Perspectivas sobre a Linguateca

Actas do encontro Linguateca: 10 anos

Luís Costa, Diana Santos e Nuno Cardoso

Linguateca, 2008

1ª Edição, Outubro de 2008.

Publicação Digital.

ISBN 978-989-20-1445-6

Prefácio

Este volume resultou da iniciativa da Linguateca de chamar a comunidade da área a quem servimos, ou seja, as pessoas envolvidas no processamento computacional da nossa língua, em linguística computacional do português, no ensino da língua ou na tradução usando meios computacionais, entre outros, a pronunciar-se sobre a nossa actividade e sobre os nossos falhanços e sucessos, por ocasião de um marco significativo na nossa existência: o terminar da Linguateca como fora concebida até aí. E perguntar, humildemente, se havia alguma maneira de continuar e/ou melhorar o que tínhamos feito, assim como pedir um balanço, pelo exterior, dos nossos contributos.

Por outro lado, comprometemo-nos a fazer, também nós, um balanço da actividade da Linguateca, quer sob uma perspectiva global da visão de cada pólo, representado pelo sénior responsável, quer sob uma perspectiva de apresentar pela primeira vez, de uma forma estruturada, todos os trabalhos de investigação (a nível de doutoramento) em curso ou terminados sob a chancela (mesmo que parcial ou não oficial) da Linguateca.

Como este volume evidencia, tivemos alguma resposta da comunidade, a qual agradecemos e muito apreciamos. Notamos também muitas ausências, embora algumas dessas possam ser explicadas numa certa medida pelo facto de a Linguateca ter três dias antes organizado o Encontro do Segundo HAREM, e termos aliás recebido um retorno muito positivo desse sector da nossa comunidade. (Veja-se o sítio desse encontro[1], e o livro em preparação a esta data (Mota e Santos, 2008)).

O encontro foi muito rico precisamente por ter congregado tanto intervenções de carácter predominantemente científico como apresentações de cariz de política científica ou de balanço (todas elas acessíveis do sítio do encontro[2]), e por ter permitido a discussão e o

[1] www.linguateca.pt/EncontroSegundoHAREM
[2] www.linguateca.pt/Linguateca10anos

debate presencial. Neste volume, contudo, optámos por apenas publicar as contribuições de carácter científico ou as que denominámos no programa do encontro por "testemunhos de convivência", ou seja, as que nos foram enviadas por colegas do Brasil ou de Portugal que quiseram homenagear e fazer a história conjunta deste período da Linguateca (os seus dez anos) e da sua instituição, grupo ou centro.

Não nos pareceu terem cabimento aqui quer o debate, por ser contingente e ainda não estar terminado, quer as apresentações dos séniores de cada pólo, visto que estas foram pedidas e desenhadas para ser mais uma autocrítica e/ou publicitação do que foi feito, com o duplo fim de estimular o debate e de informar os presentes.

Este volume foi assim organizado alfabeticamente a partir dos resumos alargados que nos foram enviados. Os editores limitaram-se a fazer recomendações de terminologia e a garantir a coerência das citações. Gostávamos de agradecer mais uma vez a todos os participantes e autores o terem respondido à chamada e nos terem facultado textos tão interessantes e enriquecedores como os que podemos apresentar aqui.

A Linguateca e, consequentemente, este encontro e este volume, foram financiados pelo governo português e pela União Europeia (FEDER e FSE), através do projecto POSC/339/1.3/C/NAC.

Os editores,

Luís Costa, Diana Santos e Nuno Cardoso

Oslo, Novembro de 2008

Capítulo 1

Utilização da programação declarativa para processamento do CETEMPúblico

Agostinho Monteiro, Júlio Barbas e Nuno C. Marques

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 1, p. 1–5, 2008

Várias entidades têm desenvolvido um substancial esforço na elaboração e processamento de corpora de texto, com utilização de diferentes linguagens, ambientes de desenvolvimento e formatos de representação. Para tal, cada entidade envolvida no processo utilizou os recursos que entendeu serem mais adequados ao seu trabalho com a consequente redundância de repositórios e eventual ambiguidade nas classificações.

Bom seria que existisse um repositório principal, de acesso livre, criteriosamente actualizado por diferentes ferramentas e validado por entidades para tal autorizadas. O trabalho desenvolvido pela Linguateca é um bom exemplo neste sentido. Nesta comunicação propõe-se um novo formato que possibilita a utilização directa da programação declarativa. Esta proposta deve ser integrada nas soluções e formatos existentes, com o objectivo de potenciar a interoperabilidade, que pode existir entre diferentes aplicações, já existentes ou a desenvolver entendidas como serviços, que utilizem e manipulem os textos. Para tal, pode-se seguir uma filosofia SOA (*Service Oriented Architecture*), cujo principal objectivo é precisamente a reutilização de aplicações já existentes e a sua integração com outras a desenvolver. Neste quadro, as aplicações já existentes devem ser entendidas como serviços, mediante disponibilização de um canal comum que liga, transparentemente, o requerente de um serviço ao seu fornecedor. A arquitectura SOA define igualmente conjuntos de métricas que devem ser aplicados aos diversos serviços, para que seja possível a análise da sua utilidade e desempenho. A Linguateca pode constituir o referido canal SOA, disponibilizando, como já o faz, o CETEMPúblico (Rocha e Santos, 2000) no formato nativo e em XML, sendo este último uma escolha consensual, por ser suficientemente expressivo, como formato de partilha de informação. No entanto, a escolha do XML deixa em aberto que tipo de ferramentas utilizar para o tratamento da informação e qual a melhor estrutura a utilizar para representar a informação a ser tratada com essas ferramentas. Para tentar responder a estes problemas, apresenta-se um formato lógico, implementado com base no PROLOG (mais especificamente no SWI-PROLOG (Wielemaker, 2008)) que resolve ambas estas questões de forma elegante e integrada. Foram utilizados como casos de estudo dois corpora de domínio público distintos: o corpus *Susanne* (Sampson, 1995) para o inglês e a *Floresta Sintá(c)tica* (Bick et al., 2007) (`www.linguateca.pt/Floresta`), mais especificamente o subconjunto anotado da Floresta também incluido no CETEMPúblico, o Bosque.

A motivação para a utilização do PROLOG e linguagens lógicas suas derivadas (por exemplo o DyALog (`dyalog.gforge.inria.fr`) advém do facto de, sendo o PROLOG uma linguagem declarativa, ser extremamente fácil descrever o conhecimento e efectuar inferências e deduções sobre esse conhecimento. Assim, a programação declarativa, na linguística computacional, desde cedo tem sido considerada uma ferramenta clássica. Por exemplo, as Gramáticas de Cláusulas Definidas[1] do PROLOG (Pereira e Shieber, 1987) são ferramentas extremamente úteis para a expressão de gramáticas e para a sua aplicação. Es-

[1] *DCG — Definite Clause Grammars*

tas gramáticas podem ser utilizadas para extrair uma representação lógica relativamente a determinados padrões de etiquetas que ocorrem em textos previamente marcados.

Contudo, apesar de o PROLOG ser excelente para a manipulação simbólica, as suas capacidades para tratamento *Input/Output* e para análise sub-simbólica eram inicialmente mais limitadas. Talvez por isso, em geral, o PROLOG tenha sido menos utilizado para representação de corpora de textos. As implementações mais recentes do PROLOG resolvem este problema de forma fácil e elegante ao possibilitar a integração com o XML.

1.1 Descrição do formato TXT/2

Na estrutura proposta (designada formato $TXT/2$), parte-se do princípio que, num texto, as frases são naturalmente divididas em palavras. No entanto, aqui, generaliza-se o conceito de palavra para estruturas de palavras (w), cada uma recursivamente composta por sequências de palavras (cw - *container of* w). Com base neste conceito de palavra/estrutura de palavras ganhamos uma maior abstracção, sem nunca perder a informação original do texto em causa. Assim, uma estrutura $TXT/2$ é constituída por estruturas de palavras (w), aglomeradas numa lista PROLOG. Esta lista representa todo o conhecimento necessário para o processamento do texto. Cada uma das palavras é caracterizada por vários atributos, salientando-se os atributos que indicam a palavra (wd) e a etiqueta dessa palavra (tag – para etiquetas morfossintácticas). Da mesma forma que se associa uma etiqueta a cada palavra, também se pode associar uma etiqueta a uma sequência de palavras cw. Nesse caso podemos igualmente ter a tag da palavra para terminais ou um atributo $gtag$ – para não terminais, tipicamente contendo a etiqueta sintáctica. Note-se que deve ser atribuído um *ID* único a cada elemento do texto. Um exemplo característico é o caso da palavra composta "Presidente da República", cuja notação (por questões de espaço, forçosamente simplificada) seria a seguinte:

```
txt(1, [w([1,wd='Presidente da República', semantic=headOfState, tag='PROP',
          cw=[w([1-1, wd='Pesidente', bw=presidente,gen=masc]),
                  w([1-2, wd='da',
                          cw=[w([1-2-1, wd='de', tag='PREP']),
                          w([1-2-2, wd='a', tag='DET', gen=fem])]]),
                    w([1-3, wd='República', gen=fem])]])
          ]).
```

No exemplo apresentado, todo o contexto (que, consoante o caso, poderá ser um texto ou simplesmente uma frase) deverá ser representado na lista PROLOG. O primeiro argumento deste termo é um *ID* único desse texto (facilitando a interligação com uma base de dados relacional). Como foi referido, a lista $TXT/2$ é composta por palavras (cada uma

com um *ID* especifico, discriminante relativamente ao texto em que se insere), inserindo os elementos cw, recursivamente, outras listas deste tipo.

Uma das grandes vantagens desta abordagem é a sua fácil e directa integração com as *D*CGs do PROLOG. Desta forma é possível representar e gerar árvores de análise e gramáticas para extracção de conhecimento dos textos anotados segundo a formatação aqui proposta. Eis o exemplo gerado para a primeira árvore no Bosque:

```
txt(s1,[
        w([s1_500, gtag=s,
         ref="CP1-1", source="CETEMPúblico n=1 sec=clt sem=92b", forest="1",
         text="Um revivalismo refrescante",
           cw=[w([s1_501, gtag=np,
                    cw=[w([s1_1, wd='Um', tag=art, label='>N', lemma=um,
                    morph='M S', extra=arti, extra2= --, extra3= --]),
                w([s1_2, wd=revivalismo, tag=n, label='H',
                    lemma=revivalismo, morph='M S', extra= --, extra2= --,
                        extra3= --]),
                        w([s1_502, gtag=adjp,
                        cw=[w([s1_3, wd=refrescante, tag=adj, label='H',
                             lemma=refrescante, morph='M S', extra= --,
                             extra2= --, extra3= --])],label='N<'])],
              label='UTT'])]])]).
```

Note-se em particular os elementos tag, terminais da gramática, classificadores, e os elementos $gtag$ que, pelas suas características não terminais, são susceptíveis de desenvolvimento arborescente, como cw, até se encontrarem elementos terminais. As anotações extra, como $label$, $lemma$ ou $morph$ são mantidas na representação proposta não se perdendo assim nenhuma informação.

Outra vantagem é a gestão de co-referências. Assim, neste formato, podem facilmente ser incluídas co-referências bem como qualquer outro tipo de anotações, atributos ou variáveis. A título de exemplo considere-se a frase "O João, que precisava de um livro, foi à biblioteca" na qual a componente relativa "que precisava de um livro" pode ser representada recursivamente.

1.2 Exemplos de aplicação e resultados

A recursividade na estrutura $TXT/2$ pode ser facilmente tratada com o PROLOG. Assim, por exemplo, para sabermos se uma palavra W está ou não marcada com uma tag (o que torna imediata a construção de um dicionário utilizando o predicado $findall/3$), basta o seguinte código PROLOG:

```
get_wd_tag(W,Tag, IDW, TXT2)  :-
          member(w( [IDW|WL]), TXT2), member(wd=W, WL), member(tag=Tag, WL).
get_wd_tag(W, Tag, IDW, TXT2) :-
          member(w([IDW|WL]), TXT2), member(cw=CW, WL),
          get_w_tag(W, Tag, CW).
```

A primeira cláusula apenas refere que uma palavra W marcada com etiqueta tag é um membro de uma palavra (ela própria membro de uma estrutura $TXT/2$). A segunda cláusula resolve a recursividade, indicando que a estrutura $TXT/2$ pode ser ela própria membro de um elemento cw.

Já o processo inverso à construção do dicionário, i.e. anotação do texto (em inglês, *POS tagging*), requer a utilização do contexto para desambiguar entre as possíveis etiquetas que uma palavra pode ter. Como foi referido, o contexto está disponível na lista $TXT/2$. Assim, tal como em qualquer outra análise sintáctica ou semântica, a facilidade de acesso directo e sequencial às palavras é essencial. Neste momento, um sistema neuro-simbólico, utilizando uma rede neuronal treinada com um conjunto de apenas 5000 palavras extraídas do Bosque Sintáctico e com auxilio de regras de desambiguação (Marques et al., 2007), já consegue obter uma precisão de classificação na ordem dos 94% no CETEMPúblico. No entanto espera-se que este valor melhore substancialmente com a adição de mais características e regras de desambiguação, visto que no *Susanne* este valor já chega aos 95%.

Uma vez que o formato proposto apresenta uma estrutura lógica (compatível com a linguagem PROLOG), é possível a sua conversão para o formato TIGER-XML (TIGER), e vice-versa. Assim, no sentido de apresentar este formato e ferramentas PROLOG associadas como possíveis serviços numa possível arquitectura SOA para a Linguateca, foi igualmente implementado um leitor de informação TIGER-XML, que possibilita a conversão de textos para o formato $TXT/2$. Utilizaram-se como base de estudo os textos em português do Bosque em formato TIGER-XML e o Susanne, convertido para formato TIGER-XML. Mas poderão igualmente ser utilizadas outras formas de converter informação para PROLOG. A título de exemplo, durante o trabalho efectuado foi comum a utilização de informação em formato *S*QL e a utilização da linguagem de reconhecimento de padrões *awk* (Aho et al., 1988).

Note-se ainda que a estrutura $TXT/2$ funciona como um armazém de anotações efectuadas sobre o texto. Isto estimula a interoperabilidade que pode haver entre diversas aplicações que utilizem e manipulem os textos, promovendo uma filosofia *S*OA. Esta filosofia pode funcionar tanto ao nível de módulos internos PROLOG, como ao nível mais geral de *W*ebServices utilizando XML. Assim, a utilização de uma estrutura deste tipo pode contribuir para a utilização de módulos com maior poder dedutivo potenciando, em simultâneo, a partilha de informação que se deseja obter com o projecto da Linguateca.

Capítulo 2

Extracção de recursos de tradução

Alberto Simões

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 2, p. 7–11, 2008

Este documento resume a dissertação na extracção de recursos de tradução (Simões, 2008) e a sua integração nos objectivos da Linguateca. A dissertação teve como principal objectivo o estudo de métodos para a extracção de recursos de tradução para a língua portuguesa, uma vez que a principal investigação na tradução automática não tem dado a atenção merecida a esta língua.

A tradução automática tem vindo a dar cada vez mais atenção aos métodos de tradução baseados em dados. Estes métodos reaproveitam as traduções que já foram realizadas (no mesmo ou noutros contextos) para realizar as novas traduções. O principal problema desta abordagem é conseguir emparelhar as traduções já realizadas com a frase a traduzir. Por exemplo, nos sistemas de tradução assistida por computador (em inglês, CAT – *Computer Assisted Translation*) é habitual que só sejam reaproveitadas frases muito semelhantes às já traduzidas. Para os sistemas de tradução automática pretende-se aumentar a aplicabilidade das frases já traduzidas, aplicando algoritmos que dividam as traduções já realizadas em segmentos mais pequenos (sintagmas ou simples segmentos de palavras paralelos) com maior reutilização (e que são chamados de *exemplos de tradução*).

Em trabalho anterior (Simões e Almeida, 2003; Simões, 2004) tinham sido estudados métodos para a extracção de dicionários probabilísticos de tradução. Estes dicionários são associações entre palavras na língua de origem com um conjunto de possíveis traduções na língua de destino juntamente com a respectiva probabilidade de tradução (ver figura 2.1).

$$\mathcal{T}\,(\text{codificada}) = \begin{cases} \text{codified} & 62.83\% \\ \text{uncoded} & 13.16\% \\ \text{coded} & 6.47\% \\ \cdots & \end{cases}$$

Figura 2.1: Extracto de um dicionário probabilístico de tradução para a palavra "codificada".

Embora extraídos automaticamente e sem garantias de grande qualidade, mostraram-se extremamente úteis para a extracção de novos recursos de tradução. Além das várias avaliações reportadas na dissertação de doutoramento, foram feitas algumas comparações (Santos e Simões, 2008) de resultados destes dicionários com dicionários de tradução de cores obtidos manualmente a partir do COMPARA (Frankenberg-Garcia e Santos, 2003).

Para a extracção dos vários recursos foram usados vários corpora. Para além do COMPARA foram utilizados o EuroParl v2 (Koehn, 2005), o JRC-Acquis (Steinberger et al., 2006), El Monde Diplomatique (Correia, 2006) e o EurLex, um corpus construido pelo Projecto Natura, com mais de um milhão de unidades de tradução. Algumas versões alinhadas destes corpora, bem como os respectivos dicionários de tradução, estão acessíveis para consulta interactiva em `linguateca.di.uminho.pt/nat/`.

Todo os recursos extraídos durante a dissertação usaram como base os dicionários pro-

babilísticos de tradução para estabelecer pontes entre palavras de duas línguas, e foram aplicaradas diferentes metodologias para a extracção de exemplos de tradução:

- o uso da hipótese das palavras-marca (*Marker Hypothesis*) como mecanismo de segmentação dos corpora paralelos, e o uso das probabilidades de tradução constantes nos dicionários probabilísticos de tradução para o alinhamento destes segmentos (Simões, 2007b).

  Este método baseia-se num conjunto de palavras (pronomes, artigos, alguns advérbios, etc) que, de acordo com Green (1979), podem ser usados como um método eficaz de segmentação:

  O João passou toda a tarde a brincar com os colegas.
  ⇓
  O João passou toda a tarde a brincar com os colegas.
  ⇓
  (*O João passou*) (*toda a tarde*) (*a brincar*) (*com os colegas.*)

  Esta abordagem já tinha sido usada para a segmentação para tradução automática (Armstrong et al., 2006) mas sem terem sido realizadas experiências com a língua portuguesa, nem usando dicionários probabilísticos de tradução para o alinhamento dos segmentos extraídos. A tabela 2.1 apresenta os exemplos (1:1) mais ocorrentes extraídos do EuroParl PT:EN com base na hipótese das palavras-marca.

Ocorrências	Português	Inglês
36886	senhor presidente	mr president
8633	senhora presidente	madam president
3152	espero	I hope
2930	gostaria	I would like
2572	o debate	the debate
2511	penso	I think
2356	está encerrado	is closed
1939	penso	I believe
1932	muito obrigado	thank
1854	em segundo lugar	secondly
$\bar{x} = 1.6654$	Total de 1 507 225 exemplos 1:1	

Tabela 2.1: Exemplos mais ocorrentes extraídos com base na hipótese das palavras marca.

- a construção de uma matriz de alinhamento para cada unidade de tradução, onde cada célula da matriz é preenchida com a probabilidade mútua de tradução entre palavras (ver figura 2.2). Nesta matriz são procuradas as células com probabilidades

mais elevadas, e que correspondem às traduções provavelmente correctas. Estas traduções são extraídas e são criados exemplos de tradução (Simões e Almeida, 2006a). Esta abordagem não é totalmente nova (Melamed, 2001), mas foi introduzido o uso de dicionários probabilísticos de tradução e o uso de padrões de alinhamento.

	discussion	about	alternative	sources	of	financing	for	the	european	radical	alliance	.
discussão	**44**	0	0	0	0	0	0	0	0	0	0	0
sobre	0	**11**	0	0	0	0	0	0	0	0	0	0
fontes	0	0	0	**74**	0	0	0	0	0	0	0	0
de	0	3	0	0	**27**	0	6	3	0	0	0	0
financiamento	0	0	0	0	0	**56**	0	0	0	0	0	0
alternativas	0	0	**23**	0	0	0	0	0	0	0	0	0
para	0	0	0	0	0	0	**28**	0	0	0	0	0
a	0	1	0	0	1	0	4	**33**	0	0	0	0
aliança	0	0	0	0	0	0	0	0	0	0	**65**	0
radical	0	0	0	0	0	0	0	0	0	**80**	0	0
europeia	0	0	0	0	0	0	0	0	**59**	0	0	0
.	0	0	0	0	0	0	0	0	0	0	0	**80**

Figura 2.2: Matriz de alinhamento.

- a extracção de exemplos, que correspondem a segmentos nominais (próximos de sintagmas nominais, e candidatos a terminologia), com base em padrões de alinhamento, que especificam as trocas de ordem de palavras que ocorrem durante a tradução (Simões e Almeida, 2008).

  Foi desenvolvida uma nova linguagem de domínio específico para a especificação de padrões com objectivos distintos das linguagens de padrões actualmente a serem usadas na área da tradução automática (Och e Ney, 2004; Sánchez-Martínez e Forcada, 2007).

  Seguem-se alguns exemplos de padrões, bem como a respectiva ilustração/interpretação na figura 2.3. Os segmentos nominais extraídos são contados. O número de ocorrências de cada par permite associar-lhe uma noção de qualidade, de acordo com a tabela 2.2. A tabela 2.3 contém algumas medidas de avaliação destes recursos. Consultar Simões (2008) para detalhes sobre a forma como a avaliação foi realizada.

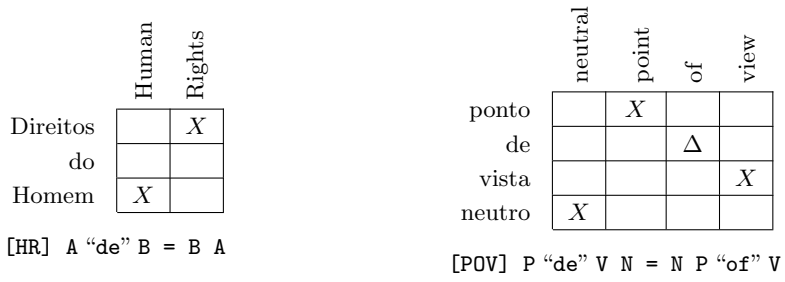

Figura 2.3: Padrões de alinhamento HR e POV.

39214	comunidades europeias	european communities
32850	jornal oficial	official journal
32832	parlamento europeu	european parliament
32730	união europeia	european union
15602	países terceiros	third countries
[...]	[...]	[...]
1	órgãos orçamentais	budgetary organs
1	órgãos relevantes	relevant bodies
1	óvulos de equino	equine ova
1	óxido de cádmio	cadmium oxide
1	óxido de estireno	styrene oxide

Tabela 2.2: Extracto das contagens de unidades nominais.

Padrão	Total	Máx.	Mediana	Min.	Precisão
A B = B A	77 497	938	2	1	86 %
A "de" B = B A	12 694	204	2	1	95 %
A B C = C B A	7 700	40	1	1	93 %
I "de" D H = H D I	3 336	21	1	1	100 %
A B C = C A B	1 466	4	1	1	40 %
P "de" V N = N P "of" V	564	6	1	1	98 %
P "de" T "de" F = F T P	360	3	1	1	96 %

Tabela 2.3: Avaliação de unidades nominais extraídas.

Para além da experimentação dos métodos, estes foram disponibilizados num pacote de ferramentas de código aberto, denominado NATools (Simões e Almeida, 2007). As ferramentas constantes neste pacote foram adaptadas para funcionarem de forma distribuída cliente/servidor (Simões e Almeida, 2006b) e de forma paralela num *cluster* computacional (Simões, 2007a). Além disso, parte destes recursos foram usados no CLEF de 2005 (Cardoso et al., 2006a).

Capítulo 3

Novas ferramentas e recursos linguísticos para a tradução automática: por ocasião d'o fim do início de uma nova era no processamento da língua portuguesa

Anabela Barreiro

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 3, p. 13–23, 2008

A comemoração dos 10 anos da Linguateca é uma data que marca o início de uma nova fase na história do processamento da língua portuguesa. Ao longo de 10 anos a Linguateca teve um papel relevante na formação de recursos humanos, nomeadamente de linguistas computacionais e revelou um empenho assumido na ampliação do número e variedade de recursos linguísticos e ferramentas partilháveis disponíveis em domínio público: mais de uma dezena de novas ferramentas públicas ao serviço da comunidade, três grandes projectos de avaliação a nível nacional, várias participações em projectos internacionais de avaliação, centenas de publicações nas mais variadas conferências e revistas da especialidade, são os resultados de uma equipa que trabalhou muito em pouco tempo.

Apesar de o meu estatuto ser, na sua globalidade, oficial e logisticamente externo à Linguateca, o meu espírito rege-se pelos ideais defendidos pela Linguateca. Quando em 1998 em Portugal se discutia o futuro do processamento computacional do português, encontrava-me a trabalhar numa empresa de tradução automática nos Estados Unidos. Tive oportunidade de vir a Lisboa aquando do debate público em torno do Livro Branco e desde então sempre acompanhei de perto e atentamente as actividades que se foram desenvolvendo ao longo destes anos, participando, ainda que de forma breve, em algumas das suas actividades, nomeadamente na criação de um programa de estágio em tradução automática, na anotação da Floresta Sintá(c)tica (Afonso et al., 2001, 2002), na organização das Morfolimpíadas e construção/revisão da lista dourada (Santos e Barreiro, 2004; Barreiro e Afonso, 2007), e no desenvolvimento de ferramentas para avaliação da tradução automática (Sarmento et al., 2007; Maia e Barreiro, 2007), entre outras. Em 2003, a Linguateca iniciou a divulgação e dinâmica de actividades de avaliação conjuntas em várias áreas do processamento de língua natural, mobilizando a comunidade para a criação de um grupo de avaliação de tradução automática, o grupo ARTUR, apresentado no AVALON 2003. Em colaboração com a Universidade do Porto, trabalhou-se no desenvolvimento de uma ferramenta automática de geração de baterias de teste (Sarmento, 2007) e num programa de categorização de erros, brevemente descrito em Santos et al. (2004). Desde essa data, a minha ligação à Linguateca estreitou-se com o meu doutoramento a ser co-orientado pela Professora Belinda Maia, responsável pelo pólo do Porto. Desde então, mais conhecimentos acerca dos problemas associados com a preservação de significado no processo de tradução foram adquiridos, incluindo os problemas levantados por expressões idiomáticas, coloquialismos e usos metafóricos, entre outros. Na etapa final deste projecto, posso testemunhar o papel dos recursos da Linguateca na minha tese e os resultados que obtive e que gostaria de apresentar e retribuir à comunidade linguística. É no espírito, ideais e prática da Linguateca, a visão da língua como um bem comum e a partilha de conhecimento e recursos para o avanço do processamento da língua portuguesa que o trabalho que a seguir é apresentado se enquadra.

3.1 Tradução automática com conhecimento linguístico parafrástico

O projecto de doutoramento que se apresenta consiste no melhoramento da tradução automática através de um conhecimento estritamente linguístico sobre paráfrases. Neste trabalho, os corpora anotados disponibilizados pela Linguateca, nomeadamente os corpora anotados do COMPARA (Frankenberg-Garcia e Santos, 2003; Santos e Inácio, 2006), serviram como ponto de partida para a inventariação de fenómenos linguísticos e de criação de algumas regras parafrásticas bilingues. Posteriormente, foram desenvolvidos recursos para a tradução automática de português para inglês com base em recursos do sistema OpenLogos (`logos-os.dfki.de`). A parte monolingue desses recursos deu origem ao Port4NooJ, um sistema baseado em ontologias lexicais, descrito em Barreiro (2007) e disponível publicamente em `www.nooj4nlp.net` e em `www.linguateca.pt/Repositorio/Port4Nooj/`. O Port4NooJ foi construído com base em dicionários e gramáticas locais, com conteúdo sintáctico e semântico, criados no ambiente de desenvolvimento linguístico NooJ (Silberztein, 2004). Os recursos linguísticos criados para o Port4NooJ já foram integrados no Corpógrafo (Sarmento et al., 2004; Maia e Matos, 2008) e estão a ser utilizados na criação de novos recursos derivados, nomeadamente um dicionário de expressões multipalavra e duas ferramentas automáticas que permitem gerar e reutilizar esses recursos, ambos geradores de paráfrases, o ReEscreve e o ParaMT, que apresentamos a seguir.

Ambos os parafraseadores integram a aplicação de conhecimento da estrutura argumental de predicados (Meyers et al., 2004b,a). A análise sintáctico-semântica é feita no âmbito do quadro teórico do léxico-gramática (Gross, 1981, 1975), que assenta nos princípios da gramática transformacional harrissiana (Harris, 1968, 1957). Para ilustrar o funcionamento dos parafraseadores são seleccionadas paráfrases de construções com verbos suporte elementares, tais como "fazer uma visita a", que podem ser parafraseados por exemplo por verbos lexicais semanticamente fortes, tais como "visitar" ou variantes estilísticas desses verbos (verbos suporte não elementares), tais como "efectuar uma visita a", entre outras. As construções com verbos suporte têm sido estudadas de modo extensivo tanto do ponto de vista teórico como prático em várias línguas incluindo o português (Ranchhod, 1990; Baptista, 2001; Chacoto, 2005) e como tal apresentam-se como um ponto de partida sólido para o parafraseamento.

3.2 ReEscreve: um parafraseador monolingue

O ReEscreve é um parafraseador multifuncional autónomo, usado para a geração de paráfrases monolingues. A interface *web* de acesso ao ReEscreve está disponível em `poloclup.linguateca.pt/Reescreve/`. O serviço público de ajuda e sugestão à escrita permite alterar, simplificar ou clarificar textos e tem aplicações, entre outras, na preparação

gosto de ver o comboio a	fazer corridas /correr	à velocidade máxima ao lon
o de cheque especial para	fazer doações /doar	às entidades que escolher. A
ores e, quando é preciso ir	fazer filmagens/filmar	fora do estúdio, às vezes fic
e queria trocar de pares e	fazer um jogo /jogar	ao melhor de três sets , mas
dra deu-me um papel para	fazer uma lista de/listar	todas as coisas boas que ex
res foram à caracterização	fazer uns retoques/retocar	, outros estão a descansar n

Figura 3.1: Reconhecimento e parafraseamento monolingue de construções com verbos suporte (construção com verbo suporte / verbo lexical equivalente).

de textos e escrita de linguagem controlada, nomeadamente na pré-edição de texto para a tradução automática. O ReEscreve foi desenhado para funcionar em modo interactivo ou em modo automático, mas, de momento, apenas o modo interactivo está disponível. Neste modo de utilização, o utilizador escolhe um texto que pretende editar e coloca-o na janela de inserção de texto. Seguidamente, selecciona um comando para obter resultados. O texto é-lhe devolvido com sugestões para a revisão. As sugestões apresentadas pelo ReEscreve surgem em paralelo com as expressões originais e o utilizador pode escolher qual a expressão que lhe agrada mais. Ao clicar na sua expressão favorita, o ReEscreve automaticamente altera (ou mantém) o texto, eliminando as expressões equivalentes não seleccionadas pelo utilizador. Para além disso, o utilizador tem a opção de comparar as diferentes frases possíveis (paráfrases), e/ou acrescentar expressões que acredite serem mais adequadas em cada contexto.

Tanto o dicionário, como a base de dados parafrástica contém apenas lemas de construções com verbos suporte. As formas flexionadas são obtidas através do sistema flexional do Port4NooJ. Pretende-se gerar futuramente paráfrases de outros tipos de expressões e oferecer um leque vasto de alternativas que o utilizador possa utilizar de acordo com o estilo que pretenda para o seu texto.

A figura 3.1 mostra uma concordância onde algumas construções com verbos suporte são reconhecidas e parafraseadas como verbos lexicais.

