Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos

Hugo Gonçalo Oliveira
CISUC, Universidade de Coimbra
hroliv@dei.uc.pt

Leticia Antón Pérez
CISUC, Universidade de Coimbra
& Universidade de Vigo
leticiaap86@gmail.com

Hernani Costa
CISUC, Universidade de Coimbra
hpcosta@dei.uc.pt

Paulo Gomes
CISUC, Universidade de Coimbra
pgomes@dei.uc.pt

Resumo

Este artigo apresenta o CARTÃO, uma nova rede léxico-semântica para o português, composta por relações extraídas a partir de três dicionários electrónicos. Após analisarmos a estrutura das definições nos três, concluímos que as mesmas regras podiam ser utilizadas para extrair relações a partir de vários dicionários. Assim, aproveitamos este facto para utilizar o mesmo conjunto de gramáticas na construção desta rede. As relações que compõem o CARTÃO são apresentadas em termos de quantidade e de acordo com o dicionário de onde foram extraídas. Verifica-se que foi possível aumentar em mais de 70% o PAPEL, uma rede semelhante já existente, o que mostra os ganhos em utilizar mais do que um recurso na construção destas redes. A cobertura do CARTÃO e os resultados da validação automática de alguns tipos de relação são aqui também apresentados e discutidos.

1 Introdução

Os dicionários são repositórios que reúnem palavras e expressões de uma língua, acompanhadas pelas definições dos seus possíveis sentidos, que por sua vez são descrições escolhidas e escritas por lexicógrafos, os especialistas neste campo. Apesar dos dicionários não estarem preparados para serem directamente utilizados como ferramentas de processamento de linguagem natural (PLN), não é de admirar que, desde cedo, tenham sido vistos como recursos fundamentais na construção automática (e manual) de bases de conhecimento lexical (veja-se, por exemplo Calzolari, Pecchia e Zampolli (1973), Chodorow, Byrd e Heidorn (1985) ou Richardson, Dolan e Vanderwende (1998)) – recursos computacionais onde itens lexicais (palavras e expressões) se encontram organizados de acordo com uma determinada teoria de semântica lexical (como a teoria apresentada por Cruse (1986)).

As bases de conhecimento revelaram-se essenciais na realização de várias tarefas PLN, incluindo a desambiguação do sentido das palavras (Agirre, Lacalle e Soroa, 2009), resposta automática a perguntas (Pasca e Harabagiu, 2001), sumarização (Plaza, Díaz e Gervás, 2010) ou tradução automática (Knight e Luk, 1994). Para o português, além de outros recursos lexicais de larga cobertura (veja-se Santos et al. (2010) para mais informação acerca deste tipo de recursos para português), existe já uma rede léxico-semântica pública extraída a partir de um dicionário, o PAPEL (Gonçalo Oliveira, Santos e Gomes, 2010). No âmbito deste trabalho, consideramos que rede léxico-semântica é uma base de conhecimento baseada em itens lexicais que, de acordo com os seus significados, podem estar ligados uns aos outros através de relações semânticas.

De forma a minimizar o problema da incompletude – há muito apontado à construção automática de bases de conhecimento lexical a partir de dicionários (veja-se Ide e Veronis (1995)) – neste trabalho utilizamos não um, mas três dicionários na criação de uma rede léxico-semântica. Ou seja, o nosso principal objectivo passa por enriquecer o PAPEL, extraído apenas de um dicionário, com informação obtida a partir de outros dois dicionários da língua portuguesa. O resultado é o CARTÃO, uma rede léxico-semântica de grandes dimensões extraída a partir de três dicionários da língua portuguesa.

Para o inglês, a WordNet de Princeton (Fellbaum, 1998) é a base de conhecimento lexical mais utilizada, e foi construída de forma manual. Por outro lado, a construção automática deste tipo de recursos é uma alternativa ao es-

força e tempo necessários para o seu desenvolvimento manual.

Neste contexto, para além da sua cobertura em termos de palavras e significados, umas das principais razões para se utilizarem dicionários é o facto de estes terem definições simples onde, normalmente, é utilizado um vocabulário controlado, o que faz com que muitas destas definições sejam quase previsíveis e, por isso, fáceis de processar automaticamente. Tendo em conta o nosso objectivo principal – extrair informação a partir de mais de um dicionário – comparamos a presença dos padrões mais frequentes nas definições de cada dicionário, e concluímos que o vocabulário estrutural é muito semelhante em todos eles. Assim, de forma a minimizar o esforço necessário para a criação manual de uma gramática para cada dicionário, reutilizamos as gramáticas em que se baseou a construção do PAPEL.

Neste artigo, depois de apresentarmos os três recursos explorados, mostramos os padrões mais frequentes e as suas ocorrências em cada dicionário. De seguida, descrevemos o procedimento utilizado para a extracção de relações semânticas. Tendo em conta cada dicionário, relatamos depois os resultados que formam o CARTÃO, que aumenta a última versão do PAPEL em mais de 70%, e damos a conhecer o (novo) Folheador, uma interface para interrogação, através da rede, dos conteúdos de recursos baseados em relações semânticas. Por fim, antes de concluir, a última secção deste artigo é dedicada à avaliação da cobertura e validação automática de algumas das relações semânticas extraídas. A primeira tarefa foi realizada através da comparação dos lemas abrangidos pelo CARTÃO e lemas abrangidos por dois *thesauri* electrónicos de larga cobertura para o português. Na segunda tarefa, compararam-se as relações de sinonímia com as relações presentes num dos *thesauri* anteriores, criado manualmente. Algumas das outras relações foram transformadas em frases que, por sua vez, foram procuradas num corpo jornalístico.

2 Trabalho relacionado

Desde cedo que os dicionários foram vistos como uma fonte de conhecimento lexical de uma língua – provavelmente a mais importante. Não é por isso surpresa que os dicionários electrónicos tenham sido também dos primeiros recursos a ser explorados na extracção automática deste tipo de conhecimento.

Os trabalhos pioneiros nesta área datam já das décadas de 1970 (Calzolari, Pecchia e Zampolli, 1973) e 1980 (Amsler, 1981), quando se começou a estudar a possibilidade de tirar partido dos dicionários na construção automática de bases de conhecimento lexical. Partindo destes primeiros estudos, surgiram também os primeiros procedimentos automáticos (Chodorow, Byrd e Heidorn, 1985), que focavam essencialmente a extracção de taxonomias. Depois de vários anos e de vários trabalhos baseados em dicionários electrónicos (e.g. Alshawi (1987), Guthrie et al. (1990), Dolan, Vanderwende e Richardson (1993), Dolan (1994)), foram apontadas algumas críticas à sua utilização na obtenção automática de informação léxico-semântica de grande cobertura (Ide e Veronis, 1995). Entre as críticas referia-se que os dicionários são inconsistentes e incompletos. No entanto, foi também mostrado que alguns dos problemas poderiam ser minimizados se fosse utilizado mais do que um dicionário. Foi também sugerido que este tipo de trabalhos poderia ser complementado com conhecimento obtido a partir de outros tipos de recurso (e.g. corpos).

Apesar das críticas em torno deste tipo de trabalho, a MindNet foi a primeira base de conhecimento lexical, completamente independente e extraída automaticamente a partir de dicionários (Richardson, Dolan e Vanderwende, 1998). Mais do que um recurso estático, a MindNet representa uma metodologia que inclui um conjunto de ferramentas para adquirir, estruturar, aceder e explorar informação semântica em texto, não só de dicionários, mas também de enciclopédias e de corpos.

Nesta altura, havia já surgido a WordNet de Princeton (Fellbaum, 1998), um recurso criado manualmente para o inglês que se revelaria como a base de conhecimento lexical mais amplamente utilizada pela comunidade de PLN. Tal como os dicionários, a WordNet, que viria a expandir-se para outras línguas[1], também é um recurso baseado em palavras e sentidos, para os quais existem definições. Por outro lado, a estrutura da WordNet encontra-se preparada a ser explorada por, ou integrada em, aplicações computacionais. As suas estruturas fundamentais são os chamados *synsets*, grupos de palavras sinónimas que representam lexicalizações de conceitos da linguagem natural. Os *synsets* podem estar ligados entre si através de relações semânticas de vários tipos (e.g. hiperonímia, parte-de).

Desde o surgimento da WordNet, começou a haver um maior interesse na aquisição automática de conhecimento que pudesse ser uti-

[1]Os projectos WordNet pelo mundo encontram-se listados na página da Global WordNet Association: `http://www.globalwordnet.org/gwa/wordnet_table.html`, de onde destacamos a WordNet.PT (Marrafa, 2002).

lizado no enriquecimento deste recurso (veja-se, por exemplo, Hearst (1998), Navigli et al. (2004), ou Toral, Muñoz e Monachini (2008)), o que levou a que a quantidade de trabalhos na extracção de conhecimento lexical a partir de dicionários diminuísse.

