DE biologia
chaminé, s. f.	parte do cachimbo onde se deita o tabaco	chaminé PARTE_DE cachimbo
director, s. m.	membro de uma direcção ou de um directório	director PARTE_DE direcção director PARTE_DE directório
cometa, s. m.	astro geralmente constituído por núcleo, cabeleira e cauda, que gravita em torno do Sol em órbita muito excêntrica	cabeleira PARTE_DE cometa núcleo PARTE_DE cometa cauda PARTE_DE cometa
deutolécito, s. m.	parte do óvulo ou do ovo animal que contém as reservas nutritivas	deutolécito PARTE_DE óvulo deutolécito PARTE_DE ovo reservas PARTE_DE deutolécito
celta, s. 2 gén.	pessoa pertencente aos Celtas	celta PARTE_DE Celtas
centáurea-maior, s. f.	planta da família das Compostas, utilizada em medicina	centáurea-maior PARTE_DE Compostas
coágulo, s. m.	porção de sangue separada do respectivo soro	sangue PARTE_DE coágulo
claquista, s. 2 gén.	pessoa que faz parte de uma claque	claquista PARTE_DE claque
armazenista, s. 2 gén.	pessoa que possui armazém ou está encarregada dele	armazém PARTE_DE armazenista
falhado, adj.	que tem falha	falha PARTE_DE_ALGO_COM_PROPRIEDADE falhado
coberto, adj.	que possui tampa ou qualquer cobertura	tampa PARTE_DE_ALGO_COM_PROPRIEDADE coberto cobertura PARTE_DE_ALGÕ_COM_PROPRIEDADE coberto
centro-europeu, adj.	relativo ou pertencente ao centro da Europa	centro-europeu PROPRIEDADE_DE_ALGO_PARTE_DE centro_da_Europa

Tabela 7: Exemplos de relações do tipo PARTE.

## FINALIDADE

Definimos que a relação FINALIDADE ocorre quando uma entidade (meio) tem ou é utilizada com determinado objectivo (finalidade). O meio pode ser um instrumento (nome) ou um procedimento (descrito sob a forma de um nome ou verbo) e a finalidade pode ser um estado (nome) ou uma acção (verbo).

No DLP são utilizadas as seguintes palavras e expressões chave para representar estas relações:

Relação	Arg 1	Arg 2
FINALIDADE_DE	nome	nome
ACCAO_FINALIDADE_DE_ALGO_COM_PROPRIEDADE	verbo	adj
ACCAO_FINALIDADE_DE_FINALIDADE_DE_ALGO_COM_PROPRIEDADE	verbo	nome
FINALIDADE_DE_ALGO_COM_PROPRIEDADE	nome	adj
FINALIDADE_DA_ACCAO	nome	verbo
MANEIRA_POR_MEIO_DE	adv	nome

Tabela 8: Relações do tipo FINALIDADE.

- Verbos relativos a utilização: **usar**, **utilizar**
- Verbos relativos a função: **servir**
- Outros verbos: **recorrer**
- Indicadores de finalidade: **finalidade**, **fim**, **objectivo**
- Expressões indicadores de meio: **por meio de**, **com o auxílio de**
- Preposição indicadora de função: **para**

### LOCAL (de origem)

A relação do tipo LOCAL ocorre sempre entre duas entidades nominais quando uma delas é um local e a outra é natural ou habita na primeira.

Os padrões utilizados para extrair esta relação do DLP são bastante simples e tiram partido da utilização das palavras **natural** e **habitante**.

Alguns exemplos desta relação encontram-se na Tabela 10.

### SINONIMIA

A relação de SINONIMIA é uma relação chave na construção de uma ontologia. Ocorre entre duas palavras e indica que ambas podem ser utilizadas para representar o mesmo conceito. As palavras têm portanto de ser da mesma categoria gramatical.

