Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

Estudando os autores:

Trabalho referente à colaboração com o RCAAP

Diana Santos e Fernando Ribeiro

Área *Linguateca*

6 de Agosto de 2010

Trabalho referente à cooperação com o RCAAP

Área *Linguateca*

Diana Santos
Fernando Ribeiro

6 de Agosto de 2010

1 RESUMO

Este documento relata o trabalho desenvolvido pela Linguateca no primeiro semestre de 2010, com o objetivo de produzir ferramentas de processamento da língua portuguesa que possam aproveitar ao projeto RCAAP, em particular à procura no portal do RCAAP.

Neste primeiro relatório, descrevemos brevemente a motivação para esta colaboração da parte da Linguateca, detalhamos o trabalho já feito, e descrevemos os serviços que pretendemos vir a desenvolver num futuro próximo.

Este documento é essencialmente para estabelecer mais claramente os objetivos da colaboração e para obter retorno por parte dos elementos do RCAAP.

2 INTENÇÕES E PRESSUPOSTOS

Ambos os projetos, Linguateca e RCAAP, se integram no objetivo global da FCCN de produzir serviços úteis à comunidade científica e potenciar a colaboração e reuso dos recursos desta.

A Linguateca encontra-se numa fase de reorganização tendo enviado, há cerca de dois anos, uma proposta em que contemplava um pólo associado ao RCAAP. Enquanto esperamos por decisão superior, contudo, decidimos apostar em algum trabalho preliminar, que terá para todos os efeitos o impacto de projeto piloto e que permitirá um trabalho mais focado caso o financiamento vier a ser concedido por parte da UMIC.

A Linguateca tem-se especializado no processamento computacional da língua portuguesa, e tem sido uma fonte de recursos e dinamização da comunidade nessa área. Contudo, passados dez anos de atividade, pareceu-nos que deveria sofrer uma mudança qualitativa de forma a alargar o seu âmbito e a dedicar-se a projetos com uma audiência e um impacto significativamente maiores, como é o caso do RCAAP.

A área da catalogação da produção científica era aliás uma das áreas a que a Linguateca se dedicou desde o princípio, tendo levado à criação de um repositório público de referências de artigos e outros documentos (no caso de serem de acesso aberto, um repositório físico também) em 1999, mais tarde automatizado e melhorado, especificamente para a língua portuguesa, no âmbito do mestrado de um antigo colaborador da Linguateca (Cabral, 2005).

Esse sistema, o SUPeRB, obrigou-nos a contatar e a lidar de perto com os problemas da catalogação, da procura, e da manutenção de um catálogo, e o seu desenvolvimento continua sendo uma das atividades da Linguateca (ver também Cabral et al. 2008a,b).

Por outro lado, a Linguateca tem alguma experiência em indexação e em recolha de informação (RI), tendo tido três doutoramentos (um ainda em curso) no seu âmbito relacionados com a procura em língua portuguesa.

Por todos esses motivos a colaboração entre o RCAAP e a Linguateca pareceu-nos a mais natural e foi a primeira a ser posta em prática (as outras duas áreas de atuação pública, para que conste, seriam o Arquivo da Web Portuguesa e a Biblioteca Nacional).

3 OBJETIVOS DE UMA PRIMEIRA COLABORAÇÃO

A nossa intenção primordial 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 isso seria uma grande melhoria, preferimos analisar primeiro a situação e os serviços já oferecidos pelo projeto RCAAP e investigar se o nosso saber-fazer poderia trazer alguma melhoria substancial.

De qualquer maneira, as ideias iniciais, apresentadas pelo Fernando Ribeiro no encontro do RCAAP em Leiria (Ribeiro & Santos, 2010), eram as seguintes:

1. Ajuda à procura através de técnicas de correção ortográfica especialmente concebidas para o efeito
2. Ajuda à catalogação através de uma identificação mais fina de autores

Além disso, esperávamos também, ao fazer uma análise da interação já havida com o sistema, poder efetuar estatísticas interessantes e até descobrir possíveis erros ou dar sugestões de usabilidade, quando tal fosse pertinente.

4 TRABALHO REALIZADO E PRIMEIROS RESULTADOS

Além da nossa familiarização com os variados membros e ferramentas usadas pelo RCAAP, e a compreensão das várias responsabilidades e das várias localizações de dados (Roma e Pavia não se fizeram num dia...), a interação com os vários atores: a KEEP, a SDUM e a FCCN/RCAAP, foi também preciso desenvolver algumas ferramentas especializadas para processamento de diários, agrupamento, e sugestões de palavras/nomes próximos.

