Extracção de relações semânticas entre palavras a partir de um dicionário: primeira avaliação

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

Diana Santos
Linguateca, SINTEF ICT, Noruega
diana.santos@sintef.no

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

Resumo

Neste artigo apresentamos o PAPEL, um recurso lexical para o português, constituído por relações entre palavras, extraídas de forma automática de um dicionário da língua geral através da escrita manual de gramáticas para esse efeito. Depois de contextualizarmos o tipo de recurso e as opções tomadas, fornecemos uma visão do processo da sua construção, apresentando as relações incluídas e a sua quantidade, correspondendo à versão 1.1. Apresentamos também uma primeira avaliação, que tomou duas formas: para as relações de sinonímia, a comparação com o TeP 2.0, um recurso publicamente acessível e de cobertura vasta; para as outras relações, interrogando corpos em português. Esta segunda forma pode ser efectuada automaticamente, ou recorrendo a avaliadores. Nesta última vertente, integrado no projecto AC/DC, é oferecido mais um serviço de validação de relações à comunidade do processamento computacional da língua portuguesa.

1 Introdução

Cada vez mais os estudos do processamento da língua exigem que haja acesso computacional a informação semântica, e é cada vez mais frequente o recurso a redes ou ontologias lexicais que tentam cobrir o panorama lexical de uma língua toda, ao invés, ou como complemento, de terminologias, cujo objectivo é descrever uma área específica do conhecimento. A ontologia lexical paradigmática é a WordNet (Fellbaum, 1998), também chamada WordNet de Princeton (WordNet.Pr), embora uma ontologia mais relacionada com o nosso trabalho seja a MindNet (Richardson, Dolan e Vanderwende, 1998).

Neste artigo apresentamos o PAPEL, Palavras Associadas Porto Editora - Linguateca, `http://www.linguateca.pt/PAPEL` (desde 17 de Agosto de 2009 livre e publicamente acessível), que é pioneiro para o português, ao tentar obter uma ontologia lexical semi-automaticamente a partir de um dicionário, o *Dicionário PRO da Língua Portuguesa da Porto Editora* (dic, 2005).

Como é notado por Sampson (2000) na sua apreciação da WordNet.Pr, é curioso que tenha sido uma abordagem manual a preferida pela comunidade do processamento de linguagem natural (PLN), mas o que é certo é que a maior parte dos projectos associados ou inspirados pela WordNet seguem uma metodologia que usa peritos para criar o recurso manualmente. Pensamos que uma das razões para isto se deve à questão dos direitos de autor, e nesse aspecto pode ser que o PAPEL seja o primeiro recurso totalmente público baseado num dicionário comercial, visto que a MindNet é propriedade de uma empresa.

Visto que não existe ainda uma terminologia completamente consensual, cumpre indicar aqui, na senda de Veale (2007), o que designamos por *ontologia lexical* de uma dada língua:

- uma estrutura de conhecimento que relaciona itens lexicais (vulgo, palavras) de uma língua entre si, por relações que têm a ver com o significado desses mesmos itens;
- uma estrutura que pretende abranger a língua toda e não conhecimento de um domínio em particular, ou seja, que não se encontre restrita a campos específicos.

Deixamos desde já bem claro que, dentro desta descrição razoavelmente abrangente, existem muitas perguntas específicas a que cada criador de recurso terá de dar uma resposta, assim como não há respostas precisas para o que é uma "palavra" (e de facto a maior parte das ontologias lexicais de que temos conhecimento

usam também expressões como nós) ou o que é a "língua toda".

De um ponto de vista operacional, é mais natural desde já afirmarmos que o PAPEL não pretende ser uma resposta definitiva a estas questões, mas sim uma abordagem concreta que se apoiou no trabalho de lexicógrafos quanto à noção de palavra/entrada e ao conjunto de itens que fazem parte da língua geral, mas que, sabendo que a língua é uma entidade claramente dinâmica, a nossa intenção é vir a expandir o PAPEL tendo em conta esse facto.

Há no entanto duas questões, completamente ortogonais, que nos parecem estabelecer uma delimitação clara na paisagem das ontologias lexicais, e sobre as quais posicionamos de imediato aqui o PAPEL:

- o carácter público ou privado de um recurso: em que o PAPEL alinha com a WordNet;
- a construção manual ou automática a partir de um dicionário com definições (ou seja, um recurso já existente), em que o PAPEL alinha com a MindNet.

Outras opções tomadas, e que nos separam de outros recursos ou abordagens, serão mencionadas à medida que as formos apresentando.

Por interessar mais à audiência deste texto, e também a nós, vamos centrar a discussão nos recursos que existem para o português de que temos conhecimento, nomeadamente a WordNet.PT (Marrafa, 2002), o TeP (Dias da Silva, Oliveira e Moraes, 2002) e a WordNet.BR, e ainda a MultiWordNet (`http://mwnpt.di.fc.ul.pt`).

É importante contudo referir que não vemos nem desenvolvemos o PAPEL[1] como sendo um competidor em relação ao trabalho já existente, mas sim como mais uma contribuição para obter informação semântica de cobertura vasta para o português.

Consideramos, de facto, que a situação ideal seria a de ter uma ontologia lexical pública para todo o português (embora naturalmente entrando em conta com as diferenças entre as variedades ou variantes da língua (Barreiro, Wittmann e de Jesus Pereira, 1996)). Em Santos et al. (2009), apresentamos uma primeira comparação entre vários recursos que sublinha a sua complementaridade.

[1]Quando o projecto de construção do PAPEL foi iniciado pela Linguateca em colaboração com a Porto Editora, após assinatura de um protocolo em Maio de 2006, não havia nenhum recurso publicamente disponível para o português. Congratulamo-nos muitíssimo pelo facto de existirem agora vários.

Nessa linha, tentaremos convencer os leitores de que as formas de avaliação que descrevemos na secção 4 constituem um bom início para uma ligação e consequente actualização de ambos os recursos envolvidos (o TeP e o PAPEL), além de apresentarmos também uma oferta de validação para outros recursos existentes ou que venham a ser desenvolvidos para o português em conjunto com a interrogação de corpos em português.

2 Contexto

Desde muito cedo que foi reconhecido que, para realizar o processamento computacional de uma língua, seria necessário o acesso a recursos de grande cobertura, como o são as ontologias lexicais, ou antes de esse termo ser cunhado, a dicionários em forma electrónica ou bases de dados lexicais, por um lado, ou bases de conhecimento sobre o mundo, por outro. Para uma excelente discussão da diferença e relação entre ontologias e bases de dados lexicais, veja-se (Hirst, 2004). Outras abordagens interessantes em relação a essa questão são (Dahlgren, 1995) e (Marcellino e Dias-da-Silva, 2009).

2.1 Modelos de ontologia lexical

2.1.1 A escola da WordNet

A WordNet.Pr (Fellbaum, 1998) é uma ontologia lexical para o inglês, construída manualmente, que procura representar a forma como o ser humano processa o vocabulário. Está disponível gratuitamente e ao longo dos anos tem sido amplamente utilizado pela comunidade do PLN. A sua estrutura mais básica é um grupo de sinónimos (do inglês, *synset*), ou seja, um conjunto de palavras que, em determinado contexto, podem ter o mesmo significado e ser utilizadas para representar o mesmo conceito. Uma rede semântica estabelece-se na WordNet.Pr através de ligações, correspondentes a relações semânticas, entre os nós, que correspondem aos grupos de sinónimos. Entre as relações cobertas encontram-se a hiponímia e a meronímia (entre substantivos) e a troponímia e a implicação (entre verbos). Há ainda a dizer que no léxico da WordNet.Pr há uma clara distinção entre nós que são substantivos, verbos, adjectivos, advérbios ou palavras gramaticais. Além de ser possível levantar gratuitamente várias versões da WordNet.Pr, através da sua página, em `http://wordnet.princeton.edu` é também possível interrogar a sua versão mais recente, 3.0 através de uma interface na rede.

Dado o enorme sucesso da WordNet.Pr, o seu modelo foi seguido para representar ontologias lexicais noutras línguas. Dessas

destacam-se as wordnets criadas para as línguas presentes no projecto EuroWordNet (Vossen, 1997), mais propriamente o holandês, castelhano, italiano, francês, alemão e estónio. A ideia do EuroWordNet foi alinhar várias wordnets com a WordNet.Pr.

