Uma incursão pelo universo das publicações em Portugal

Diana Santos
Linguateca, FCCN & Universidade de Oslo
d.s.m.santos@ilos.uio.no

Fernando Ribeiro
Linguateca, FCCN
fernando.ribeiro@fccn.pt

Resumo

Neste artigo descrevemos um projeto de colaboração entre a Linguateca e o RCAAP (Repositório Científico de Acesso Aberto de Portugal) no sentido de determinar a possibilidade de melhorar a procura no meta-repositório deste último com ferramentas de processamento da língua portugesa. Após uma breve apresentação do projeto e da sua motivação nas duas primeiras secções, na secção 3 descrevemos a quantidade de procuras a que tivemos acesso, e nas quais baseamos o estudo, assim como fazemos uma descrição do material depositado no repositório com base em oito recolhas diferentes, no que se refere ao nome dos autores. Prosseguimos descrevendo a análise e processamento dos nomes dos autores (limpeza, normalização e agrupamento), assim como a análise da população de autores nos metadados e nas procuras nas duas secções seguintes, 4 e 5. Com isso identificamos uma série de possíveis grupos de autores, e descrevemos alguns problemas encontrados. Na secção 6, a mais importante do artigo, analisamos as sessões – ou seja, sequências de procuras feitas por um mesmo utilizador a interagir no portal – para verificar se há variação, correção e alteração no nome dos autores dentro de uma sessão. As secções seguintes, 7 e 8 referem-se a assuntos relacionados com a procura em repositórios de publicações, sobre os quais se fizeram pequenas experiências piloto no âmbito do presente projeto, e que permitem ilustrar o quanto ainda estamos aquém de utilizar robustamente quer correção ortográfica quer análise de citações em ambientes realistas, mas que indicam caminhos a seguir. Acabamos a apresentação com uma discussão de possíveis formas de prosseguir, após abordar levemente trabalho relacionado na secção 9.

1 Apresentação

O projeto RCAAP[1] tem como objetivo permitir um fácil acesso às publicações públicas em Portugal. A Linguateca[2], formalmente localizada na mesma instituição infra-estrutural portuguesa, a Fundação Científica para a Computação Nacional (FCCN), iniciou um projeto conjunto com o RCAAP para investigar se as funcionalidades de processamento computacional do português podiam melhorar a experiência dos utilizadores do RCAAP e contribuir para implementar um melhor serviço.

A Linguateca (Santos, 2009), aliás, há vários anos que tem tentado incentivar a publicação em português, e tem dado apoio à catalogação e publicitação de textos na área, através entre outras coisas do SUPeRB, um sistema de catalogação de publicações com várias funcionalidades pensadas para a língua portuguesa (Cabral, 2007; Cabral, Santos e Costa, 2008).

Neste artigo apresentamos, em relação ao projeto RCAAP, e do ponto de vista da Linguateca,

- o tipo de perguntas a que tentamos responder
- uma descrição do material do RCAAP em termos de nomes de autores
- o que pretendemos fazer em futuras iterações, se for considerado útil pelos financiadores e pelos nossos parceiros no RCAAP.

2 Algumas notas de motivação

A nossa intenção primordial com uma colaboração entre a Linguateca e o RCAAP era estudar e analisar os possíveis problemas que o portal do RCAAP e a procura em geral no material por ele indexado poderia apresentar, e investigar a possível melhoria que ferramentas de processamento da língua portuguesa poderiam oferecer.

Ao contrário de afirmar peremptoriamente que faríamos isto e aquilo e que tal constituiria uma melhoria apreciável, preferimos analisar primeiro a situação e os serviços já oferecidos pelo projeto RCAAP e investigar se o nosso saber-fazer poderia de facto trazer alguma melhoria substancial.

Por nos parecer mais simples e mais rela-

[1] Repositório Científico de Acesso Aberto de Portugal, ver http://www.rcaap.pt.

[2] http://www.linguateca.pt

Anterior (1)	presente (2)	autores1	autores2	comuns	perdidos	novos
Jun 2010	15 Fev 2011	38532	42011	30230	8302	11781
15 Fev 2011	1 Mar 2011	42011	43977	41622	389	2355
1 Mar 2011	12 Abr 2011	43977	52541	43037	940	9504
12 Abr 2011	19 Maio 2011	52541	54705	52323	218	2382
19 Maio 2011	19 Jul 2011	54705	57874	54222	483	3652
19 Jul 2011	1 Ago 2011	57874	58425	57747	127	678
1 Ago 2011	5 Set 2011	58425	46003	45557	12868	446
5 Set 2011	1 Out 2011	46003	48473	43383	2620	5090

Tabela 1: Evolução do número e identidade dos autores distintos no RCAAP, comparando cada recolha com a recolha anterior

cionado com a língua portuguesa (dado que um número relativamente elevado de publicações no RCAAP estão em inglês[3]), começámos por estudar o universo dos autores, ou melhor dos seus nomes, para ver se era possível atribuir com alguma certeza várias identificações a um mesmo autor, desambiguar nomes de autores e corrigi-los. Outra motivação para a concentração específica em nomes de autores foi o facto de a correta identificação e desambiguação de pessoas na rede era na altura (e continua a ser) uma área de investigação também a nível internacional, como por exemplo o WePs (Artiles, Sekine e Gonzalo, 2008; Artiles, Gonzalo e Sekine, 2009) demonstra.

Embora a Linguateca tenha considerável experiência na área do reconhecimento de entidades mencionadas em português, devido sobretudo à organização de duas edições do HAREM – a primeira (Santos e Cardoso, 2007) e a segunda (Mota e Santos, 2008), o problema que tratamos aqui não é de facto esse: não estamos a tentar identificar que cadeias de caracteres são nomes de autores – esse é um dado da interação. Estamos sim a lidar com os problemas contíguos que são o de normalizar e validar presumíveis nomes de autores e o de melhorar a procura num universo de autores. Veja-se a secção 9 para uma discussão dos vários problemas relacionados.

3 Dados do projeto

O RCAAP agrega o material oriundo de várias bases de publicações, e tem um meta-repositório que as une e coordena, e que reúne todos os meses uma nova lista de metadados do conteúdo global no RCAAP.

Além disso, e como está mencionado na página do RCAAP que citamos em seguida, recolhe o conteúdo para facilitar o acesso.

> O RCAAP é portal agregador (meta-repositório) que reúne a descrição (metadados) dos documentos depositados nos vários repositórios institucionais em Portugal. O Portal RCAAP recolhe o texto integral para melhorar o resultado das pesquisas mas não guarda qualquer documento.
>
> Para além de poder pesquisar a produção científica portuguesa, pode optar por pesquisar também a produção científica brasileira que neste momento é composta por vários repositórios e revistas agregados no projecto OASIS.[4]
>
> (`http://www.rcaap.pt/help.jsp`)

Associado a esse meta-repositório existe um diário das pesquisas nele feitas (cujo desenho foi revisto por nós), a que nós temos acesso mensalmente. Não temos contudo acesso às pesquisas feitas em cada repositório, visto que isso é gerido (e guardado ou não) de forma distribuída pelos gestores dos diferentes repositórios.[5]

Numa primeira fase (cujos detalhes podem ser apreciados nos relatórios de Santos e Ribeiro (2010; Santos e Ribeiro (2011)), concentrámo-nos nos autores das publicações. Convém aliás indicar que todos estes dados se encontram disponíveis do sítio `http://www.linguateca.pt/colabRCAAP`, onde são atualizados todos os meses.