Convém referir a este propósito que o Fernando também teve de se familiarizar com estas técnicas visto que se encontra há relativamente pouco tempo na Linguateca e não tinha portanto ainda lidado com a maioria dos recursos e ferramentas já desenvolvidos ao longo da nossa relativamente longa história.

Nesse aspeto convém realçar a pronta ajuda do Alberto Simões que adicionou um módulo de palavras próximas ao Jspell a nosso pedido.

4.1 ESTUDO DE NOMES DE PESSOAS (AUTORES)

A primeira tarefa realizada foi a obtenção de listas de nomes próprios em português ou citados/localizados em repositórios portugueses (o RCAAP e o SUPeRB). Usámos para isso os metadados do RCAAP, o conteúdo do REB do SUPeRB (ou melhor, a lista de todos os autores incluídos no catálogo de publicações da Linguateca), uma primeira análise nossa dos diários do Apache relativos aos vários repositórios do RCAAP (ver próximo ponto), e o REPENTINO, um almanaque com nomes de entidades (incluindo nomes de pessoas) (Sarmento et al., 2006).

Para poder comparar estes recursos, e averiguar possíveis confusões entre várias formas de referir um mesmo autor/pessoa, aplicamos às várias listas os seguintes processos:

1. Normalização, de forma a transformar várias formas de grafar numa forma canónica: juntar ponto às iniciais, transformar em maiúscula inicial em alguns casos.[1]
2. Ambiguação, de forma a produzir cadeias de carateres ainda mais ambíguas e menos precisas a partir de cada identificação

Os valores resultantes dos dois processos encontram-se na tabela seguinte, a tabela 1.

[1] Note-se que em alguns casos raros, por exemplo envolvendo hífenes ou a letra E isolada, é possível que uma mesma identificação seja ambígua entre mais do que uma forma canónica. Nesse caso, criamos mais do que uma forma canónica.

Coleção	Original	Normalizado	Ambiguado
SUPeRB	3316	3328	282216
RCAAP	38538	36954	328110
REPENTINO	282222	282216	1852613

Tabela 1. Quantidade de recursos para cada coleção (29 de Junho 2010)

Um exemplo de ambos os processos pode ser encontrado na Figura 1.

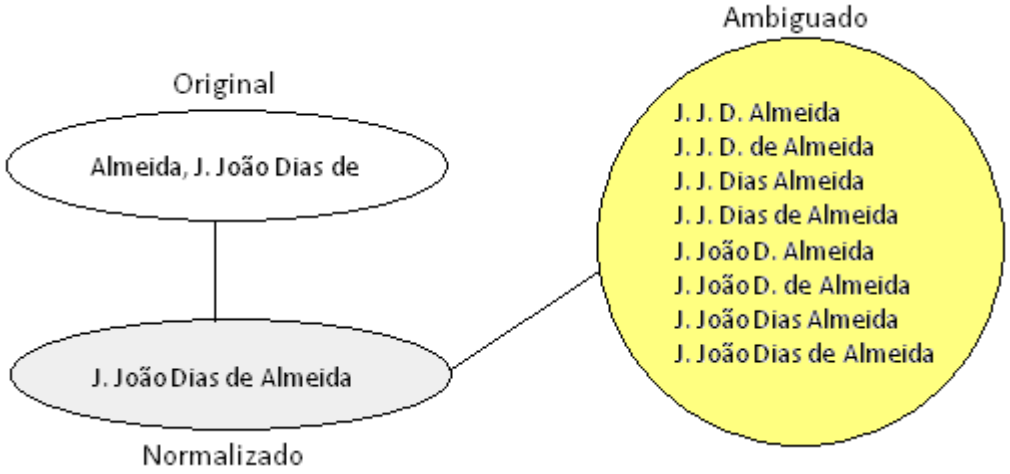

Figura 1. Exemplo do processo de normalização e ambiguação para o node de autor "Almeida,J. João Dias de".

4.2 ESTUDO DOS DIÁRIOS DE ACESSO ("LOGS")

Para compreender as necessidades e a prática dos utilizadores quisemos observar o seu comportamento de pesquisa e interação com os vários repositórios. Para isso obtivemos um grande número de diários, localizados numa máquina da FCCN, que são no formato Apache, e tentámos, primeiro, extrair todos os casos de nomes de autores.