Ainda assim, houve na última década alguns trabalhos recentes nesta área, como Nichols, Bond e Flickinger (2005), para o japonês, ou Gonçalo Oliveira, Santos e Gomes (2010) para o português. Há ainda a referir trabalhos recentes que exploram o dicionário colaborativo Wikcionário no contexto da extracção de informação. Apesar de menos popular que a sua parente Wikipédia, o Wikcionário foi já utilizado, por exemplo, no cálculo de proximidades semânticas (Weale, Brew e Fosler-Lussier, 2009)(Zesch, Müller e Gurevych, 2008), na criação de ontologias lexicais (Wandmacher et al., 2007), ou no enriquecimento de recursos léxico-semânticos existentes (Sajous et al., 2010).

3 Recursos explorados e formato dos dados

Uma razão que também contribui para que os dicionários nem sempre sejam explorados no contexto da extracção de informação é a sua disponibilidade. Os dicionários comerciais têm normalmente o seu conteúdo protegido, mesmo para fins de investigação, e nem sempre existem alternativas gratuitas.

No caso do português, o PAPEL é um recurso livre que, no entanto, resultou do processamento de um dicionário proprietário, o Dicionário PRO da Língua Portuguesa (doravante DLP) (DLP, 2005). Além da versão actual do PAPEL, o PAPEL 3.0, neste trabalho foram explorados outros dois dicionários livres, nomeadamente o Dicionário Aberto (doravante DA) (Simões e Farinha, 2011) e a versão portuguesa da iniciativa Wikcionário[2].

Nesta secção apresentamos, primeiro, os três recursos referidos anteriormente e, depois, o formato utilizado para representar os dicionários, pronto a ser processado pelos módulos que lidam com a extracção de relações.

3.1 Apresentação dos recursos

O DA é a versão electrónica de um dicionário de português cuja versão original data de 1913. A sua ortografia está actualmente a ser modernizada. O DA contém cerca de 128 mil entradas e está disponível no formato PDF e ainda em dois formatos textuais, onde se inclui uma versão em XML[3]. No entanto, neste trabalho foi utilizado o estado actual da segunda revisão da modernização do DA, gentilmente cedida pela sua equipa de desenvolvimento.

O Wikcionário é uma iniciativa mantida pela fundação Wikimedia que tem o objectivo de disponibilizar um conjunto de dicionários multilingues. Além de informação que geralmente se encontra em dicionários, tal como a categoria gramatical das palavras, a sua etimologia e pronunciação, ou traduções, algumas entradas do Wikcionário incluem informação acerca de relações semânticas relevantes para a entrada, como sinónimos, antónimos ou hiperónimos. No entanto, por se tratar de um projecto dependente de voluntários, este tipo de informação é escasso e incompleto para a versão portuguesa do Wikcionário (doravante Wikcionário.PT).

Os Wikcionários estão disponíveis em ficheiros XML, onde as entradas se encontram escritas em texto *wiki*. Neste trabalho foi utilizada a versão de 8 de Dezembro de 2011 do Wikcionário.PT, para a qual desenvolvemos um analisador para aceder à informação de cada entrada (Pérez, Gonçalo Oliveira e Gomes, 2011). A versão do Wikcionário.PT utilizada contém cerca de 210 mil entradas, das quais cerca de 115 mil estão identificadas como tendo pelo menos a definição de uma palavra portuguesa. Tratando-se de um dicionário multilingue, as restantes entradas referem-se apenas a palavras noutras línguas.

O PAPEL 3.0[4] é a mais recente versão de uma rede léxico-semântica pública, extraída de forma automática a partir do DLP. Contém cerca de 102 mil itens lexicais e 190 mil ligações entre eles, que simbolizam relações semânticas e que estão representadas através de triplos com a seguinte estrutura:

`arg1 RELACIONADO_COM arg2`
(e.g. `animal HIPERONIMO_DE cão`)

Um triplo indica que um sentido do item lexical no primeiro argumento (`arg1`) se relaciona com um sentido do item lexical no segundo argumento (`arg2`), através de uma relação identificada por `RELACIONADO_COM`.

3.2 Formato dos dados

De forma a obter informação léxico-semântica no formato descrito anteriormente a partir de outros dicionários, convertemos os seus formatos XML para um formato mais amigável, onde cada linha

[2]Ver `http://pt.wiktionary.org/`

[3]Ver `http://www.dicionario-aberto.net/`

[4]Disponível a partir do endereço `http://www.linguateca.pt/PAPEL/`

contém apenas o lema, a sua categoria gramatical e a sua definição. Veja-se o exemplo seguinte, para a palavra *coco*:

```
coco    nome    fruto gerado pelo coqueiro, muito
                usado para se fazer doces e para
                consumo de seu líquido
```

Neste formato, palavras que têm mais do que uma definição dão origem a mais de uma linha. Além disso, como o Wikcionário inclui listas de sinónimos para várias entradas, transformamos também essas listas em definições com apenas uma palavra, tal como no seguinte exemplo para a palavra *bravo*.

```
Sinónimos:    corajoso, destemido
              ⇓
bravo   adj   corajoso
bravo   adj   destemido
```

Apenas definições de categorias abertas são utilizadas, e transformadas numa notação comum: `nome` para substantivos, `verbo` para verbos, `adj` para adjectivos e `adv` para advérbios.

Intencionalmente não mantivemos informação acerca do número da acepção a que cada definição corresponde, informação que é geralmente incluída nos dicionários. Uma das razões para esta opção é, dada a ambiguidade, a impossibilidade de fazer uma correspondência clara e directa entre as ocorrências das palavras nas definições e a acepção a que dizem respeito. Além disso, a lista de acepções de uma palavra num dicionário raramente tem correspondência directa com a mesma palavra noutro dicionário, já que não existe um critério bem definido para divisão de palavras em sentidos (Dolan, 1994) (Kilgarriff, 1996) (Peters, Peters e Vossen, 1998). Os sentidos da mesma palavra podem ir desde intimamente relacionados (e.g. na polissemia ou metonímia) até totalmente não relacionados (e.g. na homonímia). Sendo assim, ao invés de desenvolver heurísticas para desambiguar as palavras nas definições (como em Navigli (2009)), e também para encontrar correspondências entre as acepções de palavras em diferentes dicionários, tratamos da mesma forma todas as ocorrências de palavras com a mesma ortografia.

Após a conversão do DA e do Wikcionário.PT, obtivemos cerca de 229 mil e 72 mil definições, respectivamente, de cada dicionário. Para além do Wikcionário ser um recurso que, apesar de estar em crescimento, ter ainda uma dimensão pequena, as suas definições são menos porque descartamos definições: (i) correspondentes apenas a palavras noutras línguas; (ii) correspondentes a palavras de categorias fechadas ou a palavras flexionadas (incluindo formas verbais); (iii) em entradas com sintaxes alternativas, não previstas pelo nosso analisador. Devido a ser criado por voluntários, nem sempre especialistas, e por não existir um padrão para o texto *wiki* das entradas, não é possível construir um analisador que preveja todas as variantes de sintaxe, nem que seja 100% livre de erros. Tal como para o Wikcionário.PT, este problema parece ser comum a outras edições do Wikcionário (veja-se por exemplo Navarro et al. (2009)).

4 Análise das regularidades nas definições

Uma das principais razões para os dicionários serem a primeira escolha na aquisição de relações semânticas está relacionada com a simplicidade e sistematicidade das suas definições, o que os torna fáceis de processar e de ser explorados na extracção automática de informação. Foi por isso que a extracção das relações do PAPEL se baseou num conjunto de gramáticas, desenvolvidas manualmente, e que incluíam padrões léxico-sintácticos que, no DLP, indicam frequentemente a presença de relações semânticas.

Além da identificação de regularidades nas definições, de forma a evitar a criação manual de uma gramática para cada dicionário, procuramos verificar se estas regularidades eram preservadas em diferentes dicionários. Assim, comparamos as quantidades dos padrões mais frequentes nas definições de cada dicionário. Os padrões considerados mais produtivos, ou seja, que são frequentes e apropriados para a exploração na extracção automática de relações, são apresentados na tabela 1, juntamente com a sua frequência em cada dicionário, bem como a relação semântica que geralmente indicam.

Esta análise permitiu-nos confirmar que a maior parte das regularidades são preservadas nos três dicionários. Desta forma é possível utilizar as mesmas regras para extrair relações semânticas a partir de todos eles, não havendo por isso necessidade de criar uma gramática específica para cada um. Assim, tendo em conta que já incluíam a maior parte dos padrões na tabela 1, foi-nos possível reutilizar as gramáticas do PAPEL na criação do CARTÃO. Consequentemente, o CARTÃO engloba também os mesmos tipos de relações semânticas que o PAPEL.

Entre as alterações mínimas que fizemos encontram-se, por exemplo:

- A utilização do padrão `o mesmo que` para extracção da relação de sinonímia.