Destacamos ainda as wordnets para a língua portuguesa, a WordNet.PT (Marrafa, 2002), para a variante europeia, e a WordNet.BR (Dias da Silva, Oliveira e Moraes, 2002), para a variante brasileira[2]. Há no entanto a lamentar que ambos os projectos tardem a tornar os seus conteúdos acessíveis para o público. Por exemplo, apesar da existência de uma interface na rede para interrogar a WordNet.PT (ou parte dela), a partir de `http://cvc.instituto-camoes.pt:8080/wordnet/index.jsp`, não é costume ser possível realizar pesquisas porque os sistema se encontra permanentemente em manutenção. Contudo, os grupos de sinónimos da WordNet.BR, bem como as relações de antonímia, encontram-se disponíveis no Thesaurus Electrônico do Português, o TeP (Maziero et al., 2008), também ele construído de acordo com os princípios da WordNet.Pr.

Inspirado pelo EuroWordNet, o projecto MultiWordNet (Pianta, Bentivogli e Girardi, 2002) procurou também alinhar várias wordnets com a WordNet.Pr, mas desta vez, ao invés de se procurar as correspondências possíveis entre as wordnets existentes nas diferentes línguas e a WordNet.Pr, a ideia foi criar novas wordnets onde fosse mantida a maior parte dos nós e relações presentes na WordNet.Pr. Desta forma, na MultiWordNet, wordnets para o italiano, o espanhol, o romeno, o hebraico, o latim e, mais recentemente, o português (`http://mwnpt.di.fc.ul.pt`) estão alinhadas com a WordNet.Pr.

2.1.2 A MindNet

Além do modelo da WordNet, outro tipo de recurso que pode ser visto como uma ontologia lexical é a base de conhecimento MindNet (Richardson, Dolan e Vanderwende, 1998).

A MindNet é mais do que um recurso estático e pode ser visto como uma metodologia que envolve um conjunto de ferramentas para adquirir, estruturar, aceder e explorar, de forma automática, informação léxico-semântica contida em texto. Como, numa fase inicial, o recurso foi construído a partir de um dicionário para a língua inglesa, a sua estrutura é baseada em entradas de dicionário. Desta forma, para cada palavra definida, além de informação típica num dicionário (e.g. informações gramaticais) existe um conjunto de registos associados aos sentidos que a palavra pode ter. Por sua vez, para cada sentido, além da definição, encontram-se ligações a outras entradas, sendo que cada ligação tem um tipo correspondente a uma relação gramatical (e.g. sujeito típico, predicado típico) ou semântica (e.g. sinónimo, hiperónimo, parte, causa, finalidade, maneira). Estas relações são extraídas com base na aplicação de regras sobre árvores sintáctico-semânticas, produzidas por um analisador sintáctico de vasta cobertura. Cada relação estabelecida tem um peso atribuído de acordo com a sua saliência.

A MindNet pode ser interrogada através do MindNet Explorer (MNEX) (Vanderwende et al., 2005), a partir do `http://stratus.research.microsoft.com/mnex/Main.aspx`. Este interface permite procurar caminhos (de relações semânticas) entre duas palavras. No entanto, apesar de existirem opções para seleccionar a categoria gramatical de cada palavra e ainda definir a procura apenas para um tipo de relações, ao mostrar os resultados, o MNEX parece não ter em conta estas opções.

2.1.3 Outros recursos semânticos

As bases de senso comum são outro tipo de recursos semânticos, sendo o recurso mais conhecido o Cyc (Lenat, 1995), uma base de conhecimento baseada em lógica de predicados de primeira ordem, que vem sendo criada de forma manual.

Outro recurso deste tipo é a ConceptNet (Liu e Singh, 2004), construído de forma automática a partir de frases semi-estruturadas. A ConceptNet utiliza uma representação semelhante à do WordNet.Pr, mas inclui conhecimento mais informal, de uma natureza mais prática e tem um maior elenco de relações (tais como propriedade de, sub-evento de, efeito de, utilizado para).

Tanto o Cyc como a ConceptNet têm associadas capacidades de raciocínio, de forma a ser possível inferir novas relações. No entanto, enquanto no Cyc o raciocínio é realizado sob representações em lógica de predicados, na ConceptNet o raciocínio é feito sobre representações em linguagem natural.

A FrameNet (Baker, Fillmore e Lowe, 1998), por seu lado, é uma rede semântica baseada no conceito de enquadramentos (em inglês, *frames*) (Fillmore, 1982). Nesta representação, cada enquadramento descreve um objecto, um

[2]Para uma comparação entre as ontologias lexicais existentes para o português, mais propriamente a WordNet.PT, a WordNet.BR e o TeP, a MultiWordNet.PT, e ainda o PAPEL, recomenda-se a leitura de (Santos et al., 2009).

evento ou um estado, que corresponde a um conceito e se pode relacionar com outros enquadramentos, através de um conjunto de relações semânticas (e.g. herança, sub-frame, causador, utiliza). Para o português existe já um projecto seguidor deste modelo de recurso, o FrameNet Brasil (Salomao, 2009), `http://www.framenetbr.ufjf.br/`.

Devemos também citar o Port4NooJ (Barreiro, 2008), um conjunto de recursos linguísticos construídos no ambiente de desenvolvimento linguístico do NooJ (Silberztein e Varadi, forthcoming 2009), que tem em vista o processamento automático do português. Estes recursos encontram-se publicamente disponíveis em `http://www.linguateca.pt/Repositorio/Port4Nooj/` e são usados em várias ferramentas públicas para o português e outras línguas. Os recursos correspondem a léxicos e a gramáticas com finalidades diversas: análise morfológica, sintáctico-semântica, desambiguação, identificação de unidades lexicais multipalavra, parafraseamento e tradução. O Port4NooJ inclui além disso uma extensão bilingue, permitindo a sua utilização em aplicações como a tradução automática do português para o inglês. As diferentes propriedades associadas aos itens lexicais contidas nos recursos provêm do OpenLogos, um sistema de tradução automática em código aberto derivado do sistema Logos (Scott, 2003), mas novas propriedades têm sido adicionadas através do NooJ e encontram-se em fase de validação.

2.1.4 Sentidos numa ontologia lexical

Enquanto que, pela escola da WordNet, cada nó da rede representa um sentido e uma "mesma" palavra pode pertencer a vários nós, que são sim as unidades básicas, no PAPEL a única distinção de sentidos feita tem a ver com a categoria gramatical, ou seja, um nó do PAPEL é uma palavra gráfica (com uma dada categoria: substantivo, adjectivo, etc.). Esta opção tem duas razões de ser: uma filosófica e outra prática. A primeira prende-se com a concepção de que a língua é soberana (Santos, 2006) e distinções de sentido são sempre imprecisas (Kilgarriff, 1996) e artificiais; veja-se (de Saussure, 1916) para a descrição de uma língua como sistema sincrónico, e (Edmonds e Hirst, 2002) sobre o problema dos quase-sinónimos. A segunda razão tem a ver com o facto de, nas definições de um dicionário, as palavras que ocorrem nas definições não aparecem indexadas pelos sentidos, tornando por isso quase impossível fazer essa identificação automaticamente.

Aliás, confrontado com o mesmo problema, no âmbito da MindNet, Dolan (1994) propôs fazer a "ambiguação" de sentidos relacionados. Desta forma, numa primeira fase de construção, a MindNet (Richardson, Dolan e Vanderwende, 1998) é uma rede entre palavras, tal e qual se encontram no dicionário, e os seus registos são relativos a palavras. Apenas numa segunda fase se procura atribuir um sentido a cada uma destas palavras, tirando partido dos campos de domínio ou de co-ocorrências nas definições.

Também a partir de uma rede onde a unidade básica é a palavra, sem qualquer distinção de sentidos, e onde as ligações, pesadas, apenas indicam a co-ocorrência em corpos, Dorow (2006) aplica algoritmos estatísticos para extrair informação semântica interessante. Por exemplo, quando dois nós não estão ligados ou têm uma ligação muito fraca (isto é, as palavras não co-ocorrem frequentemente), mas têm uma vizinhança semelhante, é provável que sejam sinónimos. Por outro lado, quando um nó é a única ligação entre duas sub-redes, é provável que se esteja perante uma palavra com dois sentidos.