Esta exploração permitiu-nos conhecer em mais profundidade a variação e diferentes metodologias e convenções usadas pelos vários repositórios, quer no que respeita às normas de catalogação e depósito seguidas (pelos administradores responsáveis) quer no que respeita a escolhas técnicas e de organização da interface, e que levam a que seja necessário um estudo e familiarização praticamente por repositório.

Por outro lado, também nos permitiu tomar contacto com algumas normas bibliográficas portuguesas, por exemplo as seguidas pela Universidade do Minho (BN, 2005), assim como levou à criação de um diário de aplicação desenhado especialmente para esta colaboração pela Keep, cujo formato apresentamos na tabela 2.

S | IP: 193.137.198.15 | Date: 12-Jul-2010 17:00:48 | Agent: 'Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_4; en-US) AppleWebKit/533.4 (KHTML, like Gecko) Chrome/5.0.375.99 Safari/533.4' | Query: 'desenvolvimento' | Results: 8965

S | IP: 188.140.82.141 | Date: 12-Jul-2010 17:10:54 | Agent: 'Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.0; Trident/4.0; SIMBAR={F70B72B7-6538-4340-85C0-2100F3CCA0ED}; SLCC1; .NET CLR 2.0.50727; Media Center PC 5.0; .NET CLR 3.5.30729; .NET CLR 3.0.30729; InfoPath.1; OfficeLiveConnector.1.5; OfficeLivePatch.1.3; Creative ZENcast v2.00.13)' | Query: 'manual de acolhimento' | Results: 559

A | IP: 193.137.198.15 | Date: 12-Jul-2010 17:02:12 | Agent: 'Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_4; en-US) AppleWebKit/533.4 (KHTML, like Gecko) Chrome/5.0.375.99 Safari/533.4' | Query: 'desenvolvimento' | Terms: NONE | From: null | To: null | Repositories: NONE | DocTypes: 'masterThesis' OR 'article' | Language: NONE | Subjects: NONE | IssueDates: '2009' OR '2008' OR '2007' | Authors: NONE | Results: 3199

A | IP: 193.137.198.15 | Date: 12-Jul-2010 17:02:35 | Agent: 'Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_4; en-US) AppleWebKit/533.4 (KHTML, like Gecko) Chrome/5.0.375.99 Safari/533.4' | Query: 'desenvolvimento' | Terms: NONE | From: null | To: null | Repositories: NONE | DocTypes: ('masterThesis' OR 'article') AND ('article') | Language: NONE | Subjects: NONE | IssueDates: ('2009' OR '2008' OR '2007') AND ('2007') | Authors: NONE | Results: 214

Figura 2 – Exemplos dos diários para as pesquisas simples (S) e avançadas (A).

Uma primeira exploração deste material (ainda e apenas de uma das três máquinas com diários) deu os resultados que apresentamos nas Tabelas 2 e 3. (Notamos, que iremos atualizando estes valores e estatíticas no futuro.)

Repositório	Pesquisas	Tipo de pesquisa			
		autor	assunto	título	simples
raberto (16 de Abril a 15 de Junho de 2010)	8220	5146	563		563
Utlr (19 de Abril a 15 de Junho de 2010)	4975	393	479		4103
bibliotecadigital (19 de Abril a 15 de Junho de	3828	794	204		2830

2010)					
rua (19 de Abril a 15 de Junho de 2010)	1539	159	196		1184
rihuc (19 de Abril a 15 de Junho de 2010)	1428	718	355		355
comum (19 de Abril a 15 de Junho de 2010)	1048	760	137		151
iconline (19 de Abril a 15 de Junho de 2010)	990	211	184		595
rispa (19 de Abril a 15 de Junho de 2010)	842	141	165		663
sapientia (19 de Abril a 15 de Junho de 2010)	607	85	371		151
rhff (19 de Abril a 15 de Junho de 2010)	572	335	46		171
ripcb (19 de Abril a 15 de Junho de 2010)	458	137	111		12
Rlneg (19 de Abril a 15 de Junho de 2010)	369	94	34		241
arca (19 de Abril a 15 de Junho de 2010)	200	51	33		81
ubithesis (1 de Abril a 15 Junho de 2010)	132	4	25		103
digituma (6 de Maio a 15 Junho de 2010)	56	27	2		27
Rips (10 de Maio a 15 Junho de 2010)	22	9	1		12
Rchporto (10 de Maio a 15 Junho de 2010)	9	9			
repul (6 de Maio a 15 Junho de 2010)	6				6

Tabela 2. Procuras nos diferentes repositórios. As datas apresentadas dizem respeito aos diários.