Actualmente a única forma utilizada para a extracção de SINONIMIA resume-se a procurar palavras cuja definição seja constituída por uma única palavra ou por uma enumeração de palavras. Na Tabela 11 encontram-se alguns exemplos de relações deste tipo extraídas:

Entrada	Definição	Relações de FINALIDADE
pente, s. m.	instrumento de ferro usado para cardar a lã	cardar_a_lã ACCAO_FINALIDADE_DE pente
cicloturismo, s. m.	actividade turística que se pratica utilizando uma bicicleta como meio de transporte	cicloturismo FINALIDADE_DE bicicleta
arrolho, s. m.	toada para adormecer as crianças	adormecer_as_crianças ACCAO_FINALIDADE_DE arrolho
acrografia, s. f.	arte de gravar em relevo sobre pedra ou metal recorrendo a ácidos	acrografia FINALIDADE_DE ácidos
comédia, s. f.	obra de ficção cuja finalidade é fazer rir	fazer_rir ACCAO_FINALIDADE_DE comédia
cooperativa, s. f.	associação que tem como objectivo a construção de habitações a custos controlados destinadas aos seus membros	construção_de_habitações FINALIDADE_DE cooperativa
acenar, v. intr.	chamar a atenção por meio de gestos	acenar ACCAO_FINALIDADE_DE gestos
enumerativo, adj.	que serve para a enumeração	enumeração FINALIDADE_DE_ALGO_COM_PROPRIEDADE enumerativo
preventivo, adj.	que tem por fim prevenir, acautelar ou impedir	prevenir ACCAO_FINALIDADE_DE_ALGO_COM_PROPRIEDADE preventivo acautelar ACCAO_FINALIDADE_DE_ALGO_COM_PROPRIEDADE preventivo impedir ACCAO_FINALIDADE_DE_ALGO_COM_PROPRIEDADE preventivo

Tabela 9: Exemplos de relações do tipo FINALIDADE.

Entrada	Definição	Relações de LOCAL
coreano, s. m.	natural ou habitante da Coreia do Norte ou da Coreia do Sul	Coreia_do_Sul LOCAL_ORIGEM_DE coreano Coreia_do_Norte LOCAL_ORIGEM_DE coreano
favelado, adj. e s. m.	habitante ou designativo de habitante de favela	favela LOCAL_ORIGEM_DE favelado
baiano, s. m.	indivíduo natural da Baía	Baía LOCAL_ORIGEM_DE baiano

Tabela 10: Exemplos de relações do tipo LOCAL.

### Outras relações

Além das relações já referidas, foram ainda extraídas relações cujo principal objectivo é a obtenção de adjectivos e advérbios. Estas relações e as categorias

Entrada	Definição	Relações de SINONIMIA
amabilidade, s. f.	afabilidade	afabilidade SINONIMO_DE amabilidade
talhar, v. tr.	gravar, cinzelar ou esculpir	esculpir SINONIMO_DE talhar
moldável, adj. 2 gén.	adaptável, flexível	flexível SINONIMO_DE moldável adaptável SINONIMO_DE moldável
sucessivamente, adv.	seguidamente	seguidamente SINONIMO_DE sucessivamente

Tabela 11: Exemplos de relações do tipo SINONIMIA.

gramaticais esperadas para os seus argumentos encontram-se representadas na Tabela 12

Directa	Arg 1	Arg 2
PROPRIEDADE_DE_ALGO_REFERENTE_A	adj	nome
PROPRIEDADE_DO_QUE	adj	verbo
MANEIRA_POR_MEIO_DE	adv	nome
NAO_HA_NA_MANEIRA	adv	nome

Tabela 12: Outras relações.

Nas duas primeiras relações apresentadas o primeiro argumento é um adjectivo que diz respeito à propriedade de um nome ou de um verbo respectivamente (o segundo argumento). Por outro lado, as duas seguintes dizem respeito a relações entre maneiras (advérbios) e nomes, representando respectivamente uma entidade que esteja associada ou que não esteja associada a essa maneira

Ainda que a utilidade destas relações só por si possa ser reduzida, o seu cruzamento com os resultados obtidos com as relações anteriormente obtidas pode vir a originar relações mais refinadas. Optámos pela sua extracção por esta razão e também por serem relações relativamente fáceis de extrair. Apesar disso, foram as relações sobre as quais investimos menos tempo, por isso será normal se forem aquelas com a maior taxa de erros. De qualquer das formas, estas relações podem no mínimo vir a ser utilizadas como uma fonte para a criação de listas dos adjectivos e advérbios definidos no DLP.

A Tabela 13 mostra alguns exemplos de relações destes tipos.

### 3.3 Visão quantitativa

Nesta secção encontram-se as quantidades das relações extraídas (Tabela 14), após algumas serem reajustadas ou descartadas de acordo após colocar frente-a-frente a classe gramatical possível para os argumentos e as classes gramaticais possíveis das palavras dos argumentos, atribuídas pelo dicionário ou pelo Jspell. Relações que incluem diferentes tipos, de acordo com a classe gramatical dos argumentos, apresentam também a contabilização de acordo com os tipos considerados (Tabelas 15, 16, 17, 18).