Padrão	Cat. gram.	Frequência			Relação
		DLP	DA	Wikcionário	
o mesmo que	Substantivo	0	10.627	1.107	Sinonímia
a[c]to ou efeito de	Substantivo	3.851	2.501	645	Causa
pessoa que	Substantivo	1.320	47	329	Hiperonímia
aquele que	Substantivo	1.148	3.357	545	Hiperonímia
conjunto de	Substantivo	1.004	316	298	Membro
espécie de	Substantivo	798	2.846	223	Hiperonímia
género/gênero de	Substantivo	29	4.148	48	Hiperonímia
variedade de	Substantivo	455	621	52	Hiperonímia
[a] parte do/da	Substantivo	445	433	107	Parte
qualidade de	Substantivo	777	775	126	Qualidade
qualidade do que é	Substantivo	663	543	105	Qualidade
estado de	Substantivo	299	223	73	Estado
natural ou habitante de/da/do	Substantivo	536	0	79	Local/Origem
instrumento[,] para	Substantivo	94	284	25	Finalidade
.. produzid[o/a] por/pel[o/a]	Substantivo	155	146	60	Produtor
o mesmo que	Verbo	0	166	97	Sinonímia
fazer	Verbo	1.680	1.294	364	Causa
tornar	Verbo	1.359	1.672	266	Causa
ter	Verbo	467	519	139	Propriedade
o mesmo que	Adjectivo	0	2.685	197	Sinonímia
relativo a/á/ao	Adjectivo	1.236	5.554	1.063	Propriedade
que se	Adjectivo	1.602	1.599	485	Propriedade
que tem	Adjectivo	2.698	4.291	477	Parte/Propriedade
diz-se de	Adjectivo	2.066	738	313	Propriedade
relativo ou pertencente	Adjectivo	1.647	9	61	Membro/Propriedade
habitante ou natural de	Adjectivo	0	0	189	Local/Origem
que não é/está	Adjectivo	485	608	98	Antonímia
de modo	Advérbio	398	2.261	109	Maneira
de maneira	Advérbio	49	9	36	Maneira
de forma	Advérbio	30	3	19	Maneira
o mesmo que	Advérbio	0	182	21	Sinonímia

Tabela 1: Padrões nas definições, frequentes e produtivos

- A possibilidade de trocar a ordem das palavras chave `natural` e `habitante` na extracção de relações Local/Origem;
- A consideração de algumas grafias na variante brasileira, que ocorrem no Wikcionário. Por exemplo, as palavras `ato` e `gênero`.

Além da utilização dos padrões estáticos representados na tabela 1, foram aplicadas mais duas regras que se revelaram bastante produtivas para extrair relações semânticas a partir dos três dicionários:

- Sinonímia pode ser extraída a partir de definições com apenas uma palavra, ou uma enumeração de palavras.
- A maior parte das definições de substantivos são estruturadas por *genus* e *differentia*, ou seja, iniciam-se com a apresentação de um género próximo, normalmente um hiperónimo do lema (eventualmente modificado por um adjectivo) e a diferença específica.[5]

[5]A excepção a estes casos é quando a palavra no início da definição é considerada uma "cabeça vazia" (*empty head*, ver Chodorow, Byrd e Heidorn (1985) ou Guthrie et al. (1990)). As cabeças vazias são substantivos que, apesar de iniciarem normalmente definições, não devem ser considerados como hiperónimos. Podem ser ignorados ou, preferencialmente, ser também explorados na extracção de hiperonímia (e.g. `espécie`, `variedade`) ou outras relações, como parte (e.g. `parte`) ou membro (e.g. `membro`, `conjunto`).

5 Aquisição automática de relações semânticas

Como já vimos na secção anterior, as regularidades nas definições de dicionário permitem que a extracção de relações semânticas se baseie num conjunto finito de regras criado manualmente. Este procedimento opõem-se aos algoritmos vulgarmente denominados de *bootstrapping*, que são habitualmente utilizados na extracção de relações semânticas a partir de texto não estruturado (veja-se, por exemplo, o trabalho de Pantel e Pennacchiotti (2006)). O funcionamento desses algoritmos baseia-se num pequeno conjunto de relações de um determinado tipo (sementes), em que existe a máxima confiança, que é utilizado para aprender novas relações.

Em relação a algoritmos de *bootstrapping*, tendo em conta que estes também aprendem

os padrões de extracção, a nossa abordagem tem a desvantagem de requerer mais tempo na construção das gramáticas e de estas não ficarem desde logo adaptáveis a todos os tipos de texto. No entanto, já vimos que no caso dos três dicionários esta desvantagem não se revela um problema. Por outro lado, a nossa abordagem permite-nos um controlo superior sobre os padrões de extracção.

O procedimento para a criação do CARTÃO é, também ele, fortemente inspirado na construção do PAPEL, relatada em Gonçalo Oliveira, Santos e Gomes (2010), e consiste, por isso, também numa fase manual e em duas automáticas. As relações semânticas, estabelecidas entre itens lexicais nas definições e o item lexical definido, são extraídas após o processamento das entradas dos dicionários, representadas no formato descrito na secção 3.2. As instâncias de cada relação são representadas como triplos, da mesma forma que no PAPEL (ver secção 3.1). Descrevemos de seguida o procedimento (exemplificado na figura 1 para um pequeno conjunto de regras e duas definições):

1. **Criação das gramáticas de extracção:** Os padrões mais produtivos são compilados manualmente em gramáticas, especialmente criadas para a extracção de relações entre itens lexicais nas definições e os lemas definidos.
2. **A própria extracção:** As gramáticas são utilizadas em conjunto com um analisador sintáctico, o PEN (Gonçalo Oliveira e Gomes, 2008), que processa as definições do dicionário, representadas no formato introduzido na secção 3.2. Apenas definições de palavras de categoria aberta são processadas. No fim, se uma definição respeita um padrão, são extraídas instâncias de relações semânticas e representadas como triplos p_1 R p_2, onde p_1 é um item lexical na definição, p_2 é o lema definido, e R é o nome da relação estabelecida entre um sentido de p_1 e um sentido de p_2.
3. **Limpeza e lematização:** Após a extracção, algumas relações apresentam argumentos inválidos, incluindo sinais de pontuação ou preposições. Nesta fase, as definições vêem a categoria gramatical dos seus elementos anotada. É para isso utilizado o anotador de categoria gramatical disponibilizado pelo projecto OpenNLP[6], utilizando os módulos para a língua portuguesa[7].

   Esta anotação não é feita antes da extracção porque os modelos do anotador foram treinados em texto de corpos, e não têm a mesma precisão na anotação de definições de dicionário. Além disso, as gramáticas do PAPEL também não consideram estas anotações.

   Depois da anotação, os triplos com argumentos inválidos são descartados. Além disso, se os argumentos dos triplos se encontrarem flexionados, e por isso não definidos directamente no dicionário, são aplicadas algumas regras de lematização.

Apesar das definições do DA se encontrarem em processo de modernização de grafia, os lemas definidos foram mantidas na sua forma original. Por isso, os triplos extraídos a partir deste recurso passaram por uma quarta fase, em que argumentos com sequências que caíram em desuso são modernizados de acordo com as sugestões de Simões, Almeida e Farinha (2010). Ainda assim, de forma a minimizar as possibilidade de gerar palavras inexistentes, de todos os triplos com argumentos alterados mantivemos apenas os 9.163 em que ambos os argumentos se encontravam também no PAPEL. Os demais, que totalizam 23.226 triplos, foram descartados.

6 Análise quantitativa

Nesta secção apresentamos as relações que fazem parte do PAPEL 3.0, e também aquelas que resultaram do processamento dos outros dois dicionários. A partir do DA foram extraídos cerca de 134 mil triplos e do Wikcionário.PT cerca de 57,3 mil. Estes foram depois juntos com os cerca de 190 mil triplos do PAPEL, de forma a constituir o CARTÃO, que aumenta o PAPEL 3.0 em 72%, em relação aos triplos, e em 52% relativamente ao número de lemas abrangidos. Os números apresentados confirmam que, apesar dos dicionários pretenderem cobrir toda a língua, acabam por ser incompletos. A melhor forma de conseguir um recurso mais abrangente é, portanto, juntar conhecimento obtido a partir de vários recursos.

6.1 As relações em números

Na tabela 2 apresentam-se os números de triplos extraídos, de acordo com o dicionário de onde são originários e com o tipo de relação, para o qual é fornecido um exemplo real. À semelhança do que acontecia no PAPEL, para cada relação existe uma relação inversa definida, que se pode obter pela troca dos argumentos e alteração do nome. Por exemplo, as relações

[6] http://incubator.apache.org/opennlp/
[7] http://opennlp.sourceforge.net/models-1.5/

1. Excerto de gramática:

```
...
RAIZ ::= HIPERONIMO_DE <&> ...
...
RAIZ ::= CABECA_VAZIA
CABECA_VAZIA ::= parte
...
RAIZ ::= ...  <&> usado <&> para <&> FAZ_SE_COM
RAIZ ::= parte <&> de <&> TEM_PARTE
RAIZ ::= ...  <&> que <&> contém <&> DET <&> PARTE_DE
...
```

2. Definições e relações extraídas:

```
candeia  nome  utensílio doméstico rústico usado para iluminação, com pavio abastecido a óleo
→ utensílio HIPERONIMO_DE candeia
→ com FAZ_SE_COM candeia
→ iluminação FAZ_SE_COM candeia
espiga  nome  parte das gramíneas que contém os grãos
→ espiga PARTE_DE gramíneas
→ grãos PARTE_DE espiga
...
```

3. Resultado da anotação, limpeza e lematização:

```
candeia  nome  utensílio#n doméstico#adj rústico#adj usado#v-pcp para#prp iluminação#n ,#punc
com#prp pavio#n abastecido#v-pcp a#prp óleo#n
→ utensílio HIPERONIMO_DE candeia
→ iluminação FAZ_SE_COM candeia
espiga  nome  parte#n de#prp as#art gramíneas#n que#pron-indp contém#v-fin os#art grãos#n
→ espiga PARTE_DE gramínea
→ grão PARTE_DE espiga
...
```

Figura 1: Exemplo do processo de aquisição de relações semânticas.

inversas de `sentimento` hiperónimo-de `afecto` e de `vírus` causador-de `doença` são respectivamente `afecto` hipónimo-de `sentimento` e `doença` resultado-de `vírus`.