Na tabela 1 apresentamos a quantidade de dados que temos, assim como a evolução em termos de número de autores (diferentes) no RCAAP. Convém talvez explicar que, se os nomes de autores também desaparecem (quando

[3]Cerca de um terço, com dois terços em português, as outras línguas tendo um peso muitíssimo reduzido. No OASIS brasileiro (o congénere do RCAAP), para comparação, a proporção é de sete vezes mais publicações em português do que em inglês.

[4]Nota dos autores do artigo: OASIS é o OASIS.BR - Portal Brasileiro de Repositórios e Periódicos de Acesso Livre, IBICT, `http://oasisbr.ibict.br/`.

[5]Sobre a forma como o projeto RCAAP funciona a nível nacional e internacional, consulte-se a documentação produzida por este (Moreira et al., 2010; Carvalho et al., 2010), assim como o sítio do projeto.

se esperaria que apenas aumentassem ao longo do tempo), isso é devido à dinâmica da publicação de acesso aberto, em que aparentemente artigos aceites em revistas ou livros deixam de poder ser obtidos em repositórios de acesso aberto.

Estes valores provêm da análise automática que fazemos aos metadados que o RCAAP nos faculta, não podendo nós saber, por exemplo, a causa da diminuição drástica de nomes de autores – e do consequente previsível decrescimento de objetos no RCAAP – entre Agosto e Setembro do presente ano de 2011.

Mas de qualquer maneira ilustram o tamanho do universo com que temos estado a lidar, assim como a dinâmica no repositório, em que mensalmente podemos dizer que se ganham geralmente dez vezes mais autores do que se perdem, mas que este último número ainda é significativo, correspondendo por exemplo no mês de Outubro de 2011 a 5,7%.

4 Agrupamento de nomes de autores

Para estudar a possibilidade de confusão entre nomes de autores diferentes, e os vários grupos que estes poderiam constituir entre si, levámos a cabo dois tipos de agrupamento diferentes.

Antes disso, procedemos a um processo extensivo de normalização e limpeza de dados, descrito pormenorizadamente em Santos e Ribeiro (2010), de que damos aqui um lamiré:

Por exemplo, nomes de autores com termos como *Universidade*, *colóquio* ou *European* são descartados; números ou datas são removidos; são adicionados pontos a iniciais sem ponto, e nomes com vírgulas são invertidos de forma a obter o nome numa forma canónica, como ilustrado na figura 1.

```
Uzan, C        -> C. Uzan
Colaço, ML     -> M. L. Colaço
Correia, P.J. -> P. J. Correia
Vacas, Joana Malta, 1977
               -> Joana Malta Vacas
Saldanha, L.[1937-1987]
               -> L. Saldanha
Falcão, Amílcar Celta
               -> Amílcar Celta Falcão
Colin, J.-P.   -> J. P. Colin
```

Figura 1: Exemplos de normalização

Depois criámos, no processo de **ambiguação**, todas as variantes possíveis de citar um dado nome em português, através da transformação em iniciais de todos ou apenas alguns constituintes (exceto o último nome), e através da omissão de partículas do nome tais como *e*, *de*, *dos*, etc., assim como a expansão ou abreviação de nomes como *Maria*, *Filho*, e *Júnior* ou *Junior* (*M.ª*, *F.*, *Jr.*) ou títulos incluídos no nome como *Pe.* (*Padre*). Nesse processo, nomes com hífens foram também tratados especialmente.

Para dar uma ideia do impacto desta fase de processamento, mencione-se que, dos 48.473 nomes de autores diferentes nos metadados do RCAAP (no mês de Outubro de 2011), após a normalização ficámos com 47.327, que foram ambiguados para 327.614, ou seja, aumentaram o seu número de um factor de 7.

Procedemos em seguida a dois tipos de agrupamento.[6]

Por um lado, transformámos todos os (nomes dos) autores na sua identificação mínima, ou seja, a inicial do primeiro nome seguido do último nome (que é infelizmente a tradição inglesa), criando assim o que chamámos **grupos de nomes máximos** no sentido de que todos os autores cobertos por esta "inicialização compulsiva" pertencem ao mesmo grupo. Obtivemos assim para o mês de Outubro de 2011, 20.970 grupos de nomes máximos, com uma média de 15,48 elementos por grupo, em que a distribuição pelo tamanho dos grupos se apresenta na tabela 2:

Tamanho do grupo	Número de grupos
1	5342
2	7543
3	374
4 (ou mais)	7898

Tabela 2: Grupos de nomes máximos

Por outro lado, criámos outra forma de agrupamento a que chamámos **grupos de nomes únicos**, em que pretendemos na medida do possível separar (e como tal identificar) univocamente autores diferentes[7]. Nesse caso, o identificador do grupo é a sequência máxima de carateres que corresponde à maneira mais precisa de identificar uma pessoa, e o resto dos membros do grupo têm de ser versões compatíveis (mas mais ambíguas) dessa cadeia de carateres.

Assim, exemplificando com o nome (nome de grupo) único José João Dias de Almeida, casos como *João David de Almeida* não lhe pertencem, mas *J. D. Almeida* pertence. Pertence, aliás, a ambos os grupos de nomes únicos

[6] De facto, na realidade fizemos três, mas não temos ainda os dados referentes aos terceiro, mais radical, em que também removemos nomes e/ou iniciais.

[7] Naturalmente que estamos conscientes de, que sendo este um processo automático baseado em muito pouco conhecimento, estamos longe de chegar a esse objetivo, como será discutido mais à frente no artigo.

José João Dias de Almeida e João David de Almeida. Todos estes nomes, por outro lado, estão agrupados no mesmo grupo de nomes máximo J. Almeida.

Tal permite-nos apontar as seguintes propriedades, distintas, destes dois tipos de agrupamento: cada nome só pode pertencer a um grupo de nomes máximo, que é de certa forma o mínimo múltiplo comum de todos esses nomes; por outro lado, quanto menos comprido e complicado um nome de autor for, mais provavelmente se poderá integrar em mais grupos de nomes únicos. Assim a caracterização estatística destes dois tipos de grupos dá-nos diferentes propriedades do que convencionámos chamar universo dos autores.

Apresentamos aqui para comparação também a distribuição, em termos de tamanho, dos 38.008 grupos de nomes únicos relativos a este mesmo mês, com uma média de 9,75 nomes por grupo, na tabela 3.

Tamanho do grupo	Número de grupos
1	5.524
2	11.380
3	1.006
4 (ou mais)	20.098

Tabela 3: Grupos de nomes únicos

Isto significa que em Outubro de 2011 havia 5.524 grupos de nomes únicos com apenas um elemento (ou seja, com uma designação que não é ambígua nesse universo), e que existiam 32.484 grupos com mais de um membro, possivelmente (idealmente) representando diferentes maneiras de a mesma pessoa assinar.