Repositório	Pesquisas	Tipo de pesquisa			
		autor	assunto	título	simples
metarepository	24102	1465	2369	1733	18535

Tabela 3. Pesquisas no portal do RCAAP (com base nos diários de 9 de Abril de 2010 a 15 de Junho de 2010)

4.3 AGRUPAMENTO DOS NOMES DOS AUTORES

A segunda análise feita foi a de agrupar os vários nomes de autores (já ambiguados) em grupos máximos que contivessem o último apelido e a primeira inicial. Este exercício, cuja descrição quantitativa se encontra na tabela 4, permitiu ter uma ideia da confusão máxima, assim como medir, para cada cadeia de caracteres, o seu grau de ambiguidade no sentido de "número de grupos a que pertence, dado o presente universo de autores".

	Nomes de autores nos metadados do RCAAP	
Nome do grupo	Nº de elementos	Elementos
A--Burenkov	1	A. A. Burenkov
A—Arantes	4	A. A. Arantes A. Adriano Arantes Artur A. Arantes Artur Adriano Arantes
A--Luís	34	A. A. C. Luís A. A. Costa Luís A. A. G. Luís A. A. Gabriel Luís A. A. da C. Luís A. A. da Costa Luís A. Alberto G. Luís A. Alberto Gabriel Luís A. António C. Luís A. António Costa Luís A. António da C. Luís A. António da Costa Luís A. Luís A. R. F. Luís A. R. Francisco Luís A. Rita F. Luís A. Rita Francisco Luís Alexandre A. C. Luís Alexandre A. Costa Luís Alexandre A. da C. Luís Alexandre A. da Costa Luís Alexandre António C. Luís Alexandre António Costa Luís Alexandre António da C. Luís Alexandre António da Costa Luís Amaral Luís Ana R. F. Luís Ana R. Francisco Luís Ana Rita F. Luís Ana Rita Francisco Luís António A. G. Luís António A. Gabriel Luís

		António Alberto G. Luís António Alberto Gabriel Luís

Tabela 4. Exemplos de grupos máximos

Tipo de ambiguidade (número de grupos)	Quantidade de casos
1	3541
2	5196
3	258
4 ou mais grupos	6607

Tabela 5. Ambiguidade de cada nome de autor contido nos metadados do RCAAP (a de 28 de Maio de 2010), por quantidade (quantos são únicos, quantos têm duas, três, etc. possibilidades).

Na próxima figura estão detalhados os casos que pertencem a quatro ou mais grupos, em escala logarítmica. Em média, cada nome de autor pertence a 22.3 grupos, e a mediana é 2541.

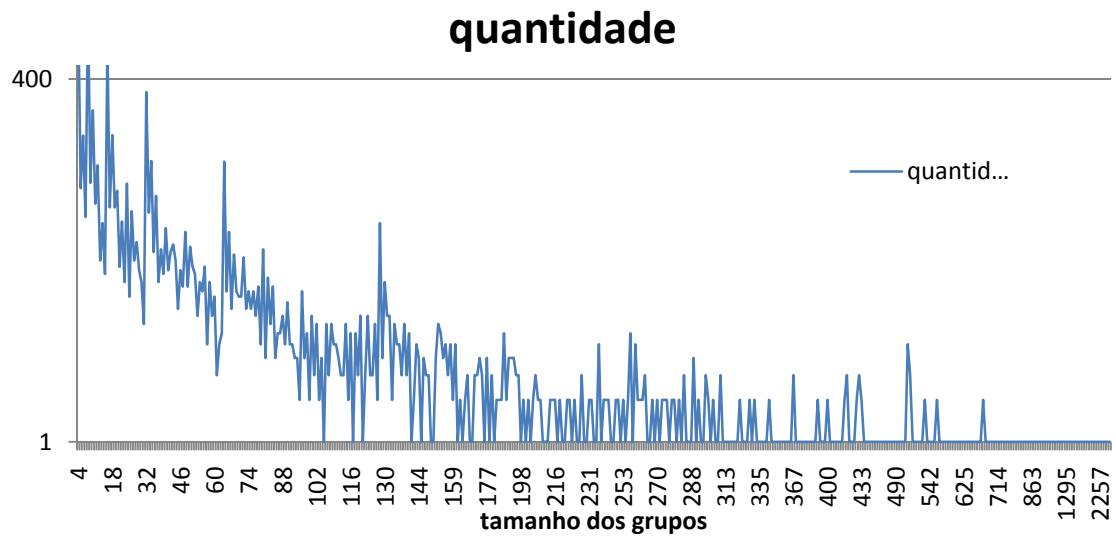

Figura 3 – Quantidade de grupos máximos com base no seu tamanho (número de elementos distintos que o compõem), para os grupos com mais de três elementos, em escala logarítmica (metadados de 28 de Maio de 2010).