Verifica-se que cerca de 40% dos triplos do CARTÃO são relações de sinonímia. As relações de hiperonímia são aproximadamente um terço. Dentro das relações restantes destacam-se os triplos referente-a, que se estabelecem entre adjectivos e outras categorias gramaticais, representando cerca de 12% dos triplos do CARTÃO.

Para dar uma ideia da contribuição de cada recurso em termos dos triplos e das suas intersecções, apresentamos a figura 2, onde é possível verificar não só a quantidade de triplos extraídos de cada recurso, mas também a quantidade extraída de dois ou dos três recursos.

A tabela 3 dá outra perspectiva na contribuição de cada dicionário para o CARTÃO. Os conjuntos de triplos extraídos de cada dicionário são comparados dois a dois através do cálculo da sua semelhança e da novidade de cada um em relação a outro, utilizando as seguintes medidas:

$$Sem(A,B) = Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad (1)$$

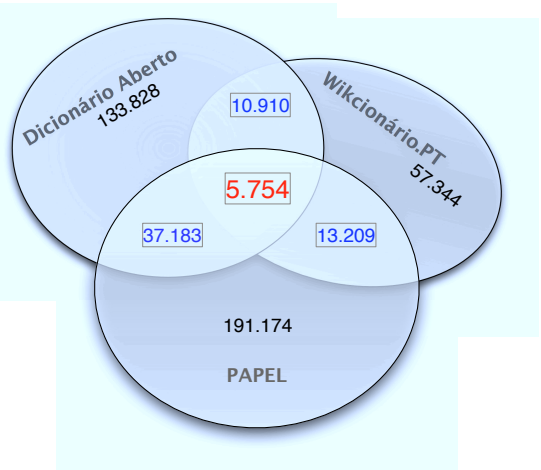

Figura 2: Intersecções dos conjuntos de triplos extraídos. A **preto** o número de triplos de cada recurso, a **azul** as intersecções dois a dois, e a **vermelho** os triplos extraídos dos três dicionários.

$$Novidade(A,B) = \frac{|A| - |A \cap B|}{|A|} \qquad (2)$$

Como seria de esperar, devido às suas diferenças de tamanho, as maiores novidades são do PAPEL e do DA em relação ao Wicionário.PT.

Relação	Args.	Quantidade				Exemplo
		PAPEL	DA	Wikcionário	Únicos	
Sinónimo-de	n,n	40,306	25.046	13.812	67.620	*alegria,satisfação*
	v,v	18.927	11.113	4.650	28.108	*esticar,estender*
	adj,adj	21.726	10.505	6.611	32.364	*racional,filosófico*
	adv,adv	1.178	1.199	277	2.286	*imediatamente,já*
Hiperónimo-de	n,n	62.591	44.777	17.068	97.924	*sentimento,afecto*
Parte-de	n,n	2.424	1.146	614	3.893	*núcleo,átomo*
	n,adj	3.033	3.414	520	5.872	*vício,vicioso*
	adj,n	43	45	16	104	*sujeito,oração*
Membro-de	n,n	5.679	928	1.161	7.328	*aluno,escola*
	n,adj	77	26	25	120	*coisa,coletivo*
	adj,n	968	80	138	1.071	*rural,campo*
Contido-em	n,n	216	124	53	381	*tinta,tinteiro*
	n,adj	176	124	34	287	*óleo,oleoso*
Material-de	n,n	335	513	146	888	*folha_de_papel,caderno*
Causador-de	n,n	951	193	317	1.423	*vírus,doença*
	n,adj	17	8	5	25	*paixão,passional*
	adj,n	494	148	173	748	*horrível,horror*
	n,v	40	17	6	60	*fogo,fundir*
	v,n	6.256	7.140	1.631	10.664	*mover,movimento*
Produtor-de	n,n	910	605	333	1.741	*oliveira,azeitona*
	n,adj	49	26	6	77	*fermentação,fermentado*
	adj,n	352	236	37	515	*fonador,som*
Finalidade-de	n,n	3.659	2.353	1.442	6.978	*sustentação,mastro*
	n,adj	56	40	9	88	*habitação,habitável*
	v,n	4.609	2.230	1.610	7.824	*calcular,cálculo*
	v,adj	236	204	27	374	*comprimir,compressivo*
Tem-qualidade	n,n	740	465	87	1.055	*mórbido,morbidez*
	n,adj	888	667	128	1.273	*assíduo,assiduidade*
Tem-estado	n,n	265	118	44	376	*exaltação,desvairo*
	n,adj	129	102	23	220	*disperso,dispersão*
Lugar-de	n,n	834	405	601	1.483	*Equador,equatoriano*
Maneira-de	adv,n	795	1.537	164	2.172	*ociosamente,indolência*
	adv,adj	345	1.624	135	1.854	*virtualmente,virtual*
Maneira sem	adv,n	116	147	16	250	*prontamente,demora*
	adv,v	6	5	3	13	*seguido,parar*
Antónimo-de	n,n	388	410	59	684	*direito,torto*
Referente-a	adj,n	6.287	5.024	1.793	10.652	*daltónico,daltonismo*
	adj,v	17.718	11,076	3.569	27.902	*musculoso,ter_músculo*
Total		191.174	133.828	57.344	326.798	

Tabela 2: Quantidades e exemplos das relações extraídas.

Ainda assim, verifica-se que todos os recursos apresentam novidades elevadas (sempre superiores a 70%) em relação a cada um dos outros.

A \ B	PAPEL		DA		Wikc.PT	
	Sem	Nov	Sem	Nov	Sem	Nov
PAPEL			0,13	0,81	0,06	0,93
DA	0,13	0,72			0,06	0,92
Wikc.PT	0,06	0,77	0,19	0,81		

Tabela 3: Semelhança (Sem) e novidade (Nov) dos recursos dois a dois, em termos de triplos

6.2 Lemas abrangidos

Fazemos também uma análise da cobertura em termos de lemas abrangidos pelo CARTÃO. A tabela 4 mostra os números de lemas diferentes nos argumentos dos triplos extraídos a partir de cada dicionário, distribuídos de acordo com a sua categoria gramatical. A maior parte dos lemas são substantivos. Depois, para o PAPEL, as categorias mais representadas são os verbos e os adjectivos, por esta ordem. Por outro lado, foram extraídos do DA e do Wikcionário.PT mais adjectivos do que verbos. O PAPEL é o recurso que fornece mais lemas ao CARTÃO, mas o DA não fica muito atrás. Os triplos extraídos do DA englobam mesmo mais substantivos e mais do dobro dos advérbios que o PAPEL.

Cat. gram.	PAPEL	DA	Wikc.PT	Total
Substantivos	55.769	59.879	23.007	89.895
Verbos	22.440	16.672	6.932	32.572
Adjectivos	22.381	18.563	7.113	29.964
Advérbios	1.376	3.073	473	3.443
Total	101.966	98.187	37.525	155,187

Tabela 4: Lemas únicos em relações semânticas

A figura 3 apresenta a contribuição e sobreposição de cada recurso em termos de lemas abrangidos. Além disso, da mesma forma que

calculámos a semelhança e novidade dos conjuntos de triplos extraídos, na tabela 5 apresentamos os mesmos valores, desta vez comparando os lemas envolvidos nos triplos de cada recurso. Tratando-se de lemas, as semelhanças são superiores e as novidades inferiores aos mesmos valores para os triplos. Ainda assim, as novidades são sempre superiores a 35%, o que mostra que para além de diferentes dicionários descrevem diferentes relações semânticas, de cada dicionário foi obtido mais de um terço de novo vocabulário relativamente aos outros.

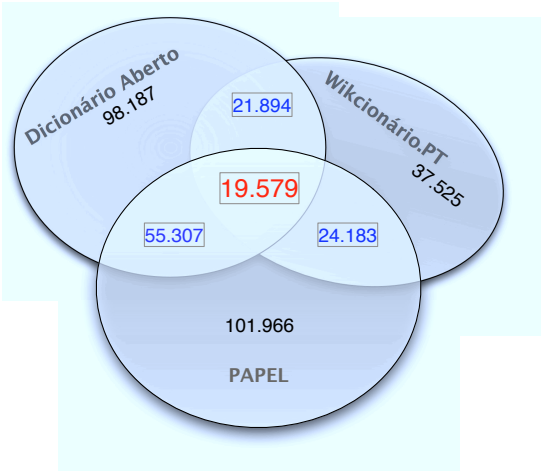

Figura 3: Intersecções dos itens lexicais extraídos. A **preto** o número de itens de cada recurso, a **azul** as intersecções dois a dois, e a **vermelho** os itens obtidos dos três recursos.