A figura 3.2 mostra uma concordância onde construções com verbos suporte que co-ocorrem com nomes predicativos ligados à área biomédica, tais como *fazer uma operação*, são reconhecidas e parafraseadas com verbos lexicais, tais como *operar*, ou variantes estilísticas léxico-sintácticas (verbos suportes não elementares) das construções com verbos suporte originais, tais como *realizar uma operação* ou *submeter-se a uma operação*. Conhecimento acerca da estrutura argumental do predicado permite a distinção de diferentes variantes estilísticas. Por exemplo, as variantes estilísticas *sujeitar-se a* e *submeter-se a* são apenas utilizadas nos casos em que o sujeito é um paciente.

nça, o cirurgião Faivre, ao	fazer uma amputação/amputar
nça, o cirurgião Faivre, ao	fazer uma amputação/efectuar uma amputação
nça, o cirurgião Faivre, ao	fazer uma amputação/realizar uma amputação
ı ser interrogadas antes de	fazer um aborto/submeter-se a um aborto
ı ser interrogadas antes de	fazer um aborto/abortar
ı ser interrogadas antes de	fazer um aborto/efectuar um aborto
ı ser interrogadas antes de	fazer um aborto/realizar um aborto
o público de saúde recusa	fazer uma operação cirúrgica/realizar uma operação cirúrgica
o público de saúde recusa	fazer uma operação cirúrgica/efectuar uma operação cirúrgica
Tiago Felizardo, vai ter de	fazer uma operação plástica depois de/sujeitar-se a uma operação plástica
Tiago Felizardo, vai ter de	fazer uma operação plástica depois de/submeter-se a uma operação plástica
Tiago Felizardo, vai ter de	fazer uma operação plástica depois de/realizar uma operação plástica
Tiago Felizardo, vai ter de	fazer uma operação plástica depois de/efectuar uma operação plástica
aber se o doente consegue	fazer uma prova de esforço/sujeitar-se a uma prova de esforço
aber se o doente consegue	fazer uma prova de esforço/submeter-se a uma prova de esforço
aber se o doente consegue	fazer uma prova de esforço/realizar uma prova de esforço
aber se o doente consegue	fazer uma prova de esforço/efectuar uma prova de esforço
ı médico também lhe pode	fazer uma prova de esforço para/realizar uma prova de esforço
ı médico também lhe pode	fazer uma prova de esforço para/efectuar uma prova de esforço
médico sempre vai querer	fazer um transplante de/realizar um transplante
médico sempre vai querer	fazer um transplante de/efectuar um transplante
omista britânico, conseguiu	fazer uma transfusão de sangue/realizar uma transfusão de sangue
omista britânico, conseguiu	fazer uma transfusão de sangue/efectuar uma transfusão de sangue
os pacientes que precisam	fazer uma transfusão de sangue/sujeitar-se a uma transfusão de sangue
os pacientes que precisam	fazer uma transfusão de sangue/submeter-se a uma transfusão de sangue
os pacientes que precisam	fazer uma transfusão de sangue/realizar uma transfusão de sangue
os pacientes que precisam	fazer uma transfusão de sangue/efectuar uma transfusão de sangue

Figura 3.2: Reconhecimento e parafraseamento de construções com verbos suporte que co-ocorrem com nomes predicativos da área biomédica (construção com verbo suporte / verbo lexical equivalente ou variante estilística)

3.3 ParaMT: um parafraseador bilingue/multilingue

O ParaMT é um parafraseador bilingue/multilingue que opera como uma função integrada em sistemas de tradução automática e é utilizado essencialmente para a geração de equivalentes de tradução (Barreiro, 2008). O processo de reconhecimento de uma construção com verbo suporte em texto é idêntico à do ReEscreve. As gramáticas locais instruem o programa a converter automaticamente a expressão da língua de partida num equivalente parafrástico na língua de chegada. Esse equivalente pode ser um verbo lexical ou uma variante estilística. A figura 3.3 mostra a tradução de uma construção com verbo suporte em português por um verbo lexical em inglês.

a fazer um estágio para	dar aulas de/teach	religião, mas não se import
m -- os filhos -- juntos e	fizeram a mudança para/change	Johannesburg, e ensinaram
. Necessitava apenas de	ter a certeza de/know	que não escapara à sua
ente hipotética. -- Deves	ter alguma ideia/know	. Dorothy andava a fazer u
não podemos deixar de	ter cautela/beware	. Pobre Caro, pensou Lync
ra dos chinelos, antes de	ter chance de/can	mudar de idéia. Como pos
ope a Jean, esta pareceu	ter dificuldade em/avoid	olhá-lo nos olhos. Deixou
ao Kiss dela. Apesar de	ter falta de/lack	amor-próprio, isso não sign
igos e imprensa estava a	ter lugar /occur	numa longa galeria com car
quiu ter filhos. -- Tens de	ter mão /control	nessa confusão toda. Sam a
spondi, minha mãe deve	ter medo de/fear	cobras. Eu disse no Gabin
da loja antes de ele	ter tempo de/could	chamar a brigada de narcó
a triste aventura havia de	ter um fim/finish	.
Ela ouvira a tia Velma	ter uma discussão com/argue	Jack acerca de mostarda r
de olhos fechados para	ter uma ideia de/know	como seria ser cego e
ter paciência.» «Voltei a	ter uma imensa vontade de/want	viver. A conversa parecia

Figura 3.3: Reconhecimento e tradução de construções com verbos suporte (construção com verbo suporte em português / verbo lexical equivalente em inglês).

3.4 Recursos e metodologia adoptados na concepção dos parafraseadores

Os corpora disponibilizados pela Linguateca foram utilizados para a pesquisa de termos e para a obtenção de exemplos ilustrativos da existência de tais construções. Para além disso, de forma a processar as construções com verbos suporte, o dicionário foi melhorado com propriedades adicionais. A acrescentar à informação mais comum de categoria gramatical e de paradigma flexional, cada entrada do dicionário inclui a descrição dos atributos sintáctico-semânticos (SynSem), bem como as propriedades distribucionais e transformacionais para as expressões com um comportamento sintáctico mais variável. As entradas apresentam propriedades como: argumentos predicativos, verbos suporte, verbos aspectuais, variantes estilísticas dos verbos suporte elementares, informação acerca dos determinantes e preposições que ocorrem com os nomes predicativos em expressões "menos variáveis" e propriedades derivacionais. A derivação é muito importante porque tem implicações não só ao nível lexical, mas também ao nível sintáctico. Muitas vezes, os sufixos derivacionais aplicam-se a palavras de uma categoria sintáctica e transformam-nas em palavras de uma categoria sintáctica diferente, mantendo a sua integridade semântica. Por exemplo, o afixo *-ção* permite transformar o verbo *adaptar* no nome *adaptação* e o afixo *-mente* permite transformar o adjectivo *rápido* no advérbio *rapidamente*. Estas transformações são extremamente importantes para as construções com verbos suporte porque permitem estabelecer gramáticas de equivalência que efectuam o mapeamento entre

(i) construções com verbos suporte como *fazer uma adaptação (de)* e o verbo lexical *adaptar*, onde o nome predicativo *adaptação* mantém uma relação semântica e morfossintáctica com o verbo *adaptar* ou (ii) construções com verbos suporte como *ter um final rápido* e a expressão verbal *terminar rapidamente*, onde o nome predicativo autónomo *final* mantém uma relação semântica com o verbo *terminar*, e o advérbio *rapidamente* mantém uma relação semântica e morfossintáctica com o adjectivo *rápido*. Assim sendo, as entradas do dicionário do Port4NooJ contêm a identificação dos paradigmas derivacionais para as nominalizações (anotação *NDRV*) e uma ligação ao(s) verbo(s) suporte(s) do nome derivado (anotação *VSUP*), como ilustra a figura 3.4.

adaptar, V+FLX=FALAR+Aux=1+INOP57+Subset132+EN=adapt+VSUP=fazer
+DRV=NDRV00:CANÇÃO +NPrep=de
favor,N+FLX=MAR+Npred+AB+state+EN=favor+VSUP=fazer+NPrep=a+VRB=ajudar
rápido, A+FLX=RÁPIDO+PV+eagerType+EN=quick+DRV=AVDRV06:RAPIDAMENTE
adoçar, V+FLX=COMEÇAR+Aux=1+OBJTRundif75+Subset604+EN=sweeten+
DRV=ADRV11:VERDE+VSUP=tornar
transplantar, V+FLX=FALAR+Aux=1+RECTR26+Subset=504+BioMed+EN=transplant+
SUBJ=AG+VSUP=fazer+DRV=NDRV79:ANO+NPrep=de+DO=BP+IO=PAT+VSTYLE=
sofrer+VSTYLE=realizar+VSTYLE=efectuar+VASP=iniciar+VASP=prosseguir+VASP=concluir
médico, N+FLX=ANO+AN+des+Med+EN=doctor
médico, N+FLX=ANO+AN+des+Med+EN=physician

Figura 3.4: Amostra do dicionário.

As nominalizações são acompanhadas pelas propriedades correspondentes ao paradigma flexional (NDRV00:CANÇÃO). Quaisquer outras restrições lexicais, tais como preposições, determinantes, ou argumentos obrigatórios, etc., são igualmente acrescentados. Os nomes predicativos autónomos (não-nominalizações), tais como *favor* são lexicalizados e classificados com a anotação *Npred* e têm associados a eles verbos suporte e outras restrições lexicais, tais como uma preposição (*NPrep*), ou um verbo lexical (*VRB*) com as mesmas características semânticas. Os adjectivos predicativos estão também classificados (ADRV) e foi estabelecida a ligação entre eles e os verbos correspondentes, tais como entre o verbo *adoçar* e o adjectivo *doce*. Foi também iniciada a atribuição de verbos suporte correspondentes a estes adjectivos. As variantes estilísticas das construções com verbos suportes elementares estão anotadas como *VSTYLE*. As variantes aspectuais estão anotadas como *VASP*. Foi iniciada a adição de argumentos sintácticos e semânticos de um predicado às entradas do dicionário. Por exemplo, na entrada lexical para o verbo *transplantar*, a propriedade *SUBJ=AG* significa que o verbo selecciona um agente como seu argumento semântico na posição sintáctica de sujeito. *SUBJ=PAT* significa que o verbo selecciona um paciente como seu argumento semântico na posição sintáctica de sujeito. O argumento sintáctico *DO=ORG* significa que o predicado selecciona um objecto directo que é um órgão humano

dar parte de fraco, V+IDIOM+FLX=PHRDAR+EN=become weak+VRB=fraquejar
dar cabo dos nervos, V+IDIOM+FLX=PHRDAR+EN=enervate+VRB=enervar
dar pontadas de dor, V+IDIOM+FLX=PHRDAR+EN=hurt+VRB=doer
bater as botas, V+IDIOM+FLX=PHRBATER+EN=die+VRB=morrer
bater na mesma tecla, V+IDIOM+FLX=PHRBATER+EN=insist+VRB=insistir
abrir o coração, V+IDIOM+FLX=PHRABRIR+EN=talk+VRB=desabafar
pôr cobro a, V+IDIOM+FLX=PHRPOR+EN=end+VRB=terminar
dar lugar a, V+IDIOM+FLX=PHRDAR+EN=lead to+EN=result in+VRB=conduzir a+VRB=resultar em
dar cabo de, V+IDIOM+FLX=PHRDAR+EN=destroy+VRB=destruir
pôr um ponto final em, V+IDIOM+FLX=PHRPOR+EN=end+VRB=acabar com

Figura 3.5: Amostra da base de dados fraseológica e parafrástica com expressões idiomáticas.

(subclasse de parte do corpo). *IO=PAT* significa que o predicado selecciona um objecto indirecto que é um paciente. *NPrep=de* significa que a construção com verbo suporte (verbo suporte mais nome predicativo) selecciona a preposição *de* (*fazer um transplante de*). Os nomes (substantivos) são classificados semanticamente. Por exemplo, o nome *médico* está classificado como um ser animado que denota uma profissão ou outra designação humana (*AN+des*), pertencente ao domínio médico (*Med*).

As construções com verbos suporte semi-cristalizadas e idiomáticas, onde o verbo suporte é a única palavra que varia em toda a expressão, são armazenadas no dicionário de expressões multipalavra e mantidas numa base de dados fraseológica juntamente com outras expressões idiomáticas. Por exemplo, em *dar parte de fraco* ou *dar cabo dos nervos*, na figura 3.5, os verbos suporte *dar* e *pôr* são marcados com uma propriedade correspondente ao paradigma flexional e as restantes palavras na expressão permanecem invariáveis. À medida que os dicionários são melhorados no que respeita à semântica e sintaxe de palavras simples, tenciona-se alargar e redefinir o papel dos dicionários electrónicos de modo a incluir entradas de expressões multipalavra, incluindo construções com verbos suporte e as suas paráfrases.

O método de reconhecimento e parafraseamento utilizado neste trabalho consiste na ligação sistemática entre palavras relacionadas semântica e/ou morfossintacticamente no dicionário electrónico através do estabelecimento de propriedades derivacionais e distribucionais. De forma a obter as paráfrases monolingues das construções com verbos suporte utilizando o NooJ, combinaram-se as propriedades formalizadas nos dicionários com as gramáticas locais. Uma das novidades deste trabalho em relação ao que já existia, é precisamente a aplicação das gramáticas locais para o reconhecimento e geração de paráfrases de construções com verbos suporte e para a tradução. De modo a estabelecer relações de equivalência morfossintáctica entre predicados nominais e verbais, utilizam-se as propriedades dos dicionários. Uma vez que todos os nomes predicativos estão classi-

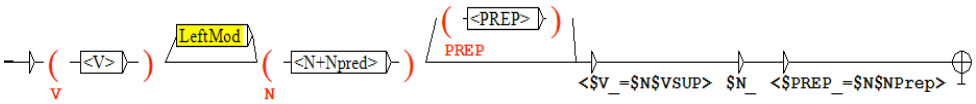

Figura 3.6: Gramática para o reconhecimento e parafraseamento de construções com verbos suporte.

ficados no dicionário como [*Npred*], esta informação lexical pode ser usada numa gramática local para a identificação do predicado numa construção com verbo suporte e aplicar esta gramática a corpora. A figura 3.6 representa uma gramática local simples usada para reconhecer e gerar construções com verbos suporte e transformá-las nas suas paráfrases verbais.

Esta gramática reconhece verbos suporte seguidos de um modificador (determinante, adjectivo, advérbio ou outros quantificadores), de um nome predicativo e opcionalmente de uma preposição. Os elementos entre parênteses ( ) são guardados em variáveis V, N ou PREP. Se uma entrada de dicionário contém uma restrição lexical, tal como NPrep=a na expressão [*dar um grande abraço a*], a construção com verbo suporte será reconhecida pela gramática e mapeada ao verbo abraçar, o lema do nome especificado na variável $N_. Os elementos a negrito <$V_=$N$VSUP>, e $PREP_=$N$NPrep> representam restrições lexicais que são apresentadas na saída, tais como a especificação do verbo suporte ou da preposição que pertence a uma construção com verbo suporte específica. O nome predicativo é identificado, mapeado com o seu derivante e exibido como um verbo. Os outros elementos da expressão são eliminados.

3.5 Avaliação quantitativa: primeiros resultados

Para a avaliação do ReEscreve, foram seleccionadas a partir do COMPARA todas as frases onde a forma infinitiva dos verbos *fazer*, *dar*, *pôr*, *tomar* e *ter* ocorria seguida de um nome ou de um determinante e um nome. Em primeiro lugar, foram classificadas manualmente estas combinações para ver se elas correspondiam a construções com verbos suporte ou não. Confirmou-se que 89% das ocorrências de *dar*, 88% de *tomar*, 77% de *pôr*, 47% de *fazer* e 20% de *ter* são verbos suporte. Isto significa que na sua globalidade, em 64,2% das vezes estes verbos são verbos suporte, o que corresponde a quase 2/3 das ocorrências. A seguir a esta contagem, foi seleccionado um sub-corpus de 500 frases obtidas de modo aleatório (100 frases para cada um dos cinco verbos seleccionados), contendo apenas construções com verbos suporte. As construções foram anotadas manualmente e os resultados comparados com os resultados obtidos automaticamente. Elaboraram-se regras de reconhecimento mais restritas para que o parafraseamento fosse mais preciso. Actualmente, são reconhecidas 62,6% de construções com verbos suporte com valores elevados em termos

	Reconhecimento Precisão	Reconhecimento Cobertura	Parafraseamento Precisão
Pôr	73/73 - 100%	73/100 - 73%	72/73 - 98,6%
Tomar	75/75 - 100%	75/100 - 75%	68/73 - 93,1%
Ter	65/65 - 100%	65/100 - 65%	59/65 - 90,7%
Dar	57/60 - 95%	57/100 - 57%	46/51 - 90,1%
Fazer	43/45 - 95,5%	43/100 - 43%	40/45 - 88,8%
Média	62,6/63,6 - **98,4**%	62,6/100 - **62,6**%	57/61 - **93,4**%

Tabela 3.1: Avaliação do reconhecimento e parafraseamento de construções com verbos suporte.

de precisão (98,4% para reconhecimento e 93,4% para parafraseamento). Os resultados do reconhecimento e parafraseamento (precisão e cobertura) do ReEscreve estão ilustrados na tabela 3.1.

3.6 Considerações finais

Os parafraseadores ReEscreve e ParaMT e os recursos linguísticos do Port4NooJ, que estão na base destas ferramentas, podem ser integrados facilmente noutros recursos da Linguateca e colocados ao serviço da comunidade. Os recursos do Port4NooJ já estão a ser utilizados no Corpógrafo, mas a sua versatilidade e detalhe linguístico, nomeadamente a informação sintáctica e semântica, são apropriados para a obtenção de concordâncias mais sofisticadas e extracção de termos, expressões multipalavra e fraseologia. Prevê-se a criação de um maior número de gramáticas de desambiguação para análise sintáctico–semântica e o desenvolvimento de dicionários mais completos e mais ricos em informação linguística. O passo seguinte será o melhoramentos dos recursos já reunidos e o desenvolvimento de novos recursos para testar e melhorar o ReEscreve, que poderão servir posteriormente, entre outras coisas, para uma anotação mais completa e rigorosa dos corpora anotados, como por exemplo, do AC/DC (Santos e Sarmento, 2002) ou para pesquisa e extracção de informação mais sofisticada do ponto de vista linguístico e com funcionalidades alargadas. E finalmente, o alargamento dos recursos, de modo a desenvolver o sistema de tradução automática já iniciado. A falta de projectos de tradução automática envolvendo o português, deixa a nossa língua desfasada da realidade da tradução automática e é necessário colmatar esta deficiência através de iniciativas como as que já foram propostas no âmbito da Linguateca.

A política de disponibilização e partilha de recursos praticada pela Linguateca, a colaboração e junção de esforços começa agora a gerar os seus primeiros frutos. É importante salvaguardar os recursos produzidos até ao momento, mantendo-os em sistemas de fácil acesso, como em código aberto. Como as peças de um quebra-cabeças que se vão unindo

para formar um todo, é necessário juntá-los para que se criem a partir deles recursos cada vez maiores, mais completos e mais enriquecidos linguisticamente. Estão criadas as infra-estruturas e reunidas as competências e condições necessárias para a criação de colaborações que possam ter objectivos concretos em relação aos actuais desafios tecnológicos de um mundo cada vez mais virado para a globalização da informação. É importante criar iniciativas semelhantes à da Linguateca, e até mesmo, há necessidade de criar um organismo ou uma sociedade internacional de análise e processamento de língua portuguesa, com actividades centradas em áreas específicas, mas sempre com uma visão global da língua. A especialização de recursos humanos em várias áreas do processamento do português, nomeadamente em entidades mencionadas, entidades geográficas, ontologias, extracção e recuperação de informação, tradução automática, entre outras, são uma mais-valia que deve ser aproveitada pela sociedade em geral, tanto para o desenvolvimento de ferramentas de utilidade pública como privada. Como legado da Linguateca, podemos contar com uma nova etapa para o futuro do processamento da língua portuguesa, com novos desafios e inúmeras oportunidades!

Capítulo 4

O corpus CONDIV e o estudo da convergência e divergência entre variedades do português

Augusto Soares da Silva

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 4, p. 25–28, 2008

Pretendemos apresentar o corpus CONDIV, em construção, e perspectivar o seu desenvolvimento como instrumento fundamental para o estudo da convergência e divergência entre as variedades europeia e brasileira do português. O CONDIV foi compilado no âmbito do projecto de investigação "Convergência e divergência no léxico do português" (2004-2006), centrado na questão diacrónica da convergência/divergência lexical entre o português europeu e o português brasileiro nos últimos 50 anos e, secundariamente, na questão sincrónica da estratificação lexical das duas variedades. Parte do corpus está disponibilizado na Linguateca, fazendo parte do projecto AC/DC (Santos e Sarmento, 2002), e espera-se que brevemente o restante também aí se possa encontrar.

O CONDIV compreende actualmente textos de três domínios – futebol, vestuário e saúde – e está estruturado na base de três variáveis: (i) geográfica (Portugal vs. Brasil), (ii) diacrónica (1950-1970-1990/2000) e (iii) estilística (jornais e revistas > etiquetas, Net fóruns > Net *chats*). Os materiais foram extraídos de três fontes: (i) jornais de desporto e revistas de moda e de saúde dos primeiros anos das décadas de 50, 70 e 90-2000; (ii) linguagem da Internet de conversação electrónica de IRC ou *chats*; e (iii) etiquetas de roupas de lojas de vestuário. Todos os materiais de (i) e (iii) foram extraídos manualmente. Os materiais do português brasileiro provêm de São Paulo e Rio de Janeiro. A extensão actual do corpus é de 5 milhões de palavras do registo formal (jornais e revistas) e 15 milhões do registo informal (*chats* e etiquetas).

O objecto de análise é a variação onomasiológica que envolve sinónimos *denotacionais* – porque são estes os que melhor revelam a própria existência e a competição de variedades de uma língua – e a sua base empírica consiste em largos milhares de observações do uso de sinónimos alternativos que designam um mesmo conceito/referente. Na fase actual, foram estudados 43 conceitos nominais dos campos lexicais do futebol (21 conceitos e 183 termos) e vestuário (22 conceitos e 264 termos).

Os resultados da investigação sociolexicológica realizada (Silva, 2005) indicam que as duas variedades divergem claramente no vocabulário do vestuário (e a distância actual entre padrão e subpadrão é aí maior na variedade brasileira), mas convergem ligeiramente no vocabulário do futebol. Os mesmos resultados apontam para a existência de mais mudanças na variedade brasileira. É também a variedade brasileira a que manifesta uma maior permeabilidade aos estrangeirismos.

Pretendemos prosseguir a investigação estendendo-a em três direcções: (i) ainda no domínio lexical, incluir outros campos lexicais e passar das palavras *de conteúdo* às palavras *funcionais*, particularmente as preposições; (ii) em direcção ao domínio gramatical, analisar variáveis não-lexicais, sobretudo sintácticas e morfológicas; e (iii) ampliar e refinar a situação estratificacional de cada variedade, acrescentando mais variáveis estilísticas, em ordem ao estudo dos indicadores de subestandardização. Para isso, é essencial aumentar quantitativa e qualitativamente o corpus existente. Dada a escassez de textos em suporte informático que cumpram o critério diacrónico do projecto e a escassez ou mesmo falta de

subsídios para a investigação, lançamos um desafio à Linguateca no sentido da convergência de projectos e recursos.

O quadro teórico da presente investigação é o da linguística cognitiva. Três razões justificam a opção pela linguística cognitiva: (i) a sua orientação *recontextualizante* (reintegrando as diferentes formas de *contexto*, excluídas pelos modelos autonomistas, particularmente o generativista), para o *significado* (incluindo o significado social) e *baseada-no-uso* (origem da própria variação linguística); (ii) a importância dada à flexibilidade e à variação semasiológica e onomasiológica (Silva, 2006a); e (iii) a assunção de que não só a capacidade para a linguagem se fundamenta em capacidades cognitivas gerais, como também todas estas capacidades são cultural e socialmente situadas. Mais especificamente, a nossa investigação insere-se no âmbito da emergente sociolinguística cognitiva (Kristiansen e Dirven, 2008; Silva, 2006b, 2008), inevitavelmente implicada por aqueles princípios. A sociolinguística cognitiva está bem colocada para ajudar a resolver duas tensões: (i) a tensão entre o *cognitivo* e o *social*, integrando no programa cognitivista a variação intralinguística (Geeraerts, 2005; Bernárdez, 2005); e (ii) a tensão entre o *cognitivo* e o *empírico*, optando pela metodologia de corpus (ou dados experimentais) e por técnicas quantitativas capazes de analisar a natureza *multivariacional* do uso linguístico (Geeraerts, 2006; Gonzalez-Marquez et al., 2007). Neste âmbito, a presente investigação apoia-se na concepção e nos métodos quantitativos da investigação desenvolvida por D. Geeraerts e sua equipa para o neerlandês da Holanda e da Bélgica (Geeraerts et al., 1999).

Para medir a convergência e a divergência entre variedades linguísticas são utilizados dois métodos quantitativos: a medida de uniformidade (U) entre as variedades e a medida do impacto de determinado traço (A) nessa uniformidade. A medida U é a medida fundamental. Envolve duas noções específicas: *perfil onomasiológico* ou conjunto de sinónimos denotacionais usados para designar determinado conceito ou função, diferenciados pela sua frequência relativa, e *uniformidade* ou medida da correspondência entre dois conjuntos de dados, definidos em termos de perfis onomasiológicos. Por exemplo, a uniformidade de um conceito entre duas amostras, em que uma contém 6 ocorrências do termo A e 4 do termo B e a outra 3 ocorrências de A e 7 de B, resulta do número de pares comuns de nomeação desse conceito (7 pares), sendo portanto U = 70%. Este resultado obtém-se somando as frequências relativas mais pequenas de cada termo alternativo: 30% de A e 40% de B. Tecnicamente, a uniformidade de um conceito é calculada pela seguinte fórmula:

$$U_Z(Y_1, Y_2) = \sum_{i=1}^{n} min(F_{Z,Y_1}(x_i), F_{Z,Y_2}(x_i))$$

Isto é, a uniformidade U para um conceito Z entre duas amostras Y_1 e Y_2 equivale à soma $\sum$ dos mínimos das frequências relativas F do termo x nos perfis onomasiológicos de Z em Y_1 e Y_2. O símbolo x_i representa os diferentes termos x_i a x_n usados nas amostras

Y para designar o conceito Z. Quando estão em causa vários conceitos, a uniformidade média é calculada em termos de média ponderada:

$$U'(Y_1, Y_2) = \sum_{i=1}^{n} U_{z_i}(Y_1, Y_2) G_{z_i}$$

A uniformidade U' para um conjunto de conceitos Z entre duas amostras Y_1 e Y_2 equivale à soma dos valores-U dos Zs ponderados pela frequência relativa G de Z dentro do conjunto total de Zs. Convergência e divergência entre duas variedades exprimem-se em aumento e diminuição de U/U', respectivamente. Os cálculos ponderados (U', A') são mais representativos, porque têm em conta a frequência relativa dos diferentes conceitos e termos em análise. A título de exemplo, a tabela 4.1 apresenta os resultados de U e U' do perfil de "avançado" no sub-corpus de jornais de desporto e relativamente a um total de 90.202 observações do uso dos termos de futebol seleccionados.

FUTEBOL	**P50**		**B50**		**P70**		**B70**		**P00**		**B00**	
atacante	101	8,8	119	36,6	50	13,6	208	73,8	42	9,7	658	96,2
avançado	820	71,6	3	0,9	175	47,4	0	0,0	240	55,4	0	0,0
avante	0	0,0	159	48,9	0	0,0	31	11,0	0	0,0	23	3,4
dianteiro	220	19,2	22	6,8	74	20,1	2	0,7	38	8,8	0	0,0
forward	1	0,1	17	5,2	0	0,0	0	0,0	0	0,0	0	0,0
ponta-de-lança	3	0,3	5	1,5	70	19,0	41	14,5	113	26,1	3	0,4
AVANÇADO	U = 16,9		U' = 0,6		U = 28,8		U' = 0,8		U = 10,1		U' = 0,4	

Tabela 4.1: Frequência (absoluta e relativa) e uniformidade U e U' do perfil onomasiológico de "avançado".

Como extensões da investigação realizada, são analisados dez perfis onomasiológicos preposicionais e três perfis onomasiológicos construcionais. Os perfis preposicionais, restringidos ao mesmo contexto sintagmático, de forma a satisfazer a condição de sinonímia denotacional, incluem casos como *falar de/sobre/acerca de/em*, *precisar/necessitar de/∅*, *ansioso de/para/por*. Os perfis construccionais incluem construções com verbos causativos e perceptivos seguidos de complemento finito ou infinitivo e construções com os adjectivos atributivos *verdadeiro*, *falso*, *bonito*, *lindo*, *recente* em posição posposta ou anteposta. A anotação do CONDIV, feita pelo analisador sintáctico automático PALAVRAS (Bick, 2000), permite a pesquisa imediata das ocorrências das referidas regências e construções, incluindo as mais complexas, como as diversas construções causativas e perceptivas seguidas de complemento infinitivo. Uma hipótese sociolinguisticamente relevante é a de que as palavras funcionais e as construções sintácticas se comportam em termos de variação linguística diferentemente dos itens lexicais de conteúdo. Os resultados obtidos, embora em número ainda reduzido, permitem confirmar a hipótese de uma divergência mais acentuada entre as duas variedades nacionais em relação a variáveis funcionais e sintácticas.

Capítulo 5

Os recursos da Linguateca ao serviço do desenvolvimento da tecnologia de fala na Microsoft

Daniela Braga e Miguel Sales Dias

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 5, p. 29–33, 2008

No contexto das comunidades científicas do processamento da linguagem natural, linguística computacional e áreas relacionadas, como o processamento da fala, é consensual dizer que o panorama do processamento computacional do português não seria definitivamente o mesmo, sem a existência do projecto e do grupo de interesse dinamizado pela Linguateca. Com efeito, apesar de o português ser uma das línguas mais faladas do mundo enquanto língua materna (com cerca de 235 milhões de falantes) e língua oficial de 8 estados independentes (Angola, Brasil, Cabo Verde, Guiné Bissau, Moçambique, Portugal, São Tomé e Príncipe, Timor), se recuarmos uma década apenas, era clara a escassez de recursos disponíveis para as comunidades científicas da linguística e da engenharia da linguagem, sobretudo em formatos inteligíveis para processamento computacional. A Linguateca veio preencher com sucesso essa lacuna, contribuindo não só para a aproximação entre as comunidades científicas portuguesa e brasileira que trabalham em processamento da linguagem natural, linguística computacional e áreas afins, como também para a divulgação de trabalhos académicos e para a disponibilização livre de recursos linguísticos, metodologias de avaliação, ferramentas computacionais e resultados de projectos de I&D, nestes domínios, que de outra forma se manteriam dispersos e dificilmente acessíveis.

Assim se compreende a surpresa e alguma consternação com que a comunidade científica, recebeu a notícia de que o projecto da Linguateca iria terminar no final do ano civil de 2008, por exaustão do respectivo financiamento público. A Linguateca tem desempenhado ao longo dos seus 10 anos de existência um papel catalisador de sinergias oriundas das várias comunidades portuguesa e brasileira que trabalham no processamento computacional do português e teve como resultado a produção de valiosos recursos linguísticos, nomeadamente, de texto processados e de motores de utilização proveitosa dos mesmos, todos eles com direitos cedidos ao domínio público e assim disponibilizado gratuitamente à comunidade científica. Para além desta, outras mais-valias saídas do projecto Linguateca, são de realçar, nomeadamente, a criação de uma comunidade científica especializada na produção de recursos linguísticos e na produção de trabalho de valor académico ao mais alto nível, a organização das campanhas mais estendidas de avaliação de recursos em português – como as Morfolimpíadas (Santos e Costa, 2003) e o HAREM (Santos e Cardoso, 2007) – ou, ainda, o apoio à participação dos recursos para o português no CLEF (Cross-Language Evaluation Forum, Forum de avaliação entre várias línguas). De destacar ainda, o trabalho tão útil à comunidade, de repositório de publicações científicas relacionadas com o processamento computacional do português que o sítio da Linguateca tem proporcionado, permitindo consultar com facilidade trabalhos de referência na área, teses de mestrado e doutoramento, etc, sendo ainda acessível através de um sistema de busca assistida, o SUPeRB (Cabral et al., 2008).

5.1 A experiência da indústria: o impacto da Linguateca no desenvolvimento de produtos na Microsoft

Mais do que destacar as virtualidades sobejamente conhecidas trazidas pela comunidade que criou e desenvolveu o universo Linguateca, gostaríamos de deixar o nosso testemunho enquanto membros integrantes da indústria, a qual entende que as sinergias estratégicas academia-indústria são factores que propiciam a inovação e a melhoria da qualidade do desenvolvimento de produtos de software.

O MLDC – Microsoft Language Development Center, sediado no Tagus Park, Porto Salvo, é um grupo de produto da Microsoft onde os autores trabalham, integrado num ambiente de desenvolvimento distribuído, que inclui, para além de Portugal, pólos em Redmond/EUA e Pequim/China. Este grupo alargado produz as tecnologias de reconhecimento e síntese de fala, utilizadas pelos diversos grupos de produto da companhia nos seus desenvolvimentos de software, tais como os grupos Live, Servidor, Cliente (Windows), Mobilidade, Automóvel e Entretenimento. O MLDC participa em todas as acividades centrais do grupo de fala, tais como a expansão das tecnologias de fala para um número elevado de línguas, onde se inclui o desenvolvimento da síntese de fala em português europeu e português do Brasil. Muito naturalmente o MLDC ficou incumbido de desenvolver as tecnologias de fala para estas duas variantes do português. Este artigo é, assim, uma oportunidade para os autores relatarem o impacto extremamente positivo dos recursos linguísticos disponibilizados pela Linguateca, no desenvolvimento de tecnologias de síntese de fala, no MLDC.

De facto, conhecedores do projecto e sítio da Linguateca, os engenheiros e linguistas do MLDC cedo começaram a utilizar de forma assídua alguns dos recursos disponíveis, para o teste e avaliação dos algoritmos de processamento da linguagem natural, que integram os sistemas de síntese de fala em português europeu e em português do Brasil. Podemos salientar a utilização dos seguintes recursos: CETEMPúblico (Rocha e Santos, 2000), CETENFolha, COMPARA (Frankenberg-Garcia e Santos, 2003) e a Floresta Sintá(c)tica (Bosque e Floresta Virgem) (Afonso et al., 2001). Todos estes recursos, com excepção do COMPARA (corpus paralelo em português e inglês nos dois sentidos), podem ser descarregados mediante o preenchimento de um formulário simples em `www.linguateca.pt`. Foram várias as tarefas executadas sobre os corpora CETEMPúblico, CETENFolha e COMPARA, de forma a torná-los adequados ao desenvolvimento de produto na Microsoft:

1. Selecção automática de frases a serem gravadas para a construção de uma base de dados de talentos de voz;
2. Selecção de casos de teste, saídos de corpora de texto real, para validação de algoritmos de:
   a. Separação de frases;

   b. Separação de palavras;
   c. Normalização de texto;
   d. Desambiguação de homógrafos;
   e. Conversão grafema-fone;

3. Obtenção e generalização de padrões para criação de regras de normalização de texto;

4. Obtenção de listas de frequência de léxico em certos domínios.

Os corpora anotados morfológica e sintacticamente, como a Floresta Sintá(c)tica, tiveram uma outra utilidade para nós. Quer o Bosque (inclui os primeiros 1000 excertos do CETEMPúblico e do CETENFolha revistos por linguistas), quer a Floresta Virgem (é composta pelo primeiro milhão de palavras do CETEMPúblico e do CETENFolha anotado automaticamente pelo analisador sintáctico PALAVRAS (Bick, 2000)), foram usados como corpus de treino do nosso analisador sintáctico automático, cujo resultado é depois utilizado pelo módulo de desambiguação de homógrafos e cuja desambiguação é também obtida através da análise do contexto sintáctico.