(Se, para efeitos de comparação, fizermos o mesmo processo aos autores brasileiros, e a todos os autores – brasileiros e portugueses – obtemos os dados das tabelas 4 e 5.)

Tamanho	Grupos em B	Grupos em B e P
1	1.960.294	2.248.567
2	111.543	12.137
3	27.801	31.508
4 (ou mais)	45.729	53.601

Tabela 4: Grupos de nomes únicos B e P

Tamanho	Grupos em B	Grupos em B e P
1	9.951	13.956
2	38.928	44.134
3	2.001	2.341
4 (ou mais)	88.359	92.007

Tabela 5: Grupos de nomes máximos B e P

Apresentamos também, para cada nome no universo dos autores, a sua ambiguidade potencial em termos dos agrupamentos de nomes únicos, ou seja, para cada nome (dos nomes que temos no nosso universo dos autores), a quantos grupos pode pertencer, na tabela 6.

Número de grupos com que emparelham	Número de casos
1	313.299
2	8016
3	2280
4 (ou mais)	3830

Tabela 6: Ambiguidade relativa a grupos de nomes únicos

Embora possamos fazer várias considerações (e estatísticas) sobre este assunto, na realidade o nosso objetivo é ver como é que o material depositado no RCAAP corresponde (ou não) ao que os utilizadores procuram (e encontram ou não).

Por isso na próxima secção apresentamos o mesmo processo (de normalização), agora referente aos nomes encontrados nos diários da procura, antes de voltar a possíveis interpretações dos valores providenciados.

5 Emparelhamento simples com as procuras

Antes de descrever este emparelhamento, apresentamos sucintamente o material de que dispomos, e que provém da análise dos diários expressamente criados para o efeito pela equipa do RCAAP (cujo formato, mais uma vez, pode ser consultado nos relatórios já mencionados).

A distribuição das consultas (desde que a elas tivemos acesso) encontra-se na figura 2, separadas entre procuras simples e avançadas (neste último caso, os utilizadores especificam campos como autor, assunto, etc).

Vemos que a maior parte dos utilizadores não usa a procura avançada, por campos, mas sim a procura simples, onde é possível também digitar autores. Por agora limitámo-nos à pesquisa avançada para obter sem esforço os autores procurados, embora uma extensão óbvia da nossa investigação seja tentar obter nomes de autores na expressão da pesquisa simples.[8]

Passemos agora ao assunto que nos interessa, nomeadamente: Quão alinhado está o universo

[8]Isto não é tão simples quanto pode parecer à primeira vista, visto que muitos apelidos em português podem ser também nomes comuns, veja-se *oliveira* ou *mota*, e que um utilizador pode procurar um autor com um nome composto "Mota Oliveira" ou um artigo com dois autores, um designado por "Mota" e outro por "Oliveira".

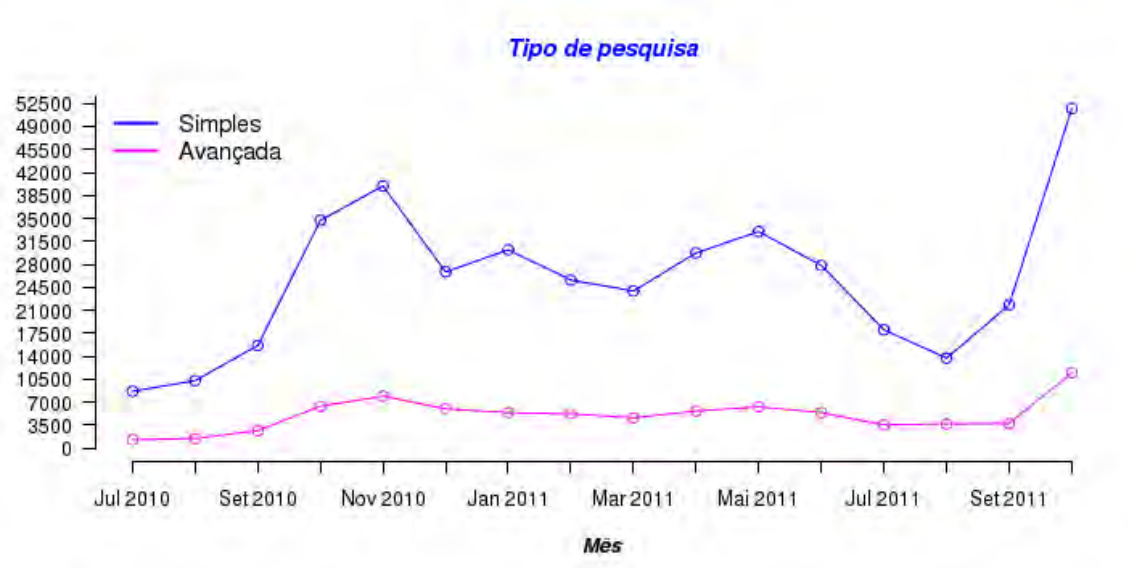

Figura 2: Número de procuras por mês

dos autores no RCAAP com o que os utilizadores procuram? Ou seja, os autores que os utilizadore procuram são os que se encontram no repositio? Para averiguar isso, começamos por caracterizar os autores procurados, verificando em que grupos se podem incluir.

Os resultados – relativos às procuras de Julho de 2010 até Outubro de 2011 inclusive – encontram-se na tabela 7.[9]

Número de grupos	Número de casos
1	3304
2	486
3	244
4 (ou mais)	1263

Tabela 7: Emparelhamento com grupos de nomes únicos

A tabela seguinte (tabela 8) mostra os mesmos valores, agora em relação aos grupos de nomes máximos:

Em primeiro lugar, apontamos que os casos concretos que correspondem a apenas um grupo – os casos em que há apenas um candidato (ambiguado ou não), e que remontam a 3304 (63%) e 3487 (66%) respetivamente – coincidem na grande maioria dos casos, nomeadamente em 3266, ou seja em 98%/94% dos casos.

[9]Poderíamos empregar vários métodos de emparelhamento, como se vê, e dá a escolher, na página do serviço, mas para o presente artigo usámos o método a que chamámos **sequência exata**, e que apenas emparelha quando a sequência procurada existe sem interrupções no grupo. Ou seja, a procura de *Fernando Ribeiro* emparelha com *José Fernando Ribeiro*, mas não com *Fernando José Ribeiro*.

Número de grupos	Número de casos
1	3487
2	527
3	240
4 (ou mais)	956

Tabela 8: Emparelhamento com grupos de nomes máximos

Tentemos explicar, com um exemplo concreto, as razões da semelhança entre os dados apresentados para cada tipo de grupos:

Considere-se o caso do nome fictício *Zacarias da Zelândia Martelo Rocambolesco*, que assumimos que, quer no grupo de nomes máximo Z. Rocambolesco, quer no grupo de nomes único Zacarias da Zelândia Martelo Rocambolesco, só apresenta um elemento, esse próprio nome (ou variantes inicializadas do mesmo).