4.4 MEDIÇÃO DO EMPARELHAMENTO DAS PROCURAS

Tendo obtido (como descrito no ponto 4.2) a lista de procuras feitas por autor, começámos por medir o grau de emparelhamento com o RCAAP, no sentido de identificar quantos

autores pedidos se encontravam nos metadados do RCAAP, quantos eram únicos e quantos eram ambíguos (após limpeza dos dados, como feito para os metadados).

Exemplo do autor pedido	Nº de grupos	Agrupamento – Metadados do RCAAP	
		Nome do grupo	Elementos do grupo
Rui Machado Gomes	1	R-Gomes	Rui Machado Gomes
Rose	2	A-Rose	A. Rose Adriana Rose
		C-Rose	C. Rose Cheryl Rose
João Rodrigues	10	A-Rodrigues	A. João Rodrigues Ana João Rodrigues
		M-Santos	4 elementos
		C-Silva	4 elementos
		M-Cardoso	8 elementos
		C-Ramos	4 elementos
		J-Simões	João Rodrigues Simões
		M-Rodrigues	M. João Rodrigues Maria João Rodrigues
		M-Oliveira	4 elementos
		A-Leal	A. João Rodrigues Leal Alberto João Rodrigues Leal
		J-Rodrigues	João Rodrigues

Tabela 6. Exemplos de emparelhamento entre as procuras e os metadados, usando os grupos máximos.

Na tabela seguinte (7) encontra-se uma primeira medição da relação entre as procuras e conteúdo do repositório, usando os diários Apache do portal do RCAAP.

Tipo de ambiguidade (número de grupos)	Quantidade de casos 962 (após limpeza) (%)
0	769 (80, 0 %)
1	55 (5,7 %)
2	21 (2,2 %)

3	20 (2,1 %)
4 ou mais grupos	96 (10,0 %)

Tabela 7. Grau de emparelhamento entre as procuras e os metadados (com base apenas nas procuras por autor usando os diários do portal do RCAAP de 19 de Abril de 2010 até 15 Junho de 2010 apenas de uma das máquinas, e nos metadados de 28 de Maio de 2010).

Na figura 4 estão detalhados os casos que pertencem a quatro ou mais grupos. Em média, cada expressão de procura por nome de autor emparelha com 44,5 grupos, e a mediana é 2541.

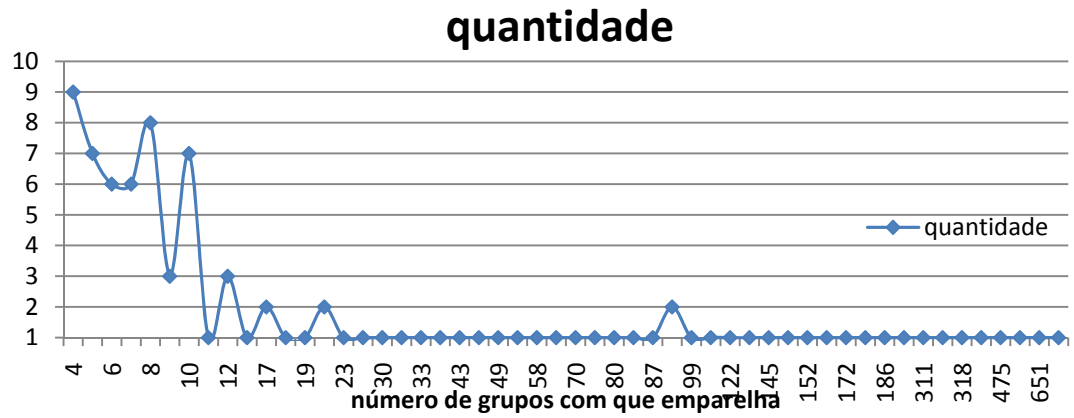

Figura 4 – Emparelhamento das expressões de pesquisa de autores no portal do RCAAP (metarepositório) com o conteúdo do repositório (em abcissas, a quantos grupos a expressão pode pertencer), referente aos metadados de 28 de Maio de 2010

Vemos por este primeiro estudo muito preliminar que 80,0% dos casos procurados não têm resultados, e que 20,0 % são ambíguos, ao considerar os grupos máximos.