Gostaríamos de salientar ainda a escassez de recursos de texto e corpora anotados morfossintacticamente de larga dimensão existentes para o português, para além daqueles que são disponibilizados pela Linguateca. Na verdade, um exercício simples de busca no catálogo da ELRA (European Language Resources Association, `catalog.elra.info`) ou da APPEN (`www.appen.com.au`), dois catálogos de recursos linguísticos de referência na indústria, mostra que não existem muitos mais recursos disponíveis no mercado para o português com qualidade e quantidade significativas, pelo que o desenvolvimento dos nossos sintetizadores nas duas variedades do português se baseou quase exclusivamente nos recursos disponibilizados pela Linguateca.

5.2 Conclusão

A cessação do financiamento público da Linguateca não deve ser encarada como o fim deste projecto, ou das comunidades que ajudou a sensibilizar e a dinamizar. Antes pelo contrário, defendemos que deve ser entendido como uma hipótese de renovação, de regeneração e renascimento. Vivemos numa sociedade de informação e comunicação em rápido processo de maturação e transformação, que necessita de sistemas de busca eficientes, de sistemas de tradução automática para se mover na babel linguística, de tecnologias de fala para facilitar a interacção com as máquinas, só para citar alguns exemplos. Toda esta panóplia tecnológica de texto e de fala assenta em corpora mais extensos, de géneros mais diversificados, com processos de busca mais eficientes. A Microsoft necessitará sempre de recursos linguísticos com qualidade (corpora de texto, corpora paralelos, léxicos fonéticos,

corpora de fala transcrito e anotado), para melhorar os seus sistemas de síntese (Braga et al., 2008) e reconhecimento de fala, os seus sistemas de busca (MSN, `www.msn.com`) e os seus tradutores automáticos (Translator, `www.windowslivetranslator.com`). Por outro lado, a Microsoft verifica que continuam a existir lacunas na oferta de recursos para o português. As actividades desencadeadas e dinamizadas pela Linguateca permitiram formar pessoas capazes de suprir essas lacunas, pessoas essas que são hoje especialistas na produção de recursos linguísticos para o português e que podem inclusive expandir esse conhecimento para a outras línguas. A indústria e a sociedade esperam assim que o ecosistema de processamento computacional do português, dinamizado pela Linguateca, se mantenha vivo e activo.

Capítulo 6

Um estudo no COMPARA: a semântica dos compostos nominais

Lílian Figueiró Teixeira e Rove Luiza de Oliveira Chishman

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 6, p. 35–41, 2008

Com os dados disponíveis no córpus paralelo COMPARA (Frankenberg-Garcia e Santos, 2002) – `www.linguateca.pt/COMPARA`, v.10.1.2 – é possível estudar uma série de fenômenos lingüísticos a partir de equivalências de tradução nas línguas portuguesa e inglesa. Neste trabalho, dedicamo-nos ao estudo da semântica dos compostos nominais, tendo como ponto de partida a sua tradução do inglês para o português. Através das linhas de concordância obtidas no sítio do COMPARA, foi possível realizar o estudo da semântica dos componentes destas construções e a identificação de padrões de tradução destes compostos. Salientamos que, por padrão, compreendemos apenas características semânticas que sejam recorrentes nos resultados de tradução na língua alvo, neste caso, no português.

Os compostos nominais são extremamente produtivos na língua inglesa, o que representa um desafio para os sistemas de análise e produção da linguagem natural, em especial para a tradução automática. A grande dificuldade encontrada por um sistema de tradução automática é o reconhecimento de mais de duas palavras como uma unidade. Uma frase como *I went to the night school* seria traduzida como *Eu fui à escola de noite*, pois *night* não seria identificado como um modificador de *school*. Este sistema não chegaria à tradução esperada: *escola noturna*.

Nosso interesse, neste trabalho, é fazer um estudo da semântica dos compostos formados por dois substantivos (NN), identificando, dentre as abordagens que se ocupam deste fenômeno, as que se prestam à sua representação. Considerando como motivação as tarefas de processamento computacional, vale salientar que algumas das dificuldades em processar os compostos estão relacionadas à complexidade deste fenômeno lingüístico, o que se evidencia na própria diversidade de tratamento teórico que o fenômeno vem recebendo. Alguns estudos elegem a teoria do Léxico Gerativo (Pustejovsky, 1995) como um modelo representativo para os compostos. É o caso de Busa e Johnston (1996), que analisaram, com base nos papéis da estrutura qualia, ocorrências nas línguas inglesa e italiana a fim de identificar os padrões semânticos dos compostos. Copestake (2003) segue na mesma linha, adaptando a estrutura qualia para uma classificação dos compostos NN, utilizando-a como base, mas incluindo outras classes que dêem conta dos dados do estudo.

Neste estudo, seguimos Barker e Szbapakowicz (1998), que consideram que um composto formado por dois substantivos (NN) em inglês apresenta um pré-modificador seguido por um substantivo núcleo. Adotamos a estrutura qualia para a interpretação dos dados, mas outras etiquetas semânticas também são consideradas, como as que foram propostas por Girju et al. (2005), tais como tempo, posse e local. Optamos por esta abordagem, já que os papéis qualia não cobrem os diferentes tipos de compostos. Em língua inglesa, geralmente o modificador é o substantivo da esquerda e o núcleo é o da direita. Em *samba school*, *samba* é o modificador e *school* o seu núcleo. A ordem muda em português, pois o modificador aparece após o seu núcleo, conforme visto em *escola de samba*.

Um outro conceito importante, quando trabalhamos com compostos, são as *core words*,

traduzidos aqui como *nódulos*. Segundo Ryder (1994), tanto o modificador quanto o núcleo podem ser o nódulo, já que é esta palavra que vai ser encontrada em outros compostos, formando o que chamamos de família de compostos. Assim, uma palavra como *school* serve de nódulo participando de diferentes compostos, tais como *grammar school*, *summer school*, *law school*, *sister school*, *pottery school* e *state school*.

6.1 Extração dos dados do COMPARA

A ferramenta de busca desenvolvida pela equipe do COMPARA se mostrou extremamente útil e capaz de fornecer os dados necessários para a realização deste estudo. Precisávamos extrair seqüências de dois substantivos em inglês seguidos pela sua tradução em português. O fato de o córpus estar etiquetado foi o que possibilitou este tipo de busca. Para obter estas informações, adotamos os seguintes passos:

1. Foi feita uma busca por linhas de concordância em que dois substantivos aparecem juntos. Consideramos tanto os substantivos no singular quanto no plural e, com a fórmula **[pos="N.*"][pos="N.*"]** digitada na busca avançada, obtivemos, como resultado, 32.216 ocorrências. A partir deste primeiro resultado, alguns nódulos recorrentes foram selecionados: *hall*, *room*, *house*, *door*, *floor*, *table*, *window* e *school*.

2. As linhas de concordância para cada combinação das palavras de busca seguidas ou antecedidas por outro substantivo foram analisadas, e os equivalentes de tradução foram identificados. Utilizamos a fórmula **[pos="N.*"& word="school"] @[pos="N.*"]** para cada busca, sendo que os diferentes nódulos eram digitados onde se encontrava a palavra *school*. Também invertemos a ordem do nódulo e selecionamos as expressões com um número maior de resultados. Um outro recurso interessante no sítio do COMPARA é a possibilidade de visualização do número de ocorrências e da lista de palavras que ocupam o lugar de N em cada fórmula. Isto é possível através dos itens "especifique os resultados" e "distribuição dos lemas" encontrados no formulário de busca avançada. A tabela 6.1 sistematiza estes primeiros resultados.

3. Como o objetivo do estudo é analisar os compostos nominais formados por dois substantivos, os compostos formados por mais de dois substantivos e os com algum elemento deverbal (como *–ing*) foram excluídos.

4. Para uma melhor visualização das opções de tradução para cada composto, os equivalentes de tradução de uma mesma expressão foram agrupados em um arquivo separado.

Composto	Ocorrências	Exemplo
N hall	67	*concert hall*
N room	78	*hotel room*
N house	226	*country house*
N door	135	*kitchen door*
N floor	54	*ground floor*
N table	94	*dinner table*
N window	55	*train window*
school N	47	*school gate*
N school	59	*summer school*

Tabela 6.1: Compostos nominais do córpus.

6.2 Análise das relações semânticas

Feita a extração dos compostos, passamos para a análise. Valemo-nos dos papéis télico e constitutivo, tal como propostos por Pustejovsky (1995) na formulação da estrutura qualia. Por papel télico, compreende-se que um dos elementos expressa a função ou propósito do composto, geralmente o modificador se presta a isto. Já o constitutivo estabelece a relação entre o todo e as suas partes. Também utilizamos as categorias de posse, local e tempo, tal como propostas por Girju et al. (2005). Nosso propósito, a partir deste estudo semântico dos compostos, foi verificar como estes sentidos vêm a se expressar nos equivalentes de tradução. A seguir, apresentamos um quadro sistematizando esta análise comparativa a partir dos papéis semânticos.

A relação entre os dois substantivos de uma expressão composta, na maioria dos casos, pode ser explicada através de dois papéis da estrutura qualia, o constitutivo e o télico. Em *dinner table*, o substantivo modificador (N1) indica o propósito desta mesa, que é o de ser utilizada durante a janta. Já o papel constitutivo estabelece a relação entre o todo e as suas partes, como em *school gate*, em que *portão* é parte de *escola*.

Analisando os equivalentes em português, identificamos diversos significados para a preposição "de" como parte de uma expressão composta. Além dos papéis constitutivo e télico, identificamos a relação de posse e outras relações como tempo e local. Sentimos a necessidade de incluir estas relações na análise, por não conseguir incluir os exemplos nos papéis e por percebermos uma relação diferente entre os substantivos. Se em *church hall* interpretamos que o salão faz parte da igreja, em *street door* não temos a mesma relação. Não se pode dizer que a porta faça parte da rua, no entanto, o que importa é o fato de alguém poder chegar até a rua ao passar por esta porta. Desta forma, a localização da porta é o que motiva a criação deste composto. Os compostos que trazem alguma informação relacionada ao tempo, como em *summer school* e *sunday school*, também não se ajustaram aos papéis estudados e mereceram uma classificação diferenciada. Entre os casos estudados,

Composto	Exemplos	Tradução	Relação semântica
N hall	concert hall	sala de concertos	papel télico
	entrance hall	átrio de entrada/entrada	papel télico
	church hall	salão de igreja	papel constitutivo
	parish hall	salão paroquial	papel constitutivo
	school hall	salão da escola/refeitório	papel constitutivo
N room	hotel room	quarto de hotel	papel constitutivo
	laundry room	quarto de engomados/lavanderia	papel télico
	emergency room	pronto-socorro	papel télico
N house	station house	delegacia	papel télico
	summer house	casa de verão	tempo
	family house	casa da família	posse
	brick house	casa de tijolo	papel constitutivo
	beach house	casa da praia	local
	hen house	galinheiro	papel constitutivo
N door	kitchen door	porta da cozinha	papel constitutivo
	trap door	alçapão	papel constitutivo
	glass door	porta de vidro/porta envidraçada	papel constitutivo
	street door	porta da rua	local
	garden door	porta que dava para o jardim	local
N floor	ground floor	andar térreo	local
	kitchen floor	chão da cozinha	papel constitutivo
	metal floor	chão metálico	papel constitutivo
N table	kitchen table	mesa da cozinha	papel constitutivo
	bedside table	mesa-de-cabeceira	local
	dinner table	mesa de jantar	papel télico
	coffee table	mesinha	papel télico
	tin table	mesa metálica	papel constitutivo
N window	kitchen window	janela da cozinha	papel constitutivo
	picture window	janela panorâmica	papel télico
	ticket window	guichê	papel télico
school N	school holiday	férias	tempo
	school gate	portão do colégio	papel constitutivo
	school report	boletim escolar	local
N school	summer school	curso de verão	tempo
	night school	escola noturna	tempo
	Sunday school	escola dominical/catequese	tempo

Tabela 6.2: Análise dos compostos.

houve apenas uma única ocorrência em que a relação de posse pudesse ser percebida: uma *family house* pode ser interpretada como uma casa que pertence à família.

Algumas vezes, os substantivos modificadores são traduzidos como um adjetivo em português. Se é possível traduzir o composto de duas formas, N de N ou N Adjetivo, os dois casos são encontrados no córpus. Geralmente o uso do adjetivo está relacionado a algum material do qual o objeto é feito. Exemplos deste caso são *metal floor* e *tin table*, cujos equivalentes de tradução são *chão metálico* e *mesa metálica*. Quando não há um adjetivo correspondente em português para o material, mantém-se a construção N de N (*brick house*). Como uma casa de tijolo possui tijolos, consideramos que o modificador representa o papel constitutivo.

Quando existe uma única palavra em português correspondente ao composto em inglês, o seu uso é preferido. Enquanto há três ocorrências para *lavanderia*, *quarto de engomados* só aparece uma única vez. Outros equivalentes são escolhidos, pois se percebe certo grau de lexicalização no seu uso. *Coffee table* foi considerado um composto télico, pois é uma mesa utilizada para servir café. No entanto, se observarmos o seu equivalente (*mesinha*), a informação mais importante aqui não é o seu uso, mas o seu tamanho.

6.3 Considerações finais

O estudo aqui empreendido e a definição de uma tipologia semântica para descrever os compostos nominais do tipo NN em inglês e seus correspondentes em português pode servir de base para pesquisas voltadas para o aprimoramento de sistemas de tradução automática. Quando padrões da língua são conhecidos, é possível identificar automaticamente os compostos e criar léxicos que possam ser usados em tarefas relacionadas ao processamento da língua natural.

O acesso a um córpus paralelo se mostrou útil para um estudo bilíngüe, podendo contribuir para outros estudos sobre diferentes fenômenos lingüísticos e inclusive multilíngües. Cumprimentamos a iniciativa dos organizadores do córpus COMPARA em compilar este material e disponibilizá-lo gratuitamente. A comunidade acadêmica carece de recursos desta qualidade e de livre acesso. Sugerimos a disponibilização de alguma ferramenta ou documento que apresente uma lista de n-gramas do córpus. Para este estudo em especial, uma lista com dois substantivos que ocorrem juntos seguidos pela sua freqüência no córpus teria ajudado.

O foco deste trabalho foi verificar as equivalências de tradução considerando o inglês como língua fonte e o português como língua alvo. No entanto, acreditamos que seja interessante, para um futuro estudo, analisar como os compostos são traduzidos do português para o inglês. Observar quais os equivalentes de tradução em inglês dos compostos formados por N de N na língua portuguesa poderia ser um propósito de estudo. Também não procuramos separar os resultados de acordo com as variantes da língua portuguesa, por-

tuguês europeu e brasileiro, pois com isso acabaríamos diminuindo os dados de estudo. No entanto, a ferramenta de busca do COMPARA permite trazer apenas os resultados de uma variante específica.

Capítulo 7

Linguateca e Processamento de Linguagem Natural na Área da Saúde: Alguns Comentários e Sugestões

Liliana Ferreira, António Teixeira e João Paulo da Silva Cunha

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 7, p. 43–48, 2008

A crescente utilização de sistemas de informação na área da saúde, levou a um aumento significativo da informação médica disponível electronicamente sob a forma de texto em linguagem natural. A necessidade de gerir e processar grandes quantidades de dados motiva o recente interesse em aproximações semânticas, cujos principais objectivos se prendem com a redução de erros médicos, a melhoria da eficiência médica e uma maior satisfação e segurança dos pacientes. As tecnologias de Rede Semântica auxiliam na obtenção destes objectivos através de múltiplas ontologias populadas, anotação semântica automática de documentos e processamento de regras, entre outros.

A experiência do Instituto de Engenharia Electrónica e Telemática de Aveiro (IEETA) no desenvolvimento de sistemas de informação na área da saúde contribuiu para o recente interesse no desenvolvimento de sistemas de processamento de português, capazes de extrair informação pertinente de um grande volume de textos médicos em linguagem natural. Um exemplo desta motivação é o projecto Rede Telemática da Saúde, RTS (Cunha et al., 2006) que pretende disponibilizar, de forma segura, o acesso a informação clínica e promover a comunicação entre profissionais de saúde credenciados, bem como envolver o cidadão na gestão da sua saúde, contribuindo para um melhor acesso aos cuidados de saúde. Esta Rede implementa um Processo Clínico Electrónico Regional resumido, que agrega informação clínica do utente, proveniente de várias fontes de informação clínica geograficamente distribuídas pelas várias instituições de saúde regionais. Entre outros, a RTS disponibiliza acesso a cartas de alta, boletins de análises clínicas, relatórios de exames de imagiologia, etc.

7.1 MedAlert

Motivado pela Rede Telemática de Saúde, está actualmente em desenvolvimento no IEETA o projecto MedAlert - Sistema de Processamento de Linguagem Médica (Ferreira et al., 2008), que tem por objectivo a utilização de técnicas de extracção automática de informação de textos médicos, de modo a inferir, de uma forma automática, irregularidades/dúvidas suscitadas pelas decisões tomadas pelos profissionais de saúde. Este sistema, que deverá tomar a forma dum módulo escalável e adaptável a diferentes configurações de sistemas de informação hospitalares, deverá usar técnicas de Processamento de Linguagem Natural (PLN) para extrair informação de um amplo conjunto de textos médicos disponibilizados pela RTS, particularmente cartas de alta e textos contendo directivas médicas. Esta informação, bem como a proveniente de recursos externos como ontologias e outras fontes de conhecimento médico, deverá ser utilizada no suporte e validação de decisões médicas.

Deste modo, tornou-se essencial o desenvolvimento de uma ferramenta capaz de extrair informação de uma forma automática a partir de texto. Na inventariação das ferramentas existentes e mais recentes na área (Oksefjell e Santos, 1998), nomeadamente das ferramentas desenvolvidas para o português, a Linguateca teve um papel preponderante.

Embora a delimitação de área, neste caso a medicina, imponha a necessidade de usar ferramentas direccionadas e o desenvolvimento de módulos específicos que identifiquem a informação relativa à aplicação em particular, o processamento inicial do texto pode ser feito com recurso a ferramentas de análise morfológica e sintáctica do português, como as que já se encontram disponibilizadas e listadas pela Linguateca. Também os manuais e a diversa literatura apresentada pela Linguateca contribuíram para uma aprendizagem mais diversificada e célere.

No entanto, a escolha das ferramentas a usar no desenvolvimento de tal sistema não recaiu sobre os recursos disponibilizados pela Linguateca, mas sim na utilização e adaptação das componentes existentes em plataformas de processamento de informação não estruturada, como por exemplo o GATE (Cunningham et al., 2002a) ou UIMA (Ferrucci e Lally, 2004). Estas plataformas permitem uma adaptação a diferentes sistemas operativos e disponibilizam as várias componentes de um sistema de processamento de linguagem natural em ambientes de desenvolvimento gráfico, facilitando a aprendizagem e a adaptação a diferentes línguas e domínios. Deste modo, considerou-se mais vantajoso e potencialmente mais rápida a utilização e adaptação de um ambiente deste tipo, do que a criação de algo de raiz.

No caso do MedAlert, começou por se desenvolver um sistema tendo por base a plataforma GATE. O GATE é uma infra-estrutura para o desenvolvimento de componentes de software, que processam linguagem natural, em desenvolvimento na Universidade de Sheffield desde 1995 e utilizado numa grande variedade de projectos. A arquitectura consiste em vários recursos de processamento independentes do domínio e aplicáveis a várias línguas, como o Atomizador e o Separador de Frases. No entanto, o processamento principal, em particular o Reconhecimento de Entidades Mencionadas, foi efectuado com recurso a almanaques e um conjunto de regras gramaticais desenvolvidas em JAPE (Java Annotations Pattern Language) (Cunningham et al., 2002b) que consideram conteúdos específicos da língua e do domínio, neste caso a medicina.

Recentemente, foi realizada uma experiência na área da vacinação com o objectivo de extrair informação do Plano Nacional de Vacinação (PNV) (DGS). A informação considerada relevante foi extraída e associada às entidades ACRONIMO, IDADE, PARTE_CORPO, DOENCA, DOSE, INTERACCAO, REACCAO e PESO de acordo com o conteúdo expresso. A figura 7.1 apresenta a interface gráfica do GATE com um excerto do PNV anotado com várias entidades e a tabela 7.1 os resultados obtidos, em termos de precisão e abrangência.

O desenvolvimento de uma arquitectura semelhante às referidas, que integre alguns dos recursos já existentes para o português, ou a produção de recursos tendo por base ambientes já existentes e que permitam a adaptação a diferentes técnicas e áreas de uma forma rápida e facilmente adaptável, permitiria estruturar os recursos actuais de uma forma mais útil. Estas arquitecturas permitem a utilização da tecnologia não só por

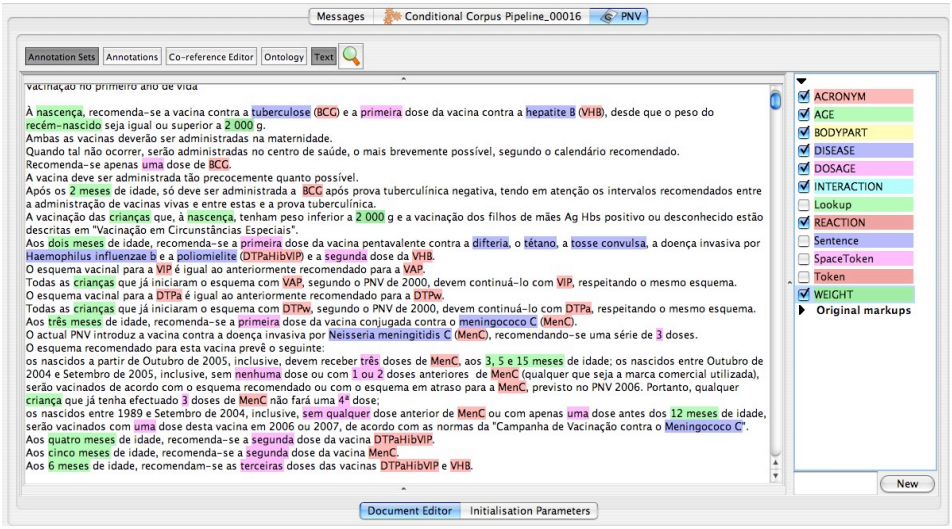

Figura 7.1: Interface gráfica do GATE com um excerto do Plano Nacional de Vacinação anotado.

	DOENCA	ACRONIMO	IDADE	PARTE_CORPO	DOSE	INTERACCAO	REACCAO	PESO
Saídas correctas	225	294	181	14	90	10	156	8
Parcialmente correctas	0	0	1	0	3	0	0	0
Em falta	0	0	6	0	0	0	0	0
Total	**225**	**294**	**188**	**14**	**93**	**10**	**156**	**8**
Abrangência	1,00	1,00	0,96	1,00	0,97	1,00	1,00	1,00
Precisão	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Medida F	**1,00**	**1,00**	**0,98**	**1,00**	**0,98**	**1,00**	**1,00**	**1,00**

Tabela 7.1: Resultados da tarefa de Extracção de Informação para cada uma das entidades definidas.

profissionais da área, mas também por grupos que procuram sistemas eficientes e prontos a usar, podendo contribuir para uma maior divulgação da área e da tecnologia. Um exemplo da utilidade destas arquitecturas constituídas por módulos é a sua aplicação no ensino, em particular no ensino de pós-graduação, pelo facto de permitir a produção de recursos independentes e facilmente integráveis e escaláveis. Embora a Linguateca disponibilize de uma forma pública os recursos criados, a sua utilização e/ou adaptação pressupõe muitas vezes a existência de informação e de conhecimento que os interessados podem não possuir, podendo tornar o processo moroso e complicado.

Ainda na área da vacinação, e no âmbito do MedAlert, foi recentemente desenvolvido trabalho no sentido de criar uma representação conceptual das directivas contidas no PNV.

Para tal, o PNV foi analisado manualmente e modelado segundo os conceitos e relacionamentos que descreve. Foi assim criada uma ontologia contendo todas as classes de vacinas, interacções e alergias descritas no PNV. A tarefa de popular a ontologia foi dividida em dois passos. Primeiro, informação automaticamente extraída do PNV foi adicionada à ontologia através da associação entre a classe e a entidade identificada. Posteriormente, adicionaram-se os relacionamentos entre as instâncias, usando uma abordagem baseada em Procura de Padrões Frequentes. Para tal, identificou-se no texto anotado padrões frequentes de entidades mencionadas, usando Procura de Regras de Associação (Agrawal e Srikant, 1994). Procurou-se identificar regras como, por exemplo, *doença* $\Rightarrow$ *vacina (80%)*, indicando, neste caso, que 4 em cada 5 vezes que uma doença é mencionada (ex. Tuberculose) é seguida pela referência a uma vacina (ex. BCG). Seguindo o exemplo, poder-se-ia concluir que a vacina mencionada, BCG, combate a doença Tuberculose, e automaticamente inferir e adicionar o correspondente triplo RDF (Lassila e Swick, 1998). Esta foi, no entanto, apenas uma primeira experiência na tentativa de automatizar o processo de adição de relacionamentos entre instâncias em ontologias. No seguimento deste trabalho de criação de ontologias para o português, o caminho usual da Linguateca de disponibilização e publicação do processo e ferramentas utilizadas, de que é exemplo o PAPEL (Oliveira et al., 2008),será certamente útil na continuação de criação das ontologias para a nossa área de aplicação.

7.2 Participação no Segundo HAREM

Recentemente, participámos pela primeira vez na avaliação conjunta de reconhecedores de entidades mencionadas organizada pela Linguateca, o Segundo HAREM (Mota e Santos, 2008). Este modelo de avaliação, em que vários grupos comparam o progresso dos seus sistemas usando uma métrica consensual(Santos, 2007), representou uma importante oportunidade para perceber quais os desafios inerentes ao reconhecimento de nomes próprios em textos não especializados na área da saúde e deste modo desenvolver diferentes técnicas de delimitação e classificação destas entidades. No caso do sistema desenvolvido em Aveiro, o desafio representou o desenvolvimento de um sistema capaz de recorrer a fontes de conhecimento externas, como a Wikipedia, de modo a melhorar a classificação e a diminuir a utilização de listas e almanaques. Este sistema, denominado REMMA – Reconhecimento de Entidades Mencionadas do MedAlert (Ferreira e Teixeira, 2008), foi desenvolvido tendo por base o sistema de processamento de linguagem não estruturada Apache UIMA. UIMA é uma plataforma para o desenvolvimento de sistemas de software capazes de analisar grandes volumes de informação não estruturada. O REMMA contém, entre outras, uma componente capaz de explorar a Wikipédia como fonte de conhecimento. A impossibilidade de construir ou aceder a um almanaque de grande dimensão e qualidade, motivou a decisão de extrair categorias e tipos através da análise da primeira frase do

artigo Wikipédia. Esta experiência, embora direccionada a textos não especializados, permitiu perceber a utilidade de tais abordagens e procurar recursos e soluções semelhantes para a área em que nos concentramos.

7.3 Conclusões e sugestões finais

Não sendo certamente possível, ou mesmo desejável, que a Linguateca desenvolva corpos e ferramentas para domínios específicos como o nosso, atrevemo-nos a sugerir que seja efectuada uma inventariação o mais completa possível de módulos e sistemas existentes e disponíveis para utilização por todos; concentração do esforço da Linguateca na criação de recursos e ferramentas ainda não disponíveis; e que haja um esforço de criação de um sistema integrado. Para facilitar a avaliação, para além dos eventos como o HAREM, seria importante a criação de directivas genéricas que facilitariam a avaliação comparativa em domínios mais específicos como o que nos interessa. Particularmente, a criação de directivas sobre construção de colecções douradas, de métricas de avaliação e possivelmente um sítio para disponibilização destes recursos possibilitaria o desenvolvimento por parte dos grupos interessados em acções de avaliação conjunta em áreas especificas, como a Medicina.

Capítulo 8

Criação e expansão de geo-ontologias, dimensionamento de informação geográfica e reconhecimento de locais e seus relacionamentos em textos

Marcirio Chaves

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 8, p. 49–59, 2008

Este artigo resume o trabalho desenvolvido ao longo de mais de quatro anos na Linguateca no âmbito do meu doutorado. Até 2004, a maior parte das fontes de dados geográficos de Portugal encontrava-se distribuída, desintegrada e desconexa. Essas fontes contêm informação complementar, heterogênea e semi-estruturada. Qualquer aplicação que necessitasse utilizá-las tinha que recorrer a diversos bancos de dados, estudar seus esquemas conceituais e traduzir a informação para um formato comum de representação, entre outras tarefas. Além disso, os dados armazenados em bancos de dados proprietários são invisíveis para aplicações da Web Semântica.

Nesse contexto havia a necessidade da criação de um modelo genérico suficiente para reunir informação geográfica de diversas fontes, de múltiplos domínios geográficos (e.g. administrativo e físico) e disponibilizá-la de forma integrada e em um formato legível por máquina. Assim, foi criada a GKB (*Geographic Knowledge Base*) (Chaves et al., 2005a,b), um sistema de gerenciamento de conhecimento geográfico, ilustrado na figura 8.1 e descrito na próxima seção.

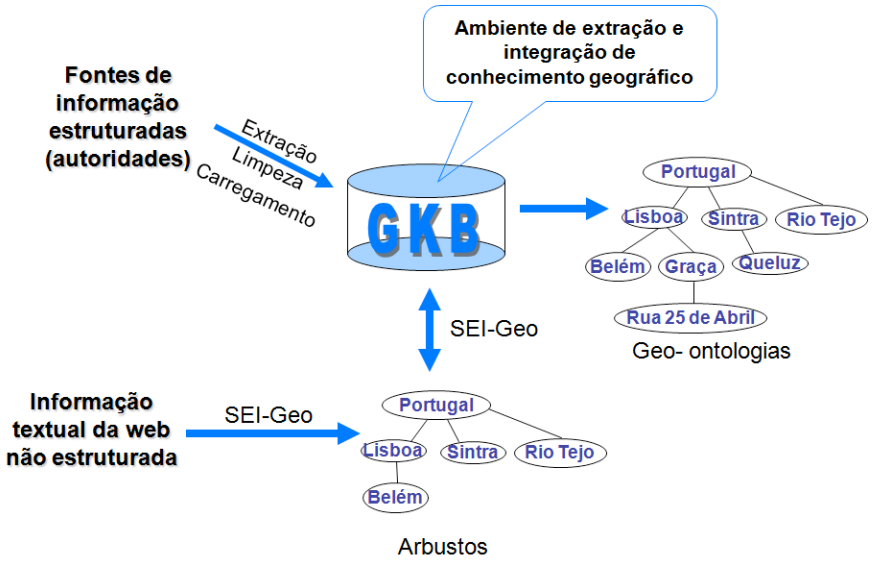

Figura 8.1: Arquitetura global do sistema de gerenciamento de conhecimento geográfico.

Este artigo está estruturado como segue: a Seção 8.1 apresenta a GKB. A Seção 8.2 descreve algumas das aplicações que utilizam as geo-ontologias geradas pela GKB. A Seção 8.3 introduz os resultados obtidos com experimentos para dimensionar a geograficidade[1] de textos em português. A Seção 8.4 descreve o sistema de extração, anotação e integração

[1] Por geograficidade entende-se a quantidade de informação geográfica presente em textos.

de conhecimento geográfico (SEI-Geo), e apresenta as avaliações realizadas com o SEI-Geo. A Seção 8.5 conclui o artigo.

8.1 Geographic Knowledge Base - GKB

A GKB é um dos componentes desenvolvidos no Pólo XLDB da Linguateca (`xldb.di.fc.ul.pt`) em colaboração com o projeto *Geographic Reasoning for Search Engines* (GREASE) (`xldb.di.fc.ul.pt/wiki/grease`), o qual pesquisa métodos, algoritmos e arquiteturas de software para atribuir âmbitos geográficos para recursos da rede e para recolher documentos usando entidades geográficas.

A GKB é um ambiente de extração e integração de conhecimento geográfico que contém informações provenientes de fontes de dados administrativas semi-estruturadas de autoridades junto com um conjunto de regras para integração de informação. A expansão do conhecimento contido na GKB ocorre com informação proveniente de textos. Esses textos são a entrada de informação para o Sistema de Extração, Anotação e Integração de Conhecimento Geográfico (SEI-Geo), que é o responsável por gerar uma representação estruturada do conhecimento geográfico extraído e integrá-lo no repositório da GKB.

A GKB suporta a definição de relacionamentos ontológicos entre entidades, tais como meronímia, sinonímia e adjacência, entre outros. A GKB também suporta relacionamentos inter-domínios, os quais são associações entre entidades de domínios diferentes. Por exemplo, o âmbito geográfico[2] de uma entidade do domínio de rede é representado como um relacionamento entre um sítio da rede (entidade do domínio da Internet) e uma região geográfica (uma entidade do domínio geográfico).

A informação armazenada no repositório da GKB pode ser exportada com uma ferramenta nomeada GOG (*Geographic Ontology Generator*). A GOG permite selecionar partes da informação armazenada na GKB, uma vez que os repositórios da GKB têm, atualmente, cerca de meio milhão de entidades e o usuário raramente quer receber toda a informação. A GOG exporta a informação no formato OWL (`www.w3.org/TR/owl-features/`), uma representação que estende o RDF (`www.w3.org/TR/REC-rdf-syntax/`) e, consequentemente, é também um formato XML. A geo-ontologia completa de Portugal (Geo-Net-PT01) contém mais de 400.000 entidades e é um recurso público disponível em `xldb.fc.ul.pt/geonetpt/`.

