Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

Estudando os nomes dos autores no RCAAP:

relatório do primeiro ano

Diana Santos e Fernando Ribeiro

Área *Linguateca*

4 de Julho de 2011

Trabalho referente à cooperação com o RCAAP

Área *Linguateca*

Diana Santos
Fernando Ribeiro

4 de Julho de 2011

1 RESUMO

Este documento relata o trabalho desenvolvido pela Linguateca de Junho de 2010 até Junho de 2011, 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 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.

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.

Contudo, a Linguateca encontra-se numa fase de reorganização tendo enviado, tanto em 2009 como em 2011, propostas em que contemplava um pólo associado ao RCAAP. Enquanto esperávamos 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.

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, embora tendo passado para uma posição francamente secundária (ver também Cabral et al. 2008a,b).

Por outro lado, a Linguateca tem alguma experiência em indexação e em recolha de informação (RI), podendo chamar à colação 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.

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 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.

De qualquer maneira, as ideias iniciais, já apresentadas 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 dos 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

Além da nossa familiarização com os variados membros (a KEEP, a SDUM e a FCCN/RCAAP) e ferramentas usadas pelo projeto RCAAP, e a compreensão das várias responsabilidades e das várias localizações de dados (Roma e Pavia não se fizeram num dia...), foi também preciso desenvolver algumas ferramentas especializadas para processamento de diários, agrupamento, e sugestões de palavras/nomes próximos.[1]

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, 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.[2]
2. Ambiguação, de forma a produzir cadeias de carateres ainda mais ambíguas e menos precisas a partir de cada identificação

Há ainda alguns aspetos a melhorar, tanto na limpeza dos nomes como na ambiguação. Por exemplo, neste momento ainda não é tratada uma entrada com múltiplas vírgulas, em que podem existir vários autores separados por vírgulas. Apenas se consideram os que têm uma vírgula a separar o primeiro dos restantes nomes. Ou seja, *Almeida, João Dias* é considerado um nome de um autor (João Dias Almeida) mas *Almeida, João Dias, Santos, Diana Maria* já não é considerado um único autor.

[1] 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.

[2] Note-se que, em alguns casos raros, por exemplo envolvendo hífenes, é possível que uma mesma identificação dê origem a mais do que uma forma canónica. Por exemplo, no caso de "Miguel Castelo-Branco", serão consideradas as formas "Miguel Castelo Branco" e "Miguel Castelo-Branco".

Nos metadados aparecem vários nomes com múltiplas vírgulas, mas não é um problema para o seu processamento uma vez que são do tipo «Nolasco, Ana Paula Branco,» e, nestes casos, só é considerada a primeira vírgula.

Na limpeza dos nomes demos especial importância ao tratamento de certas palavras individuais, como passamos a ilustrar.

Por exemplo, quando um nome tem um E maiúsculo, tal como em *Dias E Filho*, o «E» é considerado como um nome intermédio e é acrescentado um ponto para a forma canónica *Dias E. Filho*, mas se o nome for *Dias e Filho*, não há alteração, uma vez que o «e» é considerado como conjunção. Todavia se estiver grafado com um ponto a seguir ao «e», como em *Dias e. Filho* o «e.» passa a ser considerado uma inicial e o resultado da normalização será *Dias E. Filho*.

Outros casos com tratamentos especial são os «de, do, da, dos, das». Nomes que incluam «DE», por exemplo *Joaquim DE Almeida*, passam a «de» (*Joaquim de Almeida*) e nomes com hífen ou sublinhado, como é o caso de *Dias-de-Almeida* ou *Dias_de_Almeida*, passam a ser considerados como nomes distintos, um com hífen e outro sem, ou seja, *Dias de Almeida* e *Dias-de-Almeida*.

Salientamos que os nomes com hífen são uma pequena percentagem dos nomes processsados, 1046 em 52919, o que corresponde a uma percentagem de 1,98 %.

Também no processo de ambiguação dos nomes dos autores tratamos as preposições de maneira especial, simultaneamente deixando-as e retirando-as. Por exemplo, se ambiguarmos o nome *Manuel dos Santos*, o resultado da ambiguação será *Manuel Santos* e *Manuel dos Santos*.

Dados quantitativos referentes à aplicação de ambos os processos descritos às três listas de autores encontram-se na tabela 4.1.1.