A \ B	PAPEL		DA		Wikt.PT	
	Sem	Nov	Sem	Nov	Sem	Nov
PAPEL			0,38	0,46	0,21	0,76
DA	0,38	0,44			0,19	0,78
Wikc.PT	0,21	0,36	0,19	0,42		

Tabela 5: Semelhança (Sem) e novidade (Nov) dos recursos dois a dois, relativamente a lemas incluídos.

6.3 O (novo) Folheador

Em colaboração com a Linguateca[8] foi desenvolvido um interface na rede[9] onde é possível interrogar e navegar pelos conteúdos de recursos baseados em relações semânticas, representadas como triplos, ou seja, da mesma forma que no CARTÃO. A interface, apesar de ter sido criada de raíz, foi inspirada numa interface já existente, e utilizada para navegar no PAPEL, o Folheador. Como a ideia é que esta nova interface venha a substituir o (antigo) Folheador, a nova interface terá o mesmo nome.

A figura 4 mostra o novo Folheador e os primeiros resultados devolvidos para a palavra 'santo'. Tal como no antigo Folheador, é possível procurar por um item lexical para consultar todas as relações onde ele entra. No entanto, agora é possível seleccionar o tipo de relações a procurar, antes de fazer a pesquisa, ou procurar por todas as relações entre dois itens lexicais determinados. O novo Folheador está também feito de forma a permitir a navegação sobre triplos extraídos a partir de vários recursos. Para cada triplo encontrado, são apresentados identificadores dos recursos de que fazem parte ou de onde foram extraídos. O número de recursos de onde o triplo foi obtido pode ser, só por si, utilizado como um indicador de confiança.

Outra novidade importante do novo Folheador é a sua ligação a serviços de pesquisa em corpos, que oferecem uma nova dimensão a relações semânticas representadas como triplos e, consequentemente, ao CARTÃO. Actualmente é possível fazer uma ligação à interface do projecto AC/DC (Santos e Bick, 2000)(Santos, 2011), que permite consultar todas as frases de um conjunto de corpos portugueses onde dois itens lexicais, relacionados num triplo, co-ocorrem. Para alguns tipos de relação existe também uma ligação ao serviço VARRA (Freitas et al., 2010), que usa indirectamente o AC/DC para procurar ocorrências do próprio triplo nos corpos. Para tal, o VARRA transforma o triplo seleccionado num conjunto de frases em que os itens relacionados estão ligados por um padrão que normalmente indica a relação. Com base nestas duas ligações, estão actualmente a ser calculados graus de confiança para cada triplo.

Mais informação sobre o novo Folheador pode ser encontrada em Costa (2011).

7 Cobertura por outros recursos e validação automática

Esta secção é dedicada à avaliação da cobertura do CARTÃO em relação a outros recursos, e também à sua validação automática, baseada na interrogação de um corpo de notícias. A nossa opção pela validação automática deve-se não só ao facto da avaliação manual deste tipo de conhecimento ser tediosa e requerer muito tempo, mas também devido a esta última ser, geralmente, um tipo de avaliação algo subjectivo e difícil de reproduzir. Ainda assim, não descartamos, no futuro, vir a realizar uma avaliação manual de uma amostra significativa do CARTÃO. Além disso, é sempre possível utilizar o serviço

[8] `http://www.linguateca.pt/`

[9] Disponível a partir de `http://linguateca.pt/Folheador/`

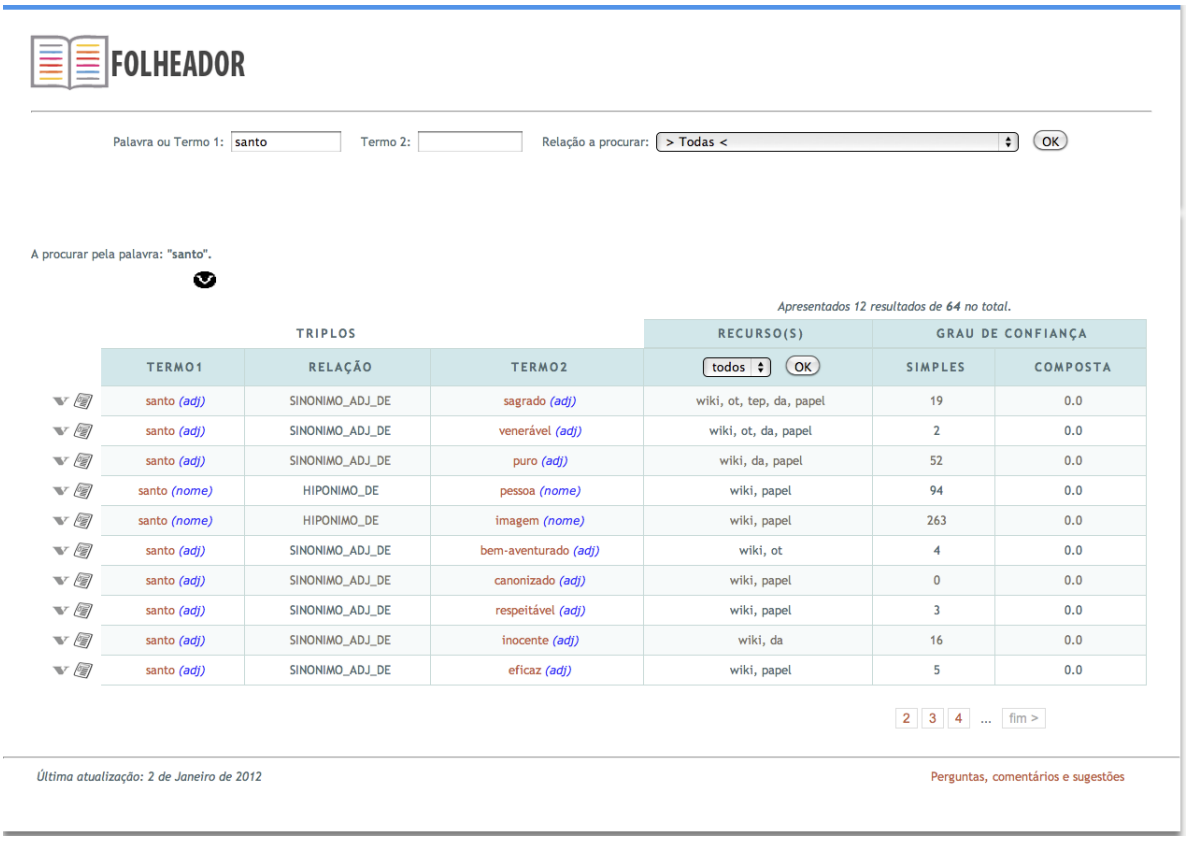

Figura 4: Primeiros resultados da pesquisa por 'santo' no Folheador.

VARRA[10] (Freitas et al., 2010) para validar triplos, também manualmente, com base no contexto em que as palavras relacionadas ocorrem.

7.1 Cobertura por *thesauri* criados manualmente

A cobertura do CARTÃO foi medida em relação a dois recursos lexicais livres para o português, criados de forma manual, nomeadamente o TeP 2.0 (Maziero et al., 2008) e o OpenThesaurus.PT[11] (OT). Estes recursos são ambos *thesauri*, organisados em *synsets*, tal como a WordNet, ainda que não possuam relações entre synsets[12]. O TeP foi criado para o português do Brasil e contém 43.666 itens lexicais, organizadas em 18.795 *synsets*. O OT é uma iniciativa colaborativa, cerca de quatro vezes mais pequena que o TeP, que contém 13.258 itens lexicais organizados em 4.102 *synsets*.

A tabela 6 apresenta a cobertura dos lemas do CARTÃO por ambos os *thesauri*. Entre cerca de 21% (substantivos no DA) e 60% (adjectivos no Wikcionário.PT) dos lemas estão abrangidos pelo TeP. Por outro lado, devido à sua dimensão, para o OT estes números ficam entre os 3% (advérbios no DA) e os 31% (adjectivos no Wikcionário.PT). Considerando apenas o TeP, existe uma maior proporção de adjectivos e advérbios cobertos, comparando com o mesmo número para os substantivos. A baixa proporção de advérbios do DA cobertos e a elevada proporção de substantivos do Wikcionário.PT cobertos são as excepções.

Os triplos do Wikcionário.PT têm a maior proporção de lemas cobertos para todas as categorias, o que se pode explicar pela natureza colaborativa deste recurso. O Wikcionário.PT ainda está em crescimento, e é criado por voluntários, normalmente não peritos, enquanto que o DLP e o DA são dicionários comerciais, criados por lexicógrafos. Assim, enquanto o DLP e o DA, para além de terem vocabulário mais comum, incluírem também definições mais formais e menos convencionais, o Wikcionário tende a utilizar vocabulário mais convencional. Há ainda a destacar que o Wikcionário.PT contém bastantes definições escritas na variante brasileira do português, que é a variante alvo do TeP, contribuindo isto também para a maior proporção de lemas cobertos do primeiro recurso pelo segundo.