[2] Nesse artigo, entende-se âmbito geográfico como a região geográfica, se ela existe, onde a média das pessoas pensa ser mais relevante para uma página, sítio ou domínio da rede. Por exemplo, o âmbito geográfico do sítio da Câmara de Lisboa (`www.cm-lisboa.pt`) é o concelho de Lisboa.

8.2 Aplicações que utilizam as geo-ontologias geradas a partir da GKB

As geo-ontologias exportadas pela GKB têm sido utilizadas por diversas aplicações que incluem: sistemas para reconhecimento de entidades mencionadas (REM), um classificador de documentos de acordo com seu âmbito geográfico, uma interface de recolha de informação para consultas geográficas e uma interface XML para consultas a almanaques geo-temporais, entre outras.

8.2.1 Sistemas de REM

CaGE: é um sistema de REM e de atribuição de âmbito geográfico a páginas da rede (Silva et al., 2006; Martins et al., 2007b). O CaGE utiliza as geo-ontologias geradas a partir da GKB nas fases de identificação e desambiguação de locais (Cardoso et al., 2006b). Martins et al. (2007b) apresentam a arquitetura do CaGE, bem como a descrição detalhada do uso das geo-ontologias.

Faísca: é um sistema de reconhecimento de locais que faz uso dos conceitos e ocorrências contidos nas geo-ontologias geradas a partir da GKB (Cardoso et al., 2008a). O Faísca não explora os relacionamentos existentes entre conceitos nas ontologias, mas utiliza os conceitos para desambiguar nomes de locais.

8.2.2 Módulos de um sistema de recolha de informação geográfica

As geo-ontologias geradas a partir da GKB têm sido utilizadas por diversos módulos do sistema de recolha de informação geográfica da Universidade de Lisboa no GeoCLEF 2007 (Cardoso et al., 2008a).

QueOnde: é um módulo que utiliza as geo-ontologias para dividir o tópico de uma consulta em três partes: 'O que', 'Relacionamento espacial' e 'Onde'. Por exemplo, para o tópico 'tráfego marítimo nas ilhas portuguesas', QueOnde consulta a geo-ontologia e verifica que 'portuguesas' é um adjetivo relativo a Portugal e que 'ilhas' é um conceito geográfico.

QuerCol: é um módulo que utiliza a Geo-Net-PT01 para fazer expansão de consulta. O QuerCol interpreta uma consulta como duas partes: 'O quê' e 'Onde'. A geo-ontologia é usada para expandir o(s) termo(s) da parte 'Onde'. Por exemplo, na consulta 'regiões vinícolas em Portugal', o módulo QuerCol expande o nome Portugal para todas as províncias, distritos, concelhos e freguesias existentes na Geo-Net-PT01 e que fazem parte de Portugal.

Outro módulo do sistema que utilizou as geo-ontologias geográficas é o sistema de reconhecimento de locais Faísca, descrito na seção anterior.

Figura 8.2: Exemplo de interface para recolha de informação geográfica usando a GKB.

8.2.3 Interface de Motor de Pesquisa Geográfica

A GKB é usada também na interface do protótipo Geotumba (`local.tumba.pt`), um sistema para recolha de informação geográfica (ver figura 8.2).

No campo `Local?` o usuário digita a região, a rua, o código postal ou outra entidade geográfica para reduzir o âmbito da consulta. Quando um nome geográfico ambíguo é detectado na consulta, Geotumba apresenta as possíveis alternativas para desambiguação da mesma. Por exemplo, o nome "rua Castelo Branco" ocorre em cinco concelhos diferentes na Geo-Net-PT01, os quais são apresentados no lado esquerdo da figura 8.2. Além da consulta por texto, o usuário pode utilizar os mapas para definir o âmbito da consulta.

8.2.4 Interface para consultas a almanaques geo-temporais

A Geo-Net-PT01 também é utilizada no projeto DIGMAP (*Discovering our Past World with Digitised Maps*, `www.digmap.eu`) (Borbinha et al.), especificamente em uma interface XML para consultas a almanaques geo-temporais. Neste serviço, a Geo-Net-PT01 é integrada com outros almanaques existentes considerando a dimensão temporal juntamente com o conteúdo geográfico dos almanaques. A figura 8.3 apresenta a interface do sistema.

Para cada local inserido pelo usuário, o sistema de consultas a almanaques geo-temporais percorre os almanaques e apresenta o nome do local juntamente com seus metadados, relacionamentos e população, entre outras informações subjacentes a cada almanaque. A informação geográfica é apresentada em diversas linguagens (e.g. XML, OWL e KML - *Keyhole Markup Language*, `www.opengeospatial.org/standards/kml/`), conforme o almanaque as disponibiliza.

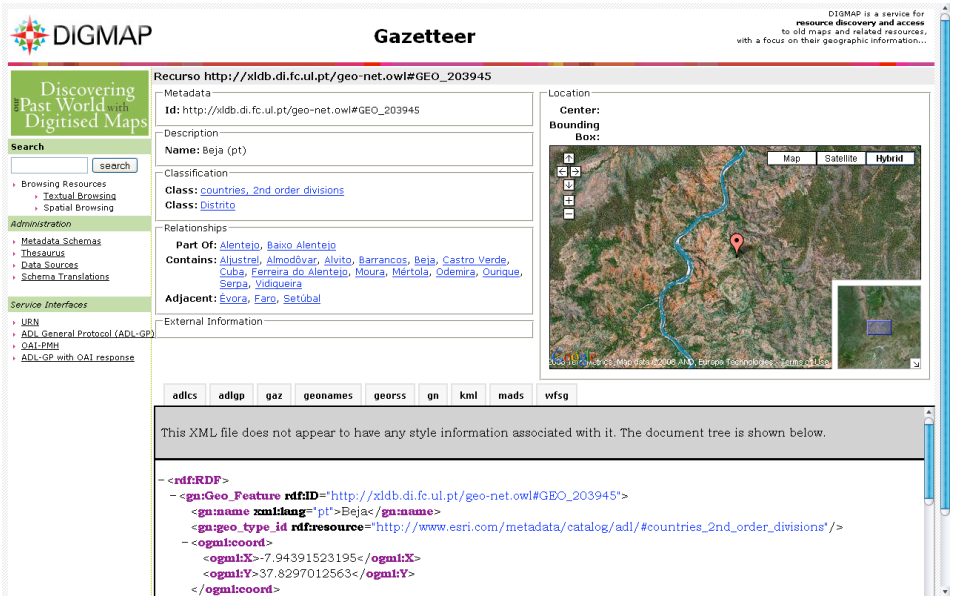

Figura 8.3: Interface para Consultas a Almanaques Geo-temporais.

No exemplo da figura 8.3, o sistema apresenta os metadados sobre o 'distrito de Beja', os quais incluem os relacionamentos de `parte-de`, `contém` e `adjacência`. Na parte inferior da figura, estão nove almanaques que contêm informação sobre o 'distrito de Beja'. No canto superior direito, o 'distrito de Beja' é ilustrado no mapa.

A Geo-Net-PT01 já foi requisitada por dezenas de investigadores, na sua maioria de Portugal e do Brasil, evidenciando o interesse da comunidade em estruturas de representação de conhecimento geográfico. A figura 8.4 apresenta a distribuição geográfica dos pedidos por países.

Por fim todo o conteúdo das geo-ontologias geradas pela GKB pode ser visualizado com a interface Geobase, apresentada na figura 8.5 (`www.tumba.pt/tumba/geobase`).

As aplicações que utilizam as geo-ontologias geradas pela GKB necessitam de informação geográfica além daquela proveniente de fontes de informação estruturadas e semi-estruturadas. Nomes históricos e alternativos de locais, por exemplo, ainda não estão na GKB, mas podem ser encontrados em textos. Programas foram implementados para dimensionar a geograficidade de textos em português e para conhecer a sobreposição da informação armazenada na GKB com a informação geográfica em textos.

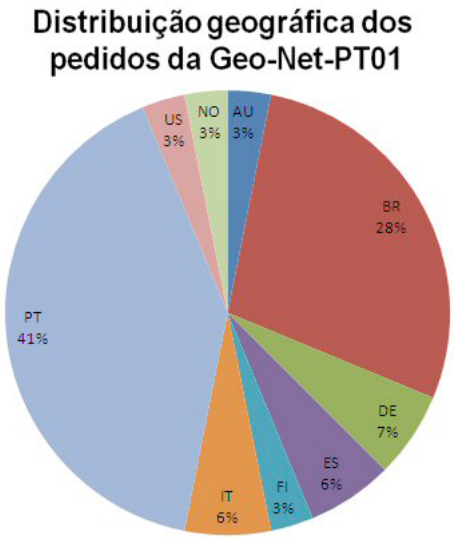

Figura 8.4: Distribuição geográfica dos pedidos da Geo-Net-PT01 por países.

Figura 8.5: Geobase: interface de visualização da Geo-Net-PT01.

8.3 Geograficidade de textos

Para verificar a geograficidade em textos da rede, foi utilizado o WPT 03, uma recolha da rede portuguesa de 2003, com 12 GB, 3.7 milhões de páginas e 1.6 bilhões de palavras (`www.linguateca.pt/WPT03`) (Cardoso et al., 2007). Aproximadamente 68.6% dessas páginas estão em português e mais de 1.5 milhões são distintas. O sistema de REM utilizado foi o SIEMÊS (Sarmento, 2006), que na avaliação conjunta do Primeiro HAREM (Santos et al., 2006) alcançou 70% de precisão e 75% de abrangência para a categoria local. Entretanto, a versão utilizada em nossos experimentos é uma versão com melhoramentos sobre aquela utilizada no Primeiro HAREM.

A caracterização da informação geográfica em textos passa pela verificação da presença de nomes de locais em nomes de pessoas e organizações. Considerando uma amostra aleatória de 32.000 documentos da rede, os resultados evidenciam que 31% das entidades mencionadas distintas da categoria pessoa e 23,43% das entidades mencionadas distintas da categoria organização contêm um nome geográfico incluído na Geo-Net-PT01.

Para investigar se o tipo de local ocorrendo em textos da rede portuguesa tinha diferentes propriedades (granularidade, geografia física (rios, montanhas, etc.)), foram verificados os tipos das entidades mencionadas da categoria local que o SIEMÊS encontrou após ser executado sobre a mesma amostra de 32.000 documentos. O resultado mostrou que 85% dos tipos de locais reconhecidos pelo SIEMÊS estão concentrados em apenas três (povoamento, endereço completo e sociedade/cultura) dos tipos de locais definidos no Primeiro HAREM. Estatísticas mais detalhadas sobre a geograficidade em textos podem ser consultadas em Chaves e Santos (2006).

Quanto aos tipos de arruamentos, os predominantes na geografia administrativa de Portugal são ruas e travessas. Somente ruas representam mais de 60% dos tipos de arruamentos do país. Rua também é o tipo de arruamento mais freqüente no WPT 03, após o tipo ambíguo acesso. Por outro lado, as travessas ocorrem com bem menos freqüência no WPT 03, sendo apenas o 28º tipo de arruamento mais freqüente.

Cera de 60% dos nomes[3] presentes na Geo-Net-PT01 estão presentes no WPT 03. Aqueles compostos por quatro palavras são os menos freqüentes, ao passo que os nomes formados por uma palavra atingem quase 80% de presença nesse corpus da rede. Outros resultados sobre a presença de informação geográfica de ontologias em textos e sobre a ambigüidade existente entre nomes de uma ontologia são descritos em (Santos e Chaves, 2006).

Após verificar que existe informação geográfica em textos suficiente para expandir geo-ontologias, foi desenvolvido o SEI-Geo.

[3] Neste contexto, todos os nomes na Geo-Net-PT01 exceto nomes de arruamentos e códigos-postais.

8.4 Sistema de Extração, Anotação e Integração de conhecimento Geográfico - SEI-Geo

O SEI-Geo foi desenvolvido no Pólo XLDB da Linguateca no âmbito do projeto GREASE e tem como objetivo reconhecer o conhecimento geográfico disponível em textos, gerar uma representação estruturada desse conhecimento e integrá-lo em geo-ontologias. O sistema é composto por dois módulos principais: o de Extração de Informação Geográfica (EIG) e o de Integração de Conhecimento Geográfico (ICG).

O EIG recebe como entrada um conjunto de textos que são segmentados em frases. O EIG contém uma quantidade abrangente de regras que indicam a presença de conceitos e relacionamentos nas frases. Tais frases, juntamente com conceitos de geo-ontologias, são a entrada de uma função que extrai frases com potencial conteúdo geográfico. Essas frases são a entrada de dois sub-módulos: o extrator de arbustos[4] e o anotador. O extrator de arbustos detecta ocorrências geográficas e relacionamentos semânticos e tem uma função de filtro, na qual o conteúdo geográfico, duplicado ou sobreposto, é eliminado. O resultado desse processo é um conjunto de arbustos que são utilizados como entrada no ICG. O anotador insere etiquetas com nomes de categoria semântica, tipo e subtipo. O anotador também possui a capacidade de reconhecer e anotar relacionamentos entre locais.

O ICG recebe os arbustos extraídos e anotados, e o conhecimento armazenado na GKB, faz a integração e retorna para a GKB o conhecimento geográfico expandido. A integração de conhecimento geográfico em geo-ontologias concentra-se em encontrar informação geográfica complementar àquela existente nas geo-ontologias e integrar essa informação no nível de granularidade mais adequado nas geo-ontologias. A integração de conhecimento geográfico com o SEI-Geo ocorre quando novos fatos geográficos são descobertos em texto.

8.4.1 Avaliação do SEI-Geo

O SEI-Geo tem sido avaliado na sua capacidade de extrair, anotar e integrar conhecimento geográfico. O SEI-Geo participou no Segundo HAREM (Mota e Santos, 2008) e conseguiu atingir resultados satisfatórios no cenário seletivo de identificação e classificação de locais. Considerando somente a medida F, o SEI-Geo foi o segundo melhor sistema nesse cenário com 0,5953, enquanto o melhor sistema atingiu 0,6246 na tarefa de classificação semântica. A participação do SEI-Geo no Segundo HAREM é descrita em Chaves (2008).

Além da tarefa de anotação de textos, o SEI-Geo foi avaliado, através de testes de mutilação, na sua capacidade de extrair locais e recompor uma geo-ontologia existente. Testes de mutilação consistem na destruição de parte de um objeto de estudo e na sua reconstrução. Especificamente quando se trata de estruturas de representação de conhecimento

[4] Um *arbusto* é composto por pelo menos duas entidades geográficas candidatas a locais e um relacionamento. Esse conjunto forma uma tripla. Não há número máximo de entidades e relacionamentos pré-definido.

	Público 1994	Público 1995	FSP 1994	FSP 1995
SEI-Geo reconstruido	148 (70,47%)	161 (76,30%)	117 (62,56%)	109 (60,55%)
ISO-3166-1 na coleção	210	211	187	180

Tabela 8.1: Resultado do teste de mutilação para países e territórios nos corpora jornalísticos.

tal como ontologias, um (ou vários) nível da hierarquia de conceitos e ocorrências é destruído e a partir de informação textual tenta-se reconstruir a informação retirada inicialmente. Para implementar esse teste, foram retiradas todas as ocorrências do tipo de entidade `ISO-3166-1` (que corresponde a países e territórios) da ontologia WGO (Martins et al., 2007a). Todos os arbustos extraídos pelo SEI-Geo que contêm o tipo de entidade `ISO-3166-1` foram enviados à geo-ontologia com o objetivo de encontrar um identificador para cada entidade geográfica reconhecida. A tabela 8.1 apresenta os resultados dos testes de mutilação usando o corpus CHAVE (Santos e Rocha, 2005).

Conforme a tabela 8.1, o Público é uma fonte mais rica em informação geográfica ao nível de países e territórios do que o Folha de São Paulo. Dos 211 países e territórios existentes na parte do Público relativa ao ano de 1995, 161 (70,30%) foram reconhecidos e representados em triplas no formato de arbusto. Das 238 ocorrências do tipo de entidade `ISO-3166-1` da WGO, 211 ocorrem nesse corpus. Um dos fatores que levam o Público a conter mais locais da WGO é os nomes de locais estarem na sua maioria descritos no português de Portugal. Exemplos desses casos encontrados no Público e ausentes no Folha de São Paulo são: 'Coreia do Sul', 'Eslovénia' e 'Ilhas Caimão'. Os resultados dos testes de mutilação indicam que o SEI-Geo é capaz reconstituir uma geo-ontologia recebendo como entrada conceitos sem ocorrências.

Quanto à expansão de geo-ontologias, o SEI-Geo recebe como entrada um corpus e geo-ontologias e devolve como resultado um conjunto de arbustos com as geo-ontologias enriquecidas com novos locais e relacionamentos reconhecidos no corpus. Se o SEI-Geo encontra uma ocorrência de um conceito e essa ocorrência já está na geo-ontologia, o resultado permite validar a ocorrência e a geo-ontologia não é expandida.

A primeira avaliação foi realizada com a parte do corpus CHAVE relativa ao ano de 1995. De um total de 50.495 arbustos, foi selecionada aleatoriamente uma amostra de 100 arbustos compostos por 143 triplas. Cada tripla dessa amostra foi avaliada manualmente de acordo com os seguintes critérios:

Integrável (I): quando as duas entidades geográficas da tripla forem realmente locais e a relação entre elas estiver correta.

Integrável com Assistência (IA): quando duas entidades geográficas forem corretas e não existir relacionamento explícito no texto ou o algoritmo não conseguiu identificar. Nesse caso o avaliador deve inserir o relacionamento correto.

Existente (E): quando as entidades geográficas e o relacionamento reconhecido entre essas entidades geográficas já está em pelo menos uma das ontologias.

Falso (F): quando no máximo uma entidade geográfica da tripla é um local ou as duas entidades geográficas não possuem relacionamento no mundo real.

Após a avaliação das 143 triplas, eu encontrei 2 I, 61 IA, 19 E e 61 F. Esses resultados indicam que a maior parte das triplas integráveis são integráveis com assistência. Ainda resta um número elevado de triplas falsas, mas esses valores já eram esperados dados os resultados da participação do SEI-Geo no Segundo HAREM.

8.5 Considerações Finais

Este artigo resumiu meu trabalho no âmbito da Linguateca ao longo dos últimos anos. A base de conhecimento geográfico armazena o conteúdo exportado como geo-ontologias que estão disponíveis publicamente. Esse conteúdo geográfico é expandido com informação textual extraída pelo SEI-Geo. O SEI-Geo foi avaliado no Segundo HAREM no que diz respeito à sua capacidade de anotação de locais e também apresentou resultados encorajadores nos testes de mutilação e expansão de geo-ontologias.

Após as geo-ontologias terem sido utilizadas por várias aplicações, torna-se essencial a criação de uma geo-ontologia mundial com nomes de locais em português, abrangendo as variantes da língua de Portugal e do Brasil. Essa nova geo-ontologia pode ser criada reutilizando o modelo-base no qual a GKB foi concebida.

Capítulo 9

Relato sobre a parceria Linguateca-NILC

Maria das Graças Volpe Nunes

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 9, p. 61–63, 2008

O NILC (Núcleo Interinstitucional de Lingüística Computacional) tem usufruído de recursos e iniciativas da Linguateca desde que tivemos contato com seus integrantes pela primeira vez, durante o IV PROPOR (Rodrigues e Quaresma, 1999).

Já no primeiro contato, ficou evidente o quanto toda a comunidade de Processamento da Língua Portuguesa ganharia com a parceria entre os grupos de pesquisadores de Portugal (e Europa) e do Brasil. Independentemente de quanto cada parte poderia, de fato, se beneficiar dos recursos gerados e distribuídos, o que tem mantido firme essa parceria é a crença na importância do trabalho comum realizado e, principalmente, da sua divulgação e ampla disponibilização.

O corpus CETENFolha (`www.linguateca.pt/CETENFolha`), com textos do jornal Folha de São Paulo, cedido pelo NILC, foi o primeiro e, talvez, o mais útil recurso compartilhado com a Linguateca. São inúmeros os pesquisadores brasileiros que se beneficiaram desse corpus, depois de ele ter sido processado, documentado e se tornado facilmente acessível pela Linguateca. Outros recursos, como a Floresta Sintá(c)tica, o REPENTINO (Sarmento et al., 2006), entre outros, têm possibilitado diversas pesquisas no NILC. De outro lado, o NILC tem disponibilizado, via Linguateca, vários recursos ali desenvolvidos: o corpus NILC (NILC/São Carlos através do AC/DC (Santos e Sarmento, 2002), corpus AmostRA (também através do AC/DC) e TeMário (através do Repositório em `www.linguateca.pt/Repositorio`). Além disso a Linguateca através do seu catálogo de recursos aponta para muitos mais desenvolvidos o NILC, tal como MacMorpho ou o sumarizador, assim como lista o material didáctico produzido pelo NILC no seu catálogo de publicações.

Nas atividades de avaliação, o NILC não tem podido participar num mesmo nível que os demais participantes, uma vez que a natureza dos recursos sob avaliação não coincidiu ainda com aquelas dos sistemas que desenvolvemos. Mas a parceria tem se mantido em outros níveis, sempre cercada pelo respeito mútuo ao trabalho desenvolvido.

O que cerca o planejamento, o desenvolvimento e a criação de recursos ou aplicativos, nesta e em outras áreas, muitas vezes está fora de nosso controle, o que acaba dificultando o compartilhamento que todos almejamos. A Linguateca tem servido como um alerta para que o desenvolvimento visando o uso comum e a qualidade atestada passem a fazer parte do trabalho do pesquisador em PLN.

A importância do processamento lingüístico na atual sociedade informatizada nos impõe uma grande responsabilidade. Alçar o português ao lugar que lhe cabe na Sociedade da Informação tem sido mais do que uma opção de pesquisa; passou a ser uma missão. Se, como cientistas, devemos investigar a melhor e mais eficiente solução para os problemas de processamento da língua, como peças de uma grande engrenagem nos cabe responder rapidamente às demandas geradas pelas mudanças tecnológicas e, principalmente, à variedade cada vez mais desafiadora da demanda gerada por novos e diferentes usuários.

Nesse cenário, são de singular importância o compartilhamento de recursos para rápidos novos desenvolvimentos, e a existência de mecanismos de controle de qualidade.

Enquanto que a decisão de compartilhar recursos depende principalmente da disposição de quem os gera, a elaboração de um sistema de avaliação é bastante mais complexa.

Promover a avaliação de recursos, nem sempre construídos sob os mesmos parâmetros e objetivos, requer uma organização cuidadosa e demorada. Uma diferença relevante entre as avaliações conduzidas pela Linguateca e as congêneres ligadas a eventos internacionais é o número de participantes. Como a comunidade de PLN/português é relativamente pequena, é compreensível que alguns dos recursos avaliados não sejam produzidos por todos os grupos. Aliás, temos aí uma contradição involuntária, já que compartilhar recursos acaba por implicar sua criação distribuída entre os grupos. Nesse sentido, seria igualmente importante a existência de um mecanismo de prospecção de recursos necessários para o português, e conseqüente incentivo ao seu desenvolvimento por um ou mais grupos sabidamente preparados para o desafio. A Linguateca tem todas as condições para servir a esse papel.

A importância do trabalho desenvolvido pela Linguateca nos seus 10 anos de existência pode ser avaliada sob diferentes aspectos e, em todos os casos, a conclusão será positiva.

Capítulo 10

Uma abordagem estatística para a identificação de colocações verbais usando o projeto AC/DC em www.linguateca.pt

Milena Uzeda Garrão e Maria Carmelita Padua Dias

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 10, p. 65–69, 2008

Tradicionalmente, pesquisas voltadas para o tratamento computacional de colocações vêm priorizando as combinações nominais ou os nomes compostos. Além de haver poucos estudos na área que se dediquem de forma sistemática às combinações verbais, existe um tipo em particular, o padrão V+SN, que se destaca das outras combinações verbais no português do Brasil (PB) e no europeu (PE) tanto pela sua freqüência quanto pelos seus alegados sub-padrões semânticos. O critério estatístico a partir de corpus adotado nesse projeto, como alternativa a uma abordagem baseada na intuição do pesquisador (Ranchhod, 2003), vem se mostrando altamente promissor no domínio das colocações do tipo V+SN (Uzeda Garrão, 2006; Uzeda Garrão e Dias, 2006).

Nossa metodologia, testada em padrões V+SN do PB pode ser resumida da seguinte forma: 1) uso de corpus etiquetado do PB como fonte de dados; 2) aplicação de um filtro para detecção de todos os padrões V+SN presentes no corpus; 3) aplicação de um teste estatístico ao filtro, chamado logaritmo de verossimilhança (Banerjee e Pedersen, 2003) para identificar as reais colocações (como "fazer parte", "tomar conta") em detrimento de combinações sintáticas casuais; 4) edição humana.

A justificativa para uma maior atenção descritiva voltada às colocações nominais em detrimento às colocações verbais se deve à importância do primeiro tipo de construção em textos de especialidade, auxiliando mais especificamente os domínios de Sumarização Automática e Recuperação de Informação. Acreditamos, contudo, que, embora se atribua à colocação verbal um papel secundário, ela vem a ser peça chave para o domínio de PLN. O padrão de combinação V+ SN ilustra claramente essa constatação, uma vez que inclui na sua estrutura um nome (SN), o que enfatiza a sua relevância também para domínios que priorizam o tratamento de colocações nominais.

10.1 Metodologia

10.1.1 O corpus utilizado: CETENFolha

O CETENFolha (Corpus de Extractos de Textos Electrónicos NILC/Folha de S. Paulo, `www.linguateca.pt/CETENFolha`) é um corpus jornalístico de cerca de 24 milhões de palavras em PB, parte integrante do corpus NILC (Pinheiro e Aluísio, 2003), ver também Aires e Aluísio (2001), que contém textos brasileiros do registro jornalístico, didático, epistolar e redações de alunos. Trata-se de uma parte de um corpus (Córpus NILC/São Carlos) com 37 milhões de palavras disponibilizado pelo projeto AC/DC (Santos e Sarmento, 2002).

De fato, a opção por um corpus de teor jornalístico tem suas implicações: a língua fica prioritariamente associada àquilo que é considerado notícia em detrimento, por exemplo, de uma conversa despretensiosa entre adolescentes. Entretanto, a escolha por esse tipo de extrato da língua também está associada à falta de um corpus mais robusto do PB. Uma outra razão da opção pelo corpus CETENFolha está no fato de, no ano de 2006, ser o único

significativo no PB disponível para *download*; e, portanto, o único passível de aplicação dos testes probabilísticos que virão mais adiante.

10.1.2 Aplicação do filtro para padrões V+SN aos verbos mais freqüentes

Com base nos 30 verbos com maior freqüência absoluta no corpus, partimos para uma restrição formal para obter os 10 verbos mais freqüentes seguidos facultativamente de determinante e obrigatoriamente de nome, formando a estrutura V+(det)+N. Tomando como exemplo o verbo *fazer*, o formalismo para tal detecção no AC/DC seria (**[lema="fazer"& pos="V"] [pos="DET.*"]? [pos="N"] [classe="JOCF"]**). A fórmula JOCF se refere à parte do corpus NILC/São Carlos que constitui o corpus CETENFolha. Obtivemos, finalmente, os 10 verbos mais freqüentes encabeçando uma estrutura V+SN. São eles: "fazer", "ter", "dar", "perder", "usar", "receber", "deixar", "tomar", "ganhar" e "criar".

Aplica-se então a todas as ocorrências desses 10 lemas no corpus, já baixado para a pesquisa, um filtro V+det+N, que foi viabilizado, nesse projeto, através de um programa, feito em linguagem Java, que recebe como entrada a ocorrência desses lemas no córpus e fornece como resultado a lista de todas as ocorrências de, por exemplo, fazer+(det)+ N (Nogueira, 2004). Somente na etapa seguinte é aplicado o teste estatístico e é estabelecida a lista das candidatas a colocações que, posteriormente, são ordenadas por freqüência.

10.1.3 A aplicação do logaritmo de verossimilhança aos padrões V+(det)+N encabeçados pelos verbos mais freqüentes no corpus

A aplicação do Logaritmo de Verossimilhança, foi disponibilizada através do pacote estatístico NSP (Banerjee e Pedersen, 2003). Após a sua aplicação é estabelecida a lista das candidatas a colocações que, posteriormente, são ordenadas por freqüência. Dentre os métodos de Testagem de Hipótese fornecidos em NSP, o Logaritmo de Verossimilhança tem por objetivo detectar se um bigrama é mais do que uma simples co-ocorrência casual na língua. Esse tipo de testagem requer a formulação de dois tipos de hipóteses formalizadas abaixo:

$$H_1 : P(w_1 \mid w_2) = P(w_1 \mid \neg w_2)$$
$$H_2 : P(w_1 \mid w_2) \neq P(w_1 \mid \neg w_2)$$

Onde H = hipótese, P = probabilidade, w = palavra

Por exemplo, assumindo que a expressão *fazer sucesso* seja uma colocação, espera-se que a hipótese de dependência H_2 : $P($ fazer $\mid$ sucesso $)$ $\neq$ $P($ fazer $\mid$ $\neg$ sucesso $)$ seja

verdadeira e que a hipótese de independência $H_1 : P(\text{ fazer} \mid \text{sucesso }) = P(\text{ fazer} \mid \neg \text{ sucesso })$ seja falsa. Portanto, o método avalia a probabilidade de H_2 ocorrer em detrimento de H_1.

10.1.4 Resultados e Edição Humana

Sob uma perspectiva quantitativa o método se revelou satisfatório. Em outras palavras, dentre as 1000 candidatas a colocações apontadas pelo método (100 de cada um dos verbos listados na seção 10.1.2, apenas 128 foram consideradas ruído. Um acerto de 87,2%. As "pseudo-colocações" extraídas pelo método, ou seja, os "deslizes" por ele cometido (12,8%), foram indicadas na listagem final da seguinte forma:

1. Erro de avaliação estrutural. Este tipo de erro pode ter sido cometido pelo método por duas razões principais: em função da etiquetagem equivocada no corpus; em função de o método ter considerado uma janela sintática menor do que a expressão representa (JAN): *ter um papel*, por exemplo, foi detectado pelo método como uma colocação do padrão procurado quando, na verdade, sua estrutura vai além de V+(det)+N.

2. Outros ruídos foram atribuídos exclusivamente ao corpus: colocações claramente datadas: como *criar a URV*, *usar a URV*, *tomar AZT*.

Há outros dois tipos de interferência na detecção de colocações que não foram considerados propriamente ruídos. São eles: recursos coesivos, como a utilização de anáfora: alguns exemplos são *fazer a denúncia*, *dar a notícia*, *ter a doença* (COE) e omissões de artigo (tanto definido quanto indefinido), características de manchetes de jornal, como Presidente da Shell ***deixa cargo*** amanhã.

A título de ilustração, a tabela 10.1 diz respeito às 100 colocações do tipo Fazer+(det)+SN mais freqüentes detectadas no corpus. Na verdade, a listagem segue até que se chegue a co-ocorrências menos freqüentes. Esse é apenas um pequeno extrato do que o teste foi capaz de gerar. Tomemos como exemplo o primeiro bigrama da lista, *fazer parte*. O número que segue à colocação (1) diz respeito à sua posição em relação às outras colocações. O segundo número (2805) se refere ao número de ocorrências no corpus.

10.2 Conclusões e trabalhos futuros

A grande vantagem deste método está no seu teor preditivo. Através dele, podemos constatar preferências de usos das expressões presentes no corpus. Portanto, o que consideramos especialmente relevante nesta abordagem com base em corpus, é que não fazemos conjecturas daquilo que ocorre e não ocorre em uma língua, pois uma perspectiva exclusivamente intuitiva pode ser muitas vezes contra-argumentada por dados reais da língua.

fazer parte,1,2805	fazer falta,17,124	fazer exames,33,86
fazer campanha,2,616	fazer acordo,18,123	fazer um discurso,34,83
fazer questão,3,485	fazer alguma coisa,19,112	fazer referência,35,82
fazer sucesso,4,289	fazer propaganda,20,112	fazer um teste,36,82
fazer compras,5,227	fazer o gol,21,107 (COE)	fazer uma avaliação,37,76
fazer papel,6,189 (JAN)	fazer greve,22,105	fazer um balanço,38,76
fazer sentido,7,177	fazer perguntas,23,104	fazer gols,39,75
fazer comício,8,176	fazer sua estréia,24,98	fazer o teste,40,74 (COE)
fazer um filme,9,151	fazer uso,25,98	fazer o pedido,41,74 (COE)
fazer um acordo,10,149	fazer filmes,26,95	fazer shows,42,74
fazer a conversão,11,143 (COE)	fazer coisas,27,95	fazer palestra,43,72
fazer um trabalho,12,142	fazer testes,28,95	fazer a ligação,44,71 (COE)
fazer mal,13,130 (JAN)	fazer uma campanha,29,92	fazer a festa,45,68
fazer sexo,14,130	fazer oposição,30,90	fazer as contas,46,68
fazer política,15,126	fazer exercícios,31,89	fazer uma pesquisa,47,67
fazer críticas,16,125	fazer um levantamento,32,88	fazer muito tempo,95,45 (JAN)

Tabela 10.1: Colocações Fazer+(det)+SN mais freqüentes no corpus.

Nosso olhar eminentemente empírico é capaz de detectar preferências de usos ao invés de intuir aquilo que pode ou não ocorrer em um corpus, com base em testes de aceitabilidade, comumente utilizados para identificar as colocações.

Em suma, acreditamos que esse trabalho tenha uma função prática e teórica para a lexicografia. Sua natureza genuinamente estatística viabiliza uma rápida construção de uma base de dados robusta das colocações verbais mais freqüentes através de evidência empírica. Uma das próximas etapas é utilizar uma ferramenta mais recente desenvolvida para identificação de colocações - Linguistics Tool (Caminada, 2008) - e comparar os resultados. Pretendemos também estender o método para detectar outros padrões de colocações freqüentes encabeçadas por verbo (ex.: V+Prep+N; V+N+prep+N) e contribuir de forma efetiva para a lexicografia computacional do PB e, futuramente, para a lexicografia do PE, trabalhando em conjunto com pesquisadores nativos da modalidade européia.