Pode-se imaginar que, num universo sempre em expansão, um dia outros autores com o primeiro nome começado pela letra Z e com o mesmo apelido virão a surgir, mas sendo este um trabalho virado para a realidade concreta e para uma situação presente e não para um universo tendente para o infinito, podemos afirmar que existirão também sempre casos não ambíguos no material, de certa forma corroborando mais uma vez a lei de Zipf (Zipf, 1949).

Seja como for, e para investigar se isso se deve a que a grande maioria das procuras por autor tem um comprimento grande (ou pequeno), tabelámos o comprimento dos nomes procurados, na tabela 9.

Ao fazer esta análise, detetámos que muitas

Palavras	Iniciais	Frequência
2	0	3822
3	0	1915
1	0	1784
4	0	1343
5	0	777
1	1	434
6	0	303
1	2	223
2	1	198
7	0	114
2	2	100
1	3	44
3	1	38
4	1	32
8	0	27
3	2	25
2	3	17
9	0	10

Tabela 9: Forma dos autores procurados: não descartamos quaisquer palavras, tal como *de* ou *e*.

das cadeias de carateres enviadas como "autores" o não são: de facto, ao analisar manualmente os primeiros 2044 casos de autores não encontrados (ordenados por ordem alfabética), encontrámos 258 casos que não eram certamente autores, e que na sua esmagadora maioria eram palavras chave.

Para indagar por outro lado até que ponto é que as procuras por autores foram bem sucedidas, precisamos de saber o número de resultados que desencadearam, e qual o comportamento do utilizador em face dos mesmos. Para isso precisamos de estudar a interação com o sistema, olhando para as sessões.

Mas antes disso, apresentamos os dados relativos às procuras no meta-repositório: das 543.684 que temos conhecimento, removendo os duplicados correspondem a 542.545 pesquisas diferentes, às quais retirámos além disso as poucas procuras automáticas (feitas por robôs), resultando em 491.376 procuras.

Dessas, apenas 65.018 (13%) não têm nenhum resultado.

Se limitarmos a nossa atenção às procuras que preencheram o campo autor (e que no total são 28.644), as que não obtiveram nenhum resultado são 11.506, ou seja, 40%.

Mas se estes valores correspondem a satisfação do utilizador, ou se a maior parte dos resultados são lixo ou pelo menos precisam de refinamento, é algo não podemos ajuizar considerando cada procura separadamente.

6 Estudo das sessões

Como matéria prima para o nosso estudo temos acesso (dados de 1 de Novembro de 2011) a 85.398 sessões diferentes, ou seja, visitas com mais de uma pesquisa); a 95.448 "utilizadores" (definidos como par IP mais browser) diferentes, representando 132.442 visitas ao RCAAP. Por **visitas** consideramos uma ou mais procuras, a que chamamos respetivamente **visitas unitárias** (46.044), ou **sessões** (com mais de uma pesquisa, portanto). Chamamos **visitas únicas** (que podem ter uma ou mais pesquisas) aos casos em que um utilizador que só comunicou uma vez com o meta-repositório do RCAAP (79.596).

Como talvez não seja preciso salientar, num sistema que não tem autenticação de utilizadores é muito difícil individualizar estes, dados os IP dinâmicos e as "proxies" das instituições, o que significa que, se conseguimos com uma certa precisão identificar sessões, o mesmo não pode ser dito de utilizadores, que repetimos, são meramente pares distintos de IP e identificador do browser.[10]

Por isso os dados mais fiáveis que temos dizem respeito às sessões, veja-se na figura 3 a distribuição das sessões em número de procuras.

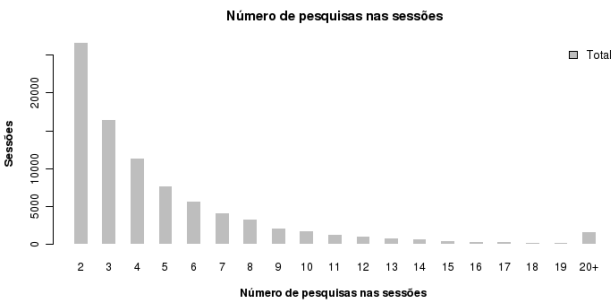

Figura 3: Número de procuras por sessão

As questões a que queremos responder, com uma análise das sessões (ou melhor, com uma análise do que os utilizadores fazem nas suas sessões), é, após consultarmos a distribuição dos resultados obtidos, conseguir pronunciar-nos sobre se o resultado é bom ou mau, ou seja, se os utilizadores conseguiram ou não obter o que pretendiam, e com quanto esforço, através de dados (indiretos) como

- Quantas vezes se refinou (até chegar a algo satisfatório, ou desistir)
- Se não se obteve nenhum resultado, isso é porque a procura foi mal feita ou porque não

[10] Para uma defesa da nossa abordagem de sessões, ver Santos e Frankenberg-Garcia (2007), por sua vez baseado em He e Göker (2000).

existia no RCAAP?

- Se se obteve resultados, mas houve mudanças à procura feita pelo utilizador, o resultado da mesma foi no sentido de aumentar, ou diminuir o número de casos encontrados?

Nas tabelas 10 e 11 contabilizamos os casos encontrados em que é possível apreender uma modificação a uma procura, respetivamente no caso de a sessão ter pesquisas em que exista o campo autor, e em todas as sessões. (A unidade são pares de consultas dentro de uma mesma sessão, e a heurística utilizada para considerar modificação em vez de nova procura não relacionada foi a existência de algo constante em duas procuras consecutivas.)

Mudança, número de resultados	Frequência
PS para PA, diminuição	5125
PS para PA, aumento	208
PA para PS, aumento	1763
PA para PS, diminuição	69
PA, aumento	2811
PA, diminuição	6762
PA, mesmo número	3081
PS para PA, mesmo número	75
PA para PS, mesmo número	37

Tabela 10: Resultado de mudanças na expressão de pesquisa em sessões incluindo o campo Autor (PS: pesquisa simples; PA: pesquisa avançada)

Por aqui vemos que na maior parte dos casos os utilizadores estão interessados, ou acabam por, obter menos resultados dos que inicialmente obtiveram, e que existe uma correlação grande entre mudar para a pesquisa avançada para restringir os resultados, e, conversamente, mudar para a pesquisa simples para obter mais resultados. Contudo, é de salientar que uma percentagem significativa de vezes os utilizadores não conseguem modificar efetivamente a pesquisa. Vemos que as generalizações descritas acima têm confirmação num universo maior, com a observação adicional de que alterações dentro da pesquisa simples tendem a ser igualmente frequentes para aumentar ou para diminuir o número de resultados, o que não seria evidente.

Convém contudo salientar que as contagens apresentadas referem-se a pares de consultas, sendo de imaginar que em alguns casos a interação à procura da pesquisa perfeita seja mais numerosa, e que o utilizador tente mais do que uma modificação, ou que apenas desista após variadas tentativas, e que dessa forma alguns dos pares não são independentes, mas sim deveriam ser contados como triplos, quádruplos, n-tuplos de modificação.