Ainda relativamente à representação dos sentidos numa ontologia lexical, os recursos que resultam de uma tradução cega de um recurso deste tipo feito para uma língua diferente, como as MultiWordNets, têm de lidar, adicionalmente às questões decorrentes da imprecisão existente na identificação de sentidos, com problemas mais específicos relacionados com a tradução. Como línguas diferentes representam diferentes realidades sociais e culturais, estas não cobrem exactamente a mesma parte do léxico e, mesmo nas partes que lhes são comuns, os vários conceitos são normalmente lexicalizados de forma diferente (Hirst, 2004). Isto leva a que, por exemplo, na MultiWordNet.PT faltem palavras para identificar alguns conceitos importados da WordNet.Pr (Santos et al., 2009), assim como muito provavelmente faltarão conceitos específicos das realidades portuguesa e brasileira.

2.2 Abordagens para a construção de uma ontologia lexical

Há basicamente três formas consagradas de construção de um recurso semântico de cobertura larga: (i) trabalho manual; (ii) processamento de corpos; e (iii) processamento de dicionários; apesar de novas ideias terem surgido nos últimos tempos, como por exemplo através da análise de logs (Costa e Seco, 2008) ou jogos colaborativos.

O PAPEL (Gonçalo Oliveira et al., 2008) seguiu a terceira via: foi construído a partir da análise automática das definições constantes numa versão electrónica do *Dicionário PRO da Língua Portuguesa* (dic, 2005). A utilização de

dicionários em formato electrónico com vista à construção de recursos lexicais iniciou-se há cerca de quarenta anos, com os estudos de Calzolari, Pecchia e Zampolli (1973) para o italiano e de Amsler (1981) para o inglês. Os autores que utilizaram dicionários apontam várias razões para a sua escolha como ponto de partida para a construção automática de uma ontologia lexical: além de serem uma enorme fonte de conhecimento lexical (Briscoe, 1991) e serem vistos como autoridades no que diz respeito ao sentido das palavras (Kilgarriff, 1997), a sua estrutura e a previsibilidade e simplicidade do vocabulário utilizado nas definições facilitam a sua utilização para a extracção e organização de informação léxico-semântica. Com base no trabalho de Amsler (1981), Chodorow, Byrd e Heidorn (1985) criaram procedimentos semi-automáticos para a extracção da relação de hiperonímia a partir de um dicionário. Alshawi (1989) desenvolveu uma gramática que tinha como único objectivo a derivação das definições de um dicionário específico, de forma a facilitar a extracção de relações que eram depois organizadas em estruturas semânticas. Montemagni e Vanderwende (1992), por outro lado, defenderam a utilização de um analisador sintáctico de grande cobertura, com o argumento de que este seria melhor para extrair informação mais específica dentro de uma definição.

Apesar de vários trabalhos com este objectivo, a MindNet (Richardson, Dolan e Vanderwende, 1998) terá sido a primeira base de dados lexical independente, criada de forma automática a partir de dicionários, mas não houve muitos continuadores nesta senda, talvez devido à análise sobre a inconsistência dos dicionários feita por Ide e Veronis (1995). Ainda assim, alguns trabalhos recentes nesta área são O'Hara (2005), Nichols, Bond e Flickinger (2005) e Zesch, Müller e Gurevych (2008), este último usando o Wikcionário[3].

Por outro lado, vários investigadores apontaram o facto de que algum conhecimento importante para o PLN não se encontrava presente em dicionários: algumas aplicações necessitam de conhecimento específico sobre determinados domínios, que é mais fácil de obter em corpos (Riloff e Shepherd, 1997; Caraballo, 1999).

Para a extracção de conhecimento que não se consegue encontrar nem em dicionário, nem em outros recursos de vasta cobertura, como qualquer WordNet já existente, iniciou-se o processamento de recursos não estruturados.

No que diz respeito à utilização de recursos estruturados (ou semi-estruturados) para extrair conhecimento léxico-semântico, nos últimos anos tem também sido dada especial atenção à utilização de recursos colaborativos, como a Wikipédia[4] ou o já referido Wikcionário, veja-se por exemplo (Medelyan et al., 2009) ou (Herbelot e Copestake, 2006).

A referência mais conhecida no que diz respeito à extracção de conhecimento léxico-semântico a partir de corpos é o trabalho de Hearst (1992), que propõe um método para identificar padrões textuais indicadores da relação de hiponímia e que aplica um conjunto de padrões para extrair automaticamente relações deste tipo. Vários trabalhos se inspiraram na abordagem de Hearst, não só para extrair relações de hiponímia (Caraballo, 1999; Freitas e Quental, 2007), mas também para extrair outros tipos de relações, como por exemplo causais (Girju e Moldovan, 2002), ou de meronímia (ou parte de) (Berland e Charniak, 1999), e mais especificamente para relações geográficas em português (Chaves, 2009).

2.3 Abordagens para a avaliação de ontologias

Brank, Grobelnik e Mladenic' (2005) apresentam quatro formas que têm sido utilizadas para avaliar ontologias de domínio: (i) avaliação manual; (ii) comparação com um recurso dourado; (iii) realização de uma tarefa independente, definida para avaliar uma ontologia; (iv) comparação com um conjunto de dados sobre o mesmo domínio.

Apesar de, regra geral, estas formas de avaliação se adaptarem a qualquer tipo de ontologia, é preciso notar que temos de distinguir entre as ontologias propriamente ditas (Gruber, 1993), que cobrem uma área específica e são baseadas numa conceptualização de um domínio, e as ontologias lexicais que, como já referimos, tentam descrever o sistema conceptual de uma língua inteira. Isto leva naturalmente a que nem todos os métodos possam ser adaptados cegamente a ontologias lexicais.

A avaliação manual é uma forma habitualmente escolhida para avaliar a qualidade de um recurso. Muitos trabalhos efectuam este tipo de avaliação – por exemplo (Riloff e Shepherd, 1997; Caraballo, 1999), ou mesmo (Richardson, Vanderwende e Dolan, 1993), no âmbito do que viria a ser a MindNet – por ser provavelmente a forma mais fiável. No entanto, está sempre dependente de trabalho por parte dos indivíduos

[3] http://wiktionary.org/

[4] http://wikipedia.org

que realizam a avaliação. De forma a minimizar o esforço necessário para avaliar manualmente uma ontologia obtida automaticamente, Navigli et al. (2004) geraram definições em linguagem natural a partir do conteúdo dessa ontologia.

Para utilizar um recurso dourado, que pode eventualmente ser outra ontologia, é necessário que exista um elevado nível de confiança na sua correcção, possivelmente por ter sido criado manualmente por peritos. A qualidade de uma ontologia pode ser assim medida através da sua comparação com um recurso dourado, de acordo com determinados critérios. Neste tipo de avaliação, Santos (2007) refere que as medidas de precisão e abrangência, tradicionalmente utilizadas em recolha de informação (Salton e McGill, 1983), têm sido extremamente populares em PLN, sendo muitas vezes propostas sem uma total compreensão das suas limitações e adequação.

Outro problema desta abordagem de avaliação é que, sendo a criação de ontologias um assunto bastante recente, nem sempre existe um recurso dourado que se adeque aos critérios da avaliação. Para o inglês, e no âmbito das ontologias lexicais, muitos autores utilizam a própria WordNet.Pr como recurso dourado na avaliação da sua ontologia (Hearst, 1992; Nichols, Bond e Flickinger, 2005).

Partindo do princípio de que uma ontologia serve para ser integrada noutras aplicações, com o objectivo de realizar determinadas tarefas, alguns autores propõem avaliar uma ontologia de forma indirecta. Desta forma a ontologia é utilizada numa aplicação para realizar uma tarefa específica, cujos resultados serão alvo de avaliação. No entanto, é necessário ter algum cuidado com as ilações tiradas deste tipo de avaliação, já que há muitas variáveis envolvidas e a qualidade dos resultados não está apenas dependente da qualidade da ontologia, mas também do resto da aplicação. (Cuadros e Rigau, 2006) realizaram uma avaliação indirecta de várias ontologias lexicais, incluindo a WordNet.Pr, no âmbito da desambiguação do sentido das palavras. Curiosamente, os recursos criados de forma automática obtiveram melhores resultados ao nível tanto da precisão como de abrangência. Outra conclusão a que chegaram foi a de que a qualidade dos resultados obtidos, ao combinar o conhecimento de todos os recursos utilizados no estudo, é muito próxima daquela que apenas selecciona o sentido mais frequente para cada palavra.