Capítulo 11

Novos rumos para a recuperação de informação geográfica em português

Nuno Cardoso

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 11, p. 71–85, 2008

A recuperação de informação (RI) tem sido uma área em franco crescimento nos últimos tempos, devido ao aumento exponencial de documentos e de serviços disponíveis através da Internet. As ferramentas de pesquisa de informação já fazem parte da nossa vida quotidiana, sendo usadas sobretudo para a procura de documentos concretos e de informação contida em documentos: motores de busca na rede, pesquisa de correio electrónico ou ferramentas de pesquisa de documentos no computador, todas estas aplicações têm como base os conceitos fundamentais de RI.

As ferramentas de RI baseiam-se na sua maioria em modelos estatísticos de termos, que estimam a relevância dos documentos para cada consulta de uma forma simples e funcional. Contudo, a incapacidade de interpretação do significado dos textos das consultas e dos documentos tem sido uma das principais limitações das ferramentas de RI, que encontram assim algumas dificuldades em encontrar documentos que satisfaçam algumas necessidades de informação mais elaboradas. Allan et al. (2003) prevêem a exaustão dos actuais modelos de RI num futuro próximo, e referem que as novas tendências de RI passarão por uma contribuição decisiva de outras áreas de investigação mais afectas ao processamento de linguagem natural, como é o exemplo da extracção de informação, sumarização de textos ou a resposta automática a perguntas, com o intuito de compreender os tópicos subjacentes às consultas do utilizador, e utilizar esse conhecimento no processo de recuperação de documentos.

Segundo Belkin (2008), os novos desafios em RI passam por dar uma maior atenção às necessidades de cada utilizador, personalizando os resultados de acordo com o seu perfil e o contexto da sua pesquisa. A pesquisa de informação deverá aplicar técnicas de tradução automática, de forma a incluir documentos escritos em várias línguas (RI multilingue) e fazer com que a língua não seja obstáculo para o acesso à informação desejada. O utilizador terá controle sobre o método de pesquisa, como por exemplo a ordenação dos resultados de acordo com uma determinada área geográfica de interesse (pesquisas com âmbito geográfico), ou a escolha do tipo de resposta pretendido (em forma de lista de documentos, resumos gerados automaticamente, ou somente a resposta exacta). Finalmente, os resultados deverão ser apresentados de acordo com o contexto da pesquisa, combinando documentos textuais, imagens, sons, vídeos ou mapas sempre que forem relevantes para ilustrar a informação pretendida.

Singhal (2008) resume esta nova fase da RI como uma mudança do ponto de vista do utilizador em relação à pesquisa de informação, onde este usa os sistemas de RI numa atitude de "dá-me o que eu quero" em vez de "dá-me o que eu disse". O futuro da investigação em RI passa inquestionavelmente pela compreensão das necessidades do utilizador e do contexto das suas pesquisas, na compreensão dos tópicos abordados nas suas línguas específicas, e no uso de novas aproximações semânticas na recuperação de documentos de forma a fornecer resultados que se adequem às características de cada pesquisa.

Neste artigo apresento a minha perspectiva sobre os novos rumos de recuperação de in-

formação, com base na investigação realizada até agora no âmbito do meu doutoramento. O meu trabalho foca a área de sistemas de recuperação de informação geográfica (RIG) para o português, nomeadamente os problemas da modelação do conhecimento geográfico, o tratamento dos textos em português para a extracção automática de pistas geográficas no texto, e a correcta interpretação e reformulação das consultas dos utilizadores com restrições geográficas. A secção 11.1 descreve a técnica de reformulação automática de consultas e a sua aplicação em RIG. A secção 11.2 caracteriza as fontes de informação que irei explorar para criar uma rede de conhecimento que permite dotar os diversos módulos desenvolvidos da informação necessária para raciocinar sobre o domínio geográfico. A secção 11.3 descreve o modelo RIG adoptado e detalha os respectivos módulos QuerCol, REMBRANDT, MG4J e RENOIR, e a secção 11.4 refere as participações em avaliações conjuntas internacionais realizadas até agora.

11.1 Compreendendo as consultas dos utilizadores

Os utilizadores interagem tipicamente com as ferramentas de RI com o intuito de realizar *pesquisas* e satisfazer uma determinada necessidade de informação. As pesquisas são compostas por uma ou mais *consultas*, ou seja, linhas de texto contendo normalmente termos-chave que procuram descrever a informação pretendida. Para cada consulta enviada, a ferramenta RI devolve uma lista de documentos ordenados de acordo com a sua pertinência em relação à consulta.

Muitas vezes o utilizador não consegue descrever convenientemente a sua necessidade de informação numa consulta. Nestes casos, ele opta por realizar consultas pequenas, cujos termos são vagos e/ou ambíguos, o que dificultará a tarefa do sistema de RI. Adicionalmente, o vocabulário usado pelo utilizador e pelos autores dos documentos para descrever os diversos assuntos pode ser diferente, existindo então uma barreira terminológica que evita que certos documentos relevantes sejam recuperados, só porque certos conceitos são descritos através de termos diferentes.

11.1.1 Reformulação automática de consultas

A reformulação automática de consultas (RAC) é uma técnica frequentemente usada para lidar com certas limitações dos modelos tradicionais de RI, nomeadamente a barreira terminológica referida anteriormente. A RAC procura reformular a consulta inicial de forma automática, adicionando termos fortemente relacionados com a pesquisa, removendo termos irrelevantes ou geradores de ruído, e atribuindo pesos de importância a cada termo (Efthimiadis, 1996). No final, a consulta reformulada deverá ser mais precisa e fiel à necessidade de informação real do utilizador, e mais robusta em relação às diferenças

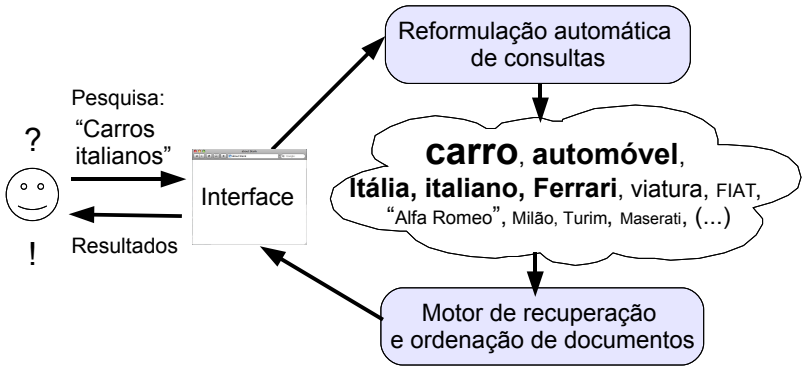

Figura 11.1: Esquema de funcionamento da reformulação automática de consultas (RAC).

de vocabulário patente entre documentos e consultas. A actuação da RAC está esquematizada na figura 11.1.

A aplicação de RAC nas pesquisas tem como objectivo representar melhor os conceitos chave através das suas várias formas textuais, algo também subjacente à filosofia das "folksonomias" (Mika, 2006, 2004), onde é normal associar uma nuvem de termos para catalogar um determinado documento, imagem ou vídeo. A nuvem de termos pode ser criada por diversos utilizadores que possuem diferentes perspectivas do documento em questão, e como tal, é frequente que as nuvens tenham bastantes termos, e inclusivé oriundos de diversas línguas.

11.1.2 Consultas de âmbito geográfico

Existe uma percentagem considerável de consultas realizadas a motores de busca que dizem respeito a determinados tópicos de interesse confinados a uma área geográfica específica (Kohler, 2003). As dificuldades nas pesquisas com âmbitos geográficos estão muitas vezes relacionadas com o facto de os nomes de locais usados serem ambíguos, e poderem designar várias entidades distintas, como é o exemplo de nomes de pessoas ("Camilo Castelo Branco") ou de nomes de empresas ("France Press"). Mesmo quando os nomes geográficos se referem a locais, podemos encontrar vários locais com o mesmo nome (por exemplo, "Cuba" refere-se a um país e a uma cidade de Portugal), ou até ser um nome usado de forma metonímica (por exemplo, usando "Bruxelas" para mencionar as instituições da União Europeia).

O objectivo da minha tese de doutoramento é a investigação de novos métodos de RAC aplicados à recuperação de informação em português com âmbito geográfico, de forma a desambiguar o significado real dos nomes geográficos nas consultas e realizar a

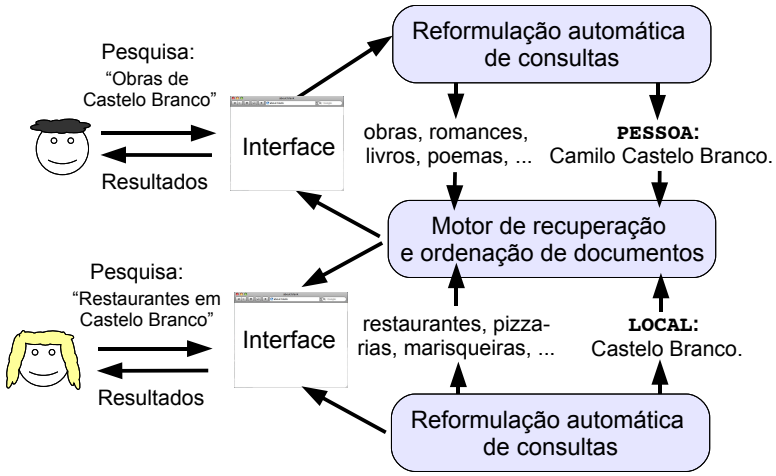

Figura 11.2: Reformulação automática de consultas para pesquisas diferentes.

reformulação de acordo com a verdadeira intenção do utilizador, fornecendo resultados de acordo com a sua área geográfica de interesse.

Um exemplo prático da aplicação do trabalho da minha tese está ilustrado na figura 11.2, onde podemos observar dois utilizadores com necessidades de informação diferentes, que formularam duas consultas diferentes nas suas pesquisas, "Obras de Castelo Branco" e "Restaurantes em Castelo Branco". Assumindo que o primeiro utilizador está interessado nas obras literárias do romancista português, e o segundo em restaurantes na cidade portuguesa[1], cabe ao sistema RIG interpretar correctamente a intenção subjacente nas duas pesquisas, e interpretar correctamente o significado de "Castelo Branco" em cada uma das consultas. O módulo de RAC deverá reajustar o seu mecanismo de reformulação de maneira a gerar consultas mais fiéis sobre a verdadeira semântica da pesquisa, em especial a consulta com âmbito geográfico na cidade de Castelo Branco. Desta forma, a recuperação de documentos terá atenção às diferenças semânticas entre as duas pesquisas, fornecendo os resultados mais relevantes para cada um dos utilizadores.

11.2 Rede de conhecimento

No contexto do meu trabalho, estou a investigar novas formas de realizar a RAC em português, aproveitando o conhecimento da língua e do significado dos termos para melhor entender as consultas. Para tal, estou a construir uma *rede de conhecimento* em português, com o objectivo de fornecer a informação necessária para que a RAC interprete convenien-

[1] Para efeitos deste exemplo, vamos considerar que estas são as reais intenções dos utilizadores, e que não estão nem interessados em obras artísticas sobre a cidade, nem sobre restaurantes relacionados de alguma forma com o romancista.

temente os conceitos envolvidos na consulta, raciocine sobre a melhor estratégia a aplicar na consulta, e obtenha consequentemente novos termos relevantes. Defino a rede de conhecimento como sendo uma rede semântica composta por diversas fontes de informação de onde é possível extrair conhecimento de uma forma objectiva e automática.

11.2.1 Fontes de informação

No âmbito do trabalho do doutoramento, estou a explorar quatro fontes de informação particularmente relevantes para a extracção de conhecimento geográfico.

i. Ontologias geográficas

A Geo-Net-PT01 é uma ontologia geográfica detalhada sobre o território português, e é usada como fonte de informação primordial para operações básicas de raciocínio geográfico (Chaves et al., 2005b). As ontologias geográficas representam o conhecimento humano sobre o domínio geográfico de uma forma hierárquica e inteligível, permitindo o acesso a conhecimento geográfico complexo, como por exemplo saber que cidades estão contidas numa região, ou quais os países atravessados por um determinado rio.

ii.Recolhas da rede

A WPT 03 é uma recolha da rede portuguesa realizada em 2003, e permite extrair informação sobre os sítios, os URL, os títulos e os resumos mais relevantes para as pesquisas realizadas pela comunidade portuguesa (Cardoso et al., 2007). Esta informação pode ser usada, por exemplo, para gerar um grafo da rede Arasu et al. (2001) e estimar a importância de cada sítio na rede, de forma a determinar se a consulta é do tipo transaccional, navegacional ou informativo (Broder, 2002), para auxiliar na detecção de consultas de cariz geográfica, ou para determinar se a consulta é vaga ou precisa,

. A caracterização das consultas é um passo importante para que seja possível ajustar a acção do módulo de RAC à pesquisa concreta, tal como evidencia Aires no seu trabalho sobre a classificação dos resultados de busca na rede portuguesa (Aires, 2005).

iii. Wikipédia

A porção portuguesa da Wikipédia, que conta em 2008 com mais de 400.000 artigos, é usada como fonte de conhecimento sobre diversos tópicos de interesse, auxiliando a interpretação das consultas dos utilizadores portugueses. Esta enciclopédia electrónica é uma referência incontornável na Internet, reunindo descrições detalhadas e bem documentadas sobre um grande número de tópicos, beneficiando das contribuições e validações de muitos utilizadores de modo a garantir a fidelidade e a organização da informação a um nível sem precedentes. As páginas da Wikipédia referentes a locais (como por exemplo rios,

países ou cidades), normalmente possuem informação adicional sobre as propriedades do local numa caixa de informação (*infobox*), como por exemplo as áreas, populações ou coordenadas respectivas, podendo ser aproveitadas para extrair conhecimento geográfico adicional para o módulo de RAC.

iv. Diários dos servidores de motores de busca

Os diários dos servidores do motor de busca *tumba!* registam as interacções entre os utilizadores e o *tumba!* (Silva, 2003). Estes diários permitem determinar as necessidades de informação mais típicas do utilizador, analisar o tipo de consultas formuladas, estudar quais as páginas visitadas ao longo da pesquisa, e analisar as estratégias de reformulação manual das consultas, até o utilizador ficar satisfeito com a pesquisa, ou desistir sem conseguir obter a informação pretendida. Os diários podem ser explorados de maneira a encontrar termos importantes a serem adicionados na RAC, ao identificar necessidades de informação semelhantes mas com consultas diferentes, ou até inferir certos focos de interesse sobre determinados tópicos a partir de determinados locais (por exemplo, pesquisas sobre o surto de determinada doença podem ser originadas a partir de um determinado local), e estudar o padrão de visualização de documentos para analisar a importância desses documentos para a respectiva área geográfica dos utilizadores.

A figura 11.3 ilustra uma forma de aplicar a rede de conhecimento formada com base nas fontes de informação apresentadas, para extrair mais conhecimento sobre o conceito "Lisboa". Um grafo da WPT 03 fornece uma lista de sítios mais relevantes sobre Lisboa, e em conjunto com os diários de registos, podem fornecer um conjunto de termos normalmente correlacionados com "Lisboa", do ponto de vista dos utilizadores do *tumba!*. A Wikipédia pode fornecer informação importante sobre a cidade, e juntamente com a ontologia geográfica, é possível determinar a semelhança de Lisboa com outras entidades geográficas (tais como freguesias, monumentos ou aeroportos), e usar essa informação para o cálculo da relevância geográfica.

11.2.2 Características das fontes de informação

A tabela 11.1 resume as características de cada uma das fontes de informação mencionadas, e refere as suas principais contribuições para a rede de conhecimento.

O acesso aos conteúdos da Wikipédia em formato compactado é livre, enquanto que o acesso a recolhas da rede é mais restritivo para fins não-académicos. O público geral normalmente não tem acesso aos diários dos servidores, por causa dos problemas relacionados com a privacidade dos utilizadores do motor de busca. Contudo, para este trabalho de investigação, é possível usar os diários dos servidores do motor de busca *tumba!*.

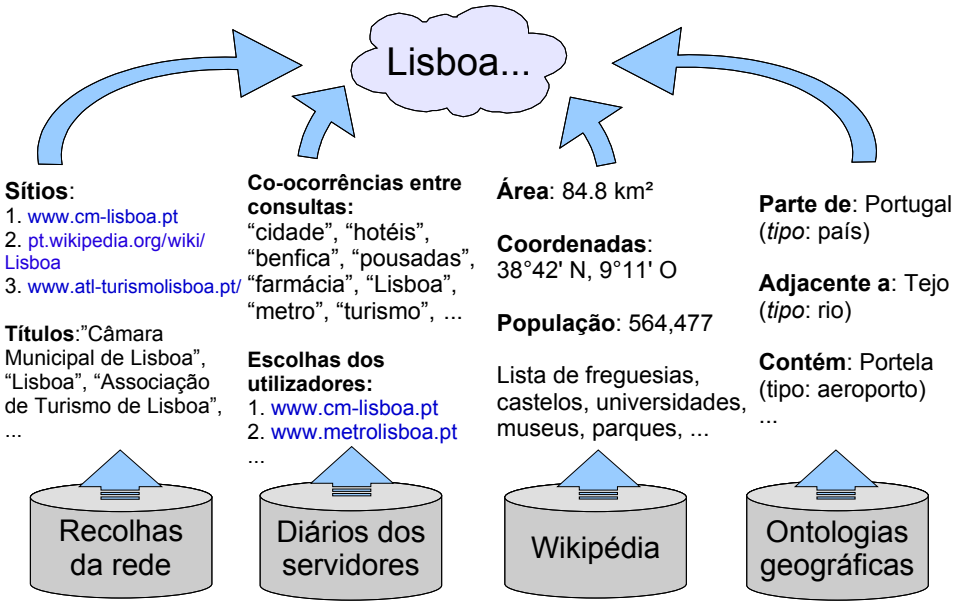

Figura 11.3: Uso da rede de conhecimento sobre o conceito "Lisboa".

No caso das ontologias geográficas, a Geo-Net-PT01 está disponível gratuitamente em `xldb.di.fc.ul.pt/geonetpt`.

A informação contida nas ontologias possui altos níveis de credibilidade, uma vez que estas são cuidadosamente revistas e validadas. A Wikipédia e a sua vasta comunidade que actualiza e verifica os seus conteúdos fazem com que seja um recurso com elevada credibilidade na sua informação. A rede, por sua vez, não possui restrições quanto à informação publicada, pelo que a sua credibilidade normalmente é estimada indirectamente através do sítio onde foi publicado, por exemplo.

As ontologias são a escolha típica para a representação fidedigna de um determinado domínio, e como tal, estão confinadas ao domínio ao qual foram projectadas. A rede e os diários dos servidores são o oposto, incluindo uma grande variedade de assuntos. A Wikipédia representa um meio termo interessante, permitindo uma organização hierárquica dos assuntos através de um leque de categorias, restringindo apenas a diversidade de assuntos com base numa política de relevância para os propósitos de uma enciclopédia da rede (ver `en.wikipedia.org/wiki/Wikipedia:List_of_policies`).

Em relação à inteligibilidade de formatos, as ontologias são o recurso mais fácil de ser usado pelos sistemas, uma vez que já vêem num formato estruturado, próprio para processamento computacional (normalmente o formato OWL/RDF). A estrutura da Wikipédia também é bastante amigável para ser analisada automaticamente, enquanto que a rede coloca bastantes desafios quanto à sua limpeza de dados. Os diários dos servidores, apesar

	Ontologias geográficas	Recolhas da rede	Wikipédia	Diários dos servidores
Acessibilidade	++	++	++	++
Credibilidade da informação	++	-	+	-
Diversidade de assuntos	-	++	+	+
Especificidade do domínio	++	-	+	- -
Inteligibilidade do formato	++	-	+	-
Actualização da informação	-	+	++	-
Conteúdos de utilizadores	- -	-	- -	++

Tabela 11.1: Características das fontes de informação.

de terem uma formatação típica com campos separados por tabulações, não possuem uma formatação padrão no que diz respeito à representação da informação sobre as interacções dos utilizadores. Os diários do *tumba!* incluem bastante informação adicional a esse nível, permitindo extrair informação sobre os hábitos de pesquisa dos utilizadores, como por exemplo estimar o tempo médio que os utilizadores dispendem nas suas pesquisas, ou agregar as várias consultas usadas para cada pesquisa (Seco e Cardoso, 2006).

A Wikipédia gera periodicamente ficheiros compactados com o seu conteúdo, em formato XML ou em SQL, e como tal, a actualização da sua informação é elevada. Apesar de teoricamente a rede estar sempre actualizada, é preciso dispender algum tempo para realizar a recolha de documentos na rede, pelo que poderá haver alguma desactualização dos conteúdos. Por outro lado, as ontologias são actualizadas com baixa frequência, uma vez que requerem a revisão e validação cuidadosa dos novos dados através de humanos peritos no domínio da ontologia.

Finalmente, a característica mais atraente dos diários dos servidores é que possuem informação sobre os tópicos de interesse dos utilizadores, enquanto que os outros recursos não possuem dados sobre os utilizadores.

11.3 Trabalho desenvolvido até ao momento

A figura 11.4 esquematiza o modelo de RIG adoptado no meu trabalho. Podemos observar que a rede de conhecimento desempenha um papel crucial, assistindo os diversos módulos com informação geográfica necessária para o desempenho das suas tarefas. O trabalho realizado até agora tem focado os seguintes três pontos:

i. Reformulação automática de consultas

A abordagem de RAC adoptada possui uma atenção especial na reformulação dos termos geográficos com a ajuda da ontologia geográfica Geo-Net-PT01. O QuerCol é um módulo desenvolvido com o propósito de investigar as melhores práticas para extrair a "geogra-

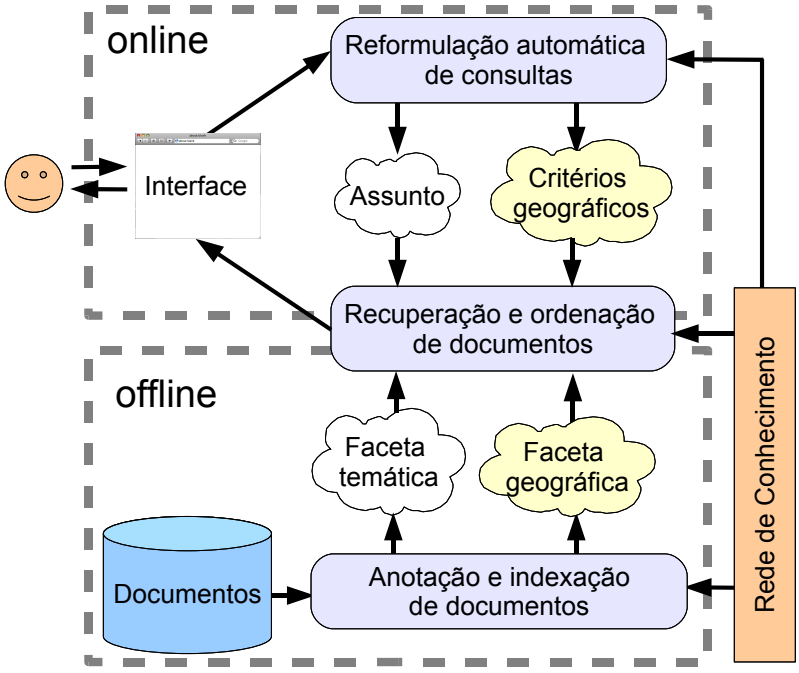

Figura 11.4: Arquitectura do sistema de RIG.

ficidade" das consultas, e de realizar a reformulação dos termos geográficos (expandindo "ilhas portuguesas" para os respectivos nomes, por exemplo), ou como lidar com relações espaciais nas consultas (por exemplo, "ao largo da costa portuguesa" torna locais como Peniche relevante, mas Évora não) (Cardoso e Silva, 2007).

ii. Anotação dos documentos

Os documentos em português são analisados automaticamente, com o intuito de extrair conteúdos de relevância geográfica e encontrar pistas que possam indicar as áreas de interesse de cada documento. O trabalho desenvolvido neste ponto está patente no REMBRANDT, um sistema de reconhecimento de entidades mencionadas vocacionado para textos em português, e que utiliza principalmente a porção portuguesa da Wikipédia como fonte de informação para poder identificar e classificar as entidades mencionadas que estão presentes no texto em português (Cardoso, 2008).

iii. Ordenação de documentos por critério geográfico

Na fase de recuperação e ordenação de documentos, procura-se conciliar os dois eixos de relevância (o assunto e a área geográfica de interesse) de forma a apresentar uma lista

final de resultados com documentos relevantes e que correspondam às expectativas do utilizador. O trabalho realizado tem focado a adaptação do MG4J (Boldi e Vigna, 2005) ao modelo de RIG.

11.3.1 QuerCol

O QuerCol é um módulo de RAC que possui duas formas de actuação: i) aplica uma técnica básica de expansão de termos intitulada de retorno de relevância cego (em inglês, *blind relevance feedback*, BRF) a todos os termos da consulta inicial (Rocchio Jr, 1971), e ii) realiza uma expansão de termos geográficos ao associar os nomes geográficos na consulta às respectivas entidades geográficas, e explorando as suas relações ontológicas com outros locais para obter mais nomes geográficos

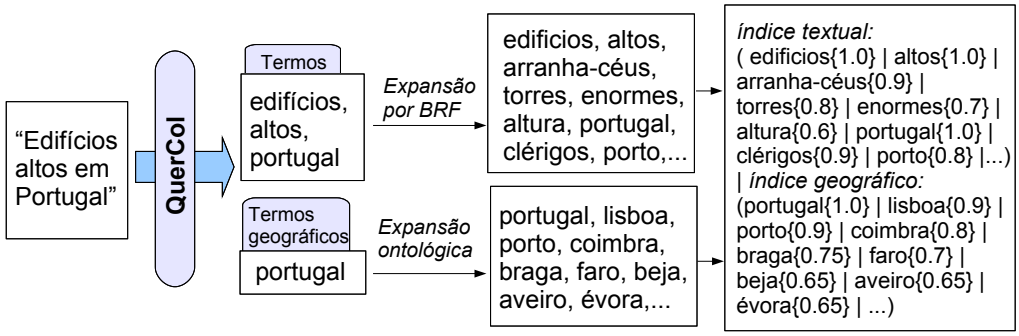

Figura 11.5: Funcionamento do QuerCol, um módulo de RAC.

A figura 11.5 ilustra o procedimento usado pelo QuerCol para reformular a consulta "Edifícios altos em Portugal". Primeiro, o QuerCol remove palavras muito frequentes da consulta (como é o caso de "em"), e reconhece "Portugal" como sendo um termo potencialmente geográfico, com a ajuda do REMBRANDT. Os termos *edifícios*, *altos* e *portugal* são enviados ao processo de BRF, utilizando o algoritmo $w_t(p_t\text{-}q_t)$ para atribuir pesos numa escala normalizada de [0,1]. (Efthimiadis, 1993) Os termos expandidos, como é exemplo "arranha-céus", são concatenados à consulta inicial através de operadores lógicos OU (|), e etiquetados de forma a serem usados posteriormente num índice textual.

Por outro lado, o termo geográfico "Portugal" é emparelhado com o conceito geográfico de 'Portugal (país)'. A expansão ontológica procura outros conceitos geográficos que estejam contidos dentro do território português, devido à relação espacial "em". As relações espaciais (por exemplo, "perto de" ou "nas costas de") e os tipos de entidades geográficas especificados (por exemplo, "praias", "montanhas" ou "universidades") são usados para conduzir a procura por mais nomes geográficos relevantes (Cardoso e Silva, 2007). Final-

mente, são atribuídos pesos aos termos geográficos, e são etiquetados para serem usados num índice geográfico.

11.3.2 REMBRANDT

O REMBRANDT (**R**econhecimento de **E**ntidades **M**encionadas **B**aseado em **R**elações e **AN**álise **D**etalhada do **T**exto, `xldb.di.fc.ul.pt/Rembrandt`) é um sistema de reconhecimento de entidades mencionadas (REM) que utiliza a Wikipédia como fonte de informação, e que explora a sua estrutura rica em categorias, ligações e redirecionamentos para classificar todo o tipo de entidades presentes no texto. Desta forma, o REMBRANDT tem acesso a conhecimento adicional sobre cada entidade mencionada (EM), o que se pode revelar útil para compreender o contexto da mensagem, detectar relações com outras EM, e usar essa informação para contextualizar e classificar EM vizinhas. Usemos como exemplo o termo "Porto", que pode ser utilizado num contexto não-geográfico, como em "António da Silva Porto". Contudo, a presença da EM "Torre de Clérigos" na mesma frase pode reforçar a confiança em que "Porto" de facto seja uma EM relativa à cidade portuguesa, devido à sua ligação com a cidade que pode ser extraída a partir da informação na sua respectiva página da Wikipédia, como é ilustrado na figura 11.6.

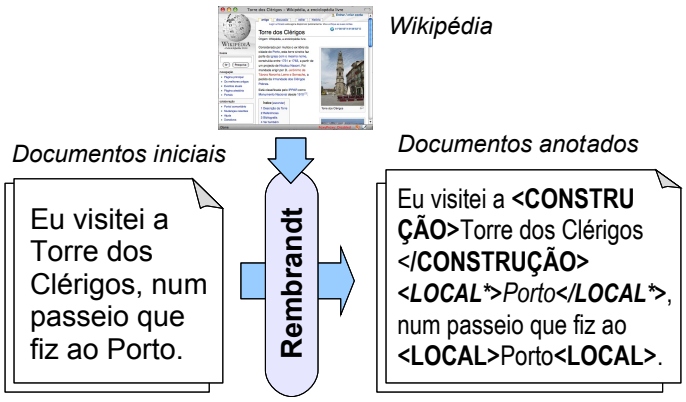

Figura 11.6: Acção do REMBRANDT na anotação de textos. Os asteriscos assinalam os locais inferidos a partir do texto.

O REMBRANDT classifica as EM de acordo com as nove categorias e as 47 sub-categorias definidas pelo Segundo HAREM, uma avaliação conjunta para sistemas de REM para textos em português (Santos et al., 2006, 2008b). As categorias principais são: `PESSOA`, `ORGANIZAÇÃO`, `LOCAL`, `TEMPO`, `VALOR`, `ABSTRACÇÃO`, `ACONTECIMENTO`, `COISA` e `OBRA`. O REMBRANDT lida perfeitamente com a vagueza intrínseca em algumas EM, ao classificá-las com mais de uma categoria ou sub-categoria. Por exemplo, a EM "Bombeiros Voluntários" pode ser considerada tanto uma organização ou um grupo de pessoas, consoante o contexto; se o

contexto não permitir destrinçar o seu verdadeiro significado, o REMBRANDT atribui as duas classificações à EM.

A estratégia do REMBRANDT baseia-se no emparelhamento de cada EM com a sua página respectiva na Wikipédia, e na análise da sua estrutura, ligações e categorias para obter mais conhecimento sobre ela. O REMBRANDT também depende de regras manuais para capturar pistas internas e externas para textos em português, tal como é descrito por McDonald (1996). As regras são usadas tanto para classificar EM que não têm correspondência na Wikipédia ou correspondem a páginas com informação insuficiente, como para corrigir o significado das EM de acordo com o contexto (por exemplo, "Rua de Portugal" designa uma rua, não um país). Adicionalmente, o REMBRANDT trata as categorias da Wikipédia como se fosse texto corrente, extraindo assim os nomes geográficos das categorias e permitindo a extracção de informação geográfica *implícita* para cada EM, como é ilustrado na figura 11.6 e descrito mais detalhadamente em Cardoso et al. (2008b).

11.3.3 MG4J

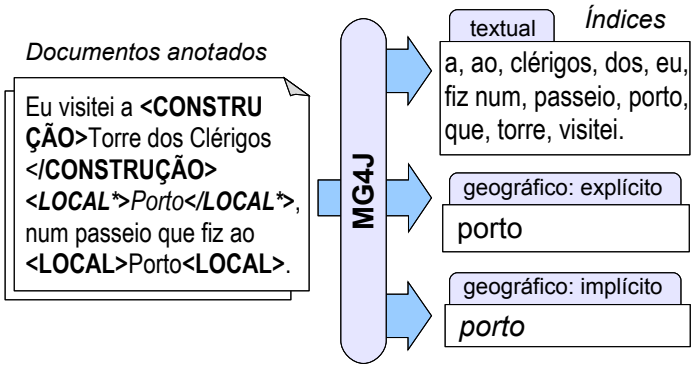

Figura 11.7: Indexação dos documentos anotados pelo MG4J. Os asteriscos assinalam os termos que serão indexados no índice geográfico implícito.

O MG4J é o módulo responsável pela indexação e ordenação dos documentos. A figura 11.7 exemplifica a indexação selectiva que o MG4J faz aos textos anotados pelo REMBRANDT. Os termos não-geográficos são indexados num índice textual, enquanto que os termos geográficos são indexados em dois índices geográficos: um índice `geográfico explícito`, que inclui EM classificadas como sendo locais geográficos, e um índice `geográfico implícito`, para os locais associados a EM que não são explicitamente locais geográficos. No caso ilustrado pela figura 11.7, podemos observar que o termo "Porto" representa o local geográfico implícito da EM "Torre dos Clérigos", e como tal é indexado no índice destinado a termos geográficos implícitos.