Entrada	Definição	Relações
areométrico, adj.	que diz respeito à areometria ou ao areómetro	areométrico PROPRIEDADE_DE_ALGO_REFERENTE_A areómetro
patológico, adj.	relativo a uma doença	patológico PROPRIEDADE_DE_ALGO_REFERENTE_A doença
viril, adj. 2 gén.	referente ao homem ou ao varão	viril PROPRIEDADE_DE_ALGO_REFERENTE_A homem viril PROPRIEDADE_DE_ALGO_REFERENTE_A varão
rápido, adj.	que se realiza em pouco tempo	rápido PROPRIEDADE_DO_QUE se_realiza_em_pouco_tempo
magro, adj.	que tem pouco peso	magro PROPRIEDADE_DO_QUE tem_pouco_peso
cruel, adj. 2 gén.	que demonstra crueldade	cruel PROPRIEDADE_DO_QUE demonstra_crueldade
devagar, adv.	sem pressa	pressa NAO_HA_NA_MANEIRA devagar
fartamente, adv.	com fartura	fartamente MANEIRA_POR_MEIO_DE fartura
proibitivamente, adv.	de modo impeditivo	proibitivamente MANEIRA_POR_MEIO_DE impeditivo

Tabela 13: Exemplos de outros tipos de relações.

Relação	Quantidade
HIPERONIMIA	64833
PARTE	15365
CAUSA	8133
PRODUTOR	1317
FINALIDADE	10331
LOCAL	768
REFERENTE	20859
SINONIMIA	86278

Tabela 14: Quantidade de relações.

Tipo	Quantidade
PARTE_DE	10688
PARTE_DE_ALGO_CÔM_PROPRIEDADE	3715
PROPRIEDADE_DE_ALGO_PARTE_DE	962

Tabela 15: Quantidades de relações do tipo PARTE.

Tipo	Quantidade
CAUSADOR_DE	1137
ACCAO_QUE_CAUSA	6426
CAUSADOR_DA_ACCAO	39
CAUSADOR_DE_ALGO_CÔM_PROPRIEDADE	16
PROPRIEDADE_DE_ALGO_QUE_CAUSA	515

Tabela 16: Quantidades de relações do tipo CAUSA.

Tipo	Quantidade
PRODUTOR_DE	937
PRODUTOR_DE_ALGO_CÔM_PROPRIEDADE	31
PROPRIEDADE_DE_ALGO_PRODUTOR_DE	349

Tabela 17: Quantidades de relações do tipo PRODUTOR.

Tipo	Quantidade
FINALIDADE_DE	2947
ACCAO_FINALIDADE_DE	5640
ACCAO_FINALIDADE_DE_ALGO_COM_PROPRIEDADE	265
FINALIDADE_DE_ALGO_COM_PROPRIEDADE	23
MANEIRA_POR_MEIO_DE	1442

Tabela 18: Quantidades de relações do tipo FINALIDADE.

## 4 Limitações e pistas para o futuro

Esta secção tem como objectivo apresentar algumas das limitações que se verificam actualmente nos conjuntos de relações extraídos, bem como dar algumas pistas e sugestões para um eventual trabalho futuro com vista à melhoria dos resultados obtidos.

### 4.1 Hiperonímia entre verbos

Actualmente apenas lidamos com a extracção de HIPERONOMIA entre nomes. A extracção de HIPERONIMIA entre verbos está muito ligada à utilização de advérbios/maneiras, como na definição:

**expurgar, v. tr. - purgar completamente.**

Neste caso temos uma relação de HIPERONIMIA entre **purgar** e **expurgar** (pugar HIPERONIMO\_DE expurgar) e temos ainda que **completamente** é uma maneira de **purgar**.

Estas relações não foram no entanto extraídas, porque não temos uma lista de verbos, nem uma lista de advérbios. A utilização dessas listas possibilitaria a extracção de grande parte destas relações, mas poderia ficar de certa forma de fora do objectivo do PAPEL, porque estaríamos a alimentar o sistema com uma parte considerável da informação que pretendemos extrair, deixando um bocado de lado a ideia de construir um recurso baseado no DLP.