Quanto à última forma de avaliação, a comparação com outros dados referentes ao mesmo domínio, Brewster et al. (2004) propõem que a adequação de uma ontologia de domínio a um dado corpo seja avaliada através do número dos termos salientes do corpo, que será sobre o domínio em questão, que também constam na ontologia. Contudo repare-se que para obter os termos salientes num dado domínio é preciso precisamente compará-lo com a linguagem geral e outros domínios, e obter os termos salientes na linguagem geral é algo que não faz muito sentido. Ainda assim, será possível medir a cobertura de um determinado corpo por um léxico, tal como Demetriou e Atwell (2001) propõem. A cobertura será medida através do número de palavras do corpo que se encontrarem no léxico.

A verdade, contudo, é que, tal como Raman e Bhattacharyya (2008) referem, a avaliação explícita de ontologias lexicais não é uma prática comum. A principal razão para esta situação será o facto de haver bastante confiança nestes recursos, que são na sua maioria criados manualmente por peritos, o que minimiza a possibilidade de existirem erros. De forma a verificar se a confiança é justificada, Raman e Bhattacharyya (2008) levaram a cabo uma validação automática dos grupos de sinónimos (*synsets*) da WordNet.Pr, utilizando um dicionário. Nesse trabalho consideraram que uma palavra estava correctamente incluída num nó da WordNet, se na sua definição forem referidas palavras dos nós hiperónimos desse nó, ou outras palavras pertencentes ao mesmo nó. Como esperado, não foram encontrados muitos erros.

Há ainda a referir um outro método de avaliação que tira partido da quantidade de texto que se consegue encontrar hoje em dia na Web, como fizeram, por exemplo, Etzioni et al. (2005) para calcular o nível de confiança de relações de hiperonímia entre classes e entidades mencionadas. Para o efeito, as relações foram transformadas em padrões textuais discriminadores (semelhantes aos de Hearst (1992)), procuraram por eles na Web e calcularam o PMI-IR (Turney, 2001) entre os padrões envolvendo a entidade e a própria entidade, ou seja, calcularam o quociente entre as ocorrências de cada um.

3 Breve apresentação do PAPEL

Nesta secção descrevemos primeiro o procedimento semi-automático utilizado para construir o PAPEL e de seguida apresentamos os conteúdos da sua versão actual, incluindo a contabilização de itens lexicais, a contabilização de relações, e ainda exemplos destas últimas.

```
PARTE{
  nome:nome * PARTE_DE:INCLUI;
  nome:adj * PARTE_DE_ALGO_COM_PROPRIEDADE:PROPRIEDADE_DE_ALGO_QUE_INCLUI;
  adj:nome * PROPRIEDADE_DE_ALGO_PARTE_DE:INCLUI_ALGO_COM_PROPRIEDADE;
}
```

Figura 1: Exemplo da descrição do grupo de relações relativas à meronímia.

3.1 Construção

De forma resumida, o processo de construção do PAPEL segue um ciclo de três passos até considerarmos ter chegado a um nível de desempenho suficiente, entrando depois no quarto e último passo.

1. **Criação de gramáticas semânticas**: foram criadas gramáticas para cada tipo de relação que se pretende extrair, por categoria gramatical (fornecida pelo dicionário). Na tabela 1 mostramos alguns dos padrões e as relações que pretendem descobrir e na figura 1 mostramos de que forma as relações que pretendemos extrair são descritas, de acordo com o grupo e especificando ainda a categoria dos argumentos e a sua relação inversa.

Padrão	Relação associada
`tipo|género|classe|forma de`	Hiperonímia
`parte|membro de`	Meronímia
`que causa|provoca|origina`	Causa
`usado|utilizado para`	Finalidade
`natural|originário de`	Local
uma palavra ou lista de palavras	Sinonímia

Tabela 1: Exemplos de padrões usados nas gramáticas.

2. **O processo de extracção**: usando um analisador automático, é feita a análise superficial das definições, a partir da qual são automaticamente extraídas relações (descritas no passo anterior) entre palavras na definição e a palavra definida, também chamada "verbete" (ver figura 2).

3. **Inspecção dos resultados**: usando um sistema de regressão para identificar mais facilmente as diferenças entre resultados anteriores, procede-se à inspecção manual dos resultados obtidos, com o eventual regresso ao primeiro passo para corrigir problemas detectados ou melhorar as gramáticas.

4. **Ajuste das relações**: aqui procura-se corrigir (ou eliminar) de forma automática relações com argumentos inválidos.

O último passo é realizado em dois tempos. Inicialmente, todas as relações são transformadas

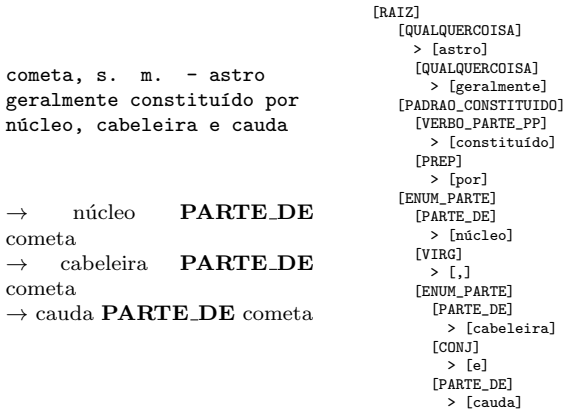


Figura 2: O resultado da análise da definição de *cometa*.

no tipo directo[5]. Por exemplo, *manga* INCLUI *punho* é convertida para *punho* PARTE_DE *manga*, e *dor* RESULTADO_DE *distensão* é transformada em *distensão* CAUSADOR_DE *dor*.

Visto que as gramáticas não fazem uma análise sintáctica das definições, não atribuindo por exemplo a classe gramatical, e que as definições do dicionário apenas incluem a classificação da vedeta, em alguns casos o processo de construção automática do PAPEL resulta em relações entre palavras de categorias erradas. É por isso preciso verificar, também de uma forma automática, esses casos, usando primeiro a própria lista de palavras/vedetas do dicionário e em seguida o analisador morfológico Jspell (Simões e Almeida, 2002). Se conseguirmos apurar que há um desajuste nas categorias mas que pode ser corrigido através da escolha de outra relação pertencente ao mesmo grupo, substituímos, senão removemos esse triplo. Por exemplo, a relação *loucura* ACCAO_QUE_CAUSA *desvario* – que pressupõe um verbo como primeiro argumento – é transformada automaticamente em *loucura* CAUSADOR_DE *desvario*, visto que ambos os argumentos são

[5] A escolha de um tipo directo e outro inverso foi arbitrariamente efectuada pelos criadores das gramáticas por um critério de naturalidade, e não de frequência, no dicionário ou em texto, e não tem qualquer consequência excepto a de facilitar a arrumação e depuração do recurso.

Categoria	Simples	Multipalavra	Total
Substantivo	51546	3330	55372
Verbo	10245	13791	24089
Adjectivo	18904	3	18933
Advérbio	1389	0	1389

Tabela 3: Distribuição dos items por categoria gramatical, no PAPEL 1.1

substantivos. Durante este processo, os casos das palavras flexionadas são também substituídos pelos seus lemas, quando essa informação é dada pelo Jspell.

3.2 Conteúdos

A versão actual do PAPEL (1.1) contém perto de 100.000 items lexicais, cujas categorias gramaticais se distribuem de acordo com a tabela 3, e perto de 200.000 relações, distribuídas de acordo com a tabela 2. A sinonímia e a hiperonímia são as relações mais frequentes, e ainda podem ser aumentadas, como discutiremos abaixo, de uma forma semelhante ao feito no ReRelEM (Freitas et al., 2009).

Como também podemos ver na tabela 3, a maior parte dos itens lexicais são expressões de uma única palavra. No entanto, o PAPEL também inclui expressões multipalavra, em casos como os seguintes:

- Substantivos seguidos das preposições *de/do/dos/da/das* e de uma outra palavra (e.g. *sistema de rodas*, *dispositivo de mira*);
- Verbos com o seu objecto directo (e.g. *abrir o apetite*, *produzir som*);

4 Avaliação do PAPEL

Aqui descrevemos uma primeira avaliação ao PAPEL, feita sobre a sua versão anterior (1.0) de duas formas diferentes: as relações de sinonímia foram comparadas com as relações representadas num thesaurus para o português, enquanto que as restantes relações, apenas entre substantivos, foram validadas através das sua transformação em padrões textuais e procura em texto por esses padrões.