11.3.4 RENOIR

Outro módulo que está a ser desenvolvido é o RENOIR (**R**EMBRANDT's **E**xtended **N**ER **O**n **I**nteractive **R**etrievals, `xldb.di.fc.ul.pt/Renoir`). O RENOIR pode ser visto como uma maneira de incorporar algumas técnicas interessantes aplicadas na área de resposta automática a perguntas (RAP), explorando não só a rede de conhecimento criada no âmbito do trabalho deste doutoramento, como também outras redes de conhecimento já extraídas e disponibilizadas, como é o caso da DBpedia (Auer et al., 2007), com o objectivo de adequar a pesquisa a um processo de interpretação das consultas e recuperando documentos com a informação pretendida.

Um exemplo que ilustra bem as motivações que norteiam o desenvolvimento do RENOIR é a realização de consultas com os termos "Castelo Branco.". Tal como foi referido anteriormente, uma pesquisa por "Obras de Castelo Branco" muito provavelmente indicia que o utilizador está à procura de documentos sobre trabalhos do romancista português. Contudo, a consulta "Restaurantes de Castelo Branco" é mais direccionada para RIG, pois Castelo Branco refere-se à cidade portuguesa e, como tal, é uma consulta de cariz geográfico.

Com o RENOIR, procura-se investigar novas formas de enriquecer as consultas de forma a introduzir etiquetas semânticas de um modo manual, supervisionado ou automático. Nos exemplos anteriores, as consultas poderiam ser reformuladas para reflectir o contexto das pesquisas, como por exemplo, "Obras de `PESSOA`:{Castelo Branco}", e "Restaurantes `LOCAL`:{Castelo Branco}". Desta forma, o sistema RIG pode adaptar a sua actuação consoante a semântica da consulta, destrinçando os significados de "Castelo Branco" nos documentos (graças às anotações do REMBRANDT) e fornecendo documentos de acordo com o contexto correcto de "Castelo Branco".

11.4 Avaliação do desempenho dos sistemas

As avaliações conjuntas constituem uma componente fundamental no processo de construção e validação dos módulos, uma vez que permitem analisar os pontos fortes e as fraquezas de cada componente, em ambientes de avaliação controlados que procuram recriar situações de pesquisas reais para as quais o sistema deverá estar devidamente preparado.

O trabalho desenvolvido no âmbito do meu doutoramento tem sido objecto de avaliação periódica, de maneira a aferir o desempenho dos protótipos e dos seus módulos constituintes na realização das tarefas a que se propõem. A participação nas pistas de avaliação é feita para as tarefas em língua portuguesa.

i. GeoCLEF

O GeoCLEF é uma pista de avaliação específica para sistemas de RIG, e que inclui o português como uma das línguas usadas nas suas tarefas de avaliação (Mandl et al., 2008). No decurso do trabalho de investigação, a participação no GeoCLEF tem fornecido resultados bastante reveladores das potencialidades e das limitações das estratégias adoptadas para cada módulo (Cardoso et al., 2008a). O estado actual dos módulos e a linha de investigação até agora seguida têm sido constantemente aperfeiçoados mediante uma análise detalhada dos resultados da avaliação, sendo que, na edição de 2008 do GeoCLEF, obtivémos resultados bastante encorajadores (Cardoso et al., 2008c).

ii. HAREM

O REMBRANDT participou no Segundo HAREM, com o propósito de reconhecer todo o tipo de EM no texto. Também participou na sub-tarefa ReRelEM, para a detecção de relações entre EM. O REMBRANDT obteve um valor de medida F de 0.567 para a tarefa genérica de REM, cotando-se como o segundo melhor sistema num total de 10, e foi o primeiro sistema classificado para o cenário de EM da categoria `LOCAL`, com uma medida F de 0.625. Na tarefa de ReRelEM, o REMBRANDT também obteve o melhor resultado entre três sistemas, com uma medida F de 0.103.

iii. GikiP

O GikiP é uma pista piloto promovida pela Linguateca sob a chancela da pista GeoCLEF, propondo aos sistemas participantes uma tarefa de procura de artigos/entradas da Wikipédia que satisfazem uma dada necessidade de informação que exija algum raciocínio geográfico (Santos e Cardoso, 2008; Santos et al., 2008a). O GikiP usou na sua tarefa de avaliação as porções portuguesa, inglesa e alemã de uma recolha da Wikipédia de 2006.

O RENOIR participou no GikiP de uma forma supervisionada, utilizando a Wikipédia e o REMBRANDT como fontes de informação e de extracção de conhecimento para assistir a sua estratégia de formulação de consultas. Apesar de o RENOIR ainda estar nos seus primeiros passos, a participação no GikiP permitiu ter uma primeira experiência de como a sua filosofia orientada a consultas semânticas poderá permitir responder a necessidades de informação elaboradas, como são os casos dos tópicos "Indique membros do círculo de Viena que nasceram fora do império austro-húngaro ou da Alemanha", ou "Locais onde Goethe viveu".

Capítulo 12

Uso de marcadores estilísticos para a busca na Internet em português

Rachel Aires

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 12, p. 87–92, 2008

Este capítulo resume a tese de doutorado *Uso de Marcadores Estilísticos para a busca na Web em português*[1] (Aires, 2005) desenvolvida por dois anos no pólo de Oslo da Linguateca e por dois anos no NILC (Núcleo Interinstitucional de Lingüística Computacional) – `www.nilc.icmc.usp.br`. O trabalho de doutorado teve como principal objetivo pesquisar uma maneira de minimizar consideravelmente um dos principais problemas dos usuários de sistemas de busca na Internet, que é ter que lidar com um grande volume de documentos irrelevantes para ter acesso à informação pretendida. Para que um documento seja relevante, não basta que ele trate do assunto procurado, é necessário ainda que dê o enfoque desejado pelo usuário. O enfoque pode ser determinado por características como, por exemplo, formalidade, objetividade e o fato de o texto ser detalhado, tratando apenas de um assunto e não de vários.

A solução explorada para auxiliar o usuário a interpretar qual o enfoque é dado por um determinado texto foi classificá-los em *gêneros*, *tipos de textos*, *necessidades de busca* e *necessidades personalizadas*. Para gerar os métodos de classificação, foram utilizados algoritmos de aprendizado de máquina, corpora em português e marcadores estilísticos.

Para a classificação em *gêneros* e em *tipos textuais*, foram utilizados os gêneros e tipos textuais do corpus Lácio-Ref (Aluísio et al., 2003). O Lácio-Ref é um corpus aberto e de referência do português contemporâneo do Projeto Lácio-Web (`www.nilc.icmc.usp.br/lacioweb/`), composto de textos em português brasileiro, tendo como característica serem escritos respeitando a norma culta. A taxonomia de gêneros do Lácio-Ref é composta por textos científicos, de referência, informativos, jurídicos, prosa, poesia, drama, instrucionais e técnico-administrativos. Entretanto a versão utilizada do corpus não contém textos do gênero de referência ou do gênero técnico-administrativo. Nos experimentos com classificação em gêneros, os gêneros poesia, prosa e drama foram reunidos em um único supergênero Literário. São 29 os tipos textuais efetivamente presentes na versão utilizada do Lácio-Ref: declaração, petição, reportagem, artigo, decreto, medida provisória, poema, resenha, edital, portaria, resolução, editorial, monografia, resumo, carta, provimento, sentença, circular, entrevista, notícia, receita, súmula, lei, ofício, regimento, crônica, livro-texto, parecer, relatório.

O esquema de classificação em *necessidades de busca* é resultado de uma análise qualitativa dos *logs* de novembro de 1999 e de julho de 2002 da máquina de busca TodoBr[2]. Os sete tipos de necessidades tratados estão descritos na tabela 12.1.

Para gerar um classificador que considerasse as necessidades acima, foi criado um corpus inicial com 511 textos coletados da Internet obedecendo a dois critérios: (i) as páginas deveriam ser escritas em português do Brasil, para que variações lexicais, morfológicas e sintáticas entre as diversas variantes não interferissem no treinamento dos classificadores,

[1] O doutorado foi financiado pela Fundação para a Computação Científica Nacional (FCCN), através da Fundação para a Ciência e Tecnologia e co-financiado pelo POSI (POSI/PLP/43931/2001).

[2] Máquina de busca do domínio .br que foi incorporada ao Google em 2005, como Google Brasil: `www.todobr.com.br`.

1. Páginas que definam alguma coisa ou ensinem como e/ou porque algo acontece. Por exemplo: o que é a aurora boreal. Para esta necessidade, os melhores resultados seriam dicionários e enciclopédias, livros didáticos, artigos técnicos e relatórios e textos do gênero informativo.
2. Páginas que ensinem como fazer algo ou como algo é feito. Por exemplo: instruções de como instalar Linux em seu computador, ou receita de um bolo. Resultados típicos seriam textos do gênero instrucional, tais como manuais, livros didáticos, receitas e também alguns artigos técnicos e relatórios.
3. Páginas que forneçam uma apresentação, apanhado ou panorama sobre um determinado assunto. Por exemplo, um panorama sobre a literatura americana no século XX. Nesse caso, os melhores textos seriam dos gêneros instrucionais, informativo e científico, por exemplo, reportagens.
4. Páginas com notícias. Por exemplo: uma notícia sobre um atentado. As melhores respostas seriam textos do gênero informativo, como, por exemplo, notícias em jornais e revistas.
5. Páginas que forneçam informações sobre uma pessoa, empresa, instituição ou organização. Por exemplo: páginas pessoais, páginas com informações para contato (com currículo, telefone, endereço). Respostas típicas seriam páginas pessoais e institucionais.
6. Uma página específica que o usuário quer visitar, mas não se lembra da URL. Nesse caso, os resultados poderiam ser de qualquer tipo textual ou gênero.
7. Páginas que forneçam algum serviço online. Por exemplo: lojas virtuais, serviço dos correios para acompanhamento de envio de encomendas. As melhores respostas, nesse caso, seriam textos comerciais (empresas ou indivíduos oferecendo produtos e serviços).

Tabela 12.1: Descrição dos sete tipos de necessidades.

e (ii) as páginas selecionadas deveriam ser de diversas fontes e assuntos, já que textos de uma mesma fonte ou área podem ter estilo próprio (por exemplo, textos da Folha de São Paulo e textos médicos), sendo que o que pretendíamos investigar eram os marcadores de estilo relacionados ao propósito do texto. Nessa primeira versão do corpus, não foi considerado o fato de que um mesmo texto pode atender a mais de uma necessidade. Por isso, todos os textos foram revistos e novos textos foram incluídos para cada uma das combinações dos tipos de necessidades encontradas. A versão final do corpus (denominado Yes,User!) contém 1703 textos extraídos da Internet brasileira classificados conforme as necessidades de usuários a que satisfazem.

Investigou-se também a possibilidade de permitir a criação de esquemas de categorização pelo próprio usuário para suas *necessidades personalizadas*. Nessa opção, o usuário fornece exemplos de textos de um problema com o qual lida freqüentemente em suas bus-

cas na Internet, e o sistema, através de marcadores estilísticos, gera um esquema de classificação novo para aquele usuário. Os exemplos devem ser de problemas binários (de duas classes), que estejam relacionados a tipos de texto, assim como as sete necessidades citadas anteriormente. Por exemplo, no caso de um advogado, distinguir entre textos técnicos sobre direito e textos voltados para o público comum, como a sentença dada para uma determinada ação e uma página informal sobre os direitos do consumidor, respectivamente. Essa abordagem não serve para problemas de classificação relacionados ao assunto, como, por exemplo, distinguir entre textos científicos que falam sobre problemas do coração da área de cardiologia e textos de outras áreas médicas que também falem sobre problemas do coração.

Para testar a busca personalizada foram inicialmente utilizados dois corpora, um criado em um mestrado do ICMC (Martins Junior e Moreira, 2004), e outro criado para o trabalho de doutorado para testes. O propósito do primeiro corpus é distinguir se uma página contém descrições de produtos à venda ou não e é composto por 1252 páginas (723 exemplos positivos e 529 negativos). O segundo corpus é composto por 200 páginas relacionadas ao domínio de direito; tem o propósito de distinguir entre páginas de direito para pessoas da área (advogados, juízes, etc.) e textos para pessoas em geral, sendo formado por 100 exemplos positivos e 100 negativos.

A avaliação da busca personalizada foi feita também com corpora criados por seis usuários, cinco portugueses e um brasileiro, dos quais dois têm formação em letras e quatro em computação. Foi solicitado a cada um por e-mail que descrevessem o problema que seria tratado por seus corpora e que criassem cada um, um corpus com 200 textos, sendo 100 exemplos positivos e 100 negativos. Foram criados sete corpora em resposta à solicitação por e-mail. Tanto os corpora[3] criados para o trabalho de doutorado, como o protótipo de ferramenta de busca na rede criado para utilização nos testes com os diversos esquemas de classificação[4] estão disponíveis no sítio da Linguateca. A descrição dos problemas tratados em cada um dos corpora personalizados é apresentada na tabela 12.2.

Cinco conjuntos de marcadores estilísticos foram utilizados para a criação dos diversos tipos de classificadores. O primeiro foi criado com base nos trabalhos de Biber (1988) e Karlgren (2000). Para sua criação foram consideradas intuições lingüísticas e foi dada preferência a marcadores que pudessem ser calculados sem a ajuda de qualquer tipo de analisador, como etiquetadores morfossintáticos e sintáticos. Esse conjunto contém 46 marcadores e é formado por estatísticas baseadas em palavras, como número de palavras longas; estatísticas baseadas no texto como um todo, como número de frases; e outras estatísticas, como número de advérbios de lugar.

[3] Os corpora estão disponíveis em `www.linguateca.pt/Repositorio/YesUser/`.

[4] O protótipo e seu código fonte encontram-se disponíveis em `www.linguateca.pt/Repositorio/leva-e-traz/`.

Problema 1. Obter textos teóricos em HTML sobre filosofia da linguagem e sobre os principais pensadores e não textos (também em HTML) que apresentem programas de cursos, colóquios, conferências, livros, etc. sobre este tópico.
Problema 2. Obter textos teóricos em HTML sobre língua portuguesa e sobre os principais pensadores e não textos (também em HTML) que apresentem programas de cursos, colóquios, conferências, livros, etc. sobre este tópico.
Problema 3. Diferenciar textos que apresentem fatos sobre Fado de textos que emitam opiniões. No primeiro caso estão textos contendo informação histórica, biografias, notícias, etc. No segundo, entrevistas, críticas a discos e espetáculos, etc.
Problema 4. Encontrar textos que sejam uma descrição sobre determinado tema de História Geral. Entretanto, páginas com eventos, conferências, catálogos de livros não interessam, bem como informações sobre cursos de História, links para páginas de História ou ementas de disciplinas. Além disso, relatos de pessoas sobre seu gosto pela História também não são de interesse.
Problema 5. Distinguir glossários, receitas e técnicas sobre culinária japonesa de anúncios de livros, informação nutricional, críticas a restaurantes, páginas de restaurantes, informação sobre alimentação, cursos, festivais gastronômicos ou culturais sobre o mesmo tema.
Problema 6. Textos que interessam são história/fatos sobre surrealismo, como "Salvador Dali e o Surrealismo", "Manifesto do Surrealismo" e "Enciclopédia Universal Multimídia On-line". Blogs, exposições ou opiniões como "BdE - Blogue de Esquerda (II) 80 ANOS DE SURREALISMO", "Adelto Gonçalves,- comemorações", "A estranha sombra do surrealismo português", não interessam.
Problema 7. Documentos relevantes são: 1 - Documentos que explicam os princípios físicos que permitem que os aviões voem; 2 - Explicações técnicas de partes de componentes de aviões, tais como altímetros, tipos de motores ou rotores de helicópteros, etc.; 3 - História da aviação - biografia de pioneiros da aviação, os avanços aeronáuticos ao longo do tempo; 4 - História dos aviões - História de certos aviões importantes para a historia, as suas características, o motivo do seu desenvolvimento, o seu impacto na história da aviação. Documentos não-relevantes são: 1 - Notícias relacionadas com compras de aviões e empresas de aviação comercial; 2 - Notícias e descrições detalhadas de acidentes aéreos; 3 - Relatos de desvio de aviões e terrorismo aéreo; 4 - Opiniões sobre pilotagem, histórias e relatos de clubes de aviação, diversos documentos sobre psicologia do avião, deveres dos pilotos, etc.

Tabela 12.2: Descrição informada pelos usuários para as sete necessidades personalizadas tratadas.

O segundo conjunto de marcadores utilizado é composto por cinco funções para medir a riqueza de vocabulário propostas por Stamatatos et al. (2000).

O terceiro conjunto de marcadores é composto pelas 62 palavras mais freqüentes do corpus de necessidades Yes,User!: eliminando-se as *stopwords*, verbos auxiliares, advérbios, palavras relacionadas a domínios e agrupando algumas das palavras mais freqüentes como um único marcador. Esse conjunto é utilizado apenas nos experimentos com a classificação em sete necessidades de busca, pois possui marcadores dependentes dessa tarefa, como, por exemplo, número de ocorrência das palavras "download" e "kb".

Foi também utilizado um conjunto formado por 15 marcadores sintáticos selecionados com base em intuição lingüística, que foram calculados com ajuda do etiquetador sintático PALAVRAS (Bick, 2000), além de outro formado por 27 marcadores de aparência gráfica (layout). Exemplos de marcadores sintáticos são a porcentagem de sujeitos pronominais e o número de orações subordinadas e de marcadores de aparência gráfica são características de documentos HTML que indicam decisões como fonte utilizada, espaçamento e cor.

Capítulo 13

Listas de frequência de palavras como marcadores de estilo no reconhecimento de autoria

Rui Sousa Silva

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 13, p. 93–97, 2008

O estudo e análise do discurso tem sido objecto de diferentes teorias e abordagens (Coulthard, 1977; Dijk, 1997; Fairclough e Wodak, 1997; Sinclair, 1991), desde a análise da interacção entre o discurso e a sociedade e a análise crítica do discurso (Dijk, 1997; Fairclough e Wodak, 1997) à análise do discurso enquanto realização linguística (Coulthard, 1977; Sinclair, 1991), passando pelo estudo da relação entre a linguística e a lei como forma de linguística forense (análise forense do discurso) (Coulthard e Johnson, 2007; Shuy, 2006).

A análise forense do discurso, enquanto ramo da linguística aplicada, possui aplicações diversificadas, entre as quais: a identificação de autoria; a identificação do modo (no sentido de Halliday); a tradução e a interpretação jurídica; a transcrição de declarações e depoimentos; o estudo da linguagem e discurso dos tribunais; o estudo de direitos linguísticos; a análise de declarações; a fonética forense; e o estudo do estatuto textual. Neste artigo, debruçamo-nos sobre a primeira: a identificação de autoria. Recorrendo à análise da utilização da linguagem pelo autor e das informações que essa análise transmite ao analista acerca do escritor, linguisticamente (Olsson, 2004), procuramos determinar o perfil de autoria textual, isto é, identificar o autor com base numa análise contrastiva de um corpo de textos limitado (Coulthard e Johnson, 2007; Olsson, 2004).

Para determinar este perfil, não podemos limitar-nos à utilização de dados puramente estatísticos dos próprios textos estudados, uma vez que o contexto sociocultural e a realidade extra-textual influenciam a forma de falar e de escrever dos falantes de uma determinada língua; num mesmo país ou cultura, diferentes pessoas, com acesso diferente a educação, formação e informação, têm formas semelhantes de produção textual. O sociolecto (i.e., a variedade de uma língua característica de uma determinada classe ou estatuto social) pode restringir a gama possível de autores, mas não é um factor decisivo. A análise estatística dos dados constitui, assim, um dos métodos utilizados, mas não o único. Daí o recurso à análise forense do discurso como forma de equacionar os dados mais relevantes do corpo de textos.

Considerando todos estes princípios, teremos, então, que procurar identificar o idiolecto de cada um dos autores, isto é, presumindo que todos os falantes nativos de uma língua possuem uma versão distinta e individual da língua que falam e escrevem, teremos que procurar no texto marcadores que apontem para a selecção individual de aspectos linguísticos genéricos (Coulthard e Johnson, 2007). Socorremo-nos, para o efeito, de três princípios da estilística forense: o princípio de que o estilo individual de cada autor é determinado pela escolha (Hänlein, 1998); o grau em que o autor tende para determinadas formas de "expor as coisas" (McEnery e Wilson, 1996); e, finalmente, o pressuposto de que é necessário identificar um conjunto agregado e único de marcadores, presentes individualmente noutros autores (McMenamin, 2002). Reconhecendo a validade e a fiabilidade de marcadores como o formato do texto, a utilização de números/símbolos, abreviaturas, pontuação, maiúsculas/minúsculas, ortografia, formação lexical, sintaxe, discurso, er-

ros e correcção, utilização da voz activa e passiva, entre outros, focmo-nos, neste estudo, nas expressões e palavras de elevada frequência, no sentido de verificar a sua utilidade e aplicabilidade como marcador de discurso no reconhecimento de autoria em português, a exemplo do que acontece para outras línguas (Hänlein, 1998)[1].

Com base nos estudos em linguística com corpos (Biber et al., 2000; McEnery e Wilson, 1996), criámos um corpo de 84 textos escritos pelos cronistas António Barreto e José Pacheco Pereira, com 107.360 átomos, publicados no jornal Público entre Janeiro e Dezembro de 2007. Recorrendo ao Corpógrafo (Sarmento et al., 2004; Maia e Matos, 2008), analisámos a frequência de expressões com um comprimento de quatro gramas (i.e., tetragramas) utilizadas pelo autor uma única vez (*hapax legomena*) e a frequência de expressões que ocorrem mais vezes nos textos do mesmo autor (*hapax dislegomena*). Depois de proceder à extracção dos tetragramas mais frequentes, procedemos à sua classificação, manualmente, segundo uma taxonomia de 15 classes, conforme proposto por Sousa Silva (2006): *especificação*, *explicação*, *exemplificação*, *comparação*, *contraste*, *generalização*, *correcção*, *preparação*, *inclusão*, *concessão*, *restrição*, *enumeração*, *propósito*, *negação*, *justificação*. Os resultados desta análise, apresentados nas tabelas 13.1 (com uma ordenação por classe semântica) e 13.2 (com uma ordenação por frequência decrescente de utilização), mostram que os dois autores recorrem a estratégias semânticas de produção textual diferentes. Os valores classificados como ruído resultam de n-gramas obtidos com caracteres não reconhecidos – e, por isso, considerados erros.

Comparando a utilização das classes pelos dois autores, verificamos, conforme apresentado na tabela 13.1, que os dois autores recorrem com uma frequência idêntica a estratégias de *correcção*, *negação* e *restrição*, utilizando, porém, de forma distinta as restantes classes:

A interpretação que fazemos dos dados obtidos permite-nos constatar que, enquanto António Barreto recorre a expressões com um valor semântico que lhe permitem ser mais claro, directo e focalizado, José Pacheco Pereira apresenta características de uma produção textual mais vaga, hesitante e inconstante -– frequentemente conotada com uma literacia elitista.

Para verificar os resultados do presente estudo, analisámos dois textos escritos pelos dois autores, publicados no jornal Público em 2008. A metodologia adoptada consiste na aplicação de um "teste cego" (isto é, com textos cuja autoria foi tornada anónima), com o objectivo de confrontar os textos com as conclusões do estudo do corpo de textos. Considerando que estes dois textos são demasiado pequenos para uma análise estatística (cerca de mil átomos por texto), procurámos traços individuais marcantes em cada um deles, nomeadamente a frequência das palavras utilizadas no corpo de textos recolhidos em 2007

[1] Neste contexto, entendemos "palavras" no sentido que lhe foi atribuído por Halliday (1994) de "wordings", ou seja, são palavras as sequências gramaticais, ou "sintagmas", constituídas por elementos de três tipos: elementos lexicais (tais como verbos e nomes), elementos gramaticais (tais como artigos e determinantes), e elementos intermédios (tais como preposições) – todos eles elementos que constituem os n-gramas.

António Barreto			José Pacheco Pereira		
Classe	Total	%	Classe	Total	%
comparação	20	5,13	comparação	50	10,22
concessão	8	2,05	concesão	13	2,66
contraste	41	10,51	contraste	17	3,48
correcção	0	0,0	correcção	0	0,0
enumeração	24	6,15	enumeração	63	12,88
exemplificação	9	2,31	exemplificação	11	2,25
explicação	18	4,62	explicação	91	18,61
generalização	18	4,62	generalização	8	1,64
inclusão	16	4,10	inclusão	4	0,82
justificação	0	0,0	justificação	10	2,04
negação	0	0,0	negação	0	0,0
preparação	10	2,56	preparação	8	1,64
propósito	8	2,05	propósito	6	1,23
restrição	0	0,0	restrição	0	0,0
especificação	218	55,90	especificação	208	42,54
Total	**390**	**100,0**	**Total**	**489**	**100,0**
ruído	**0**		**ruído**	**1**	

Tabela 13.1: Lista comparativa de classes semânticas utilizadas pelos autores (ordenadas por classe semântica).

António Barreto			José Pacheco Pereira		
Classe	Total	%	Classe	Total	%
especificação	218	55,90	especificação	208	42,54
contraste	41	10,51	explicação	91	18,61
enumeração	24	6,15	enumeração	63	12,88
comparação	20	5,13	comparação	50	10,22
explicação	18	4,62	contraste	17	3,48
generalização	18	4,62	concesão	13	2,66
inclusão	16	4,10	exemplificação	11	2,25
preparação	10	2,56	justificação	10	2,04
exemplificação	9	2,31	generalização	8	1,64
concessão	8	2,05	preparação	8	1,64
propósito	8	2,05	propósito	6	1,23
correcção	0	0,0	inclusão	4	0,82
justificação	0	0,0	correcção	0	0,0
negação	0	0,0	negação	0	0,0
restrição	0	0,0	restrição	0	0,0
Total	**390**	**100,0**	**Total**	**489**	**100,0**
ruído	**0**		**ruído**	**1**	

Tabela 13.2: Lista comparativa de classes semânticas utilizadas pelos autores (ordenadas por frequência).

António Barreto			José Pacheco Pereira	
−	comparação	≠	+	comparação
−	concessão	≠	+	concessão
+	contraste	≠	−	contraste
−	**correcção**	=	−	**correcção**
−	enumeração	≠	+	enumeração
+	exemplificação	≠	−	exemplificação
−	explicação	≠	+	explicação
+	generalização	≠	−	generalização
+	inclusão	≠	−	inclusão
−	justificação	≠	+	justificação
−	**negação**	=	−	**negação**
+	preparação	≠	−	preparação
+	propósito	≠	−	propósito
−	**restrição**	=	−	**restrição**
+	especificação	≠	−	especificação

Tabela 13.3: Comparação das classes semânticas utilizadas pelos dois autores.

(e que aqui utilizamos como corpo de referência, isto é, com o corpo de textos com o qual comparamos os textos A e B). A lista de frequência de palavras dos textos anónimos referidos como "Texto A" e "Texto B" mostra que, enquanto o Autor A utiliza com maior frequência as expressões "acima de tudo," e "o que significa que", o Autor B utiliza expressões como "a verdade é que", "ao mesmo tempo que" e "assim como o de". Contrastando estes resultados com os resultados obtidos na análise do corpo de textos utilizado no estudo, verificamos que as expressões utilizadas pelos autores A e B correspondem, respectivamente, a José Pacheco Pereira e António Barreto.

Este estudo permite, assim, comprovar que existem diferenças semânticas significativas, mesmo tratando-se de autores que escrevem com uma regularidade semelhante para um mesmo público, sob orientações editoriais idênticas. Poderemos, por isso, interpretar os dados obtidos como sendo um marcador de autoria válido e fiável em português, a exemplo do que acontece com outras línguas (como é o caso do inglês). Poderemos, por isso, constatar que, uma vez que cada autor possui um idiolecto próprio (Coulthard e Johnson, 2007), com marcas de autoria distintas, diferentes textos, produzidos por diferentes autores, recorrem à utilização de elementos idiossincráticos e padrões linguísticos distintos.

Em conclusão, esta análise demonstra a utilidade das listas de frequência de palavras como critério de reconhecimento de autoria em português.

Capítulo 14

Conversor de grafemas para fones baseado em regras para português

Sara Candeias e Fernando Perdigão

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 14, p. 99–104, 2008

Esta apresentação tem por objectivo descrever um sistema de conversão automática de grafema para fone (GR2PH) para o português de Portugal. Para o desenvolvimento do GR2PH está a ser usado o corpus de unidades acentuais (palavras) em língua portuguesa SPEECHDAT (SPEECHDAT), disponibilizado pela Universidade do Minho (proveniente da colaboração entre a Linguateca e o Projecto Natura). A avaliação do GR2PH fará uso do vocabulário da base de dados SPEECHDAT bem como de outros corpora de teste já usados por diversos investigadores a trabalhar neste domínio. A anotação fonética de corpora em língua portuguesa seria um interessante recurso linguístico a tornar público na Linguateca. Este recurso poderia ficar disponível, depois de avaliado e validado o sistema.

A crescente procura de soluções baseadas em produtos de tecnologia da fala tem sido uma motivação para o desenvolvimento de sistemas capazes de estabelecer um interface Homem-Máquina mais natural, como são exemplos as práticas subjacentes a áreas do ensino/aprendizagem do português e da linguística clínica.

A consciencialização da necessidade destes produtos mobilizou ao desenvolvimento do GR2PH, que convertesse, de forma automatizada, corpora grafados em corpora notados foneticamente.

O GR2PH, do qual fazem parte os subsistemas 'divisor de sílabas' e 'marcador de tonicidade', é aquele para o qual o conhecimento linguístico contribui com um maior impacto.A estratégia adoptada para o GR2PH baseia-se em regras linguísticas cotejadas na estrutura da língua portuguesa. Para o desenvolvimento quer do sistema que transmuta grafema em fone, quer dos sistemas intermédios para divisão silábica e para marcação de sílaba tónica, foi usado o corpus de unidades acentuais (perto de 680000) em língua portuguesa, disponibilizado como recurso nascido da colaboração entre a Linguateca e o Projecto Natura. Na verdade, o acesso a este recurso resultou numa mais valia ao desempenho do(s) sistema(s) que se pretendia(m) desenvolver, e os testes que foram sendo feitos, mesmo de forma faseada, mostraram-se basilares na fase de estruturação da arquitectura do(s) próprio(s) sistema(s), complementares e final.

Para o português de Portugal, alguns transcritores de grafema para fone baseados em regras surgem descritos em Almeida e Simões (2001); Braga e Resende Jr (2007); Teixeira et al. (2006); Gouveia et al. (2000); Viana e Andrade (1985). Para a implementação das regras, em certos grupos, é reconhecida a importância da identificação da unidade silábica (Almeida e Simões, 2001; Braga e Resende Jr, 2007; Teixeira et al., 2006; Gouveia et al., 2000); noutros, é usada a informação da tonicidade da vogal (Almeida e Simões, 2001; Braga e Resende Jr, 2007; Viana e Andrade, 1985). A indispensabilidade de desenvolvermos um novo sistema de conversão GR2PH para o português de Portugal advém de factores como a escassa partilha dos algoritmos dos sistemas já implementados (dos quais poder-se-ia partir para um esforço de melhoramento do sistema) e dos resultados dos testes de desempenho provenientes de estudo comparativos. Este artigo apresenta uma tessitura alternativa de

Convenções	Significado
C	consoante
V	vogal
.	divisor de sílaba
’	marcador de tonicidade
#	fronteira final de UA
|	ou

Tabela 14.1: Convenções usadas nas regras para implementação.

regras linguísticas a serem aplicadas no GR2PH para o português de Portugal, aliando a pertinência da informação linguística de regras de silabificação e de marcação de tonicidade. Resultando o sistema final da configuração de dois subsistemas perspectivados em regras inerentes à língua, o esforço do investimento tem por objectivo a viabilidade de um conversor capaz de uma eficácia que torne dispensável o recurso a dicionários de excepções. A arquitectura do GR2PH é resultado da complementaridade da aplicação do conhecimento linguístico e da ciência de engenharia, parceria esta que se traduz num diálogo necessário a uma execução que se pretende optimizada e eficaz.

14.1 Arquitectura do sistema de conversão GR2PH

O GR2PH recorre ao uso de sistemas intermédios, como o de separação da unidade acentual (UA, palavras) em sílabas e o de marcação de sílaba tónica (e consequente delimitação de sílaba(s) pré-tónica(s) e de sílaba(s) pós-tónica(s)). A vantagem desta abordagem explica-se pelo facto de ela permitir resolver a quase totalidade de casos de escolha fonética que não seria a acertada se resultasse apenas da inserção dos fones (nomeadamente vocálicos) considerados a partir de inventários fonéticos não diferenciados, isto é, não ponderados nem silabicamente nem atendendo à tonicidade em âmbito contextual de UA.

Todas as regras foram implementadas inicialmente em Matlab e foram testadas no vocabulário da base de dados SPEECHDAT (SPEECHDAT) e no corpus de unidades acentuais disponibilizado pela Linguateca/Projecto Natura.

Esta segunda parte apresenta as especificidades dos subsistemas de divisão silábica, de marcação de tonicidade e do transcritor, de forma a se ter uma visão global do sistema geral de conversão GR2PH. Na tabela 14.1 figuram as convenções usadas nas regras para implementação.

14.1.1 Subsistema de divisão silábica

A estrutura deste subsistema assenta a) num modelo de regras de divisão de base ortográfica, b) na consideração de vogal como núcleo de sílaba e c) na consideração de alguns dígrafos como grafema singular (‘ch’, ‘ss’, ‘lh’, ‘gu’+‘i’|‘e’, ‘qu’+ ‘i’|‘e’, etc.). O algoritmo

Sequência	Exemplo	Sequência	Exemplo	Sequência	Exemplo
CCVCC	trans.cre.ver	CVCC	subs.cre.ver	VC	ac.tu.ar
CCVVC	grãos	CVVC	mães	VV	eu
CVCCC	tungs.té.ni.o	VCVC	achar	V	á.gua
CCCV	stre.sse	VVC	aus.cul.tar	CVV	pai
CCVC	trás	VCC	abs.tra.ir	CVC	a.cam.par
CCVV	grão	CCV	a.cre	VC	ac.tu.ar

Tabela 14.2: Lista dos padrões de sequências de grafemas a formar sílaba em português de Portugal.

do 'divisor de sílabas' reproduz uma busca feita por padrões de até 5 grafemas, resultando em 18 possíveis encontros de sequências que formam sílaba em português de Portugal (tabela 14.2). As regras foram distribuídas por dois grandes grupos para cada padrão de sequência de grafemas, isto é, considerando se na sílaba da UA a analisar é pertinente a informação dos 4 caracteres ou de mais que os 4 caracteres da sequência. Nesta repartição, surgem regras explícitas que apresentam um tipo repetido subsequente da iteração de sequências, como é exemplo a sequência VV presente nos padrões CCVV, CVVC, CVV e VVC. Na tabela 14.3, a título de exemplificação de procedimentos, surgem descritas regras para o padrão CVVC.