A tabela 7 mostra a cobertura de cada triplo de sinonímia pelo TeP – se o TeP tiver um *synset*

[10] Ver `http://www.linguateca.pt/VARRA/`

[11] `http://openthesaurus.caixamagica.pt/`

[12] Na verdade, o TeP contém ligações que representam relações antonímia, mas que não foram utilizadas neste trabalho.

Cat. gram.	TeP						OT					
	PAPEL		DA		Wikc.PT		PAPEL		DA		Wikc.PT	
Substantivos	13.137	23,6%	12.701	21,2%	8.079	35,1%	5.736	10,3%	5.532	9,2%	4.440	19,3%
Verbos	6.029	26,9%	5.835	35,0%	3.138	45,3%	2.731	12,2%	2.644	15,9%	1.977	28,5%
Adjectivos	9.104	40,7%	8.264	44,5%	4.265	60,0%	3.249	14,5%	2.846	15,3%	2.256	31,7%
Advérbios	574	41,7%	683	22,2%	264	55,8%	94	6,8%	94	3,1%	79	16,7%

Tabela 6: Cobertura de lemas por thesauri criados manualmente.

que contenha ambos os argumentos de um triplo de sinonímia, consideramos que o triplo é abrangido pelo TeP. A proporção de triplos cobertos é apresentada para todos os triplos de sinonímia do recurso em questão (Total), bem como considerando apenas os triplos em que ambos os argumentos do triplo existem no TeP (ArgsNoTeP). Decidimos omitir os mesmos dados para o OT por se tratar de um recurso demasiado pequeno. A cobertura da sinonímia de acordo com a categoria gramatical é consistente para os três recursos – mais elevada para sinonímia entre verbos, seguida pela sinonímia entre adjectivos. À semelhança da cobertura dos lemas, a proporção de triplos de sinonímia cobertos pelo TeP é também maior para o Wikcionário.PT.

Para além de darem uma ideia acerca da cobertura do CARTÃO, estes números mostram que os *thesauri* disponíveis e criados manualmente podem ser uma fonte adicional de relações de sinonímia. Além do mais, por se encontrarem em recursos criados manualmente, a confiança na qualidade destas relações é elevada.

7.2 Validação com base na interrogação de um corpo

O procedimento de validação que vamos apresentar de seguida é inspirado num procedimento já utilizado para validar as relações semânticas do PAPEL (ver Gonçalo Oliveira, Santos e Gomes (2010)). Baseia-se num conjunto de padrões discriminadores, indicadores de relações semânticas em texto, e procura por ocorrências desses padrões a ligar os argumentos das relações semânticas extraídas.

Contudo, os resultados apresentados não devem ser confundidos com a precisão das relações extraídas, tendo em conta que:

- Um corpo é um recurso com conhecimento limitado;
- Há imensas formas de exprimir uma relação semântica em texto, o que torna impossível a codificação de todos os padrões e variações possíveis;
- Alguns tipos de relação são específicos dos dicionários, e não é expectável que estejam explícitas em texto de corpos. Isto acontece, por exemplo, para relações entre substantivos e verbos, que implicam a nominalização do verbo, tal como em `aumentar` causador-de `aumento`.
- Alguns estudos (Dorow, 2006) mostram que palavras sinónimas não co-ocorrem frequentemente em corpos, especialmente na mesma frase[13]. Esta ideia vai ao encontro do pressuposto de um sentido por discurso (Gale, Church e Yarowsky, 1992), dado que, principalmente em textos especializados, o autor tenderá a utilizar sempre a mesma palavra para se referir ao mesmo conceito. Também por isso realizamos previamente a validação das relações de sinonímia.

Apesar disto, estes resultados dão-nos uma ideia da utilização das relações extraídas em texto não estruturado. Mais do que isso, se for utilizado o mesmo conjunto de padrões e o mesmo corpo, os resultados são um indicador que pode ser utilizado na comparação de recursos baseados em relações semânticas, e que pode informar acerca da aplicabilidade das suas relações.

Dadas as limitações referidas, apenas foram validados quatro tipos de relações, todos eles entre substantivos. Nesta validação utilizamos o corpo jornalístico CETEMPúblico (Rocha e Santos, 2000) (Santos e Rocha, 2001), onde procuramos por relações de hiperonímia, parte-de, membro-de e finalidade-de, extraídas a partir dos três dicionários utilizados. Apesar de versões anteriores do PAPEL terem já sido validadas através de um procedimento semelhante, repetimos essa validação seguindo os mesmos critérios para os três recursos, de forma a tornar possível uma comparação directa dos resultados. A lista de padrões discriminadores utilizada foi construída com base nos padrões léxico-sintácticos utilizados nas validações anteriores do PAPEL e ainda dos padrões do serviço VARRA.

A tabela 8 apresenta os resultados da validação automática dos triplos extraídos dos três dicionários. A tabela mostra o número, e respectiva proporção, de todos os triplos dos quatro tipos validados cujos argumentos co-ocorrem

[13] Palavras sinónimas tenderão antes a ocorrer em contextos semelhantes.

Cat. gram.	PAPEL			DA			Wikc.PT		
	Cobertos	Total	ArgsNoTeP	Cobertos	Total	ArgsNoTeP	Cobertos	Total	ArgsNoTeP
Substantivos	11.920	30,0%	56,2%	6.821	27,2%	41,4%	4.126	29,9%	50,4%
Verbos	10.063	53,1%	83,5%	5.927	53,3%	76,2%	2.532	54,3%	78,5%
Adjectivos	8.506	39,2%	69,7%	4.891	46,6%	66,9%	2.903	43,9%	71,8%
Advérbios	267	22,7%	38,1%	208	17,3%	27,6%	131	32,9%	47,3%

Tabela 7: Cobertura da sinonímia pelo TeP

em pelo menos uma frase do corpo (ArgsCooc). Mostra-se ainda o número de triplos suportados pelo corpo e a proporção dos triplos cujos argumentos co-ocorrem a que esse número corresponde (Suportados).

Constata-se que a proporção de triplos cujos argumentos co-ocorrem nunca é mais de 37,5% (hiperonímia no Wikcionário.PT), nem menor de 17,5% (hiperonímia no DA). Curiosamente os valores máximo e mínimo obtêm-se para o mesmo tipo de relação, mas recurso diferente. Tanto para o PAPEL, como para o Wikcionário.PT, a proporção de relações membro-de cujos argumentos co-ocorrem no CETEMPúblico é inferior à mesma proporção para as demais relações, nos mesmos recursos. Isto poderá depois contribuir para que seja a relação com mais triplos suportados.

Nos três recursos, a proporção de triplos suportados no corpo é sempre mais elevada para a relação membro-de e mais baixa para finalidade. Acreditamos que a baixa proporção de relações de finalidade suportadas se deve ao facto desta relação não estar tão bem definida semanticamente como as outras três. Além disso, existirão mais padrões discriminadores para esta relação, e os padrões utilizados serão menos frequentes e com mais variações.

Dos triplos de hiperonímia e parte-de do PAPEL e do DA cujos argumentos co-ocorrem no CETEMPúblico, cerca 30% são suportados. Mais uma vez, devido ao seu tamanho e natureza colaborativa, as proporções mais elevadas são obtidas pelo Wikcionário.PT, De forma a dar uma ideia mais clara daquilo em que consistiu a avaliação, a tabela 9 mostra exemplos de frases que suportam triplos do CARTÃO. Nas mesmas frases, os padrões discriminadores encontram-se a negrito.

8 Notas finais

Neste artigo apresentamos o CARTÃO, uma rede léxico-semântica de grandes dimensões, extraída automaticamente a partir de três dicionários da língua portuguesa. Após analisarmos a estrutura das definições nos dicionários utilizados, verificamos que podíamos tirar partido das mesmas regras, baseadas em padrões léxico-sintácticos, para extrair relações semânticas a partir dos três.

Além de utilizar as mesmas gramáticas, a construção do CARTÃO é inspirada na construção do PAPEL, uma rede léxico-semântica pública, também extraída de um dicionário, e incluída no CARTÃO. A versão do PAPEL utilizada neste trabalho é o PAPEL 3.0, a sua versão mais recente, desenvolvida em paralelo com o resto do trabalho aqui apresentado. Analisando os resultados de extracção, mostramos a contribuição de cada dicionário para o CARTÃO, e verificamos que este recurso aumenta o PAPEL 3.0 em mais de 70%. Tendo em conta que os dicionários pretendem abranger toda a língua, é um aumento significativo, e confirma que há vantagens na utilização de mais de um dicionário neste tipo de trabalho.

A cobertura do CARTÃO foi avaliada através da comparação dos lemas que este inclui com os lemas em *thesauri* portugueses de larga cobertura, livres e criados manualmente. Além disso, algumas relações entre substantivos foram validadas automaticamente com base na sua ocorrência num corpo jornalístico.