4.1 Avaliação da sinonímia

Dado que o Thesaurus Eletrônico para o Português do Brasil (Maziero et al., 2008) (Tep) pode ser levantado na rede, usámo-lo como recurso de referência para validar as relações de sinonímia, embora estejamos conscientes das várias diferenças entre as variantes. O TeP 2.0 contém 19.888 nós, ou seja grupos de unidades lexicais com o mesmo sentido, correspondendo a 44.678 unidades lexicais ao todo.

Para que a avaliação pudesse prosseguir sem enviesamento, começámos por retirar da comparação as entradas do Tep que não estivessem presentes no PAPEL assim como todas os casos de relações do PAPEL que contivessem argumentos ausentes do Tep. Ficámos assim apenas com 68% do nosso material, e com apenas 35% das possíveis 405.026 relações do Tep[6]. A comparação de ambos os conjuntos de relações produziu os seguintes resultados: 50% das nossas relações estavam presentes no Tep, e 39% das relações do Tep estavam presentes no PAPEL.

Embora estes valores possam ser surpreendentes, convém relembrar que as nossas relações tinham de ser encontradas directamente no dicionário, e não foram portanto ainda alvo de qualquer raciocínio. Em particular, a relação de transitividade parece ser óbvia: A SINONIMO_DE B $\wedge$ B SINONIMO_DE C $\rightarrow$ A SINONIMO_DE C. Após aplicação desta relação (uma vez só), obtivemos, dos 80.432 sinónimos iniciais, 689.073 sinónimos derivados. Claro está que, como as definições (e as nossas regras) não separam entre sentidos distintos de uma mesma palavra, esta expansão poderá levar a muitas relações infelizes, tal como *queda* SINONIMO_DE *ruína* $\wedge$ *queda* SINONIMO_DE *habilidade* $\rightarrow$ *ruína* SINONIMO_DE *habilidade*. Após esta expansão, e como esperado, o número de casos atestado no Tep caiu para 14%, contudo, 90% das relações no Tep puderam ser encontradas no PAPEL. Fica assim demonstrado que a combinação dos dois recursos permite não só melhorar ambos como separar o trigo do joio e mesmo alertar automaticamente para palavras com vários sentidos.

4.2 Avaliação das demais relações

Em relação às outras relações, e na impossibilidade de comparar automaticamente com outros recursos para o português, tivemos de desenvolver uma metodologia diferente, inspirada nos vários trabalhos de extracção automática de relações semânticas em texto, ou de validação das mesmas em texto.

Para os nossos testes usámos o CETEMPúblico (Rocha e Santos, 2000), através da interface do projecto AC/DC[7], que nos permitiu além disso acesso às frequências dos lemas respectivos. O trabalho

[6]Para conversão do Tep todos os elementos de um grupo de sinónimos foram considerados como pertencendo a uma relação de sinonímia com todos os outros elementos do mesmo grupo.

[7]Sobre este projecto da Linguateca, consultar (Santos e Bick, 2000; Santos e Sarmento, 2003; Costa, Santos e Rocha, 2009; Santos, 2009) para mais informações.

Grupo	Nome	Args.	Qnt.	Exemplos
Sinonímia	SINONIMO_N_DE	n,n	37.259	(*auxílio*, *contributo*)
	SINONIMO_V_DE	v,v	21.534	(*tributar*, *colectar*)
	SINONIMO_ADJ_DE	adj,adj	19.073	(*flexível*, *moldável*)
	SINONIMO_ADV_DE	adv,adv	1.169	(*após*, *seguidamente*)
Hiperonímia	HIPERONIMO_DE	n,n	61.477	(*planta*, *salva*)
Parte	PARTE_DE	n,n	9.970	(*cauda*, *cometa*)
	PARTE_DE_ALGO_COM_PROP	n,adj	3.806	(*tampa*, *coberto*)
	PROP_DE_ALGO_PARTE_DE	adj,n	900	(*celular*, *célula*)
Causa	CAUSADOR_DE	n,n	1.010	(*fricção*, *assadura*)
	CAUSADOR_DE_ALGO_COM_PROP	n,adj	17	(*paixão*, *passional*)
	PROP_DE_ALGO_CAUSADOR_DE	adj,n	498	(*reactivo*, *reacção*)
	ACCAO_QUE_CAUSA	v,n	6.399	(*limpar*, *purgação*)
	CAUSADOR_DA_ACCAO	n,v	39	(*gases*, *fumigar*)
Produtor	PRODUTOR_DE	n,n	885	(*romãzeira*, *romã*)
	PRODUTOR_DE_ALGO_COM_PROP	n,adj	34	(*sublimação*, *sublimado*)
	PROP_DE_ALGO_PRODUTOR_DE	adj,n	359	(*fotógeno*, *luz*)
Finalidade	FINALIDADE_DE	n,n	2.878	(*defesa*, *armadura*)
	FINALIDADE_DE_ALGO_COM_PROP	n,adj	38	(*reprodução*, *reprodutor*)
	ACCAO_FINALIDADE_DE	v,n	5.185	(*fazer_rir*, *comédia*)
	ACC_FINALIDADE_DE_ALGO_COM_PROP	v,adj	284	(*corrigir*, *correccional*)
Localização	LOCAL_ORIGEM_DE	n,n	816	(*Japão*, *japonês*)
Maneira	MANEIRA_POR_MEIO_DE	adv,n	1.113	(*timidamente*, *timidez*)
	MANEIRA_SEM	adv,n	121	(*devagar*, *pressa*)
	MANEIRA_SEM_ACCAO	adv,v	11	(*assiduamente*, *faltar*)
Propriedade	PROP_DE_ALGO_REFERENTE_A	adj,n	3.520	(*dinâmico*, *movimento*)
	PROP_DO_QUE	adj,v	17.246	(*familiar*, *ser_conhecido*)

Tabela 2: As relações do PAPEL 1.1

realizado tem de ser considerado preliminar, já que, devido a limitações de ocorrência de muitas das unidades lexicais nos corpos que usámos, não tivemos possibilidade de as testar. Com efeito, não só muitas das palavras no PAPEL eram demasiado raras ou especializadas, como cedo nos demos conta que em texto jornalístico seria quase impossível encontrar num mesmo contexto (numa mesma frase) pares ou relações como *liquidar* ACCAO_QUE_CAUSA *liquidação*, *fósforo* PARTE_DE_ALGO_COM_PROPRIEDADE *fosforoso*, visto que são característicos de texto dicionarístico ou enciclopédico.

Restringimos assim o processo de validação, em primeiro lugar, apenas a relações entre substantivos, e, além disso, retirámos do teste as relações que envolvessem palavras cujos lemas estivessem ausentes do CETEMPúblico. Mesmo assim, e por questões de sobrecarga do serviço, para as duas relações mais populosas do PAPEL, hiperonímia e meronímia, ainda escolhemos uma amostra aleatória de relações a testar, correspondente respectivamente a 8% e 63% dos casos. Os resultados encontram-se na tabela 4.

Cerca de 20% destas relações parecem ser validadas ou confirmadas pelo corpo, enquanto que a percentagem é menor para as outras relações. Estes resultados parecem-nos satisfatórios, tendo em conta que: o corpo é bastante pequeno; os padrões usados foram muito simples (em texto real há uma miríade de outras possibilidades de indicar uma relação); e os nossos valores não se encontram demasiado longe daqueles apresentados na literatura de confirmação.

De qualquer maneira, e para mostrarmos que esta confirmação está longe de ser definitiva ou mesmo conclusiva, na tabela 5 apresentamos alguns exemplos, quer de confirmação certa quer de espúria (ou seja parecem confirmar mas não o fazem). Casos que não foram confirmados embora existam ambas as palavras no CETEMPúblico são por exemplo: *fruto* HIPERONIMO_DE *alperce*, *algoritmia* PARTE_DE *matemática*, *ausência* CAUSADOR_DE *saudade*, *tamareira* PRODUTOR_DE *tâmara*, *aquecimento* FINALIDADE_DE *salamandra*.

5 Ferramentas

Esta secção apresenta duas ferramentas associadas ao PAPEL, para a sua exploração e validação.

5.1 Folheador

O Folheador é uma interface na rede desenvolvida para navegar num conjunto de relações, como as do PAPEL, depois de carregadas numa base de dados. Este sistema encontra-se actualmente instalado no URL `http://sancho.dei.uc.pt/folheador/` e permite fazer procuras na versão 1.1 do PAPEL.