Mudança, número de resultados	Frequência
PS para PA, diminuição	39547
PS para PA, aumento	5313
PA para PS, aumento	8744
PA para PS, diminuição	1253
PA, aumento	33689
PA, diminuição	54415
PA, mesmo número de resultados	23815
PS para PA, mesmo número	512
PA para PS, mesmo número	1574
PS, aumento	12328
PS, diminuição	12833
PS, mesmo número	18268

Tabela 11: Mudanças na expressão de pesquisa em geral

Refizemos assim as contagens. No caso de sessões (de pesquisa avançada) com nome de autor, verificámos quantas vezes este foi mudado (mas não radicalmente alterado), em casos que não são apenas pares mas sim sequências de duas, três ou mais pesquisas.

Em termos das sessões correspondentes, na tabela 12 mostramos o tamanho em número de pesquisas consecutivas de uma modificação à pesquisa.

Número de modificações	Frequência
1	1375
2	532
3	247
4	99
5	61
6	30
7	18
8 (ou mais)	46

Tabela 12: Tamanho da interação à volta de uma expressão de pesquisa incluindo o campo Autor

E na tabela 13 indicamos, em termos do resultado (do primeiro para o último caso), os casos em que houve aumento dos resultados, diminuição ou manteve-se o número. Aqui como nos resultados anteriores, é a redução do número

Mudança nos resultados	Frequência
Aumento	314
Diminuição	1447
Mesmo número	647

Tabela 13: Resultado da interação à volta de uma expressão de pesquisa incluindo o campo Autor

de resultados que parece ser o objetivo mais importante da reformulação de consultas.

A um nível mais geral, no entanto, o que queremos compreender, é, além de saber quantos autores são ambíguos (ou seja, passíveis de terem várias interpretações), que foi o que tentámos averiguar na secção anterior,

- quantas vezes os autores "ambíguos" têm de ser re-procurados, ou seja a sua especificação refinada na expressão da procura?
- qual é a melhor maneira de os distinguir para um utilizador? Por assunto(s), por data das publicações, pela instituição associada à publicação, pelos seus co-autores?

Para isso, tivemos de observar individualmente uma amostra de casos (pares de consultas relacionadas, em sessões com campo autor) em que houve diminuição de resultados, para ver se essa redução se deveu a uma maior desambiguação do nome do autor. A distribuição das diferentes ações após analisar 60 casos é apresentada na tabela 14. Em alguns (poucos) casos, mais do que uma destas modificações foi executada pelo utilizador. Nesses casos, contabilizámos cada mudança separadamente.

Mudança na pesquisa	Frequência
Aumento no nome do autor	5
Adição de autor	26
Adição de autor adicional	6
Adição de tipo ou data	12
Mudança para o campo autor	8
Adição de assunto	3
Mudança no assunto	4
Diminuiçao de termos	3

Tabela 14: Análise fina das alterações à pesquisa que levaram a uma diminuição de resultados

Embora o número de casos seja pequeno demais para permitir generalizações, é interessante realçar que há diferenças apreciáveis entre colocar um nome de autor na pesquisa simples ou precisá-lo (por vezes repetindo-o) como nome de autor: os oito casos analisados demonstram reduções de 594 para 31, 86 para 34, 37 para 3 e 97 para 4. Isto parece indicar que um serviço que identificasse cadeias de carateres na procura simples como possíveis autores (constando no repositório), e apresentasse primeiro essa resposta, eventualmente seguida de procura em texto livre, poderia imediatamente evitar algumas reformulações.

Por outro lado, também é interessante verificar que na maior parte das procuras analisadas os utilizadores pareciam ter grande conhecimento do que estavam à procura (sabendo nomes de co-autores, assuntos e mesmo datas). Talvez por isso apenas oito casos redundaram em resultados nulos.

Outra coisa que nos surpreendeu, foi o facto de muitas vezes os utilizadores reformularem a procura mesmo no caso de obterem poucos resultados (menos de dez). Tal leva a suspeitar que, ao contrário de minimizar o seu trabalho, os utilizadores preferem por vezes resultados sem ruído.

Como foi descrito, de uma forma um tanto provocatória, em Santos (2011), para avaliar se vale a pena implementar um sistema que permita uma identificação mais fina e rigorosa de autores (com possível necessidade de mais escolhas e cliques do utilizador), muito provavelmente apenas se conseguirão obter resultados conclusivos após a própria implementação e subsequente comparação entre duas versões.

Isto porque, quando se está em presença de um problema de engenharia (e não puramente científico) é preciso pesar os contras da insatisfação ou descontentamento que os erros de um tal sistema podem também causar. Concretamente, se em 5% dos casos (um em vinte) essa distinção não era relevante para o utilizador, o preço de ter de decidir não seria contraproducente em relação à experiência do utilizador? Algo que pelo menos na situação presente da disciplina da usabilidade, apenas é possível medir com testes com utilizadores.

Os dados de que dispomos por agora – e lembramos que este é um projeto ainda em progresso – parecem apontar para que uma muito pequena percentagem de procuras muda o autor. Das 75.651 sessões, apenas 4.148 contêm algo no campo autor, e dessas apenas cerca de 2.000 o altera.

De qualquer maneira, podemos tambem apresentar, como dados úteis para futuro estudo, a forma dos nomes dos autores procurados e que resultaram em resultados nulos, na tabela 15.

7 Correção de nomes de autores

Um outro problema sobre que nos debruçámos – e que parece receber confirmação de premência dado o grande número de resultados nulos nas pesquisas feitas – corresponde ao utilizador ter digitado o nome de um autor incorretamente.

Embora a correção ortográfica seja uma das áreas mais antigas e mais utilizadas do PLN, a transformação de nomes próprios tem particularidades próprias: em primeiro lugar, porque mesmo numa língua como o português com

Palavras	Iniciais	Frequência
2	0	2269
1	0	1515
3	0	1103
4	0	616
5	0	351
1	1	217
6	0	149
1	2	119
2	1	95
7	0	66
2	2	61
3	1	37
3	2	21
4	1	19
1	3	17
8	0	13
+	+	68

Tabela 15: Forma dos autores procurados que deram resultados nulos

pouca variabilidade de grafia, nomes próprios provenientes de estádios anteriores da língua, e mesmo de palavras estrangeiras (apelidos), são frequentes. Em segundo lugar porque a correção de uma palavra ou grupo de palavras não pode ser feita com auxílio do contexto (como por exemplo com palavras correntes quando diferentes grafias implicam diferentes categoria gramaticais).

Finalmente, só queríamos corrigir para nomes que se encontrassem já no RCAAP. Não seria de qualquer ajuda obrigar um utilizador a corrigir *Dulcinela Alves* para *Dulce Alves* para depois informar que não havia nenhuma obra dessa autora no RCAAP.

Assim, utilizámos o Jspell (Almeida e Pinto, 1994; Almeida e Pinto, 1995; Simões e Almeida, 2002) criando um novo dicionário apenas com os nomes de autores presentes no RCAAP, e com uma sintaxe especial (visto que inicialmente os dicionários do Jspell não tinham sido desenhados para palavras com pontos e com mais do que uma palavra). Assim, *A. Bonfante* passa a *A@%Bonfante*.