14.1.2 Subsistema de marcação de tonicidade

Na estruturação deste subsistema, toda a unidade (palavra) foi considerada acentual (UA) e, por isso, não foram admitidos segmentos desprovidos de tonicidade (Candeias, 2007). O algoritmo de marcação da sílaba tónica funciona com regras instituídas a partir da divisão silábica. Admitiu-se o acento tónico como o acento da UA (o acento principal), pelo que, nesta estrutura, não se considerou pertinente marcar os acentos secundários. Na tabela 14.4 figuram regras de marcação de sílaba tónica.

14.1.3 Subsistema de transcrição para fones

Para a anotação fonética, seguimos o alfabeto SAMPA para o português (SAMPA), sem o recurso a extensões como seria o caso das «oclusivas orais sonoras» «fricatizadas», traço que advém da posição em início de sílaba e intervocálica. Ainda que se tenha em vista a construção de um sistema de síntese futuro, o que leva a ter em conta, entre outros aspectos, a natureza particular de cada som em contexto de co-articulação e/ou de sandhi, o facto deste mapeamento da transmutação grafema–fone ir ser adicionado a um modelo acústico baseado em trifones, anula a necessidade de uma anotação fonética mais estreita. Com este mesmo princípio, não foram consideradas como «semiconsonânticas» 'j' e 'w' as unidades vocálicas grafadas 'i'(ou 'e') e 'u' (ou 'o') dos ditongos ditos crescentes (pre-

	C	**V**	**V**	**C**	Grafema final da UA	Grupo silábico	Exemplo
Sequência		a|e|o|u a|e|o	i u	l|r|m|s|j	**V**	**CVV.C**	*pau*.lada, *mou*.ro, *tei*.ma, *lou*.sa, *bei*.jo
		ã|õ ã	e o	≠s|#			*mãe*.zinha, *mão*, ta.*lão*
	g|q	u	a|o		**V**		*quo*.ciente, *gua*.rida, *qua*.se, *qua*.lidade
		a|e|o|u a|e|o	i u	l|z	#	**CV.VC**	*pa*.ul, *ra*.iz
		a|e|o|u a|e|o	i u	r|m	**C**|#		*ca*.ir, *ru*.im, *co*.imbra
		a|e|o|u	i	nh	**V**		*ba*.inha, ta.*bu*.inha, *mo*.inho
		a|e|o|u a|e	i u	n	**C**		*re*.incide, tran.*se*.unte
		a|e|o|u a|e|o	i u	s	**C**|#	**CVVC.**	*cais*, *faus*.to, a.*zuis*, *bois*
		ã|õ ã	e o	s			*mãos*, *pães*
	g|q	u	a|o	l|n|r	**C**|#		*qual*, *qual*,quer, *guar*.da, *quan*.do
	por defeito					**CV.VC**	*be*.ata, *fi*.os

Tabela 14.3: Ilustração de algoritmo de divisão silábica para o padrão de grafemas CVVC.

	Regra	Marcador de tonicidade	Exemplo
1.	Se na sílaba existirem vogais com acento gráfico	sílaba em questão	a.'*ná*.li.se
2.	Se na sílaba não existirem vogais sem acento gráfico		
2.1.	Se a UA tiver 1 sílaba	sílaba em questão	'*voz*
2.2.	Se a UA tiver ≥ 2 sílabas		
2.2.1.	Se for a última sílaba da UA com estrutura de a|e|i|o|u + l|r|z i|u + ∅|s i + m	sílaba em questão	pa.'*ul* ra.'*iz* *ca*.'ir an.'*dou*, ca.pi.'*tais* pe.'*ru*, pe.'*rus* ru.'*im*
2.2.2.	por defeito	penúltima sílaba	a.na.'*li*.se

Tabela 14.4: Algoritmo de marcação de sílaba tónica.

Fone	Posição de tonicidade	Posição silábica	Exemplos
o∼		+ m|n (mesma sílaba)	’*o*m.bro → *o*∼bru; p*o*n.tu.’al → p*o*∼tual
w∼		ã + (mesma sílaba)	’cã*o* → k6∼*w*∼; cã*o*.’zi.nho → k6∼*w*∼ziJu
o	tónica	+ nh (sílaba seguinte)	ri.’s*o*.nho → riz*o*Ju
O	tónica	+ x (sílaba seguinte)	pa.ra.’d*o*.xo → p6r6d*O*ksu
o	tónica	+ i (mesma sílaba)	’*o*i.to → *o*jtu
o	tónica	+ r (mesma sílaba e final de UA)	pa.ssa.’d*o*r → p6s6d*o*r
O	tónica	+ r (mesma sílaba)	’c*o*r.ta → k*O*rt6
o	tónica	+ a (sílaba seguinte e final de UA)	’t*o*.da → t*o*d6
O	tónica	por defeito	’*o*.de → *O*d@; ’c*o*.rre → k*O*R@
O	átona	(inicial de UA) + r	*O*r.ga.’ni.za → *O*rg6niz6
u	átona	+ r (mesma sílaba)	c*o*r.’tar → k*u*rtar
O	átona	(inicial de UA)	*o*.’ní.ri.co → *O*niriku
u	átona	o (sílaba anterior) +	co.*o*. pe.ra.’ção → ku*u*p@r6s6∼w∼
u	átona	(final de UA)	’fi.lh*o* → fiL*u*
O	átona	+ c|p (mesma sílaba)	*o*c.’ta.vio → *O*taviu; *o*p.’ção → *O*ps6∼w∼
u	átona	por defeito	p*o*.’ção → p*u*s6∼w∼

Tabela 14.5: Ilustração de algoritmo de conversão do grafema ‘o’ para fones.

sentes em relógio e em área, em suave e em nódoa). O algoritmo da conversão do grafema em fone funciona a partir das sílabas com ‘marcação de tonicidade’. Isto é, a partir de um contexto-base, resultam casos de grafemas admitidos à conversão em fones que consideram a pertinência de informação da a) posição de tonicidade e da b) posição no âmbito da sílaba (na qual é pertinente o comportamento fonético dados os grafemas vizinhos). Na tabela 14.5 são exemplificados os algoritmos de conversão do grafema ‘o’ para os fones [o∼], [w∼], [o], [O] e [u], que resultam da atenção aos parâmetros descritos.

A análise e verificação de muitas regras foi conseguida por análise exaustiva ao corpus de UAs disponibilizado pela Universidade do Minho. Transcrições ou pronunciações alternativas não são consideradas neste sistema, como é o caso de homógrafos heterófonos.

14.2 Conclusão e trabalho futuro

Até esta fase, a forma gráfica convertida automatizadamente em forma fonética foi avaliada com referência à anotação manual. Dispomos apenas do vocabulário associado à base de dados SPEECHDAT como material de teste, embora a avaliação com este corpus não esteja ainda concluída, especialmente devida à discordância encontrada na conversão das semiconsoantes dos ditongos crescentes. Uma forma alternativa de fazer a avaliação do sistema consiste em comparar os resultados de vários sistemas de conversão – pelo menos um é de domínio público (Almeida e Simões, 2001) –, contando e analisando as diferenças encontradas. Como trabalho futuro, pretendemos construir uma aplicação on-line de conversão de grafemas para fones bem como de um corpus anotado foneticamente.

Capítulo 15

Dez anos de convivência: um apanhado geral quanto ao uso dos recursos da Linguateca no Programa de Pós-Graduação em Ciência da Computação da PUCRS - Brasil

Vera Lúcia Strube de Lima

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 15, p. 105–107, 2008

Nota dos editores: Este capítulo mantém-se aqui POR INSISTÊNCIA DOS EDITORES, embora a autora nos tivesse INICIALMENTE pedido para retirar, visto que os editores não tinham explicado que os resumos seriam mais tarde publicados neste formato.

No momento em que a Linguateca completa o seu décimo aniversário, parece-nos interessante relatar a importância que ela representa na tarefa de ensino do Processamento da Língua (ou Linguagem) Natural, especialmente junto ao Programa de Pós-Graduação em Ciência da Computação (PPGCC) da Pontifícia Universidade Católica do Rio Grande do Sul, Brasil. Nosso Programa, criado em 1994, como Programa de Pós-Graduação em Informática, e oferecendo então o curso de Mestrado em Informática, passou em 1999 à denominação atual e em 2004 a oferecer o curso de Doutorado. Desde os anos 90, o Programa conta com disciplinas de Inteligência Artificial e Processamento da Língua Natural. Mais recentemente, em 2006, passou-se a oferecer a disciplina de Inteligência Computacional, em lugar de Inteligência Artificial, e a de Aplicações do Processamento da Língua Natural, mais voltadas ao público que recebemos, o qual é proveniente das áreas de Sistemas de Informação, Interface Humano-Computador e outras, além dos interessados em Processamento da Língua Natural, propriamente ditos. Nesse período que decorre desde a implantação do Programa, as disciplinas de Processamento da Língua Natural ou Aplicações do Processamento da Língua Natural foram oferecidas ao todo mais de 12 vezes, atendendo a um contingente de mais de 60 alunos regulares do PPGCC, e também alunos provenientes do Programa de Pós-Graduação em Letras de nossa Universidade, ou mesmo interessados em geral, na qualidade de alunos especiais.

O amadurecimento na proposta de tarefas e de temas de estudos a estes alunos seguiu, acreditamos, o próprio ritmo de evolução da Linguateca. Inicialmente, nossos trabalhos com os alunos focalizavam a tradução de materiais de base, compondo a fundamentação para a disciplina, além de pequenos exercícios. Com o passar do tempo, e com uma base de leitura mais consolidada, fomos alçando vôos um pouco mais altos: juntamente com os textos fundamentais, passamos a trabalhar textos mais atualizados e publicações voltadas a problemas específicos do português. Mais recentemente, passamos a explorar recursos e ferramentas através de exercícios comparativos, possibilitando ao aluno maior fluência nas diferentes técnicas empregadas para a solução de problemas da área. Especialmente nos últimos 10 anos, essa evolução gradual coincidiu com a consolidação da Linguateca.

Ainda cabe mencionar que, nesse cenário profícuo que se foi delineando, pudemos exercer também um papel de produtores de recursos, e não apenas de consumidores: essa singela participação se deu com a disponibilização da coleção Folha-RiCol (`www.linguateca.pt/Repositorio/Folha-RIcol/`), que pretende ser um benchmark para testes de ferramentas de categorização/classificação de textos em nossa língua e foi produzida no contexto de uma dissertação de mestrado por nosso grupo (`www.inf.pucrs.br/~linatural/index.htm`).

Outra forma singular de interação ocorreu com a possibilidade de participação de alunos, que puderam realizar estadas de curta duração na Linguateca, experimentando o ritmo impresso e a energia sempre dedicada aos trabalhos.

Enfim, dez anos passados, ao olhar para traz creio que possamos afirmar: sem a Linguateca, não seríamos os mesmos.

Capítulo 16

10 anos de Linguateca - depoimento

Violeta Quental

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 16, p. 109–113, 2008

A Linguateca teve um papel pioneiro ao organizar e disponibilizar um importante acervo de recursos e ferramentas de análise computacional do português e de bibliografia correspondente. Ao longo desses últimos 10 anos, muitos foram os alunos de graduação e pós-graduação da PUC-Rio que, em algum momento de sua pesquisa, usaram os serviços da Linguateca, principalmente o projeto AC/DC (Santos e Sarmento, 2002) e o catálogo de publicações. A facilidade de uso da ferramenta de busca em corpora, o acesso gratuito, a variedade de recursos disponibilizados pelo portal foram e continuam sendo da maior importância para o ensino e para a pesquisa sobre o português. Para nós também, professores e pesquisadores, a Linguateca foi uma iniciativa extremamente útil e inspiradora – e como tal permanece, esperamos que por muito tempo ainda.

Nessa sessão de depoimentos, mais do que relatar as pesquisas feitas no âmbito da Universidade, pretendo comentar a experiência de elaboração do mini-dicionário Caldas Aulete (Geiger, 2004), em versão resumida dirigida ao público escolar. O público-alvo do dicionário definiu algumas das políticas de sua elaboração: a cobertura do léxico, de tamanho pequeno — cerca de 20.000 mil verbetes -–, com um número de acepções para cada palavra limitado aos seus usos mais frequentes e atuais; a ordem de apresentação das acepções e das propriedades sintático-semânticas dos vocábulos escolhida a partir de pesquisa em corpora; exemplos e abonações breves e em linguagem "simples".

Essa filosofia guiou a metodologia adotada pelos autores e a busca dos significados em uso, das regências verbais, dos exemplos, foi muito facilitada pelo uso dos recursos do projeto AC/DC, através de consulta por concordância e distribuição em corpora. Outros corpora foram também utilizados, especialmente o acervo literário digitalizado da editora, além da busca na rede, para busca de usos mais informais da língua, tentando evitar o viés fortemente jornalístico do CETEMPúblico (Rocha e Santos, 2000) e do NILC/São Carlos, principais recursos do AC/DC que utilizamos.

Essas escolhas de usar dados de corpus e de avaliar sua frequência já podem ser consideradas tradicionais na literatura lexicográfica. Para dicionários que se dirigem ao público escolar, aos aprendizes de uma língua, Rundell (1999) e Duran e Xatara (2006) propõe que se enfoquem os usos mais comuns, típicos e frequentes. Em nosso caso, para definir a frequência de uma determinada acepção de uma palavra, contávamos apenas com a leitura dos resultados de busca em concordância. Pode-se imaginar que, para palavras muito frequentes, essa leitura resulta em uma estimativa bastante imprecisa das acepções mais comuns. Tentamos evitar esse efeito, nos resultados do NILC/São Carlos - o mais usado por seu tamanho e por representar o português do Brasil - observando exemplos de cadernos variados do jornal: Brasil, Cotidiano, Dinheiro, Ilustrada, etc.

Já para a definição da regência verbal preferencial, por exemplo, a ferramenta oferece a busca por distribuição, que garante, em geral, resultados mais exatos. Logicamente, a teoria gramatical que fundamenta a anotação é relevante para a avaliação dos resultados.

Apresento a seguir um exemplo de busca da regência preferencial do verbo *pedir*: vê-se

que, no corpus NILC/São Carlos, a regência transitiva direta é muito mais frequente, se comparada à frequência da regência indireta.

Procura: **[lema="pedir"] [func=«ACC.*"]**.
Pedido de uma concordância em contexto
Corpus: NILC/São Carlos anotado v. 4.5
(3564 ocorrências.)

Procura: **[lema="pedir"] [func=«DAT.*"]**.
Pedido de uma concordância em contexto
Corpus: NILC/São Carlos anotado v. 4.5
(4 ocorrências.)

Observando os exemplos, podemos notar que aquilo que o corpus marca com a função dativa, no entanto, se restringe aos pronomes oblíquos. Vejamos:

Procura: **[lema="pedir"] [func=«DAT.*"]**.
Pedido de uma concordância em contexto
Corpus: NILC/São Carlos anotado v. 4.5
4 ocorrências.

Concordância

Procura: **[lema="pedir"] [func=«DAT.*"]**.

par=118459: Um porteiro veio humildemente **pedir me** que me retirasse, oferecendo me com estúpida e revoltante aparência de benignidade a vil quantia, por que eu pagara o meu bilhete; resisti e furioso disse uma injúria ao mísero porteiro.

par=119045: – **Pedes me** uma segunda luneta mágica que te será fatal como a primeira.

par=119749: De súbito chegou se a mim um mancebo com o semblante abatido, e repassado de dor, e mal podendo falar, expôs me a sua situação que era das mais pungentes sem dúvida, e acabou, **pedindo me** o óbulo da minha caridade para enterrar o filhinho, o filho único, que deixara em casa morto no colo da consternada esposa.

par=119796: Ela tinha parado e olhava me provocadora, insolente, como a **pedir me** jantar...

Assim, o exemplo 1 a seguir, em que aparece o sintagma "ao papa", não é apresentado se pedimos busca por complemento dativo, embora tenha o mesmo valor semântico de "lhe".

1. *par=1805*: Henrique IV foi, então, **pedir perdão** ao papa, em Canossa (1077).

Este resultado deriva de classificação de Bick (2000), conforme podemos observar nos exemplos abaixo, de saídas do analisador PALAVRAS:

2.
pedir [pedir] **V** INF @IMV
perdão [perdão] **IN** @<ACC
a [a] <sam-> **PRP** @<PIV
o [o] <artd> <-sam> **DET** M S @>N
papa [papa] **N** M S @P<
3.
pedir [pedir] **V** INF @IMV
perdão [perdão] **IN** @<ACC
a [a] **PRP** @<PIV
ele [ele] **PERS** M 3S NOM/PIV @P<
4.
pedir- [pedir] **V** INF @IMV
lhe [ele] **PERS** M/F 3S DAT @<DAT
perdão [perdão] **IN** @ADVL

Para o dicionário, adotamos a classificação de regência indireta para sintagmas pronominais ou sintagmas nominais preposicionados com a mesma semântica de destinatário ou beneficiário. Para termos certeza, assim, dos resultados de freqüência, teríamos de voltar à leitura dos trechos.

Outra questão interessante para a elaboração dos verbetes diz respeito à escolha de exemplos para ilustrar uma acepção. É tradição nos dicionários buscar abonações em textos literários. Nem sempre, para o aprendiz, esta é a melhor escolha, já que muitas das vezes a linguagem literária apresenta dificuldades a mais para quem desconhece o(s) significado(s) de uma palavra. Nem sempre também o exemplo do corpus é o mais esclarecedor.

Duran e Xatara (2006) discutem a questão a partir de 2 exemplos:

1. Faz mal molhar as plantas com sol quente. (Borba, 2002)
2. Molhou os pés no mar. (Houaiss, 2001)

O exemplo (1) foi extraído de corpus e o (2) elaborado por lexicógrafos. Embora não tenha nenhuma crítica a tais exemplos no contexto em que aparecem, se tivéssemos que escolher um para compor um dicionário de português para estrangeiros, o exemplo (2) seria mais adequado que o (1), uma vez que nele um dos complementos do verbo molhar (no mar) apresenta o traço (+ líquido), enquanto no exemplo (2) esse traço não aparece e a construção "molhar com sol quente" poderia gerar dúvidas no aprendiz que não conhece ainda o significado do verbo.

De modo geral, não utilizamos exemplos extraídos dos corpora por esse motivo e também pela necessidade de redigir exemplos curtos, uma restrição de formato de dicionário pequeno, mas nos inspiramos neles para construir nossos próprios exemplos.

Outra questão que surge sempre quando consultamos corpora não analisados sintática e semanticamente é a da homonímia. À época da confecção do dicionário, muitos enganos ainda apareciam nos resultados. A título de brincadeira, uso o exemplo que me foi fornecido pela Profa. Maria Carmelita Dias, de consulta para o verbo "gerar", por seu lema, que retornou vários trechos sobre o ator Richard Gere. Hoje esta mesma consulta não traria esse resultado, mas ainda aparece como erro o nome de um restaurante famoso – o Gero.

Por fim, a questão mais sensível: a decisão de ir contra a visão diacrônica, talvez contestável em outro contexto que não o de um pequeno dicionário para uso escolar. Um exemplo de decisões que tomamos: o verbo "azeitar" que, nos dicionários mais conhecidos, têm como sua primeira acepção algo como "Temperar com azeite; pôr azeite em". Esta acepção, nos corpora consultados, não apareceu uma única vez, mas é provavelmente a historicamente mais antiga. Seguindo a filosofia de obedecer ao uso, propusemos como primeira e única acepção a que surge nos corpora: "Passar óleo (em uma engrenagem, máquina etc.) ; LUBRIFICAR."

Todas as decisões que a elaboração de um dicionário envolve são difíceis e sempre há um contra-exemplo para considerar. Sem dúvida, a contribuição dos recursos de consulta a corpora da Linguateca constituíram uma ajuda inestimável para a discussão lexicográfica.

Referências

(Afonso et al., 2001) Susana Afonso, Eckhard Bick, Renato Haber e Diana Santos. Floresta sintá(c)tica: um treebank para o português. Em Anabela Gonçalves e Clara Nunes Correia, editoras, *Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística (APL 2001)*. Lisboa, Portugal. 2-4 de Outubro de 2001. p. 533–545. APL. `http://www.linguateca.pt/documentos/AfonsoetalAPL2001.pdf`. 14, 31

(Afonso et al., 2002) Susana Afonso, Eckhard Bick, Renato Haber e Diana Santos. Floresta sintá(c)tica: a treebank for Portuguese. Em Manuel González Rodrigues e Carmen Paz Suarez Araujo, editores, *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas de Gran Canaria, Espanha. 29-31 de Maio de 2002. p. 1698–1703. ELRA. `http://www.linguateca.pt/documentos/AfonsoetalLREC2002.pdf`. 14

(Agrawal e Srikant, 1994) Rakesh Agrawal e Ramakrishnan Srikant. Fast algorithms for mining association rules. Em Jorge B. Bocca, Matthias Jarke e Carlo Zaniolo, editores, *Proceedings of the 20th International Conference Very Large Data Bases (VLDB'94)*. Santiago de Chile, Chile. 12-15 de Setembro de 1994. p. 487–499. Morgan Kaufmann. 47

(Aho et al., 1988) Alfred V. Aho, Brian W. Kernighan e Peter J. Weinberger. *The AWK Programming Language*. Addison-Wesley. 1988. 5

(Aires e Aluísio, 2001) Rachel Aires e Sandra Aluísio. Criação de um Corpus com 1.000.000 de Palavras Etiquetado Morfossintaticamente. Relatório Técnico NILC-TR-01-8. Núcleo Interinstitucional de Lingüística Computacional. 2001. 66

(Aires, 2005) Rachel Virgínia Xavier Aires. *Uso de marcadores estilísticos para a busca na Web em português*. Tese de doutoramento. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Agosto de 2005. `http://www.linguateca.pt/documentos/TeseDoutRachelAires.pdf`. 76, 88

(Allan et al., 2003) James Allan, Jay Aslam, Nicholas Belkin, Chris Buckley, Jamie Callan, Bruce Croft, Sue Dumais, Norbert Fuhr, Donna Harman, David J. Harper, Djoerd Hiemstra, Thomas Hofmann, Eduard Hovy, John Lafferty Wessel Kraaij, Victor Lavrenko,

David Lewis, Liz Liddy, R. Manmatha, Andrew McCallum, Jay Ponte, John Prager, Dragomir Radev, Philip Resnik, Stephen Robertson, Roni Rosenfeld, Salim Roukos, Mark Sanderson, Rich Schwartz, Amit Singhal, Alan Smeaton, Howard Turtle, Ellen Voorhees, Ralph Weischedel, Jinxi Xu e Cheng Xiang Zhai. Challenges in Information Retrieval and Language Modeling: Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002. *SIGIR Forum*. 2: 31–47. 2003. 72

(Almeida e Simões, 2001) José João de Almeida e Alberto Simões. Text to speech – "A rewriting system approach". *Procesamiento del Lenguaje Natural*. 27:247–255. Setembro de 2001. `http://alfarrabio.di.uminho.pt/~albie/publications/text2speech.pdf`. 100, 104

(Aluísio et al., 2003) Sandra M. Aluísio, Gisele M. Pinheiro, Marcelo Finger, Maria das Graças Volpe Nunes e Stella E.O. Tagnin. The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. Em Dawn Archer, Paul Rayson, Andrew Wilson e Tony McEnery, editores, *Proceedings of the 2nd International Conference on Corpus Linguistics Corpus Linguistics (CL2003)*. Lancaster, Reino Unido. 28-31 de Março de 2003. p. 14–21. `http://www.nilc.icmc.usp.br/lacioweb/downloads/CL2003Lacio.zip`. 88

(Arasu et al., 2001) Arvind Arasu, Junghoo Cho, Hector Garcia-molina, Andreas Paepcke e Sriram Raghavan. Searching the Web. *ACM Transactions on Internet Technology*. 1:2–43. 2001. 76

(Armstrong et al., 2006) Stephen Armstrong, Marian Flanagan, Yvette Graham, Declan Groves, Bart Mellebeek, Sara Morrissey, Nicolas Stroppa e Andy May. MaTrEx: machine translation using examples. Em *TC-STAR OpenLab Workshop on Speech Translation*. Trento, Itália. 31 de Março-1 de Abril de 2006. 9

(Auer et al., 2007) Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak e Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. Em Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Diana Maynard Peter Mika, Riichiro Mizoguchi, Guus Schreiber e Philippe Cudré-Mauroux, editores, *6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007, Proceedings*. 2007. p. 722–735. Springer. 84

(Banerjee e Pedersen, 2003) Satanjeev Banerjee e Ted Pedersen. The Design, Implementation, and Use of the Ngram Statistics Package. Em Alexander Gelbukh, editor, *Proceddings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2003)*. Cidade do México, México. 16-22 de Fevereiro de 2003. p. 370–381. Springer. 66, 67

(Baptista, 2001) Jorge Baptista. *Sintaxe dos Predicados Nominais construídos com o Verbo-suporte SER DE*. Tese de doutoramento. Universidade do Algarve. 2001. 15

(Barker e Szbapakowicz, 1998) Ken Barker e Stan Szbapakowicz. Semi-Automatic Recognition of Noun Modifier Relationships. Em *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*. Montreal, Canadá. 10-14 de Agosto de 1998. p. 96–102. ACL / Morgan Kaufmann. 36

(Barreiro, 2008) Anabela Barreiro. ParaMT: a Paraphraser for Machine Translation. Em António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira e Paulo Quaresma, editores, *Proceedings of the 8th International Conference on Computational Processing of the Portuguese Language (PROPOR'2008)*. Aveiro, Portugal. 8-10 de Setembro de 2008. p. 202–211. Springer. 17

(Barreiro, 2007) Anabela Barreiro. Port4NooJ: Portuguese Linguistic Module and Bilingual Resources for Machine Translation. Em *Proceedings of the 2007 International NooJ Conference*. Barcelona, Espanha. 7–9 de Junho de 2007. Cambridge Scholars Publishing. 15

(Barreiro e Afonso, 2007) Anabela Barreiro e Susana Afonso. Construção da lista dourada para as primeiras Morfolimpíadas do português. Em Diana Santos, editora, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. p. 107–118. IST Press. 20 de Março de 2007. 14

(Belkin, 2008) Nicholas J. Belkin. Some (what) Grand Challenges for Information Retrieval. Em Craig MacDonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven e Ryen W. White, editores, *Advances in Information Retrieval, 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings*. 2008. p. 1. Springer. 72

(Bernárdez, 2005) Enrique Bernárdez. Social cognition: variation, language, and culture in a cognitive linguistic typology. Em Francisco J. Ruiz de Mendoza e Sandra Peña Cervel, editores, *Cognitive Linguistics: Internal Dynamics and Interdisciplinary Interaction*. p. 191–222. Mouton de Gruyter. 2005. 27

(Biber, 1988) Douglas Biber. *Variation across speech and writing*. Cambridge University Press. Cambridge, Reino Unido. 1988. 90

(Biber et al., 2000) Douglas Biber, Susan Conrad e Randi Reppen. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press. Cambridge, Reino Unido. 2000. 95

(Bick, 2000) Eckhard Bick. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Tese de doutoramento. Universidade de Aarhus, Dinamarca. Aarhus University Press. Novembro de 2000. 28, 32, 92, 112

(Bick et al., 2007) Eckhard Bick, Diana Santos, Susana Afonso e Rachel Marchi. Floresta Sintá(c)tica: Ficção ou realidade? Em Diana Santos, editora, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. p. 291–300. IST Press. 20 de Março de 2007. 2

(Boldi e Vigna, 2005) Paolo Boldi e Sebastiano Vigna. MG4J at TREC 2005. Em *Proceedings of the The 14th Text REtrieval Conference (TREC 2005)*. 2005. p. 4. 81

(Borba, 2002) F. S. Borba, editor. *Dicionário de usos do português contemporâneo do Brasil*. Ática. 2002. 112

(Borbinha et al.,) José Luís Borbinha, Gilberto Pedrosa, Diogo Reis, João Luzio, Bruno Martins, João Gil e Nuno Freire. DIGMAP - Discovering Our Past World with Digitised Maps. Em László Kovács, Norbert Fuhr e Carlo Meghini, editores, *Research and Advanced Technology for Digital Libraries, 11th European Conference, ECDL 2007, Budapest, Hungary September 16-21, 2007, Proceedings*. LNCS. Springer. 53

(Braga e Resende Jr, 2007) Daniela Braga e Fernando Gil Vianna Resende Jr. Módulos de Processamento de Texto Baseados em Regras para Sistemas de Conversão Texto-Fala em Português Europeu. Em *XXII Encontro Nacional da Associação Portuguesa de Linguística*. 2007. p. 141–155. APL. 100

(Braga et al., 2008) Daniela Braga, Pedro Silva, Manuel Ribeiro, Mário Henriques e Miguel Sales Dias. HMM-based Brazilian Portuguese TTS. Em António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira e Paulo Quaresma, editores, *Proceedings of the 8th International Conference on Computational Processing of the Portuguese Language (PROPOR'2008)*. Aveiro, Portugal. 8-10 de Setembro de 2008. Springer. 33

(Broder, 2002) Andrei Broder. A Taxonomy of Web Search. *SIGIR Forum*. 36(2):3–10. 2002. 76

(Busa e Johnston, 1996) Federica Busa e Michael Johnston. Qualia Structure and the Compositional Interpretation of Compounds. Em Evelyne Viegas, editora, *Proceedings of the ACL SIGLEX Workshop on Breath and Depth of Semantic Lexicons*. 1996. p. 167–187. Kluwer. 36

(Cabral et al., 2008) Luís Miguel Cabral, Diana Santos e Luís Fernando Costa. SUPeRB: Building bibliographic resources on the computational processing of Portuguese. Em *PROPOR 2008 Special Session: Applications of Portuguese Speech and Language Technologies*. Aveiro, Portugal. 10 de Setembro de 2008. `http://www.linguateca.pt/documentos/CabralSantosCostaMLDC08.pdf`. 30

(Caminada, 2008) Nuno Caminada. *Identificação Automática de Expressões Cristalizadas Preposicionais em Corpora da Língua Portuguesa*. Tese de doutoramento. Instituto Militar de Engenharia, Rio de Janeiro, Brasil. 2008. 69

(Candeias, 2007) Sara Candeias. *Vocalismo dos "clíticos", Sistema fonológico da Beira Interior e algumas considerações sintáctico-semântica*. Tese de doutoramento. Departamento de Línguas e Cultura da Universidade de Aveiro. 2007. 102

(Cardoso, 2008) Nuno Cardoso. REMBRANDT - Reconhecimento de entidades mencionadas Baseado em relações e análise detalhada do texto. Em Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca. 2008. 80

(Cardoso e Silva, 2007) Nuno Cardoso e Mário J. Silva. Query Expansion through Geographical Feature Types. Em *4th Workshop on Geographic Information Retrieval (GIR 2007)*. Lisboa, Portugal. 9 de Novembro de 2007. ACM press. `http://www.geo.unizh.ch/~rsp/gir07/`. 80, 81

(Cardoso et al., 2006a) Nuno Cardoso, Leonardo Andrade, Alberto Simões e Mário J. Silva. The XLDB Group participation at CLEF 2005 ad hoc task. Em Carol Peters, Frederic Gey, Julio Gonzalo, Henning Müeller, Gareth J.F. Jones, Michael Kluck, Bernardo Magnini e Maarten de Rijke, editores, *Acessing Multilingual information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Vienna, Austria, September 2005. Revised Selected papers*. p. 54–60. Volume 4022 de *LNCS*. Springer. 2006. 11

(Cardoso et al., 2006b) Nuno Cardoso, Bruno Martins, Leonardo Andrade, Marcirio Silveira Chaves e Mário J. Silva. The XLDB Group at GeoCLEF 2005. Em Carol Peters, Frederic Gey, Julio Gonzalo, Henning Müeller, Gareth J.F. Jones, Michael Kluck, Bernardo Magnini e Maarten de Rijke, editores, *Acessing Multilingual information Repositories:6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Vienna, Austria, September 2005. Revised Selected papers*. p. 997–1006. Volume 4022 de *LNCS*. Springer. 2006. 52

(Cardoso et al., 2007) Nuno Cardoso, Bruno Martins, Daniel Gomes e Mário J. Silva. WPT 03: a primeira colecção pública proveniente de uma recolha da web portuguesa. Em Diana Santos, editora, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. p. 279–288. IST Press. 20 de Março de 2007. 56, 76

(Cardoso et al., 2008a) Nuno Cardoso, David Cruz, Marcirio Chaves e Mário J. Silva. Using Geographic Signatures as Query and Document Scopes in Geographic IR. Em Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Doug W. Oard, Anselmo Peñas, Vivien Petras e Diana Santos, editores, *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest,*

Hungary, September 19-21, 2007, Revised Selected Papers. p. 802–810. Volume 5152 de *LNCS*. Springer. 2008. 52, 85

(Cardoso et al., 2008b) Nuno Cardoso, Mário J. Silva e Diana Santos. Handling Implicit Locality Evidence for Geographic IR. Em *Proceedings of the 17th Conference on Information and Knowledge Management (CIKM 2008)*. Napa Valley, CA, EUA. 26-30 de Outubro de 2008. ACM. 83