Outra alternativa passaria por utilizar um padrão como **\*r \*mente** que nos daria todos os verbos seguidos de um advérbio. Até ao momento esta opção não foi seguida porque o PEN não suporta expressões regulares e a introdução desta funcionalidade acabou por não ser realizada devido a assumida falta de tempo.

Uma terceira alternativa passaria por construir uma lista de advérbios através da recolha de todas as entidades em relações, na posição de advérbios.

## 4.2 Cruzamento com hiperónimos

O cruzamento do actual conjunto de relações com os hiperónimos dos seus argumentos (obtidos através da extracção de HIPERONIMIA) poderiam ser ainda mais especificados.

- No caso da relação de CAUSA, poderia eventualmente interessar separar as relações de acordo com o tipo de causa e resultado (um agente, um sintoma, um evento, um fenómeno, ...).
- A partir das relações CAUSA e PRODUTOR, poderia interessar separar aquelas cujo CAUSADOR/PRODUTOR é um procedimento daquelas onde é uma entidade.
- No que diz respeito às relações PARTE, o cruzamento com os hiperónimos (e eventualmente até com outras relações) poderia facilitar a separação dos vários tipos de meronímia.
- No que toca às relações de FINALIDADE, este cruzamento poderia ser utilizado para separar relações onde o meio é um instrumento de relações onde o meio é um procedimento.

Estas distinções não foram realizadas até ao momento não só porque os seus indicadores não são facilmente distinguíveis ou são mesmo completamente vagos a nível sintáctico, mas também porque não estamos completamente certos de que todas estas subdivisões sejam necessárias ou suficientemente interessantes e úteis.

A título de exemplo, partindo do princípio que a partir de uma análise dos hiperónimos dos argumentos de uma relação do tipo PARTE é possível obter a classe dos argumentos, será possível aplicar regras como: *Se ambos os argumentos forem locais, estamos perante a subdivisão de um local ou se ambos os argumentos forem do mesmo tipo, podemos estar perante a subdivisão de uma área abstracta ou uma componente de um objecto, dependendo do tipo.*

## 4.3 Melhor extracção de locais

De todas as relações extraídas, a relação LOCAL é aquela que utiliza menos padrões. Apesar dos resultados serem animadores a sua extracção tem uma limitação que passa pela detecção de locais identificados por mais de uma palavra, não ligadas entre preposições (como por exemplo Estados Unidos ou Serra Leoa). A detecção deste tipo de locais poderia ser facilitada de uma de duas formas:

- Utilização de uma lista com nomes de locais
- Utilização de um detector de entidades mencionadas

A primeira hipótese seria sem dúvida muito simples, mas ao optar por ela estaríamos mais uma vez a deixar de lado o objectivo de extrair o máximo de informação possível a partir do DLP, já que estaríamos a alimentar o sistema com metade da informação que pretendíamos a extrair (os locais).

A segunda hipótese seria bastante mais trabalhosa e até ao momento, tendo em conta o objectivo do PAPEL, nunca foi uma prioridade. Talvez nem fosse preciso um detector de entidades mencionadas muito sofisticado e bastasse uma regra para detectar sequências de palavras iniciadas por maiúscula. No entanto esse tipo de regras não é possível no PEN e a sua implementação não foi realizada, mais uma vez por falta de tempo.

## 4.4 Tratamento da SINONÍMIA

### Relação especial

A relação de SINONIMIA é diferente das restantes e central na criação de uma ontologia. A ser utilizada na construção de uma rede desambiguada seria necessário que fosse a primeira a ser tratada de forma a serem criados conjuntos de sinónimos que representam conceitos.

Apesar do tratamento da SINONÍMIA poder tirar partido da consulta de outras relações (que podem mesmo ajudar na desambiguação), o tratamento dessas outras relações só deve ser realizado com os conceitos já estabelecidos. Para fazer a correspondência entre palavras relacionadas e o respectivo conceito, seria necessário recorrer a uma estratégia de desambiguação que poderia passar por uma análise das palavras na definição em questão, frases exemplo, domínio... a nível de categoria gramatical essa desambiguação já é realizada praticamente por completo pelas gramáticas construídas.

### Outras formas de obtenção

Além da forma actualmente utilizada para obter sinónimos, existem naturalmente outras estratégias:

- Procurar palavras cuja definição seja exactamente igual;
- Tirar partido da divisão em acepções que existe na estrutura do DLP.
- Cruzar as relações de sinonímia com outros tesouros do português, como o TeP<sup>4</sup>.