Coleção	Original	Normalizado	Ambiguado
SUPeRB	3316	3328	282216
RCAAP	54705	52919	450462
REPENTINO	282222	282216	1852613

Tabela 4.1.1. Quantidade de recursos para cada coleção (19 de Maio 2011)

Um exemplo do resultado de ambos os processos (Normalizado e Ambiguado) pode ser observado na Figura 4.1.1.

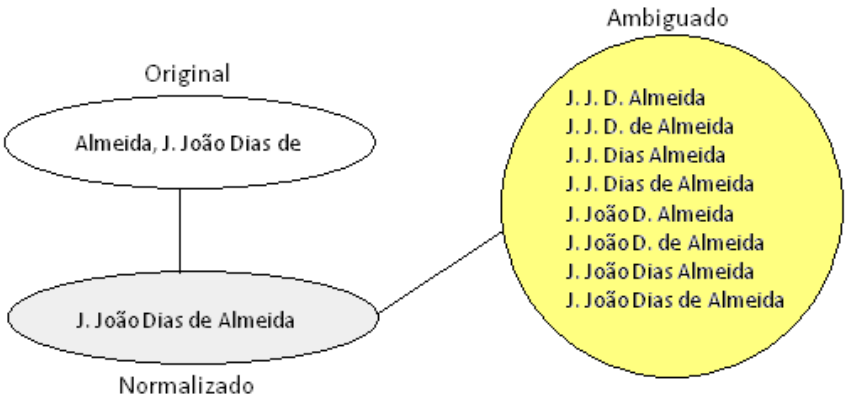

Figura 4.1.1. Exemplo do processo de normalização e ambiguação para o nome *Almeida, J. João Dias de*.

Note-se que, enquanto o REPENTINO é um recurso estático e o SUPeRB é atualizado muito lentamente, os metadados do RCAAP[3] deveriam ser atualizados mensalmente. Contudo, apenas a partir de Fevereiro de 2011 se obteve uma nova versão dos metadados com uma periodicidade mensal.[4] A tabela seguinte mostra a evolução dos dados do RCAAP.

Coleção - RCAAP	Original	Normalizado	Ambiguado
Junho 2010 (PT)	38538	36954	328110
Fevereiro 2010 (PT)	42011	41094	306523
Fevereiro 2010 (BR)	179531	167277	1039940
Fevereiro 2010 (PT-BR)	221204	207289	1335729
Março (PT)	43984	42846	315633
Abril (PT)	52547	50852	427870
Maio (PT)	54705	52919	450462

Tabela 4.1.2. Quantidade de nomes em cada período (últimos valores de 19 de Maio de 2011). A junção de dados do Brasil e Portugal pode levar a uma sobreposição indevida de nomes que serão considerados como um só embora correspondam a pessoas diferentes, por isso não a efetuamos.

[3] Os metadados são gerados pelo próprio RCAAP através das ferramentas de recolha dos metadados que posteriormente nos disponibilizam para análise.
[4] A cooperação entre o RCAAP e o seu equivalente brasileiro permite que novos metadados sejam disponibilizados. No entanto, devido a problemas técnicos da parte brasileira, não podemos ainda obter mensalmente estes metadados.

Os metadados estão a ser recolhidos numa base regular (excluindo os da parte brasileira), de modo a aferirmos da evolução de novos nomes, no que aos autores diz respeito.

A tabela seguinte compara os valores das quantidades de nomes presentes nos metadados e a sua evolução.

Data da recolha	Quantidade	Nomes iguais	«Removidos»	«Novos»
Junho de 2010	38538	30230	8302	11781
Fevereiro de 2011	42011			
Fevereiro de 2011	42011	41622	389	2355
Março de 2011	43984			
Março de 2011	43984	43037	940	9504
Abril de 2011	52547			
Abril de 2011	52547	52323	218	2382
Maio de 2011	54705			

Tabela 4.1.3. Evolução dos nomes nos metadados do RCAAP.

Na tabela 4.1.3, a coluna «data da recolha» indica a altura em que se fez a compilação dos metadados no repositório do RCAAP. Na coluna «quantidade» é apresentado o número de nomes distintos que se obteve para cada versão dos metadados. Na coluna «nomes iguais» está indicado quantos nomes nas duas versões dos metadados são iguais, isto é, estão presentes em ambas as recolhas dos metadados. Na coluna «removidos» estão os valores que apareciam na primeira versão e deixaram de aparecer na segunda versão. Na coluna «novos» estão os nomes que só aparecem na segunda versão.