O seu funcionamento é muito simples: basta procurar por uma palavra e o sistema responde com uma lista de todas as relações onde essa palavra entra. Se a palavra tiver mais de uma

Relação	Relações c/ args no CETEMPúblico	%	Amostra	%	Encontradas	%
Hiperonímia	40,079	63%	3,145	8%	560	18%
Meronímia	3,746	35%	2,343	63%	521	22%
Causa	557	50%	557	100%	20	4%
Produtor	414	44%	414	100%	12	3%
Finalidade	1,718	59%	1,718	100%	173	10%

Tabela 4: Resultados da validação das relações excepto sinonímia.

Relação	Certa?	Justificação
língua HIPERONIMO_DE *italiano*	Sim	*As **línguas latinas, como o italiano** ou o português, tornam-se mais fáceis por causa das vogais.*
arbusto PARTE_DE *floresta*	Sim	*A **floresta é um conjunto de árvores, arbustos** e ervas de várias qualidades e tamanhos.*
cólera CAUSADOR_DE *diarreia*	Sim	*A **cólera provoca fortes diarreias** e vómitos e pode levar à desidratação e, consequentemente, à morte em poucas horas.*
oliveira PRODUTOR_DE *azeitona*	Sim	*Também a quantidade e tamanho das **azeitonas produzidas por uma oliveira** biológica é inferior, já que não são utilizados compostos de azoto que ajudam a planta a crescer.*
recrutamento FINALIDADE_DE *inspecção*	Sim	*Menos de metade dos jovens entre os 20 e os 22 anos apresentaram-se às **inspecções para recrutamento**, revelou o ministro da Defesa.*
músico PARTE_DE *música*	Não	*... um espectáculo baseado na obra "Cantos de Maldoror", de Lautréamont, com **música composta pelo músico** inglês Steven Severin...*
fim FINALIDADE_DE *sempre*	Não	*Sicilia aponta **sempre para o fim** do dia, para o fim da luz.*

Tabela 5: Exemplos de validação através do CETEMPúblico.

categoria gramatical possível, as relações são separadas de acordo com a categoria gramatical. Além disso, é possível filtrar o resultado por tipo de relação e ainda, ao clicar nas palavras em argumentos das relações apresentadas, verificar todas relações que envolvam estas palavras. Na Figura 3 é apresentada uma imagem do Folheador, depois de procurar pela palavra *vencedor*.

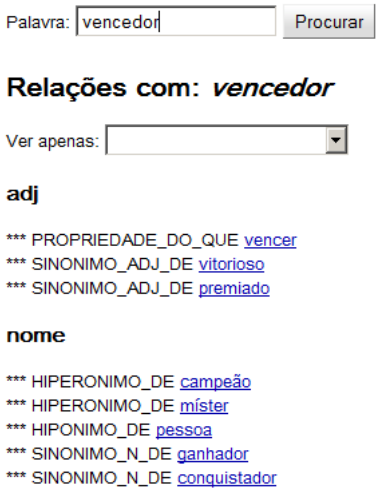


Figura 3: Resultados para a procura pela palavra *vencedor*, no Folheador.

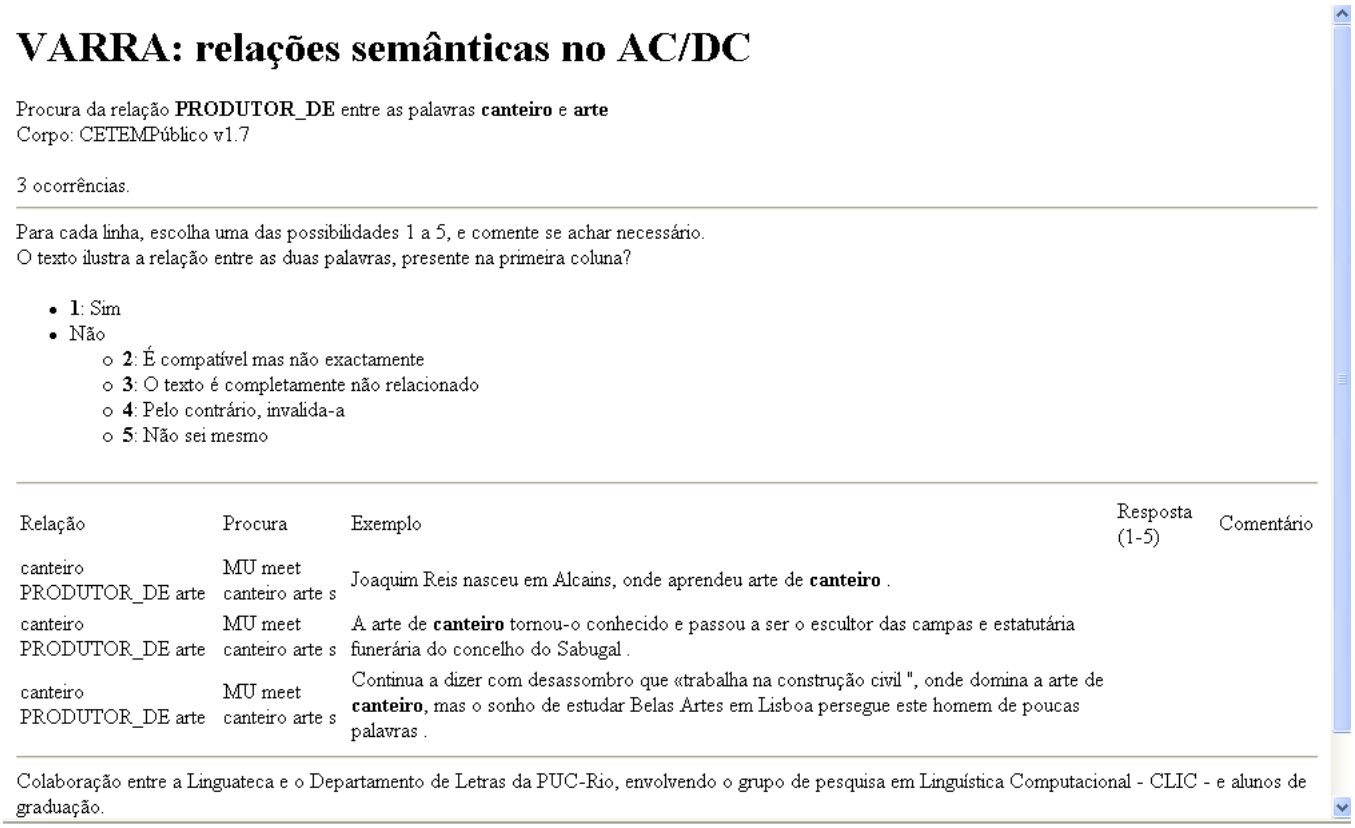
5.2 VARRA

Para permitir uma validação mais pormenorizada das relações presentes no PAPEL – e possivelmente noutros recursos – desenvolvemos o VARRA (Validação, Avaliação e Revisão de Relações no AC/DC), em conjunto com o projecto AC/DC, de forma a obter julgamentos mais completos em relação à seguinte questão: dado um triplo e uma possível frase que o ilustra e consequenetemente valida, obtida automaticamente dos corpos do AC/DC, pedimos às pessoas que escolham uma das seis possíveis alternativas:

1. Relação claramente incorrecta. Passe à frente
2. Relação possivelmente correcta. O texto ilustra a relação entre as duas palavras?
   (a) Sim
   (b) Não... É compatível mas não exactamente.
   (c) Não... O texto é completamente não relacionado.
   (d) Não... Pelo contrário, invalida-a.
   (e) Não sei

Esse serviço[8], ilustrado na figura 4 encontra-se presentemente em fase de teste, e pretendemos alargá-lo de forma a que sirva também para avaliar outro tipo de recursos e de padrões de procura, além de poder ser usado pedagogicamente na formação de alunos na área de linguística com corpos.

[8] acessível de `http://lusiadas.linguateca.pt/acesso/avalia_papel.php`.



VARRA: relações semânticas no AC/DC

Procura da relação **PRODUTOR_DE** entre as palavras **canteiro** e **arte**
Corpo: CETEMPúblico v1.7

3 ocorrências.

Para cada linha, escolha uma das possibilidades 1 a 5, e comente se achar necessário.
O texto ilustra a relação entre as duas palavras, presente na primeira coluna?