Criámos tambem algumas regras padrão de confusão de letras e sons em português (tais como entre "ç" e "ss", por exemplo), que é uma funcionalidade do Jspell que permite fazer correção baseada em regras.

Também incluímos como correções possíveis nomes em que existem variantes com acento e sem acento, tal como *Raúl* e *Raul*, *Luís* e *Luís*, *Marcirio* e *Marcírio*, ou nomes que têm consoantes mudas e que é possível que uma pessoa que os procure não saiba qual a ortografia com que um dado autorgrafa o nome em questão, tal como *Christina* e *Cristina*, *David*, *Davi* e *Davide*, ou *Raquel* e *Rachel*, *Estela*, *Stella*, e *Estella*. Isto é especialmente necessário para nomes brasileiros, visto que no Brasil não existe uma lista de nomes autorizados, como em Portugal, e portanto a variação é incomparavelmente maior.

Finalmente, esta versão também remove ou adiciona acentos e cedilhas para remediar o problema de utilizadores sem teclado com diacríticos – ou sem paciência para os incluir, habituados que estão aos motores de procura internacionais os ignorarem.

De momento, dos 6600 casos de nomes de autores procurados mas não encontrados, é possível apenas corrigir 692 com a presente versão.

8 Experiências preliminares de medição de impacto

Embora tal não esteja diretamente relacionado com o que nos propusemos fazer nesta primeira colaboração com o RCAAP, não há dúvida de que o ter acesso (se for possível) a uma base tão grande de publicações permite um conjunto de outros estudos de interesse sociológico, de política de investigação, e de relacionamento entre a comunidade de investigadores, estudos aliás que têm sido moda a nível internacional nos últimos tempos, cf. Tsatsaronis et al. (2011) e de que são além disso prova os serviços Arnetminer - Instant Social Graph Search[11], Microsoft Academic Research[12], Google Scholar[13] e Scholarometer[14].

Em particular, entre estas questões existe a funcionalidade, oferecida por alguns sistemas, de estudar a referência/fator de impacto de uma dada publicação, ou de um autor (Couto et al., 2009). Estas medidas são contudo pobres no sentido de que um verdadeiro impacto mede-se pelo uso e possível replicação de uma obra e não na simples menção, por vezes crítica, muitas vezes nem lida, de autores "célebres" ou que seja conveniente citar na comunidade em questão.

Mais relevante, portanto, é a capacidade de estudar a pragmática das próprias referências: ou seja, quantas meramente mencionam ou exemplificam, quantas criticam ou apontam fraquezas, e quantas finalmente usam a obra citada como ponto de partida.

[11] `http://www.arnetminer.org/`
[12] `http://academic.research.microsoft.com/`
[13] `http://scholar.google.com/`
[14] `http://scholarometer.indiana.edu/`

Tomando por exemplo a questão interessante de avaliar a contribuição de Eckhard Bick para o processamento computacional da língua portuguesa, e admitindo que o uso do PALAVRAS (que tem uma citação fixa) é algo que pode ser relativamente fácil de considerar como um indicador desse impacto, já Santos (2011) discutiu o diferente peso que deveria ser dado a formas – fictícias – de o referir, tal como

- *O nosso sistema é semelhante a Bick(2000)*
- *Outros sistemas, tal como Bick(2000)*
- *Ao contrário de Bick(2000)*
- *Usámos Bick(2000)*
- *O nosso sistema é baseado em Bick(2000)*
- *A nossa pesquisa usa o AC/DC*

Usando o GoogleScholar e a funcionalidade de ver o impacto, fizemos um pequeno teste com as referências feitas aos artigos da primeira autora (consulta feita em 22 de Outubro de 2011), que resultou em 81 artigos em forma eletrónica (de notar que a página de impacto refere também artigos e entradas que não se encontram acessíveis electronicamente, e que não foram contados artigos em que ela figurasse como autora, nem mesmo de alunos ou colaboradores seus).

Uma primeira análise manual (com a intenção, naturalmente, de desenvolver requisitos para um programa que o fizesse automaticamente no futuro) indicou

- (Uma grande maioria de) casos em que a citação aparece associada ao nome de um recurso ou de uma iniciativa (em que o HAREM (Santos e Cardoso, 2007; Mota e Santos, 2008), a propósito, é de longe o mais citado)
- 20 casos onde vem associada a uma afirmação ou descrição de trabalho (mais 3 que simplesmente remetem para um artigo para ver questões gerais, tais como *Para os diversos problemas relativos á divulgación de corpus a través da web, véxase Santos (1999)*)
- 3 casos onde é criticado ou menorizado ou contrastado negativamente, tais como *Similar comparisons had already been made (.. ; Santos and Ranchhod, 2002; ...), but in general they were not focused on free software.*
- 4 casos onde apenas aparece na lista de referências sem ser referido no texto
- 2 casos onde nova terminologia é-lhe associada (por outras palavras, é indicada como a origem de um dado termo), tal como *see also Santos 1996 and what she calls "acquisitions").*

Além de aproveitarmos a ocasião para indicar que consideramos que o impacto das diversas publicações foi maior do que a sua citação (contagem de citações) testemunha, dado o número de visitas ao sítio da Linguateca e o levantamento massiço dos recursos criados, notamos de qualquer maneira que estas referências são diferentes entre si em sentido e em peso.

É além disso extraordinário o facto de haver imensos erros nas citações analisadas – erros na ortografia de nomes (tal como *Aries* para *Aires*, *Dianna* em vez de *Diana*), mas sobretudo datas e referências completas – que muitas vezes variam o ano, ou variam a própria referência que seria apropriada. Isso só (quase) a própria autora pode distinguir, mas não deixa de ser digno de menção.

9 Trabalho relacionado

Tanto quanto sabemos, Spink e Jansen (2004) foram dos primeiros a debruçar-se sobre a procura de nomes de pessoas na rede, e nessa altura consideraram que não constituía uma fatia suficientemente importante das procuras.

Quase dez anos mais tarde, o surgimento de muitos sistemas e serviços baseados precisamente num tipo especial de nomes de pessoas (académicos), como os mencionados na secção anterior, parece indicar, das duas uma, ou que houve uma mudança nos hábitos e na cultura internáutica mundial, ou que, mesmo que a nível de motores de procura genéricos esses pedidos não sejam maioritários, fazem-se certamente em serviços especiais. Além de bases de dados de teses e dissertações públicas, e de repositórios institucionais, são também rotineiramente executadas em serviços, ou sítios, como a Wikipédia, como a Linguateca aliás espera poder estudar em breve com o Págico[15] (Costa, Mota e Santos, 2012; Mota, 2012; Simões, Mota e Costa, 2012).