Visto que a existência de nomes desaparecidos pode não ser evidente para os leitores, convém explicar que um dado documento pode deixar de estar em acesso aberto a partir da altura em que é, por exemplo, aceite para publicação numa revista.

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. Embora tivéssemos começado por tentar observar os diários do Apache, para tentar extrair todos os casos de nomes de autores, fomos alertados pela KEEP para o facto de que o *dspace*, o sistema subjacente ao projecto RCAAP, usa o programa *tomcat* e tem portanto os seus próprios diários.

Apercebemo-nos também de que não seria tão simples ter acesso à interação de todos os utilizadores, visto que o RCAAP é um projeto distribuído e as várias instituições geram os

seus próprios repositórios, por isso numa primeira fase ficámo-nos pelas interacção com o meta-repositório, cujos servidores estão localizados na KEEP.

Seja como for, 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 figura 4.2.1.

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 4.2.1 – Exemplos de entradas no diário para as pesquisas simples (S) e avançada (A).

Note-se, contudo, que pesquisas marcadas como avançadas no RCAAP podem ser derivações de pesquisas simples anteriores. Para a nossa análise reconsiderámos como «pesquisa simples» uma pesquisa que, no diário do portal RCAAP, esteja definida como «pesquisa avançada» mas em que apenas exista o campo «Query», estando os outros campos vazios.

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 exemplificação se encontra na tabela 4.3.1, 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	82	A. A. C. Luís A. A. Costa Luís A. A. da C. Luís A. A. da Costa Luís A. A. Gabriel Luís A. A. G. Luís A. Alberto Gabriel Luís A. Alberto G. Luís A. António C. Luís A. António Costa Luís A. António da C. Luís A. António da Costa Luís A. Bessa Luís A. B. Luís Agustina Bessa Luís Agustina B. Luís A. I. L. Luís A. I. Lopes Luís A. Isabel L. Luís A. Isabel Lopes Luís

		A. L. E. de Jesus Luís A. L. E. de J. Luís A. L. E. Jesus Luís A. L. E. J. Luís A. L. Emidia de Jesus Luís A. L. Emidia de J. Luís A. L. Emidia Jesus Luís A. L. Emidia J. 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 A. Luís Amaral Luís Ana I. L. Luís Ana I. Lopes Luís Ana Isabel L. Luís Ana Isabel Lopes Luís Ana L. E. de Jesus Luís Ana L. E. de J. Luís Ana L. E. Jesus Luís Ana L. E. J. Luís Ana L. Emidia de Jesus Luís Ana L. Emidia de J. Luís Ana L. Emidia Jesus Luís Ana L. Emidia J. Luís A. L. Luís A. Lúcia E. de Jesus Luís A. Lúcia E. de J. Luís A. Lúcia E. Jesus Luís A. Lúcia E. J. Luís A. Lúcia Emidia de Jesus Luís A. Lúcia Emidia de J. Luís A. Lúcia Emidia Jesus Luís A. Lúcia Emidia J. Luís A. Lúcia Luís Ana Lúcia E. de Jesus Luís Ana Lúcia E. de J. Luís Ana Lúcia E. Jesus Luís Ana Lúcia E. J. Luís Ana Lúcia Emidia de Jesus Luís Ana Lúcia Emidia de J. Luís Ana Lúcia Emidia Jesus Luís Ana Lúcia Emidia J. Luís Ana Lúcia Luís Ana R. F. Luís

		Ana R. Francisco Luís Ana Rita F. Luís Ana Rita Francisco Luís António A. Gabriel Luís António A. G. Luís António Alberto Gabriel Luís António Alberto G. Luís A. R. F. Luís A. R. Francisco Luís A. Rita F. Luís A. Rita Francisco Luís

Tabela 4.3.1. Exemplos de grupos máximos

Nos metadados correspondentes a 19 de Maio de 2011, havia 450.261 casos de nomes normalizados diferentes, constituindo 20.970 grupos. Cada nome apenas pertence a um, e um só, grupo máximo: Uma vez que os agrupamentos são dados pela inicial do primeiro nome e pelo último nome, não é possível que existam casos em que um «A. Abreu» vá parar a outro agrupamento que não seja o «A. Abreu». Assim, um nome não pode aparecer em dois grupos máximos distintos.