4.5 AGRUPAMENTO POR AUTORES ÚNICOS

Aproximando-nos da questão inversa, que é a obtenção do autor certo, e não do grupo de autores com que é confundido (o chamado grupo máximo), procedemos a um outro tipo de agrupamento, que não contém autores "contraditórios" no mesmo grupo, ou seja "João Almeida" e Joaquim José Almeida" não pertencem ao mesmo grupo, embora o elemento "J. Almeida" pertença ainda aos dois.

Na tabela 8, mostramos um exemplo de vários grupos correspondentes a autores únicos, que provêm do mesmo grupo máximo J.-ALMEIDA. Neste caso e ao contrário da estratégia anterior, os grupos são identificados pela cadeia de carateres mais longa do grupo.[2]

	Nomes de autores nos metadados do RCAAP	
Grupo	**Nº elementos**	**Elementos**
José Joaquim de Almeida	10	J. Almeida J. de Almeida J. J. Almeida J. J. de Almeida J. Joaquim Almeida J. Joaquim de Almeida José J. Almeida José J. de Almeida José Joaquim Almeida José Joaquim de Almeida
José João Almeida	6	J. Almeida J. J. Almeida J. João Almeida João Almeida José J. Almeida José João Almeida

Tabela 8. Exemplos de grupos por autor único.

A tabela 9 e figura 5 repetem a caracterização do material da tabela 5 e figura 3, com os novos grupos.

Tipo de ambiguidade (número de grupos)	Quantidade de casos
1	3805
2	10150
3	936
4 ou mais grupos	19106

[2] Neste momento estamos a trabalhar com todo o resultado da ambiguação, mas iremos também tentar no futuro contabilizar grupos só com os elementos dos metadados como aparecem exatamente, ou seja, criar grupos únicos sem ter alargado o universo através do processo de ambiguação.

Tabela 9. Ambiguidade de cada nome contido nos metadados do RCAAP de 28 de Maio 2010, por quantidade (quantos são únicos, quantos têm duas, três, etc. possibilidades), após agrupamento por autor único.

Na próxima figura estão detalhados os casos que pertencem a quatro ou mais grupos, em escala logarítmica. Em média, cada designação de autor pode pertencer a 11,5 grupos, e a mediana é 571,5.

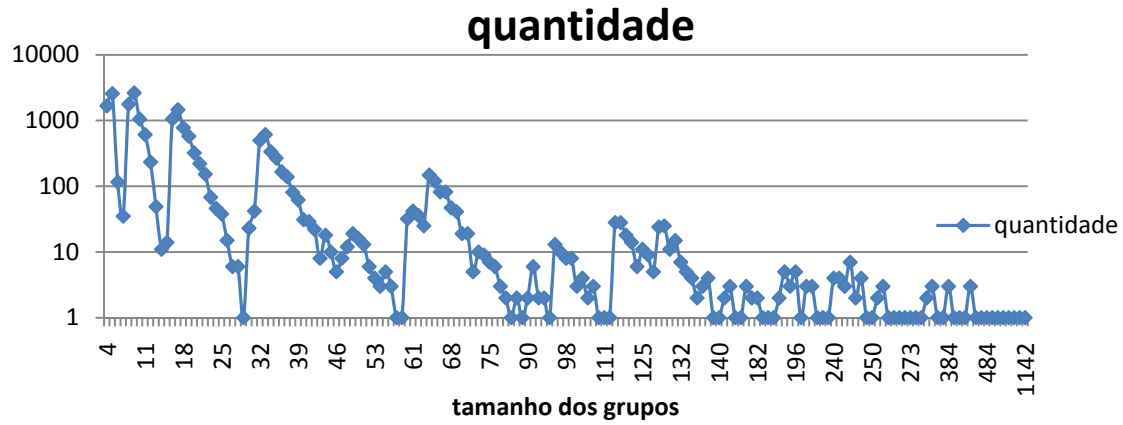

Figura 5 – Quantidade de grupos de autor único com base no seu tamanho (número de elementos distintos que o compõem), para os grupos com mais de três elementos, em escala logarítmica, a partir dos metadados de 28 de Maio de 2010

A tabela 10 e a figura 6 descrevem, analogamente à tabela 7 e à figura 4, o grau de emparelhamento entre as procuras observadas e o material no RCAAP após este novo método de agrupamento.

Tipo de ambiguidade (número de grupos com que emparelha)	Quantidade de casos 962 (após limpeza) (%)
0	769 (80, 0 %)
1	52 (5,4 %)
2	27 (2,9 %)
3	13 (1,3 %)
4 ou mais grupos	100 (10,4 %)

Tabela 10. Emparelhamento das expressões de pesquisa de autores no portal do RCAAP (metarepositório) com o conteúdo do repositório de 28 de Maio de 2010

Na próxima figura estão detalhados os casos que pertencem a quatro ou mais grupos, em escala logarítmica. Em média, cada expressão de procura por nome de autor emparelha com 16,81 grupos, e a mediana é 326,5.