Entre outras tarefas, no futuro pretendemos, por exemplo, refinar algumas relações (e.g. finalidade-de) e rever o contributo de alguns dos padrões utilizados. De forma a termos uma informação mais clara sobre a qualidade dos vários tipos de relação, estamos a ponderar a realização de uma avaliação manual de uma parte do CARTÃO. Esta avaliação poderá mesmo ser integrada no projecto VARRA, que contribuirá ainda na atribuição de graus de confiança a cada triplo, e na investigação das formas em que palavras relacionadas co-ocorrem no texto de corpos. Além do trabalho aqui descrito, utilizando procedimentos semelhantes, o CARTÃO pode ainda ser enriquecido com relações léxico-semânticas obtidas a partir de outros recursos, incluindo não apenas dicionários, mas também *thesauri*, ou mesmo a Wikipédia (veja-se Herbelot e Copestake (2006) para o inglês, ou Gonçalo Oliveira, Costa e Gomes (2010), para o português).

Desde o início do projecto PAPEL que a opção foi construir um recurso lexical em que os lemas não estivessem divididos em sentidos. Esta opção é justificada, inicialmente, pela ar-

Relação	PAPEL				DA				Wikcionário.PT			
	ArgsCooc		Suportados		ArgsCooc		Suportados		ArgsCooc		Suportados	
Hiperonímia	13.724	21,9%	4.098	29,7%	7.846	17,5%	2.255	28,7%	6.405	37,5%	2.086	32,6%
Parte-de	573	23,6%	186	32,5%	247	21,6%	81	32,8%	226	36,8%	94	41,6%
Membro-de	1.089	19,2%	464	42,6%	303	32,7%	109	36,0%	317	27,3%	147	46,4%
Finalidade	1.017	27,8%	164	16,1%	473	20,1%	65	13,7%	498	34,5%	75	15,1%

Tabela 8: Validação automática do CARTÃO

Relação	Suporte
língua hiperónimo-de *alemão*	*As iniciativas deste gabinete passam geralmente pela promoção de conferências, exposições, workshops e aulas de* ***línguas, como o inglês, alemão*** *ou japonês.*
ciência hiperónimo-de *paleontologia*	*A* ***paleontologia é uma ciência*** *que depende do que se descobre.*
rua parte-de *quarteirão*	*De resto, o* ***quarteirão formado pelas ruas*** *de São João e de Mouzinho da Silveira está, por esse motivo, assente em estacas de madeira...*
mão parte-de *corpo*	*As* ***mãos são a parte do corpo*** *mais atingida (29,7%).*
pessoa membro-de *comissão*	*A* ***comissão é constituída por pessoas*** *que ficaram marcadas pela presença de Dona Amélia: ...*
lobo membro-de *alcateia*	*Mech e os seus colegas constataram que alguns dos cheiros contidos nas marcas de urina servem para os* ***lobos de uma alcateia*** *saberem por onde andou o lobo que deixou as marcas ...*
transporte finalidade-de *embarcação*	*... onde foi descoberto o resto do casco de uma* ***embarcação presumivelmente utilizada no transporte*** *de peças de cerâmica ...*
espectáculo finalidade-de *anfiteatro*	*Sobre a hipótese da construção de «stands» de artesanato e de um* ***anfiteatro para espectáculos****, a edilidade portuense diz ainda não estar nada decidido.*

Tabela 9: Frases exemplo, que suportam relações semânticas.

tificialidade da divisão em sentidos (Kilgarriff, 1996), que é muitas vezes diferente de lexicógrafo para lexicógrafo. Outra razão é, na prática, a inexistência de sinónimos. Existem sim quase-sinónimos, que suscitam questões interessantes, e que também devem ser exploradas. Além disso, em linguagem natural, o estudo da vagueza é tão ou mais importante que o estudo da ambiguidade, ou seja, muitos dos casos mais interessantes e frequentes são casos de vagueza, o que favorece a opção em não se separar os lemas, muitas vezes de forma arbitrária, em sentidos.

Por outro lado, reconhecemos que, ainda que artificial, em muitas tarefas PLN seja útil a existência de um recurso em que as palavras estejam separadas nos seus sentidos mais típicos, tal como acontece numa *wordnet*. Assim, no projecto Onto.PT (Gonçalo Oliveira e Gomes, 2010) (Gonçalo Oliveira e Gomes, 2011c), onde se integra o desenvolvimento do CARTÃO, pretendemos manter o mesmo recurso livre e estruturado de duas formas alternativas – rede baseada em lemas, como o CARTÃO, e ontologia semelhante à *wordnet*, o Onto.PT propriamente dito – para que os investigadores utilizem aquela que lhes for mais útil.

Muito resumidamente, para a construção automática do Onto.PT é necessário passar por três fases. A primeira, passa pela extracção de relações semânticas a partir de recursos textuais, o que dá origem exactamente ao CARTÃO. De seguida, o objectivo é identificar automaticamente *synsets* no meio das relações de sinonímia (Gonçalo Oliveira e Gomes, 2011a). Por fim, procuram-se integrar, também automaticamente, os restantes triplos, através da associação dos lemas dos seus argumentos a *synsets* adequados (Gonçalo Oliveira e Gomes, 2011b). O resultado é uma ontologia lexical para o português, com uma estrutura semelhante a uma *wordnet*, ou seja, uma rede constituída por relações semânticas entre *synsets*.

As relações que compõem o PAPEL podem ser livremente descarregadas a partir do sítio da Linguateca, enquanto que as relações extraídas a partir do DA e do Wikcionário.PT podem ser descarregadas a partir do sítio do projecto Onto.PT, respectivamente:

- `http://www.linguateca.pt/PAPEL/`
- `http://ontopt.dei.uc.pt/`

Agradecimentos

Gostaríamos de agradecer à Diana Santos pela sugestão do nome para o recurso, pela orientação no desenvolvimento do Folheador e pela argumentação da utilidade de recursos lexicais em que não é feita a distinção entre sentidos.

Agradecemos à Linguateca por ter financiado o desenvolvimento do novo Folheador, através da

contratação do Hernani Costa.

Agradecemos também à Cláudia Freitas por, em conjunto com a Diana Santos, ter participado activamente na discussão opções para o PAPEL 3.0, e ao Alberto Simões por nos ter cedido a mais recente revisão do DA modernizado.

Por fim, agradecemos aos revisores pelos seus valiosos comentários, que contribuíram para melhorar este este trabalho.

Hugo Gonçalo Oliveira é apoiado pela bolsa de doutoramento da FCT SFRH/BD/44955/2008, co-financiada pelo FSE.

Referências

Agirre, Eneko, Oier Lopez De Lacalle, e Aitor Soroa. 2009. Knowledge-based WSD on specific domains: performing better than generic supervised WSD. Em *Proceedings of 21st International Joint Conference on Artifical Intelligence (IJCAI)*, pp. 1501–1506, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Alshawi, Hiyan. 1987. Processing dictionary definitions with phrasal pattern hierarchies. *Computational Linguistics*, 13(3-4):195–202.

Amsler, Robert A. 1981. A taxonomy for english nouns and verbs. Em *Proceedings of 19th annual meeting on Association for Computational Linguistics*, pp. 133–138, Morristown, NJ, USA. ACL Press.

Calzolari, Nicoletta, Laura Pecchia, e Antonio Zampolli. 1973. Working on the italian machine dictionary: a semantic approach. Em *Proceedings of 5th Conference on Computational Linguistics*, pp. 49–52, Morristown, NJ, USA. ACL Press.

Chodorow, Martin S., Roy J. Byrd, e George E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. Em *Proceedings of 23rd annual meeting on Association for Computational Linguistics*, pp. 299–304, Morristown, NJ, USA. ACL Press.

Costa, Hernani. 2011. O desenho do novo folheador. Relatório técnico, Linguateca, Dezembro, 2011. `http://www.linguateca.pt/Equipa/Hernani/HernaniCostarelFolheador.pdf`.

Cruse, D. A. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

2005. *Dicionário PRO da Língua Portuguesa*. Porto Editora, Porto.

Dolan, William, Lucy Vanderwende, e Stephen D. Richardson. 1993. Automatically deriving structured knowledge bases from online dictionaries. Em *Proceedings of the 1st Conference of the Pacific Association for Computational Linguistics*, PACLING'93, pp. 5–14.

Dolan, William B. 1994. Word sense ambiguation: clustering related senses. Em *Proceedings of 15th International Conference on Computational Linguistics*, COLING'94, pp. 712–716, Morristown, NJ, USA. ACL Press.

Dorow, Beate. 2006. *A Graph Model for Words and their Meanings*. Tese de doutoramento, Institut fur Maschinelle Sprachverarbeitung der Universitat Stuttgart.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Freitas, Cláudia, Diana Santos, Hugo Gonçalo Oliveira, e Violeta Quental. 2010. VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC. Em *Livro do IX Encontro de Linguística de Corpus*, ELC 2010.

Gale, William A., Kenneth W. Church, e David Yarowsky. 1992. One sense per discourse. Em *Proceedings of the HLT'91 workshop on Speech and Natural Language*, pp. 233–237, Morristown, NJ, USA. ACL Press.