Quantidade de nomes que existe num grupo	Quantidade de casos
1	4554
2	7224
3	312
4 ou mais nomes	8880

Tabela 4.3.2. Tamanho dos grupos máximos no RCAAP (de 19 de Maio de 2011), por quantidade (quantos grupos têm 1, 2, 3 ou mais nomes).

Foram obtidos 20.970 grupos e cada grupo pode conter entre 1 e N nomes (elementos). Em média, cada grupo de autores contém 21,47 nomes, e a mediana é 2. Pode verificar-se que há mais grupos que apresentam dois nomes, seguido dos grupos com apenas um nome. No que diz respeito a quatro ou mais nomes apenas se apresenta o somatório dos valores obtidos para os grupos, isto é, existem 8.880 grupos que contêm quatro ou mais nomes (e que são a maioria). A figura seguinte mostra a distribuição do número de nomes nos grupos.

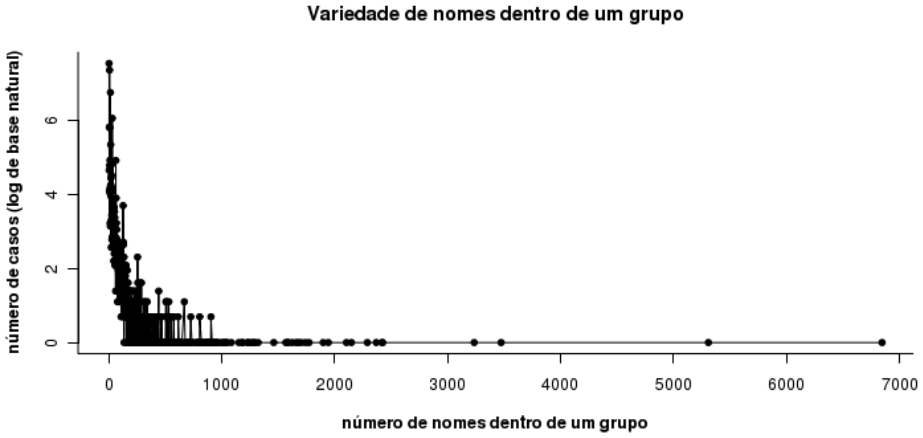

Figura 4.3.2 – Variedade dos nomes dentro de um grupo (em abcissas o número de grupos com 4, 5 ou mais nomes), referente aos metadados de 19 de Maio de 2011).

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).

Neste trabalho, a análise do emparelhamento baseia-se na correspondência exata do nome com os nomes nos agrupamentos, ou seja, o nome que se obtém do diário irá emparelhar apenas se a sequência de caracteres for exatamente a mesma. Como exemplo, se procurar «Dias Almeida» só se obterá resultados caso o nome no grupo contenha exatamente a sequência «Dias Almeida». Este ponto será explicado mais adiante neste relatório.

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	3	A-Rose	A. Rose Adriana Rose
		C-Rose	C. Rose Cheryl Rose
			M. R. Rose

		M. Rose	
João Rodrigues	13	A-Rodrigues	A. João Rodrigues Ana João Rodrigues Ana-João Rodrigues
		A-Gonçalves	A. João Rodrigues Gonçalves António João Rodrigues Gonçalves
		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-Gomes	João Rodrigues Gomes
		M-Jesus	8 elementos
		J-Rodrigues	João Rodrigues

Tabela 4.4.1. Exemplos de emparelhamento entre as procuras e os metadados, usando os grupos máximos.

Na tabela seguinte (4.4.2) encontra-se uma primeira medição da relação entre as procuras e conteúdo dos metadados, usando os diários do portal do RCAAP.

Tipo de ambiguidade (número de grupos)	Quantidade de casos 8362 (após limpeza) (%)
0	4324 (51.7 %)
1	2602 (31,1 %)
2	402 (4,8 %)
3	205 (2,5 %)
4 ou mais grupos	829 (9,9 %)

Tabela 4.4.2. 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 1 Junho de 2011, e os metadados de 19 de Maio de 2011).

Na figura 4.4.2 estão detalhados os casos que pertencem a quatro ou mais grupos. Em média, cada expressão de procura por nome de autor emparelha com 4,14 grupos, e a mediana é 1.