Figura 6 – Emparelhamento das expressões de procura por autor em função do número diferente de grupos por autor único (em abcissas, a quantos grupos a expressão pode pertencer) em relação ao conteúdo do repositório de 28 de Maio de 2010

Vemos por este primeiro estudo muito preliminar que 80,0% dos casos procurados não têm resultados, e que 14,6 % são ambíguos, ao considerar os grupos por autor único.

4.6 SEMELHANÇA DE AUTORES

Em paralelo, adaptámos um dicionário para o Jspell (Almeida & Pinto, 1994, Simões & Almeida, 2002) de forma a conter nomes de autores e desenvolvemos regras de proximidade entre nomes.

Na Figura 7 apresentamos um excerto do dicionário criado, e na figura 8 alguns exemplos de regras de confusão possível para além da simples troca de letras.

A@%Bonfante/#tc/

Anido%Nayade%Freire/#tc/

Figura 7. Exemplo do novo dicionário do Jspell contemplando nomes próprios, onde % corresponde a um espaço e @ a um ponto.

ss ç

ç ss

ch x

x ch

Figura 8. Exemplos de regras de confusão para a língua portuguesa

Este módulo permitirá ainda calcular ainda outro tipo de agrupamento, por autores tipograficamente semelhantes, quando tivermos completado um conjunto de regras abrangentes para os casos de nomes de autores.

Na figura 9 ilustramos alguns casos em que este Jspell personalizado conseguiu obter um autor parecido com algum já existente nos metadados, nos casos em que esse autor não existia literalmente.

David A. Clarke
Sugestão: David T. Clarke

J. A. Alemida:
Sugestão: A. A. Almeida
Sugestão: B. A. Almeida
Sugestão: C. A. Almeida
Sugestão: F. A. Almeida
Sugestão: S. A. Almeida
Sugestão: T. A. Almeida
Sugestão: V. A. Almeida
Sugestão: J. B. Almeida
Sugestão: J. C. Almeida
Sugestão: J. D. Almeida
Sugestão: J. E. Almeida
Sugestão: J. F. Almeida
Sugestão: J. J. Almeida
Sugestão: J. L. Almeida
Sugestão: J. P. Almeida
Sugestão: J. R. Almeida

Figura 9. Exemplos de sugestões de autores próximos nos metadados do RCAAP, obtidos com a versão muito preliminar do Jspell que estamos a melhorar

4.7 FILTRAGEM DOS DIÁRIOS

Esta tarefa é mais do que uma limpeza preliminar dos diários, porque é de certa forma um corretor ou validador das procuras.

Ou seja, por vezes os utilizadores incluem mais do que um autor no campo autor, ou mesmo títulos dos artigos: nesse caso, um programa de interface inteligente que pudesse imediatamente traduzir para, ou procurar diretamente, o que o utilizador queria seria uma mais-valia evidente.

Assim, estamos a tentar criar um detetor de problemas de fácil resolução, tal como autores múltiplos, ou títulos óbvios, e removemos naturalmente logo estes "autores falsos" das listas de autores obtidas dos diários.

Na figura 10 apresentamos alguns exemplos verídicos de casos que não são autores:

> "G. Or Clowns As a Treatment For Preoperative Anxiety In Children Golan"
> "Institute Of Medicine."
> "Interpersonal Orientation Scale"
> "Johnson Am Macdowall W, Wellings K, Mercer Ch, Nanchahal K, Copas Aj, McManus S, Fenton Ka, Erens B."

Figura 10: Exemplos de falsos autores que foram encontrados por nós nos diários

4.8 IDENTIFICAÇÃO DE SESSÕES

Os estudos anteriores sobre diários que referimos acima referem-se apenas a consultas individuais, mas todos sabemos que para estudos mais completos interessa perceber a sequência de ações de um utilizador, para o que o conceito de sessão é essencial.

Em breve iremos estudar estas questões até para ver se existe correção e outras tentativas por parte do utilizador típico do RCAAP ou se este desiste à primeira resposta vazia.

Pretendemos também ver se o problema da clarificação do autor (quando há mais do que um autor) é algo que preocupa os utilizadores ou se eles preferem fazer a filtragem manualmente.