Gonçalo Oliveira, Hugo e Paulo Gomes. 2008. Utilização do (analisador sintáctico) PEN para extracção de informação das definições de um dicionário. Relatório técnico, CISUC, November, 2008. PAPEL Tech Report 3, `http://linguateca.dei.uc.pt/papel/GoncaloOliveiraetal2008relPAPEL3.pdf`.

Gonçalo Oliveira, Hugo, Hernani Costa, e Paulo Gomes. 2010. Extracção de conhecimento léxico-semântico a partir de resumos da Wikipédia. Em *Actas do II Simpósio de Informática (INFORUM 2010)*, pp. 537–548. Universidade do Minho.

Gonçalo Oliveira, Hugo e Paulo Gomes. 2010. Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. Em *Proceedings of 5th European Starting AI Researcher Symposium (STAIRS 2010)*. IOS Press.

Gonçalo Oliveira, Hugo e Paulo Gomes. 2011a. Automatic discovery of fuzzy synsets from dictionary definitions. Em *Proceedings of 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1801–1806, Barcelona, Spain. IJCAI/AAAI.

Gonçalo Oliveira, Hugo e Paulo Gomes. 2011b. Ontologising relational triples into a portuguese thesaurus. Em *Proceedings of the 15th Portuguese Conference on Artificial Intelligence (EPIA 2011)*, pp. 803–817, Lisbon, Portugal, October, 2011. APPIA.

Gonçalo Oliveira, Hugo e Paulo Gomes. 2011c. Onto.PT: Construção automática de uma ontologia lexical para o português. Em Ana R. Luís, editor, *Estudos de Linguística*, volume 1. Imprensa da Universidade de Coimbra, Coimbra. No prelo.

Gonçalo Oliveira, Hugo, Diana Santos, e Paulo Gomes. 2010. Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. *Linguamática*, 2(1):77–93, May, 2010.

Guthrie, L., B. Slator, Y. Wilks, e R. Bruce. 1990. Is there content in empty heads? Em *Proceedings of the 13th International Conference on Computational Linguistics*, volume 3 of *COLING'90*, pp. 138–143, Helsinki, Finland.

Hearst, Marti A. 1998. Automated Discovery of WordNet Relations. Em *(Fellbaum, 1998)*. pp. 131–151.

Herbelot, Aurelie e Ann Copestake. 2006. Acquiring ontological relationships from wikipedia using RMRS. Em *Proceedings of ISWC 2006 Workshop on Web Content Mining with Human Language Technologies*.

Ide, N. e J. Veronis. 1995. Knowledge extraction from machine-readable dictionaries: An evaluation. Em *Machine Translation and the Lexicon, LNAI*. Springer.

Kilgarriff, A. 1996. Word senses are not bona fide objects: implications for cognitive science, formal semantics, nlp. Em *Proceedings of 5th International Conference on the Cognitive Science of Natural Language Processing*, pp. 193–200.

Knight, Kevin e Steve K. Luk. 1994. Building a large-scale knowledge base for machine translation. Em *Proceedings of 12th national conference on Artificial intelligence (vol. 1)*, AAAI '94, pp. 773–778, Menlo Park, CA, USA. American Association for Artificial Intelligence.

Marrafa, Palmira. 2002. Portuguese Wordnet: general architecture and internal semantic relations. *DELTA*, 18:131–146.

Maziero, Erick G., Thiago A. S. Pardo, Ariani Di Felippo, e Bento C. Dias-da-Silva. 2008. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. Em *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pp. 390–392.

Navarro, Emmanuel, Franck Sajous, Bruno Gaume, Laurent Prévot, ShuKai Hsieh, Tzu Y. Kuo, Pierre Magistry, e Chu R. Huang. 2009. Wiktionary and NLP: Improving synonymy networks. Em *Proceedings of Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pp. 19–27, Suntec, Singapore. ACL Press.

Navigli, Roberto. 2009. Using cycles and quasi-cycles to disambiguate dictionary glosses. Em *Proceedings of the 12th Conference on European chapter of the Association for Computational Linguistics*, EACL'09, pp. 594–602, Athens, Greece.

Navigli, Roberto, Paola Velardi, Alessandro Cucchiarelli, e Francesca Neri. 2004. Extending and enriching Wordnet with OntoLearn. Em *Proceedings of 2nd Global WordNet Conference (GWC)*, pp. 279–284, Brno, Czech Republic. Masaryk University.

Nichols, Eric, Francis Bond, e Dan Flickinger. 2005. Robust ontology acquisition from machine-readable dictionaries. Em *Proceedings of 19th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1111–1116. Professional Book Center.

Pantel, Patrick e Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. Em *Proceedings of 21st International Conference on Computational Linguistics and 44th annual meeting of the Association for Computational Linguistics*, pp. 113–120, Sydney, Australia. ACL Press.

Pasca, Marius e Sanda M. Harabagiu. 2001. The informative role of WordNet in open-domain question answering. Em *Proc. NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pp. 138–143, Pittsburgh, USA.

Pérez, Leticia Anton, Hugo Gonçalo Oliveira, e Paulo Gomes. 2011. Extracting lexical-semantic knowledge from the portuguese wiktionary. Em *Proceedings of the 15th Portuguese Conference on Artificial Intelligence (EPIA 2011)*, pp. 703–717, Lisbon, Portugal, October, 2011. APPIA.

Peters, Wim, Ivonne Peters, e Piek Vossen. 1998. Automatic Sense Clustering in EuroWordNet.

Em *Proceedings of 1st International Conference on Language Resources and Evaluation*, LREC'98, pp. 409–416, Granada, May, 1998.

Plaza, Laura, Alberto Díaz, e Pablo Gervás. 2010. Automatic summarization of news using wordnet concept graphs. *International Journal on Computer Science and Information System (IADIS)*, V:45–57.

Richardson, Stephen D., William B. Dolan, e Lucy Vanderwende. 1998. Mindnet: Acquiring and structuring semantic information from text. Em *Proceedings of 17th International Conference on Computational Linguistics*, COLING'98, pp. 1098–1102.

Rocha, Paulo Alexandre e Diana Santos. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. Em Maria das Graças Volpe Nunes, editor, *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*, pp. 131–140, São Paulo, 19-22 de Novembro, 2000. ICMC/USP.

Sajous, Franck, Emmanuel Navarro, Bruno Gaume, Laurent Prévot, e Yannick Chudy. 2010. Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary. Em *Advances in Natural Language Processing, 7th International Conference on NLP (IceTAL)*, volume 6233 of *LNCS*, pp. 332–344, Reykjavik, Iceland. Springer.

Santos, Diana. 2011. Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. *OSLa: Oslo Studies in Language*, 3(2):113–128, Junho, 2011. Volume edited by J.B.Johannessen, Language variation infrastructure.

Santos, Diana, Anabela Barreiro, Cláudia Freitas, Hugo Gonçalo Oliveira, José Carlos Medeiros, Luís Costa, Paulo Gomes, e Rosário Silva. 2010. Relações semânticas em português: comparando o TeP, o MWN.PT, o Port4NooJ e o PAPEL. Em A. M. Brito, F. Silva, J. Veloso, e A. Fiéis, editores, *Textos seleccionados. XXV Encontro Nacional da Associação Portuguesa de Linguística*. APL, pp. 681–700.

Santos, Diana e Eckhard Bick. 2000. Providing Internet access to Portuguese corpora: the AC/DC project. Em *Proc. of the 2nd International Conf. on Language Resources and Evaluation (LREC)*, pp. 205–210.

Santos, Diana e Paulo Rocha. 2001. Evaluating CETEMPúblico, a free resource for Portuguese. Em *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pp. 442–449. ACL Press, 9-11 July, 2001.

Simões, Alberto e Rita Farinha. 2011. Dicionário Aberto: Um novo recurso para PLN. *Viceversa*, (16):159–171, December, 2011.

Simões, Alberto, José João Almeida, e Rita Farinha. 2010. Processing and extracting data from dicionário aberto. Em *Proceedings of International Conference on Language Resources and Evaluation*, LREC 2010, Malta.

Toral, Antonio, Rafael Muñoz, e Monica Monachini. 2008. Named entity wordnet. Em *Proc. International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco. ELRA.

Wandmacher, Tonio, Ekaterina Ovchinnikova, Ulf Krumnack, e Henrik Dittmann. 2007. Extraction, evaluation and integration of lexical-semantic relations for the automated construction of a lexical ontology. Em *3rd Australasian Ontology Workshop (AOW)*, volume 85 of *CRPIT*, pp. 61–69, Gold Coast, Australia. ACS.

Weale, Timothy, Chris Brew, e Eric Fosler-Lussier. 2009. Using the wiktionary graph structure for synonym detection. Em *Proceedings of 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, People's Web '09, pp. 28–31, Stroudsburg, PA, USA. ACL Press.

Zesch, Torsten, Christof Müller, e Iryna Gurevych. 2008. Using wiktionary for computing semantic relatedness. Em *Proceedings of 23rd National Conference on Artificial Intelligence (AAAI)*, pp. 861–866. AAAI Press.