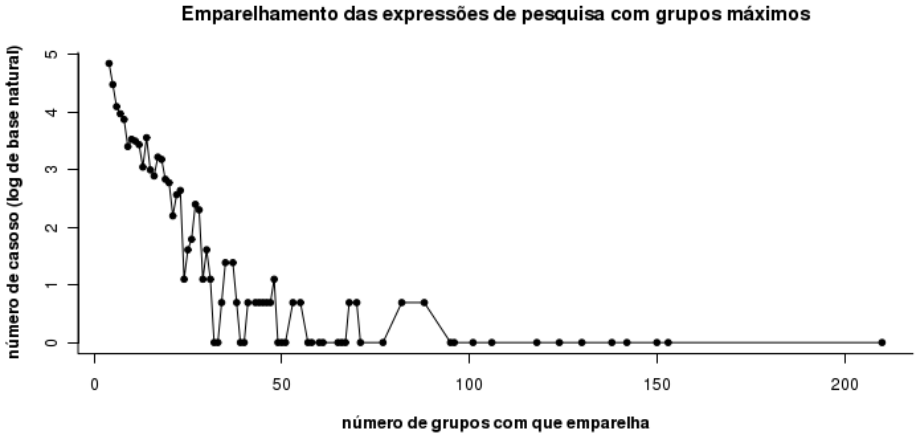

Figura 4.4.2 – Emparelhamento das expressões de pesquisa de autores no portal do RCAAP (meta-repositório) com o conteúdo do repositório (em abcissas, a quantos grupos a expressão pode pertencer), referente aos metadados de 19 de Maio de 2011.

Vemos por este primeiro estudo que 51,7 % dos casos procurados não têm resultados, e que 17,2% são ambíguos – se considerarmos os grupos máximos.

4.5 AGRUPAMENTO POR AUTORES ÚNICOS (MÍNIMOS)

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 4.5.1, 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.[5]

[5] Neste momento estamos a trabalhar com todo o resultado da ambiguação, mas iremos também 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.

	Nomes de autores nos metadados do RCAAP	
Grupo	**Nº elementos**	**Elementos**
José Joaquim de Almeida	12	J. Almeida J. de Almeida J. J. Almeida J. J. de Almeida J. Joaquim Almeida J. Joaquim de Almeida José Almeida José de Almeida José J. Almeida José J. de Almeida José Joaquim Almeida José Joaquim de Almeida
José João Dias de Almeida	34	J. Almeida J. J. Almeida J. João Almeida João Almeida José J. Almeida José João Almeida outros

Tabela 4.5.1. Exemplos de grupos por autor único.

Como se pode verificar, os nomes e os agrupamentos podem alterar-se de acordo com novas versões dos metadados. E, neste particular, um grupo que antes era único pode deixar de o ser, e os seus elementos pertencerem a um outro grupo. No exemplo da tabela anterior, *José João Almeida* deixou de representar um grupo e passou a pertencer ao grupo «José João Dias de Almeida». Outra situação que pode ocorrer é aparecerem mais nomes no grupo ao adicionarem-se mais publicações com novos autores.

A tabela 4.5.1 e figura 4.5.1 descrevem a ambiguidade dos nomes de autores constantes nos metadados, considerando estes grupos (por autores únicos).

Tipo de ambiguidade (número de grupos)	Quantidade de casos
1	429790
2	11766
3	3150
4 ou mais grupos	5550

Tabela 4.5.2. Ambiguidade de cada nome contido nos metadados do RCAAP de 19 de Maio 2011, em termos de a quantos grupos (por autor único) pode pertencer (quantos só pertencem a um, quantos têm duas, três, etc. possibilidades)

A tabela 4.5.2. mostra que os nomes na sua maioria não são ambíguos. Os nomes com mais de uma possibilidade são aqueles que podem pertencer a vários grupos, como por exemplo *J. Almeida* que pertence não só ao grupo do *JOSÉ JOAQUIM DE ALMEIDA* como ao grupo *JOSÉ JOÃO DIAS DE ALMEIDA* (e possivelmente a outros).

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 1,14 grupos, e a mediana é 1 (dados relativos à tabela 4.5.2).