Numa primeira fase, vamos contar o tipo de sessões por número de procuras, tempo entre a primeira e a última, e se há semelhança entre as procuras numa mesma sessão. Mais tarde, iremos tentar investigar mais em pormenor qual a intenção do utilizador , num subconjunto de casos a determinar, quando tivermos processado uma maior quantidade de diários.

4.9 IDENTIFICAÇÃO DO QUE O UTILIZADOR REALMENTE QUER

Idealmente, uma interface amigável compreende o que o utilizador quer e tenta proporcionar os resultados com um mínimo de interação.

A intenção última deste trabalho é conseguir melhorar a experiência de interação dos utilizadores do RCAAP, ajudando e forncendo sempre alguma informação, de uma forma cooperativa, mas dando sempre a escolha ao utilizador, assim como a explicação do processamento do sistema quando não transparente.

Assim, prevemos que, quanto maior for o número de registos e de material incorporado no RCAAP, mais possibilidade haverá de ajudar o público com sugestões relevantes, e mais sentido fará proceder a uma escolha criteriosa de resultados quando o número for

demasiado ou houver razões para suspeitar de erro ou incompreensão por parte do utilizador..

5 PRÓXIMOS PASSOS

Quando este trabalho estiver realizado, teremos implementado duas funções, que serão testadas por nós localmente no SUPeRB: uma de crítica às procuras quando conseguirmos detetar problemas; outra de sugestão de refinamento ou de esclarecimento no caso de procuras ambíguas ou com resultados a mais.

Estas funções (associadas naturalmente a bibliotecas e recursos que serão atualizados semanalmente, com base nos novos diários e nos novos metadados) serão depois enviadas para o projeto RCAAP que experimentará utilizá-las no seu ambiente de teste e depois de produção se assim o entenderem.

Um segundo passo (que, imaginamos, só poderá ter lugar em 2011) será fazermos o mesmo processo para a procura livre e para a procura por título. Isso sem prejuízo fe contemplarmos, quer no nosso estudo, quer nas funções que iremos desenvolver, todos os casos que conigamos detetar na procura livre que se refiram de facto a uma procura por autor.

6 REFERÊNCIAS

Almeida, José João & Ulisses Pinto. "Jspell — um módulo para análise léxica genérica de linguagem natural". In *Actas do X Encontro Nacional da Associação Portuguesa de Linguística* (Évora, 6-8 de Outubro de 1994), pp. 1-15.

Biblioteca Nacional, Divisão da PorBase, Área de Normalização. *Recomendações para a Construção de Registos de Autoridade de Autor Pessoa Física.* Biblioteca Nacional, Lisboa, 2005, 2ª edição.

Cabral, Luís Miguel, Diana Santos & Luís Fernando Costa. "SUPeRB: Building bibliographic resources on the computational processing of Portuguese". In Daniela Braga, Miguel Sales Dias & António Teixeira (eds.), *PROPOR 2008 Special Session: Applications of Portuguese Speech and Language Technologies (full proceedings)* (Curia, Portugal, 10 de Setembro de 2008).

Cabral, Luís Miguel, Diana Santos & Luís Fernando Costa. "SUPeRB - Gerindo referências de autores de língua portuguesa". In *VI Workshop Information and Human Language Technology (TIL'08)* (Vila Velha, ES, Brasil, 28-29 de Outubro de 2008).

Cabral, Luís Miguel. SUPeRB - **S**istema **U**niformizado de **Pe**squisa de **R**eferências **B**ibliográficas. Tese de Mestrado. Faculdade de Engenharia da Universidade do Porto. Março 2007.

Ribeiro, Fernando & Diana Santos. "Colaboração entre a Linguateca e o RCAAP: Primeiros passos". Apresentação no *Encontro do RCAAP* (Leiria, 22 de Março de 2010).

Sarmento, Luís, Ana Sofia Pinto & Luís Cabral. "REPENTINO - A Wide-Scope Gazetteer for Entity Recognition in Portuguese". In Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira & Maria Carmelita Dias (eds.), *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006 (PROPOR'2006)* LNAI 3960, 13-17 de Maio de 2006, Berlin/Heidelberg : Springer Verlag, pp. 31-40.

Simões, Alberto Manuel & José João Almeida. "`jspell.pm` -- um módulo de análise morfológica para uso em Processamento de Linguagem Natural". In Anabela Gonçalves & Clara Nunes Correia (eds.), *Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística (APL 2001)* (Lisboa, 2-4 de Outubro de 2001), Lisboa: APL, 2002, pp. 485-495.

ÍNDICE