- **1**: Sim
- Não
  - **2**: É compatível mas não exactamente
  - **3**: O texto é completamente não relacionado
  - **4**: Pelo contrário, invalida-a
  - **5**: Não sei mesmo

Relação	Procura	Exemplo	Resposta (1-5)	Comentário
canteiro PRODUTOR_DE arte	MU meet canteiro arte s	Joaquim Reis nasceu em Alcains, onde aprendeu arte de **canteiro** .		
canteiro PRODUTOR_DE arte	MU meet canteiro arte s	A arte de **canteiro** tornou-o conhecido e passou a ser o escultor das campas e estatutária funerária do concelho do Sabugal .		
canteiro PRODUTOR_DE arte	MU meet canteiro arte s	Continua a dizer com desassombro que «trabalha na construção civil ", onde domina a arte de **canteiro**, mas o sonho de estudar Belas Artes em Lisboa persegue este homem de poucas palavras .		

Colaboração entre a Linguateca e o Departamento de Letras da PUC-Rio, envolvendo o grupo de pesquisa em Linguística Computacional - CLIC - e alunos de graduação.

Figura 4: Exemplo de resultados da invocação do VARRA

6 Considerações finais

Apresentámos neste artigo um novo recurso lexical para o português, o PAPEL, que pode ser levantado integralmente no endereço acima citado, junto com ampla documentação sobre o mesmo. Também apresentámos algumas ferramentas relacionadas com este recurso.

Salientamos novamente que o PAPEL não pretende ser um recurso final, mas sim um ponto de partida para futuros projectos, que o poderão enriquecer recorrendo a outras fontes de informação. Por exemplo, e visto que o TeP foi criado à mão, um processo relativamente fácil de melhorar o PAPEL seria apenas juntar-lhe (ao PAPEL) as relações de sinonímia obtidas por transitividade (as 12%) que eram validadas no TeP, ou pelo menos a informação adicional de "concordância" com outros recursos.

Pretendemos no futuro continuar a melhorar o PAPEL através da obtenção de novos dados na rede assim como aperfeiçoar a validação dos já presentes. Uma melhoria óbvia que em breve implementaremos é a associação de um grau de certeza/validação a cada triplo.

A primeira avaliação do PAPEL, relatada aqui, apesar de bastante preliminar, pode ser interessante como exemplo de avaliação, também para outros recursos. Esperamos que em breve possamos também referir trabalho de outros investigadores a usar e a melhorar este recurso, que é para ser propriedade comum de todos os investigadores e desenvolvedores na área do processamento da língua portuguesa.

Agradecimentos

Agradecemos à Cláudia Freitas a colaboração preciosa no desenho do sistema VARRA, e a todos os co-autores do artigo de comparação de ontologias: Anabela Barreiro, Cláudia Freitas, José Carlos Medeiros, Luís Costa e Rosário Silva, as discussões férteis e o trabalho realizado. Agradecemos também ao grupo de R&D da Porto Editora a colaboração na criação do PAPEL, ao CLIC e à Violeta Quental a colaboração com a PUC-Rio, assim como ao Nuno Seco a sua anterior participação no projecto.

O projecto PAPEL foi desenvolvido no âmbito da Linguateca, co-financiada pelo Governo Português, pela União Europeia (FEDER e FSE), sob o contrato POSC/339/1.3/C/NAC, pela UMIC e pela FCCN. Hugo Gonçalo Oliveira é actualmente financiado pela FCT, bolsa SFRH/BD/44955/2008.

Referências

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

Alshawi, H. 1989. Analysing the dictionary definitions. *Computational lexicography for natural language processing*, pp. 153–169.

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

Baker, Collin F., Charles J. Fillmore, e John B. Lowe. 1998. The berkeley framenet project. Em *Proc. 17th Intl. Conference on Computational linguistics*, pp. 86–90, Morristown, NJ, USA. Association for Computational Linguistics.

Barreiro, Anabela. 2008. Port4NooJ: an open source, ontology-driven Portuguese linguistic system with applications in machine translation. Em *(Silberztein e Varadi, forthcoming 2009)*.

Barreiro, Anabela, Luzia Helena Wittmann, e Maria de Jesus Pereira. 1996. Lexical differences between European and Brazilian Portugueses. *INESC Journal of Research and Development*, 5(2).

Berland, Matthew e Eugene Charniak. 1999. Finding parts in very large corpora. Em *Proc. 37th Annual Meeting of the ACL on Computational Linguistics*, pp. 57–64, Morristown, NJ, USA. Association for Computational Linguistics.

Brank, Janez, Marko Grobelnik, e Dunja Mladenic'. 2005. A survey of ontology evaluation techniques. Em *Proc. Conference on Data Mining and Data Warehouses (SiKDD)*.

Brewster, Christopher, Harith Alani, Srinandan Dasmahapatra, e Yorick Wilks. 2004. Data-driven ontology evaluation. Em *Proc. the Language Resources and Evaluation Conference (LREC)*, pp. 164–168, Lisbon, Portugal. European Language Resources Association.

Briscoe, Ted. 1991. Lexical issues in natural language processing. Em E. Klein e F. Veltman, editores, *Natural Language and Speech: Symposium Proc.* Springer, Berlin, Heidelberg, pp. 39–68.

Calzolari, Nicoletta, Laura Pecchia, e Antonio Zampolli. 1973. Working on the italian machine dictionary: a semantic approach. Em *Proc. 5th conference on Computational linguistics*, pp. 49–52, Morristown, NJ, USA. Association for Computational Linguistics.

Caraballo, Sharon A. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. Em *Proceedings of the 37th annual meeting of the ACL on Computational Linguistics*, pp. 120–126, Morristown, NJ, USA. Association for Computational Linguistics.

Chaves, Marcirio Silveira. 2009. *Uma Metodologia para Construção de Geo-Ontologias*. Tese de doutoramento, Faculdade de Ciências, Universidade de Lisboa, Setembro, 2009. `http://www.linguateca.pt/documentos/TeseDoutMarcirioChaves2009.pdf`.

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

Costa, Luís, Diana Santos, e Paulo Alexandre Rocha. 2009. Estudando o português tal como é usado: o serviço AC/DC. Em *The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)*, 8-11 de Setembro, 2009.

Costa, Rui P. e Nuno Seco. 2008. Hyponymy extraction and web search behavior analysis based on query reformulation. Em *Proc. 11th Ibero-American Conference on Artificial Intelligence (IBERAMIA)*, LNAI, pp. 332–341. Springer Verlag.

Cuadros, Montse e German Rigau. 2006. Quality assessment of large scale knowledge resources. Em *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 534–541, Sydney, Australia, July, 2006. Association for Computational Linguistics.

Dahlgren, Kathleen. 1995. A linguistic ontology. *Int. J. Hum.-Comput. Stud.*, 43(5-6):809–818.

de Saussure, Ferdinand. 1916. *Cours de Linguistique Générale*. Payot, Paris.

Demetriou, George e Eric Steven Atwell. 2001. A domain-independent semantic tagger for the study of meaning associations in english text. Em *Proceedings of 4th International Workshop on Computational Semantics (IWCS)*.

Dias da Silva, Bento C., Mirna Oliveira, e Helio Moraes. 2002. Groundwork for the Development of the Brazilian Portuguese Wordnet. Em Nuno Mamede e Elisabete Ranchhod, editores, *Proc. Advances in Natural Language Processing: 3rd International Conference*, LNAI, pp. 189–196. Springer Verlag, 23-26 de Junho, 2002.

Dolan, William B. 1994. Word sense ambiguation: clustering related senses. Em *Proc. the 15th conference on Computational linguistics*, pp. 712–716, Morristown, NJ, USA. Association for Computational Linguistics.

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

Edmonds, Philip e Graeme Hirst. 2002. Near-synonymy and lexical choice. *Computational Linguistics*, 28(2):105–144.

Etzioni, Oren, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, e Alexander Yates. 2005. Unsupervised named-entity extraction from the web: an experimental study. *Artificial Intelligence*, 165(1):91–134.

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

Fillmore, Charles J. 1982. Frame semantics. Em Linguisticsocietykorea, editor, *Linguistics in the morning calm*. Seoul: Hanshin Publishing Co.

Freitas, Cláudia e Violeta Quental. 2007. Subsídios para a elaboração automática de taxonomias. Em *XXVII Congresso da SBC - V Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pp. 1585–1594.