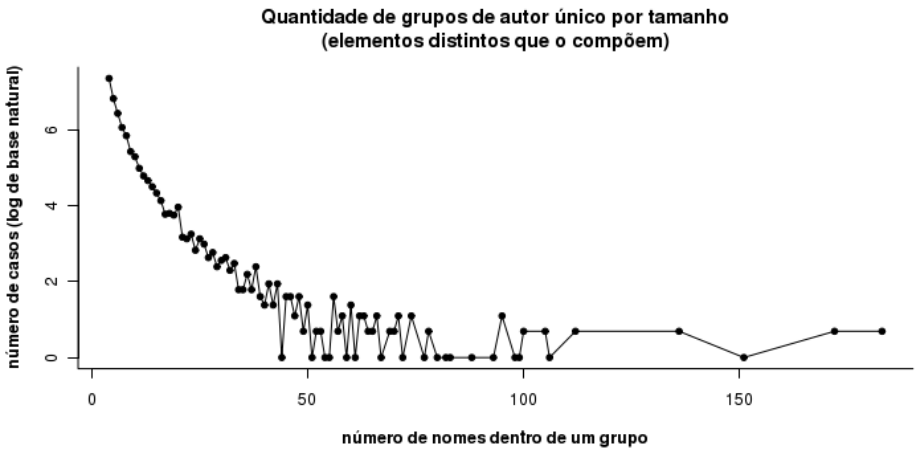

Figura 4.5.1 – Quantidade de grupos de autor único por tamanho (número de elementos distintos que o compõem), para os grupos com mais de três elementos, em escala logarítmica, a partir dos metadados de 19 de Maio de 2011.

A tabela 4.5.3, por seu lado, apresenta os resultados para o número de nomes que existem em cada grupo.

Quantidade de nomes que existe num grupo	Quantidade de casos
1	4673
2	11186
3	867
4 ou mais nomes	25684

Tabela 4.5.3. Tamanho dos grupos por autor único do RCAAP (de 19 de Maio de 2011), ou seja, quantos grupos têm 1, 2, 3 ou mais nomes.

Neste tipo de agrupamento, foram obtidos 42.410 grupos (cada grupo pode conter entre 1 e N nomes (elementos)). Em média, cada grupo de autores contém 12,13 nomes, e a mediana é 5.

A figura 4.5.2 mostra a quantidade de nomes por grupo.

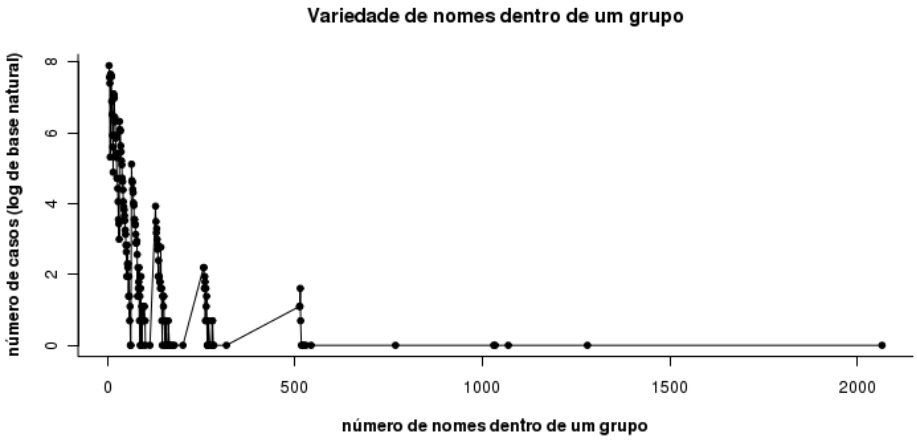

Figura 4.5.2 – Variedade dos nomes dentro de um grupo (em abcissas o tamanho dos grupos com 4, 5 ou mais nomes), referente aos metadados de 19 de Maio de 2011).

A tabela 4.5.4 e a figura 4.5.3 descrevem, analogamente à tabela 4.4.2 e à figura 4.4.2, 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 8362 (após limpeza) (%)
0	4255 (50,8 %)
1	2463 (29,5 %)
2	368 (4,4 %)
3	191 (2,3 %)
4 ou mais grupos	1085 (13,0 %)

Tabela 4.5.4. Emparelhamento das expressões de pesquisa de autores no portal do RCAAP com o conteúdo do repositório de 19 de Maio de 2011

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 19,27 grupos, e a mediana é 1.