Assim, existem variados artigos (Han et al., 2004; Han e Zhao, 2009; Ferreira et al., 2010; Gong e Oard, 2009; Kern, Zechner e Granitzer, 2011) que se preocupam com a desambiguação dos autores na procura na rede, o que é um objetivo (decomponível em vários problemas distintos) diferente do que abordámos aqui. Assim, podemos apreciar e ver formas de solucionar, na procura na rede

[15] `http://www.linguateca.pt/Pagico/`

- muitos autores com o mesmo nome, que se queiram individualizar para o cálculo de citações, e para encontrar a pessoa certa
- muitas formas de descrever o mesmo autor que se querem juntar, pelas mesmas razões do que o ponto anterior
- o facto de os sistemas de autoridade bibliográfica eles mesmos conterem mutas incorreções
- o facto de uma procura por autores na rede vai produzir um conjunto de respostas em que muitas delas são já o resultado de sistemas agregadores (Tan, Kan e Lee, 2006)
- o facto de autores prolíficos poderem publicar em várias áreas (Sun et al., 2011)

sendo que uma das propostas mais recentes (Qian et al., 2011) advoga o uso de colaboração homem-máquina para resolver o problema.

Outros autores (Dervos et al., 2006; Sun et al., 2011; Tsatsaronis et al., 2011), pelo contrário, dedicam-se a artigos científicos, embora os métodos e as bases documentais sejam bastante diferentes. Por exemplo, Tsatsaronis et al. (2011) categorizam os autores em termos da sua produtividade e do número de co-autores com que publicam, enquanto Han et al. (2004) reparam que o nome da revista em que publicam é a melhor forma de desambiguar autores através de aprendizagem automática.

Está claro que neste aspeto um dos trabalhos pioneiros é o do CiteSeer, continuado pelo CiteSeer X[16], embora arquivos internacionais como o DBLP[17] e a ACL[18], também tenham desenvolvido projetos meritórios e que são usados no dia a dia dos informáticos e linguistas computacionais. Note-se a esse respeito também a Biblioteca Digital Brasileira de Computação[19], que tem, na nossa opinião, uma interação muito feliz em termos de usabilidade precisamente na navegação de autores.

Diferentemente do nosso objetivo, os autores acima citados preocupam-se primordialmente com o estabelecimento de identificações únicas e com a desambiguação entre vários autores que publicam com o mesmo nome, usando geralmente mais informação que o simples nome. O nosso projeto tem sido muito mais modesto e a sua principal preocupação é averiguar o comportamento dos utilizadores para a estudar a própria necessidade deste tipo de processo, no universo português.[20]

Convém, a esse propósito, referir que a tese de mestrado de Luís Miguel Cabral (Cabral, 2007) foi um trabalho valioso, em português, sobre os variados problemas da gestão de um catálogo de referências bibliográficas.

10 Observações finais

Este projeto começou há um ano e meio e tem sido desenvolvido a meio tempo, e o processo de configurar o ımodus faciendi, ao mesmo tempo estabelecendo uma cooperação produtiva com várias outras instituições e grupos (KEEP, Serviços de documentação da Universidade do Minho, e o grupo paralelo do RCAAP na FCCN) com outras prioridades e objetivos, levou também o seu tempo, o que nos leva a afirmar que apenas podemos concluir que nos encontramos no início de um trabalho que, a ser continuado, permite muito mais estudos.

Por um lado, é óbvio que deveremos proceder à análise de outras vertentes da procura: assunto, título, e composição em geral da procura (quais os campos utilizados e como), que poderão ser igualmente elucidativos quer sobre a atividade dos utilizadores como como melhorar o ambiente de procura. A análise de co-autores, e de auto-citações se tivermos acesso aos próprios artigos, permitirá uma especificação muito mais fina dos autores do lado do RCAAP.

Com efeito, e como já discutido informalmente em Santos (2011), o trabalho até agora feito sobre a possibilidade de melhorar a identificação dos nomes dos autores não parece produzir uma melhoria considerável na usabilidade (e consequente satisfação) dos que pesquisam no portal do RCAAP, visto que a parte teoricamente melhorável corresponde a uma fatia percentualmente pequena dos casos.

Contudo, se isso pode indicar que o processamento da língua, mesmo em casos muito específicos, ainda não é suficientemente robusto para ser utilizado na prática, o problema pode ser o oposto – o de que a área ou questão dos nomes dos autores é mínima em relação à melhoria que poderíamos obter se processássemos o texto todo e não só essa parte semi-estruturada e bastante rígida que são os nomes.

É preciso também mencionar que o sistema de

[16] `http://csxstatic.ist.psu.edu/about`

[17] `http://www.informatik.uni-trier.de/~ley/db/`

[18] `http://www.aclweb.org/`, veja-se sobretudo a ACL Anthology.

[19] `http://www.lbd.dcc.ufmg.br/bdbcomp/`

[20] A este respeito, é interessante observar que os dados dos repositórios brasileiros com que o RCAAP funciona já desambiguam o autor colando-lhe ao nome a instituição a que pertence, o que pode indicar tanto uma quantidade muito maior de objetos como uma situação linguística em que os nomes são mais pequenos e mais ambíguos.

procura usado no RCAAP (pelo menos o usado na procura simples que é a mais utilizada) é baseado no modelo "saco de palavras", ou seja, não foi desenhado especialmente para o tipo de objetos e procuras que serve.

Um dos grandes desafios da Linguateca é conseguir convencer – com dados práticos e não apenas recorrendo a retórica – os desenvolvedores informáticos de que o "modelo único de recolha de informação" não é sempre a melhor maneira de obter resultados, e que conhecimento da estrutura do universo dos objetos e da forma de os referir poderá produzir resultados mais adequados.

No fundo, não é mais do que isso o que pretendemos conseguir com o trabalho aqui iniciado. Mas não podemos deixar de reconhecer que estamos muito longe de o concluir, e que pouco mais podemos apresentar do que pistas para futuras ações.

Agradecimentos

O trabalho aqui descrito enquadra-se no âmbito da Linguateca, co-financiada desde o seu início pelo Governo Português, pela União Europeia (FEDER e FSE), sob o contrato POSC/339/1.3/C/NAC, pela UMIC e pela FCCN, e presentemente pela Fundação da Ciência e da Tecnologia (FCT) e pela Fundação para a Computação Científica Nacional (FCCN).

Agradecemos vivamente aos membros do projeto RCAAP a colaboração e a cedência dos materiais e do acesso aos seus diários. Sem essa colaboração o trabalho não teria sido possível.

Agradecemos também ao Alberto Simões a inestimável ajuda que nos deu na adaptação do Jspell, assim como todo o apoio sobre o programa.

Referncias

Almeida, José João e Ulisses Pinto. 1994. Manual de Utilizador do JSpell. Relatório técnico, Departamento de Informática, Universidade do Minho. `http://www.di.uminho.pt/~jj/pln/jspellman.ps.gz`.

Almeida, José João e Ulisses Pinto. 1995. Jspell – um módulo para análise léxica genérica de linguagem natural. Em *Actas do X Encontro Nacional da Associação Portuguesa de Linguística*, pp. 1–15, 6-8 de Outubro de 1994, 1995.