Freitas, Cláudia, Diana Santos, Cristina Mota, Hugo Gonçalo Oliveira, e Paula Carvalho. 2009. Detection of relations between named entities: report of a shared task. Em *Proc. NAACL-HLT Workshop, Semantic Evaluations: Recent Achievements and Future Directions*, Junho, 2009.

Girju, Roxana e Dan Moldovan. 2002. Text mining for causal relations. Em Susan M. Haller e Gene Simmons, editores, *Proc. 15th Intl. Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pp. 360–364.

Gonçalo Oliveira, Hugo, Diana Santos, Paulo Gomes, e Nuno Seco. 2008. PAPEL: a dictionary-based lexical ontology for Portuguese. Em António Teixeira et. al, editor, *Proc. of Computational Processing of the Portuguese Language, 8th Intl. Conf. (PROPOR)*, volume 5190 of *LNAI*, pp. 31–40. Springer Verlag.

Gruber, Thomas R. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. Em *Proc. the 14th conference on Computational linguistics*, pp. 539–545, Morristown, NJ, USA. Association for Computational Linguistics.

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

Hirst, Graeme. 2004. Ontology and the lexicon. Em Steffen Staab e Rudi Studer, editores, *Handbook on Ontologies*, International Handbooks on Information Systems. Springer-Verlag, pp. 209–230.

Ide, Nancy e Jean Veronis. 1995. Knowledge extraction from machine-readable dictionaries: An evaluation. Em Petra Steffens, editor, *Machine Translation and the Lexicon, LNAI*. Springer Verlag.

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

Kilgarriff, Adam. 1997. Ï don't believe in word senses*:. Computing and the Humanities*, 31(2):91–113.

Lenat, Douglas B. 1995. Cyc: a large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Liu, H. e P. Singh. 2004. Conceptnet: A practical commonsense reasoning toolkit. *BT Technology Journal*, 22(4):211–226.

Marcellino, Erasmo Roberto e Bento Dias-da-Silva. 2009. Sistematização linguístico-computacional do léxico do domínio conceitual indústria do bordado de ibitinga. Em *The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)*, 8-11 September, 2009.

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.

Medelyan, Olena, David Milne, Catherine Legg, e Ian H. Witten. 2009. Mining meaning from wikipedia. *Intl. Journal of Human-Computer Studies*, May, 2009.

Montemagni, Simonetta e Lucy Vanderwende. 1992. Structural patterns vs. string patterns for extracting semantic information from dictionaries. Em *Proc. the 14th conference on Computational linguistics*, pp. 546–552, Morristown, NJ, USA. Association for Computational Linguistics.

Navigli, Roberto, Paola Velardi, Alessandro Cucchiarelli, e Francesca Neri. 2004. Quantitative and qualitative evaluation of the ontolearn ontology learning system. Em *Proc. 20th Intl. conference on Computational Linguistics*, Morristown, NJ, USA. Association for Computational Linguistics.

Nichols, Eric, Francis Bond, e Dan Flickinger. 2005. Robust ontology acquisition from machine-readable dictionaries. Em Leslie Pack Kaelbling e Alessandro Saffiotti, editores, *IJCAI*, pp. 1111–1116. Professional Book Center.

O'Hara, Thomas Paul. 2005. *Empirical Acquisition of Conceptual Distinctions via Dictionary Definitions*. Tese de doutoramento, NMSU CS, August, 2005.

Pianta, E., L. Bentivogli, e C. Girardi. 2002. Multiwordnet: developing an aligned multilingual database. Em *1st Intl. Conference on Global WordNet*.

Raman, J. e Pushpak Bhattacharyya. 2008. Towards automatic evaluation of wordnet synsets. Em Attila Tanács, Dóra Csendes, Veronika Vincze, Christiane Fellbaum, e Piek Vossen, editores, *Proceedings of the 4th Global WordNet Conference (GWC 2008)*, Szeged, Hungary. University of Szeged, Department of Informatics.

Richardson, Stephen D., William B. Dolan, e Lucy Vanderwende. 1998. Mindnet: acquiring and structuring semantic information from text. Em *Proc. 17th Intl. Conference on Computational linguistics*, pp. 1098–1102, Morristown, NJ, USA. Association for Computational Linguistics.

Richardson, Stephen D., Lucy Vanderwende, e William Dolan. 1993. Combining dictionary-based and example-based methods for natural language analysis. Em *Proc. 5th Intl. Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 69–79, Kyoto, Japan.

Riloff, Ellen e Jessica Shepherd. 1997. A corpus-based approach for building semantic lexicons. Em *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pp. 117–124.

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)*, pp. 131–140, São Paulo. ICMC/USP.

Salomao, Maria M. M. 2009. Framenet Brasil: Um trabalho em progresso. *Calidoscópio*, 7(2).

Salton, G. e M. J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Sampson, Geoffrey. 2000. Review of (Fellbaum, 1998). *International Journal of Lexicography*, 13(1):54–59.

Santos, Diana. 2006. What is natural language? Differences compared to artificial languag es, and consequences for natural language processing, 15 May, 2006. Palestra convidada no SBLP2006 e no PROPOR'2006, `http://www.linguateca.pt/Diana/download/SantosPalestraSBLPPropor2006.pdf`.

Santos, Diana. 2007. Evaluation in natural language processing, 6-17 August, 2007. `http://www.linguateca.pt/Diana/download/EvaluationESSLLI07.pdf`.

Santos, Diana. 2009. Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties, 17-18 September, 2009. `http://www.hf.uio.no/tekstlab/rilivs/slides/SantosRILiVS2009workshop.pdf`.

Santos, Diana, Anabela Barreiro, Luís Costa, Cláudia Freitas, Paulo Gomes, Hugo

Gonçalo Oliveira, José Carlos Medeiros, e Rosário Silva. 2009. O papel das relações semânticas em português: Comparando o TeP, o MWN.PT e o PAPEL, 22-24 de Outubro, 2009.

Santos, Diana e Eckhard Bick. 2000. Providing Internet access to Portuguese corpora: the AC/DC project. Em Maria Gavrilidou, George Carayannis, Stella Markantonatou, Stelios Piperidis, e Gregory Stainhauer, editores, *Proc. of the 2nd International Conference on Language Resources and Evaluation (LREC)*, pp. 205–210.

Santos, Diana e Luís Sarmento. 2003. O projecto AC/DC: acesso a corpora/disponibilização de corpora. Em Amália Mendes e Tiago Freitas, editores, *Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística (APL 2002)*, pp. 705–717, Lisboa, 2-4 de Outubro de 2002, 2003. APL.

Scott, Bernard. 2003. The logos model: An historical perspective. *Machine Translation*, 18:1–72.

Silberztein, Max e Tamas Varadi, editores. forthcoming 2009. *Proceedings of the 2008 International NooJ Conference (NooJ'08)*. Cambridge Scholars Publishing, 8-10 June, forthcoming 2009.

Simões, Alberto M. e J.J. Almeida. 2002. Jspell.pm – um módulo de análise morfológica para uso em processamento de linguagem natural. Em *Actas do XVII Encontro da Associação Portuguesa de Linguística*, pp. 485–495, Lisboa.

Turney, Peter D. 2001. Mining the web for synonyms: PMI–IR versus LSA on TOEFL. Em Luc De Raedt e Peter Flach, editores, *Proceedings of 12th European Conference on Machine Learning (ECML-2001)*, volume 2167, pp. 491–502. Springer-Verlag.

Vanderwende, Lucy, Gary Kacmarcik, Hisami Suzuki, e Arul Menezes. 2005. Mindnet: An automatically-created lexical resource. Em *Proc. of HLT/EMNLP on Interactive Demonstrations*. The Association for Computational Linguistics.

Veale, Tony. 2007. Enriched lexical ontologies: Adding new knowledge and new scope to old linguistic resources, 6-17 August, 2007. `http://afflatus.ucd.ie/papers/Essilli_EnrichedLexiOnto.pdf`.

Vossen, Piek. 1997. Eurowordnet: a multilingual database for information retrieval. Em *Proc. the DELOS workshop on Cross-Language Information Retrieval*, Zurich.

Zesch, Torsten, Christof Müller, e Iryna Gurevych. 2008. Using wiktionary for computing semantic relatedness. Em *AAAI'08: Proceedings of the 23rd national conference on Artificial intelligence*, pp. 861–866. AAAI Press.