---

<sup>4</sup><http://www.nilc.icmc.usp.br/tep2/index.htm>

A segunda opção não foi seguida porque existe uma grande dificuldade em fazer corresponder palavras relacionadas que ocorrem nas definições com a acepção correspondente. Como esta limitação ocorre sempre que determinada palavra possa ter mais um sentido/acepção, seria necessário a utilização de regras mais complexas para efectuar a desambiguação. Ainda assim, se a divisão das acepções no DLP fosse seguida "às cegas" poderia levar a uma divisão excessiva dos possíveis sentidos da mesma palavra e seria também necessário proceder a uma "ambiguação".

O cruzamento com outros recursos, ainda que mais uma vez se esteja a sair fora do âmbito do DLP, poderia não só completar/melhorar os conjuntos de sinónimos (vulgarmente chamados de *synsets*) obtidos, como poderia também ser uma valiosa ferramenta de avaliação. As definições utilizadas nos thesaurus poderiam também ser utilizadas como auxílio a uma estratégia de desambiguação.

## 4.5 Substituições

Na criação de conjuntos de sinónimos, para que o sistema saiba que está a lidar com instâncias da mesma palavra, seria necessário realizar a sua lematização e, nesse caso, não nos ocorre mesmo nenhuma alternativa que não seja a utilização de um recurso externo. As palavras que não estiverem na forma do lema, deverão ser substituídas pelo seu lema.

Outro tipo de substituições que também poderá ser feito no que diz respeito ao tratamento da sinonímia está relacionado com a utilização de determinadas palavras funcionais que estão claramente ligadas a um tipo de entidade, como por exemplos os pronomes demonstrativos *aquilo* e *aquele* que estão claramente ligados a coisas e pessoas, respectivamente. Tendo isto em conta, pode ser interessante manter relações entre determinada entidade e uma destas palavras.

## 4.6 Correção de relações

Apesar da maior parte dos padrões conseguir prever as categorias gramaticais das palavras relacionadas, isto pode não acontecer em todas as situações. Além disso a procura de enumerações pode levar à extracção de relações completamente erradas, muitas vezes até com palavras funcionais. Para corrigir estas situações poderia ser utilizado um etiquetador morfológico que fizesse pelo menos as seguintes:

- Verificar se alguma das possíveis categorias gramaticais para as palavras relacionadas respeita efectivamente as categorias esperadas para os

argumentos da relação. Em caso negativo, procurar se, dentro do mesmo tipo de relações, existe alguma especificação adequada para as categorias dos argumentos e substituir o tipo de relação.

- Se algum dos argumentos for de uma categoria gramatical fechada (palavra funcional), descartar a relação.

Outra alternativa à utilização do etiquetador, ainda que tivesse uma função semelhante menos abrangente, seria a criação de uma lista de palavras funcionais. Essa lista poderia ser utilizada para percorrer todas as relações e descartar aquelas onde um dos argumentos fosse uma dessas palavras.

## 5 Agradecimentos

Este relatório foi escrito no âmbito da Linguateca, financiada pela Fundação para a Ciência e Tecnologia e pela União Europeia através dos projectos POSI/PLP/43931/2001 e POSC 339/1.3/C/NAC.

Agradecemos também a colaboração valiosa da Diana Santos pela orientação e sugestões relativamente ao conteúdo do relatório, ao Nuno Seco que criou o PEN e iniciou o projecto, à Cláudia Freitas pela disponibilidade e apoio dado na realização desta fase do trabalho e ainda ao Núcleo de Investigação e Desenvolvimento da Porto Editora, que nos prestou o apoio necessário do lado da Porto Editora.

# Referências

- [1] *Dicionário PRO da Língua Portuguesa*. Porto Editora, Porto, 2005.
- [2] Hugo Gonçalo Oliveira and Paulo Gomes. Utilização do (analisador sintáctico) PEN para extracção de informação das definições de um dicionário. Technical report, CISUC, Dezembro 2008.
- [3] Hugo Gonçalo Oliveira, Paulo Gomes, and Diana Santos. PAPEL - Trabalho relacionado e relações semânticas em recursos semelhantes. Technical report, Departamento de Engenharia Informática, FCTUC, CISUC, Dezembro 2007.
- [4] Hugo Gonçalo Oliveira, Diana Santos, Paulo Gomes, and Nuno Seco. Extracção de relações a partir de dicionários: Breve história. Technical report, CISUC, SINTEF, Agosto 2008.