Artiles, Javier, Julio Gonzalo, e Satoshi Sekine. 2009. WePS 2 Evaluation Campaign: Overview of the Web People Search Clustering Task. Em *18th International World Wide Web Conference*, 20-24 de Abril, 2009.

Artiles, Javier, Satoshi Sekine, e Julio Gonzalo. 2008. Web People Search - Results of the first evaluation and the plan for the second. Em *17th International World Wide Web Conference*, pp. 1071–1072, 21-25 April, 2008.

Cabral, Luís Miguel. 2007. SUPeRB - Sistema Uniformizado de Pesquisa de Referências Bibliográficas. Tese de Mestrado, Faculdade de Engenharia da Universidade do Porto, Março, 2007. `http://www.linguateca.pt/documentos/DissertacaoLuisCabral-SUPeRB.pdf`.

Cabral, Luís Miguel, Diana Santos, e Luís Fernando Costa. 2008. SUPeRB - Gerindo referências de autores de língua portuguesa. Em *VI Workshop Information and Human Language Technology (TIL'08)*, 28-29 de Outubro, 2008.

Carvalho, José, João Mendes Moreira, Eloy Rodrigues, e Ricardo Saraiva. 2010. O Repositório Científico de Acesso Aberto de Portugal : origem, evolução e desafios. Em Maria João Gomes e Flávia Rosa, editores, *Repositórios institucionais : democratizando o acesso ao conhecimento*, pp. 127–152. EDUFBA.

Costa, Luís, Cristina Mota, e Diana Santos. 2012. SIGA, a Management System to Support the Organization of Information Retrieval Evaluations. Em *Proceedings of PROPOR'2012*. Springer. No prelo.

Couto, Francisco M., C. Pesquita, T. Grego, e Paulo Veríssimo. 2009. Handling self-citations using Google Scholar. *Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics*, 13(1).

Dervos, Dimitris A., Nikolaos Samaras, Georgios EvangelisSundis, Jaakko P. Hyvarinen, e Ypatios Asmanidis. 2006. The Universal Author Identifier System(UAI Sys). Em *Proceedings 1st International Scientific Conference*.

Ferreira, Anderson A., Adriano Veloso, Marcos André, e H. F. Laender. 2010. Effective Self-Training Author Name Disambiguation in Scholarly Digital Libraries. Em *Proceedings of the 10th annual joint conference on Digital libraries*, pp. 39–48. ACM, 21-25 de Junho, 2010.

Gong, Jun e Douglas W. Oard. 2009. Determine the Entity Number in Hierarchical Clustering

for Web Personal Name Disambiguation. Em *18th International World Wide Web Conference*, 20-24 de Abril, 2009.

Han, Hui, Lee Giles, Hongyuan Zha, Cheng Li, e Kostas Tsioutsiouliklis. 2004. Two Supervised Learning Approaches for Name Disambiguation in Author Citations. Em *JCDL'2004*, pp. 296–305, 7-11 Junho, 2004.

Han, Xianpei e Jun Zhao. 2009. CASIANED: Web Personal Name Disambiguation Based on Professional Categorization. Em *18th International World Wide Web Conference*, 20-24 de Abril, 2009.

He, Daqing e Ayse Göker. 2000. Detecting session boundaries from web user logs. Em *In Proceedingsof of the BCS-IRSG 22nd Annual Colloquium on Information Retrieval Research*, pp. 57–66.

Kern, Roman, Mario Zechner, e Michael Granitzer. 2011. Model Selection Strategies for Author Disambiguation. Em *22nd International Workshop on Database and Expert Systems Applications*, pp. 155–159, 29 de Agosto a 2 de Setembro, 2011.

Moreira, João Mendes, José Carvalho, Ricardo Saraiva, e Eloy Rodrigues. 2010. Repositório Científico de Acesso Aberto de Portugal : uma ferramenta ao serviço da ciência portuguesa. Em *Congresso nacional de bibliotecários, arquivistas e documentalistas, 10, Guimarães, Portugal, 2010 Políticas de informação na sociedade em rede : actas*. APBD.

Mota, Cristina. 2012. Resultados págicos: participação, resultados e recursos. *Linguamática*, 4(1). No prelo.

Mota, Cristina e Diana Santos, editores. 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*. Linguateca, 31 de Dezembro, 2008.

Qian, Yanan, Yunhua Hu, Jianling Cui, Qinghua Zheng, e Zaiqing Nie. 2011. Combining Machine Learning and Human Judgment in Author Disambiguation. Em *20th ACM Conference on Information and Knowladge Management*, 24-28 de Outubro, 2011.

Santos, Diana. 2009. Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. *Linguamática*, 1(1):25–59, Maio, 2009.

Santos, Diana. 2011. Compreensão de linguagem natural: voltando à carga, 18 de Julho, 2011. Palestra convidada na Universidade de Aveiro, `http://www.linguateca.pt/Diana/download/SantosAveiro2011.pdf`.

Santos, Diana e Nuno Cardoso, editores. 2007. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, 12 de Novembro, 2007.

Santos, Diana e Ana Frankenberg-Garcia. 2007. The corpus, its users and their needs: a user-oriented evaluation of COMPARA. *International Journal of Corpus Linguistics*, 12:335–374, Maio, 2007.

Santos, Diana e Fernando Ribeiro. 2010. Estudando os autores: Trabalho referente à colaboração com o RCAAP. Relatório técnico, Linguateca, FCCN, 6 de Agosto, 2010. `http://www.linguateca.pt/Diana/download/SantosRibeiroRelRCAAP6Ago2010.pdf`.

Santos, Diana e Fernando Ribeiro. 2011. Estudando os nomes dos autores no RCAAP: relatório do primeiro ano. Relatório técnico, FCCN, 4 de Junho, 2011. `http://www.linguateca.pt/documentos/SantosRibeiroEstudandoNomesAutoresRCAAP.pdf`.

Simões, Alberto, Cristina Mota, e Luís Costa. 2012. A wikipédia em português no Págico: adaptação e avaliação. *Linguamática*, 4(1). No prelo.

Simões, Alberto Manuel e José João Almeida. 2002. jspell.pm – um módulo de análise morfológica para uso em Processamento de Linguagem Natural. Em *Actas da Associção Portuguesa de Linguística (APL2001)*.

Spink, Amanda e Bernard J. Jansen. 2004. Searching for people on Web search engines. *Journal of Documentation*, 60(3):266–278.

Sun, Xiaoling, Jasleen Kaur, Lino Possamai, e Filippo Menczer. 2011. Detecting Ambiguous Author Names in Crowdsourced Scholarly Data. Em *SocialCom 2011*.

Tan, Yee Fan, Min-Yen Kan, e Dongwon Lee. 2006. Search engine driven author disambiguation. Em *JCDL'2006*, June, 2006.

Tsatsaronis, George, Iraklis Varlamis, Sunna Torge, Matthias Reimann, Kjetil Norvig, Michael Schroeder, e Matthias Zschunke. 2011. How to Become a Group Leader? or Modeling Author Types Based on Graph Mining. Em *International Conference on Theory and Practice of Digital Libraries*, Setembro, 2011.

Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Addison-Wesley Press, Cambridge, Mass.