(Cardoso et al., 2008c) Nuno Cardoso, Patrícia Sousa e Mário J. Silva. The University of Lisbon at GeoCLEF 2008. Em *Working Notes for the CLEF 2008 Workshop*. Aarhus, Dinamarca. 17-19 de Setembro de 2008. 85

(Chacoto, 2005) Lucília Chacoto. *O Verbo Fazer em Construções Nominais Predicativas*. Tese de doutoramento. Universidade do Algarve. 2005. 15

(Chaves et al., 2005a) Marcirio Chaves, Bruno Martins e Mário J. Silva. GKB - Geographic Knowledge Base. Relatório Técnico DI/FCUL TR–05–12. Departmento de Informática, Faculdade de Ciências da Universidade de Lisboa. Julho de 2005. `http://www.linguateca.pt/documentos/gkb_technical_report.pdf`. 50

(Chaves e Santos, 2006) Marcirio Silveira Chaves e Diana Santos. What Kinds of Geographical Information Are There in the Portuguese Web? Em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira e Maria Carmelita Dias, editores, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 13-17, 2006, Proceedings*. p. 264–267. Volume 3960. Springer. 2006. `http://www.linguateca.pt/documentos/Poster_ChavesSantosPROPOR2006.pdf`. 56

(Chaves et al., 2005b) Marcirio Silveira Chaves, Mário J. Silva e Bruno Martins. A Geographic Knowledge Base for Semantic Web Applications. Em C. A. Heuser, editor, *20º Simpósio Brasileiro de Banco de Dados (SBBD 2005)*. Uberlândia, MG, Brasil. 3-7 de Outubro de 2005. p. 40–54. `http://www.linguateca.pt/documentos/ChavesetalSBBD2005.pdf`. 50, 76

(Chaves, 2008) Marcírio Chaves. Geo-ontologias para reconhecimento de relações entre locais: a participação do SEI-Geo no Segundo HAREM. Em Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca. 2008. 57

(Copestake, 2003) Ann Copestake. Compounds revisited. Em *2nd International Workshop on Generative Approaches to the Lexicon (GL'2003)*. Geneva, Suíça. 15-17 de Maio de 2003. 36

(Correia, 2006) Ana Teresa Varajão Moutinho Pereira Correia. Colaboração na constituição do corpus paralelo Le Monde Diplomatique (FR-PT). Relatório de estágio, Universidade do Minho. Dezembro de 2006. 8

(Coulthard, 1977) Malcolm Coulthard. *An Introduction to Discourse Analysis*. Longman. Londres, Reino Unido. 1977. 94

(Coulthard e Johnson, 2007) Malcolm Coulthard e Alison Johnson. *An Introduction to Forensic Linguistics: Language in Evidence*. Routledge. Londes, Reino Unido e Nova Iorque, EUA. 2007. 94, 97

(Cunha et al., 2006) João Paulo Silva Cunha, Isabel Cruz, Ilídio Oliveira, António Sousa Pereira, César Telmo Costa, Ana Margarida Oliveira e Amândio Pereira. The RTS project: Promoting secure and efective clinical telematic communication within the Aveiro region. Em *Proceedings of the eHealth 2006 High Level Conference and Exhibition*. Málaga, Espanha. 10-12 de Maio de 2006. p. 1–10. 44

(Cunningham et al., 2002a) Hamish Cunningham, Diana Maynard, Kalina Bontcheva e Valentin Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Em *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*. Filadélfia, PA, EUA. 6-12 de Julho de 2002. p. 168–175. Association for Computational Linguistics. 45

(Cunningham et al., 2002b) Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan e Cristian Ursu. The GATE User Guide. `http://gate.ac.uk/`. 2002. 45

(DGS, 2006) Direcção geral da saúde. Programa Nacional de Vacinação. 2006. `http://www.min-saude.pt/portal/conteudos/informacoes+uteis/vacinacao/vacinas.htm`. 45

(Dijk, 1997) Teun A. van Dijk. Discourse as Interaction in Society. *Discourse Studies: A Multidisciplinary Introduction - Discourse as Social Interaction*. 2:1–37. 1997. 94

(Duran e Xatara, 2006) Sanches Magali Duran e Claudia Maria Xatara. A Metalexicografia Pedagógica. *Cadernos de Tradução*. 2:41–66. 2006. `http://www.cadernos.ufsc.br/online/cadernos18/magali_xatara.pdf`. 110, 112

(Efthimiadis, 1993) Efthimis N. Efthimiadis. A User-centered Evaluation of Ranking Algorithms for Interactive Query Expansion. Em Robert Korfhage e Edie M. Rasmussen e Peter Willett, editores, *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)*. Pitsburgo, PA, EUA. 27 de Junho a 1 de Julho de 1993. p. 146–159. 81

(Efthimiadis, 1996) Efthimis N. Efthimiadis. Query expansion. *Annual Review of Information Systems and Technology*. 31:121–187. 1996. 73

(Fairclough e Wodak, 1997) Norman Fairclough e Ruth Wodak. Critical Discourse Analysis. *Discourse Studies: A Multidisciplinary Introduction - Discourse as Social Interaction*. 2: 258–284. 1997. 94

(Ferreira e Teixeira, 2008) Liliana Ferreira e António Teixeira. REMMA - Reconhecimento de entidades mencionadas do MedAlert. Em Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca. 2008. `http://linguateca.dei.uc.pt/harem/encontro/remma.pdf`. 47

(Ferreira et al., 2008) Liliana Ferreira, António Teixeira e João Paulo da Silva Neto. Ontology-driven Vaccination Information Extraction. Em *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2008)*. Barcelona, Espanha. 12-13 de Junho de 2008. p. 94–103. 44

(Ferrucci e Lally, 2004) David Ferrucci e Adam Lally. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*. 10(3-4):327–348. 2004. 45

(Frankenberg-Garcia e Santos, 2002) Ana Frankenberg-Garcia e Diana Santos. COMPARA, um corpus paralelo de português e de inglês na Web. *Cadernos de Tradução*. IX(1):61–79. 2002. `http://www.linguateca.pt/documentos/Frankenberg-GarciaSantosCadTrad.pdf`. 36

(Frankenberg-Garcia e Santos, 2003) Ana Frankenberg-Garcia e Diana Santos. Introducing COMPARA, the Portuguese-English parallel translation corpus. Em Federico Zanettin, Silvia Bernardini e Dominic Stewart, editores, *Corpora in Translation Education*. p. 71–87. St. Jerome Publishing. 2003. `http://www.linguateca.pt/documentos/Frankenberg-GarciaSantos2000.pdf`. 8, 15, 31

(Geeraerts, 2005) Dirk Geeraerts. Lectal variation and empirical data in Cognitive Linguistics. Em Francisco J. Ruiz de Mendoza e Sandra Peña Cervel, editores, *Cognitive Linguistics: Internal Dynamics and Interdisciplinary Interaction*. p. 163–189. Mouton de Gruyter. 2005. 27

(Geeraerts, 2006) Dirk Geeraerts. Methodology in Cognitive Linguistics. Em Gitte Kristiansen, Michel Achard, René Dirven e Francisco J. Ruiz de Mendoza, editores, *Cognitive Linguistics: Current Applications and Future Perspectives*. p. 21–49. Mouton de Gruyter. 2006. 27

(Geeraerts et al., 1999) Dirk Geeraerts, Stefan Grondelaers e Dirk Speelman. *Convergentie en Divergentie in de Nederlandse Woordenschat. Een onderzoek naar kleding- en voetbaltermen*. Meertens Instituut. 1999. 27

(Geiger, 2004) Paulo Geiger. *Minidicionário Caldas Aulete*. Editora Nova Fronteira. 2004. 110

(Girju et al., 2005) Roxana Girju, Dan Moldovan, Marta Tatu e Daniel Antohe. On the semantics of noun compounds. *Computer Speech and Language*. 19:479–496. Março de 2005. 36, 38

(Gonzalez-Marquez et al., 2007) Monica Gonzalez-Marquez, Irene Mittelberg, Seana Coulson e Michael J. Spivey. *Methods in Cognitive Linguistics*. John Benjamins. 2007. 27

(Gouveia et al., 2000) Paulo D.F. Gouveia, João P.R. Teixeira e Diamantino R.S. Freitas. Divisão Silábica Automática do Português Escrito e Falado. Em Maria das Graças Volpe Nunes, editora, *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*. São Paulo, Brasil. 19-22 de Novembro de 2000. p. 65–74. ICMC/USP. 100

(Green, 1979) Thomas R. G. Green. The Necessity of Syntax Markers: Two Experiments With Artificial Languages. *Journal of Verbal Learning and Behaviour*. 18:481–496. 1979. 9

(Gross, 1975) Maurice Gross. *Méthodes en Syntaxe - Régime des constructions complétives*. Hermann. 1975. 15

(Gross, 1981) Maurice Gross. Les bases empiriques de la notion de prédicat sémantique. *Formes Syntaxiques et Prédicat Sémantiques, Langages*. 63:7–52. 1981. 15

(Halliday, 1994) M. A. K. Halliday. *An Introduction to Functional Grammar*. Edward Arnold. Londres, Reino Unido. 1994. 95

(Harris, 1957) Zellig Harris. Co-occurrence and transformation in linguistic structure. *Language*. 33:293–340. 1957. 15

(Harris, 1968) Zellig Harris. *Mathematical Structures of Language*. Wiley. 1968. 15

(Hänlein, 1998) Heike Hänlein. *Studies in Authorship Recognition - A Corpus-based Approach*. Peter Lang. Francoforte, Alemanha. 1998. 94, 95

(Houaiss, 2001) Antonio Houaiss, editor. *Dicionário Houauss da Língua Portuguesa*. Editora Objectiva. 2001. 112

(Karlgren, 2000) Jussi Karlgren. *Stylistic Experiments for Information Retrieval*. Tese de doutoramento. Universidade de Estocolmo. 2000. 90

(Koehn, 2005) P. Koehn. Europarl: A parallel corpus for statistical machine translation. Em *Proceedings of the Tenth Machine Translation Summit (MT-Summit X)*. Ilha de Phuket, Tailândia. 12-16 de Setembro de 2005. p. 79–86. 8

(Kohler, 2003) Janet Kohler. Analysing Search Engine Queries for the Use of Geographic Terms. Tese de mestrado. Universidade de Sheffield, Reino Unido. 2003. 74

(Kristiansen e Dirven, 2008) Gitte Kristiansen e René Dirven. *Cognitive Sociolinguistics:Language Variation, Cultural Models, Social Systems*. Mouton de Gruyter. 2008. 27

(Lassila e Swick, 1998) Ora Lassila e Ralph Swick. Resource Description Framework (RDF) Model and Syntax. W3C, World Wide Web Consortium. 1998. `http://www.w3.org/TR/WD-rdf-syntax/`. 47

(Maia e Barreiro, 2007) Belinda Maia e Anabela Barreiro. Uma experiência de recolha de exemplos classificados de tradução automática de inglês para português. Em Diana Santos, editora, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. p. 205–216. IST Press. 20 de Março de 2007. 14

(Maia e Matos, 2008) Belinda Maia e Sérgio Matos. Corpógrafo V4 - Tools for Researchers and Teachers using Comparable Corpora. Em Pierre Zweigenbaum, Éric Gaussier e Pascale Fung, editores, *LREC 2008 Workshop on Comparable Corpora*. Marraquexe, Marrocos. 31 de Maio de 2008. p. 79–82. European Language Resources Association (ELRA). `http://www.linguateca.pt/documentos/MaiaMatosW12LREC08.pdf`. 15, 95

(Mandl et al., 2008) Thomas Mandl, Fredric Gey, Giorgio Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker e Xing Xie. GeoCLEF 2007: The CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. Em Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Doug W. Oard, Anselmo Peñas, Vivien Petras e Diana Santos, editores, *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*. p. 745–772. Volume 5152 de *LNCS*. Springer. 2008. 85

(Marques et al., 2007) Nuno C. Marques, Sebastian Bader, Vitor Rocio e Steffen Hölldobler. Neuro-Symbolic Word Tagging. Em *Proceedings of 13th Portuguese Conference on Artificial Intelligence (EPIA'07)*. Guimarães, Portugal. 3-7 de Dezembro de 2007. IEEE. 5

(Martins et al., 2007a) Bruno Martins, Nuno Cardoso, Marcirio Silveira Chaves, Leonardo Andrade e Mário J. Silva. The University of Lisbon at GeoCLEF 2006. Em Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke e Maximilian Stempfhuber, editores, *Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers*. p. 986–994. Volume 4730 de *LNCS*. Springer. 2007. 58

(Martins et al., 2007b) Bruno Martins, Mário Silva e Marcirio Chaves. O sistema CaGE no HAREM - reconhecimento de entidades geográficas em textos em língua portuguesa. Em Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. p. 97–112. 12 de Novembro de 2007. `http://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap08-SantosCardoso2007-Martinsetal.pdf`. 52

(Martins Junior e Moreira, 2004) José Martins Junior e Edson dos Santos Moreira. Using Support Vector Machines to Recognize Products in E-commerce Pages. Em *Proceedings of the 22nd IASTED International Multi-Conference on Applied Informatics*. Fevereiro de 2004. p. 212–217. 90

(McDonald, 1996) David D. McDonald. Internal and external evidence in the identification and semantic categorization of proper names. Em I. Boguraev e J. Pustejovsky, editores, *Corpus processing for lexical acquisition*. p. 21–39. 1996. 83

(McEnery e Wilson, 1996) Tony McEnery e Andrew Wilson. *Corpus Linguistics: An Introduction*. Edinburgh University Press. Edimburgo, Reino Unido. 1996. 94, 95

(McMenamin, 2002) Gerald R. McMenamin. *Forensic Linguistics: Advances in Forensic Stylistics*. CRC Press. Boca Raton e Nova Iorque, EUA. 2002. 94

(Melamed, 2001) I. Dan Melamed. *Empirical methods for exploiting parallel texts*. MIT Press. 2001. 10

(Meyers et al., 2004a) Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekeley, Veronique Zielinska e Brian Young. The Cross-Breeding of Dictionaries. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa e Raquel Silva, editoras, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*. Lisboa, Portugal. 26-28 de Maio de 2004. 15

(Meyers et al., 2004b) Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young e Ralph Grishman. Annotating noun argument structure for NomBank. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa e Raquel Silva, editoras, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*. Lisboa, Portugal. 26-28 de Maio de 2004. 15

(Mika, 2004) Peter Mika. Social Networks and the Semantic Web. Em *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*. Pequim, China. 20-24 de Setembro de 2004. p. 285–291. 74

(Mika, 2006) Peter Mika. Ontologies Are Us: A Unified Model of Social Networks and Semantics. Em V. Richard Benjamins e Mark A. Musen Yolanda Gil, editor, *The Semantic*

Web : ISWC 2005, 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, November 6-10, 2005, Proceedings (ISWC'2005). 2006. p. 522–536. Springer. 74

(Mota e Santos, 2008) Cristina Mota e Diana Santos, editoras. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: Actas do Encontro do Segundo HAREM*. Linguateca. 2008. i, 47, 57

(Nogueira, 2004) Cícero Nogueira. Algoritmo para extração de Combinações do tipo V+ (det)+N. Programa feito em Java. 2004. 67

(Och e Ney, 2004) Franz Josef Och e Hermann Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*. 30:417–449. 2004. 10

(Oksefjell e Santos, 1998) Signe Oksefjell e Diana Santos. Breve panorâmica dos recursos de português mencionados na Web. Em Vera Lúcia Strube de Lima, editora, *III Encontro para o Processamento Computacional do Português Escrito e Falado (PROPOR'98)*. Porto Alegre, RS, Brasil. 3-4 de Novembro de 1998. p. 38–47. `http://www.linguateca.pt/documentos/recursos.pdf`. 44

(Oliveira et al., 2008) Hugo Gonçalo Oliveira, Paulo Gomes e Diana Santos. PAPEL: a dictionary-based lexical ontology for Portuguese. Em António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira e Paulo Quaresma, editores, *Proceedings of the 8th International Conference on Computational Processing of the Portuguese Language (PROPOR'2008)*. Aveiro, Portugal. 8-10 de Setembro de 2008. p. 31–40. Springer. 47

(Olsson, 2004) John Olsson. *Forensic Linguistics: An Introduction to Language, Crime and the Law*. Continuum. Londres, Reino Unido. 2004. 94

(Pereira e Shieber, 1987) Fernando C. N. Pereira e Stuart M. Shieber. Prolog and Natural Language Analysis - Digital Edition. Microtome Publishing. 1987. `http://www.mtome.com/Publications/PNLA/pnla.html`. 2

(Pinheiro e Aluísio, 2003) Gisele Montilha Pinheiro e Sandra Maria Aluísio. Córpus NILC: descrição e análise crítica com vistas ao projeto Lacio-Web. Relatório Técnico NILC–-TR–03–03. Núcleo Interinstitucional de Lingüística Computacional. Fevereiro de 2003. 66

(Pustejovsky, 1995) James Pustejovsky. *The Generative Lexicon*. MIT Press. 1995. 36, 38

(Ranchhod, 1990) Elisabete Ranchhod. *Sintaxe dos predicados nominais com Estar*. INIC. 1990. 15

(Ranchhod, 2003) Elisabete Ranchhod. O Lugar das Expressões Fixas na Gramática do Português. 2003. 66

(Rocchio Jr, 1971) J. J. Rocchio Jr. Relevance Feedback in Information Retrieval. Em Gerald Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*. 1971. p. 313–323. 81

(Rocha e Santos, 2000) Paulo Alexandre Rocha e Diana Santos. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. Em Maria das Graças Volpe Nunes, editora, *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*. São Paulo, Brasil. 19-22 de Novembro de 2000. p. 131–140. ICMC/USP. `http://www.linguateca.pt/documentos/RochaSantosPROPOR2000.pdf`. 2, 31, 110

(Rodrigues e Quaresma, 1999) Irene Rodrigues e Paulo Quaresma, editores. Évora, Portugal. 20-21 de Setembro de 1999. 62

(Rundell, 1999) Michael Rundell. Recent trends in publishing monolingual learners' dictionaries. Em R. R. K. Hartmann, editor, *Thematic Network Projects, Sub-project 9 : Dictionaries - Dictionaries in Language Learning, Final Report Year Three*. p. 83–98. 1999. 110

(Ryder, 1994) Mary Ellen Ryder. *Ordered Chaos: The Interpretation of English Noun-Noun Compounds*. University of California Press. 1994. 37

(SAMPA,) Speech Assessment Methods Phonetic Alphabet. `http://www.phon.ucl.ac.uk/home/sampa/portug.htm`. 102

(Sampson, 1995) Geoffrey Sampson. *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Clarendon Press. 1995. 2

(Santos, 2007) Diana Santos, editora. *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press. 20 de Março de 2007. `http://www.istpress.ist.utl.pt/lavaliacaoconjunta.html`. 47

(Santos e Barreiro, 2004) Diana Santos e Anabela Barreiro. On the problems of creating a consensual golden standard of inflected forms in. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa e Raquel Silva, editoras, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*. Lisboa, Portugal. 26-28 de Maio de 2004. p. 483–486. `http://www.linguateca.pt/Diana/download/SantosBarreiroLREC2004.pdf`. 14

(Santos e Cardoso, 2007) Diana Santos e Nuno Cardoso, editores. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca. 12 de Novembro de 2007. `http://www.linguateca.pt/aval_conjunta/LivroHAREM/Livro-SantosCardoso2007.pdf`. 30

(Santos e Cardoso, 2008) Diana Santos e Nuno Cardoso. GikiP: Evaluating geographical answers from Wikipedia. Em *5th Workshop on Geographic Information Retrieval (GIR'2008)*. Napa Valley, CA, EUA. 30 de Outubro de 2008. ACM. 85

(Santos e Chaves, 2006) Diana Santos e Marcirio Chaves. The place of place in geographical IR. Em *3rd Workshop on Geographic Information Retrieval (GIR'2006)*. Seattle, WA, EUA. 10 de Agosto de 2006. p. 5–8. `http://www.linguateca.pt/Diana/download/acetSantosChavesGIR2006.pdf`. 56

(Santos e Costa, 2003) Diana Santos e Luís Costa. Morfolimpíadas - Apresentação detalhada da metodologia e dos problemas identificados. Em *Encontro de Avaliação Conjunta de Sistemas de Processamento Computacional do Português (AvalON'2003)*. Universidade do Algarve, Faro. 28 de Junho de 2003. `http://www.linguateca.pt/documentos/SantosMorfolimpiadasAvalon2003.pdf`. 30

(Santos e Inácio, 2006) Diana Santos e Susana Inácio. Annotating COMPARA, a grammar-aware parallel corpus. Em Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik e Daniel Tapias, editores, *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*. Génova, Itália. 22-28 de Maio de 2006. p. 1216–1221. `http://www.linguateca.pt/Diana/download/SantosInacioLREC2006.pdf`. 15

(Santos e Rocha, 2005) Diana Santos e Paulo Rocha. The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. Em Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck e Bernardo Magnini, editores, *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*. p. 821–832. Volume 3491 de *LNCS*. Springer. 2005. `http://www.linguateca.pt/documentos/SantosRochaCLEF2004Springer2005.pdf`. 58

(Santos e Sarmento, 2002) Diana Santos e Luís Sarmento. 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)*. Porto, Portugal. 2-4 de Outubro de 2002. p. 705–717. APL. `http://www.linguateca.pt/documentos/SantosSarmentoAPL2002.pdf`. 22, 26, 62, 66, 110

(Santos e Simões, 2008) Diana Santos e Alberto Simões. Portuguese-English word alignment: some experiments. Em *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. Marraquexe, Marrocos. 28-30 de Maio de 2008. European Language Resources Association (ELRA). `http://www.linguateca.pt/documentos/SantosSimoesLREC2008.pdf`. 8

(Santos et al., 2004) Diana Santos, Belinda Maia e Luís Sarmento. Gathering empirical data to evaluate MT from English to Portuguese. Em Lambros Kranias, Nicoletta Calzolari, Gregor Thurmair, Yorick Wilks, Eduard Hovy, Guðrún Magnúsdóttir, Anna Samiotou e Khalid Choukri, editores, *Proceedings of LREC 2004 Workshop on the Amazing Utility of Parallel and Comparable Corpora*. Lisboa, Portugal. 25 de Maio de 2004. p. 14–17. `http://www.linguateca.pt/documentos/SantosMaiaSarmentoAmazing2004.pdf`. 14

(Santos et al., 2006) Diana Santos, Nuno Seco, Nuno Cardoso e Rui Vilela. HAREM: An Advanced NER Evaluation Contest for Portuguese. Em Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik e Daniel Tapias, editores, *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*. Génova, Itália. 22-28 de Maio de 2006. p. 1986–1991. `http://www.linguateca.pt/Diana/download/SantosSecoCardosoVilelaLREC2006.pdf`. 56, 82

(Santos et al., 2008a) Diana Santos, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling e Yvonne Skalban. Getting geographical answers from Wikipedia: the GikiP pilot at CLEF. Em Francesca Borri, Alessandro Nardi e Carol Peters, editores, *Cross Language Evaluation Forum: Working Notes for the CLEF 2008 Workshop*. Aarhus, Dinamarca. 17-19 de Setembro de 2008. p. s/pp. `http://www.linguateca.pt/Diana/download/SantosetalWNCLEF2008.pdf`. 85

(Santos et al., 2008b) Diana Santos, Cláudia Freitas, Hugo Gonçalo Oliveira e Paula Carvalho. Second HAREM: new challenges and old wisdom. Em António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira e Paulo Quaresma, editores, *Proceedings of the 8th International Conference on Computational Processing of the Portuguese Language (PROPOR'2008)*. Aveiro, Portugal. 8-10 de Setembro de 2008. p. 212–215. Springer. 82

(Sarmento, 2006) Luís Sarmento. SIEMÊS - A Named Entity Recognizer for Portuguese Relying on Similarity Rules. Em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira e Maria Carmelita Dias, editores, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 13-17, 2006, Proceedings*. p. 90–99. Volume 3960. Springer. 2006. `http://www.linguateca.pt/documentos/SarmentoSiemes2006.pdf`. 56

(Sarmento, 2007) Luís Sarmento. Ferramentas para experimentação, recolha e avaliação de exemplos de tradução automática. Em Diana Santos, editora, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. p. 193–203. IST Press. 20 de Março de 2007. 14

(Sarmento et al., 2004) Luís Sarmento, Belinda Maia e Diana Santos. The Corpógrafo - a Web-based environment for corpora research. Em Maria Teresa Lino, Maria Fran-

cisca Xavier, Fátima Ferreira, Rute Costa e Raquel Silva, editoras, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*. Lisboa, Portugal. 26-28 de Maio de 2004. p. 449–452. `http://www.linguateca.pt/Diana/download/SarmentoMaiaSantosLREC2004.pdf`. 15, 95

(Sarmento et al., 2006) Luís Sarmento, Ana Sofia Pinto e Luís Cabral. REPENTINO - A Wide-Scope Gazetteer for Entity Recognition in Portuguese. Em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira e Maria Carmelita Dias, editores, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 13-17, 2006, Proceedings*. p. 31–40. Volume 3960. Springer. 2006. `http://www.linguateca.pt/documentos/SarmentoPintoCabral2006PROPORSpringer.pdf`. 62

(Sarmento et al., 2007) Luís Sarmento, Anabela Barreiro, Belinda Maia e Diana Santos. Avaliação de Tradução Automática: alguns conceitos e reflexões. Em Diana Santos, editora, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. p. 181–190. IST Press. 20 de Março de 2007. 14

(Seco e Cardoso, 2006) Nuno Seco e Nuno Cardoso. Detecting user sessions in the tumba! web log. Relatório técnico. Março de 2006. `http://eden.dei.uc.pt/~nseco/tumba.pdf`. 79

(Shuy, 2006) Roger W. Shuy. *Linguistics in the Courtroom: A Practical Guide*. Oxford University Press. 2006. 94

(Silberztein, 2004) Max Silberztein. NooJ: A Cooperative, Object-Oriented Architecture for NLP. *INTEX pour la Linguistique et le traitement automatique des langues*. 2004. Cahiers de la MSH Ledoux. 15

(Silva, 2005) Augusto Soares da Silva. Convergência e divergência no léxico do Português Europeu e do Português Brasileiro: resultados do estudo sobre termos de futebol e de moda. Em Joaquim Barbosa e Fátima Oliveira, editores, *Textos seleccionados do XXI Encontro da Associação Portuguesa de Linguística*. 28–30 de Setembro de 2005. p. 633–646. Colibri. 26

(Silva, 2006a) Augusto Soares da Silva. *O Mundo dos Sentidos em Português: Polissemia, Semântica e Cognição*. Almedina. 2006. 27

(Silva, 2006b) Augusto Soares Silva. Sociolinguística cognitiva e o estudo da convergência/divergência entre o Português Europeu e o Português Brasileiro. *Veredas: Revista de Estudos Lingüísticos*. 10. 2006. `http://www.revistaveredas.ufjf.br`. 27

(Silva, 2008) Augusto Soares da Silva. *Sociolinguística cognitiva, lexicologia quantitativa e variação do Português*. Tese de doutoramento. Universidade Católica de Braga. 18 de Junho de 2008. 27

(Silva, 2003) Mário J. Silva. The Case for a Portuguese Web Search Engine. Em Pedro Isaías, editor, *Proceedings of the IADIS International Conference WWW/Internet 2003 (ICWI 2003)*. Faro, Portugal. 5-8 de Novembro de 2003. p. 411–418. IADIS. 77

(Silva et al., 2006) Mário J. Silva, Bruno Martins, Marcirio Silveira Chaves, Ana Paula Afonso e Nuno Cardoso. Adding Geographic Scopes to Web Resources. *CEUS - Computers Enviroment and Urban Systems*. 30(4):378–399. 2006. 52

(Simões, 2004) Alberto Simões. Parallel Corpora word alignment and applications. Tese de mestrado. Faculdade de Engenharia da Universidade do Minho. 2004. `http://alfarrabio.di.uminho.pt/~albie/publications/msc.pdf`. 8

(Simões, 2007a) Alberto Simões. Makefile::Parallel - Uma ferramenta para paralelização de processos. Em *V Simpósio Doutoral da Linguateca 2007*. Faculdade de Ciências, Universidade de Lisboa. 4 de Outubro de 2007. `http://www.linguateca.pt/documentos/MakefileParallel-AlbertoSimoes.pdf`. 11

(Simões, 2007b) Alberto Simões. Segmentação bilingue com base na marker hypothesis. Em César Analide, Paulo Novais e Pedro Henriques, editores, *Simpósio Doutoral em Inteligência Artificial (SDIA 2007)*. Guimarães, Portugal. 3–7 de Dezembro de 2007. p. 135–144. `http://ambs.perl-hackers.net/publications/sdia07.pdf`. 9

(Simões, 2008) Alberto Simões. *Extracção de Recursos de Tradução com base em Dicionários Probabilísticos de Tradução*. Tese de doutoramento. Faculdade de Engenharia da Universidade do Minho. Março de 2008. `http://www.linguateca.pt/documentos/SimoesPHD.pdf`. 8, 10

(Simões e Almeida, 2003) Alberto Simões e José João Almeida. NATools – A Statistical Word Aligner Workbench. *Procesamiento del Lenguaje Natural*. 31:217–226. Setembro de 2003. `http://alfarrabio.di.uminho.pt/~albie/publications/sepln2003.pdf`. 8

(Simões e Almeida, 2006a) Alberto Simões e José João Almeida. Combinatory Examples Extraction for Machine Translation. Em *EAMT 11th Annual Conference*. Oslo, Noruega. 19-20 de Junho de 2006. p. 27–32. `http://alfarrabio.di.uminho.pt/~albie/publications/eamt06.pdf`. 10

(Simões e Almeida, 2006b) Alberto Simões e José João Almeida. NatServer: A Client-Server Architecture for building Parallel Corpora applications. *Procesamiento del Lenguaje Natural*. 37:91–98. Setembro de 2006. `http://alfarrabio.di.uminho.pt/~albie/publications/sepln06.pdf`. 11

(Simões e Almeida, 2007) Alberto Simões e José João Almeida. Parallel Corpora based Translation Resources Extraction. *Procesamiento del Lenguaje Natural*. 39:265–272. Setembro de 2007. `http://www.sepln.org/revistaSEPLN/revista/39/32.pdf`. 11

(Simões e Almeida, 2008) Alberto Simões e José João Almeida. Bilingual Terminology Extraction based on Translation Patterns. *Procesamiento del Lenguaje Natural*. 41:281–288. Setembro de 2008. `http://alfarrabio.di.uminho.pt/~albie/publications/sepln08.pdf`. 10

(Sinclair, 1991) John M. Sinclair. *Corpus, Concordance, Collocation*. Oxford University Press. 1991. 94

(Singhal, 2008) Amit Singhal. Web Search: Challenges and Directions. Em Craig MacDonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven e Ryen W. White, editores, *Advances in Information Retrieval, 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings*. 2008. p. 2. Springer. 72

(Sánchez-Martínez e Forcada, 2007) Felipe Sánchez-Martínez e Mikel L. Forcada. Automatic induction of shallow-transfer rules for open-source machine translation. Em *Proceedings of the The 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI'2007)*. Skövde, Suécia. 2007. p. 181–190. 10

(Sousa Silva, 2006) Rui Sousa Silva. Performance and Individual Act Out: The Semantics of (Re)Building and (De)Constructing in Contemporary Artistic Discourse. Tese de mestrado. Faculdade de Letras da Universidade do Porto. 2006. 95

(SPEECHDAT, 1998) Portuguese SpeechDat (II) FDB-4000, European Language Resources Association. 1998. `http://www.elda.org/catalogue/en/speech/S0092.html`. 100, 101

(Stamatatos et al., 2000) Efstathios Stamatatos, George Kokkinakis e Nikos Fakotakis. Automatic text categorization in terms of genre and author. *Computational Linguistics*. 26: 471–495. 2000. 92

(Steinberger et al., 2006) Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis e Dániel Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. Em Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik e Daniel Tapias, editores, *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*. Génova, Itália. 22-28 de Maio de 2006. p. 2142–2147. 8

(Teixeira et al., 2006) António Teixeira, Catarina Oliveira e Lurdes Moutinho. On the Use of Machine Learning and Syllable Information in European Portuguese Grapheme-Phone Conversion. p. 212–215. Volume 3960. Springer. 2006. 100

(TIGER, 2007) TIGER PROJECT Linguistic Interpretation of a German Corpus. 2007. `http://www.ims.uni-stuttgart.de/projekte/TIGER/`. 5

(Uzeda Garrão, 2006) Milena de Uzeda Garrão. *O Córpus não mente jamais: Sobre a identificação e uso de combinações multivocabulares do tipo "verbo + sintagma nominal"*. Tese de doutoramento. Pontifícia Universidade Católica do Rio de Janeiro. 2006. 66

(Uzeda Garrão e Dias, 2006) Milena de Uzeda Garrão e Carmelita Dias. The corpus never lies: a statistical approach for the identification of verbal collocations. Em *Proceedings of Collocations and Idioms 1: Papers from the First Nordic Conference on Syntactic Freezes*. Joensuu, Finlândia. 19-20 de Maio de 2006. 66

(Viana e Andrade, 1985) Maria do Céu Viana e Ernesto de Andrade. CORSO I: um conversor de texto ortográfico em código fonético para o português. Relatório Técnico n.º 6. Grupo de Fonética e Fonologia, Centro de Linguística da Universidade de Lisboa. 1985. 100

(Wielemaker, 2008) Jan Wielemaker. SWI-Prolog 5.6 Reference Manual. 2008. `http://gollem.science.uva.nl/cgi-bin/nph-download/SWI-Prolog/refman/refman.pdf`. 2

Índice