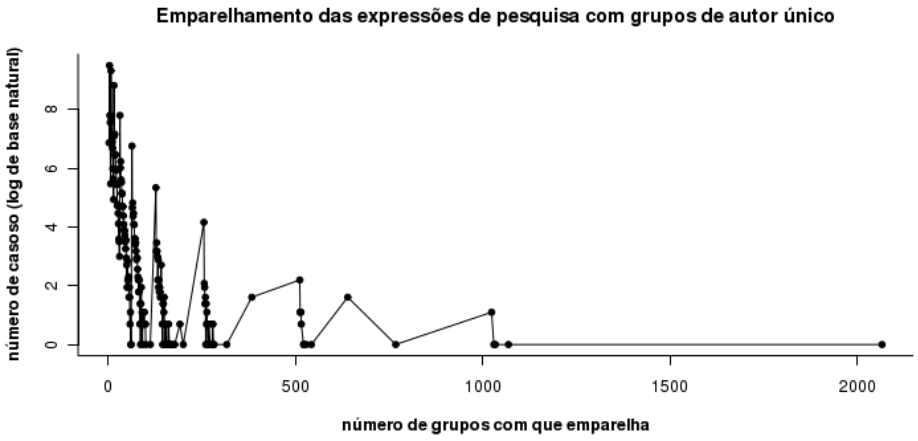

Figura 4.5.3 – Emparelhamento das expressões de procura por autor em função do número diferente de grupos por autor único (em abcissas, a quantos grupos a expressão pode pertencer) em relação ao conteúdo do repositório de 19 de Maio de 2011

Vemos que 50,8 % dos casos procurados não têm resultados, e que 19,7 % são ambíguos, ao considerar os grupos por autor único.

Todos os gráficos e tabelas apresentados anteriormente foram obtidos com base no emparelhamento da sequência exata do nome com os elementos nos grupos, isto é, se procurar por «João Almeida», obter-se-ão apenas resultados em que a sequência presente no nome seja exactamente *João Almeida*, ou seja, *João A. Almeida* já não é considerado. Contudo, incluíram-se os casos onde possa existir as palavras «a, e, da, de, …», o que significa que também emparelham com *João* ***de*** *Almeida*. Apesar de não termos ainda efectuado nenhum estudo sobre outro tipo de emparelhamento, o algoritmo está preparado para essa possibilidade.

Existe uma versão na rede do trabalho desenvolvido para a colaboração entre a Linguateca e o RCAAP em http://www.linguateca.pt/colabRCAAP/, onde se pode explorar estes

tipos de associação escolhendo o menu Agrupamentos. Para se determinar a que agrupamento um determinado nome pertence (ou não) dever-se-á introduzir em primeiro lugar o nome que vamos analisar no campo correspondente. É feita a uniformização, tal como descrito na secção sobre a limpeza, e em seguida é feita a análise de acordo com o tipo de emparelhamento pretendido e em que agrupamento (grupos máximos ou por autor único) se fará a procura. Existem vários agrupamentos para várias datas de recolha dos metadados do RCAAP, que se mantêm para estudar a evolução da população dos nomes.

Vejamos o seguinte exemplo:

Se introduzirmos *Almeida, José J* como nome de pesquisa este nome será convertido para *José J. Almeida*. Caso fosse introduzido *Almeida José, J.* o nome a pesquisar será *J. Almeida José*. Assim, é preciso alguma atenção na colocação de vírgulas para que o resultado seja o desejado.

Na tabela seguinte estão os vários tipos de emparelhamento suportados, apesar do presente relatório apenas apresentar os resultados obtidos com a sequência exata. O nome escolhido para os resultados apresentados na tabela é o *José J. Almeida* e com os agrupamentos de 19 de Maio de 2011.

Emparelhamento	Grupo máximo	Grupo nome único (mínimo)
Sequência exata	J. Almeida – 1 • José J. Almeida	José Joaquim de Almeida -1 • José J. Almeida José João Dias de Almeida – 1 • José J. Almeida
Sequência exata com expansão de iniciais.	J. Almeida – 3 • José J. Almeida • José Joaquim Almeida • José João Almeida	José Joaquim de Almeida – 2 • José J. Almeida • José Joaquim Almeida José João Dias de Almeida – 2 • José J. Almeida • José João Almeida
Sequência com nomes pelo meio sem expansão de iniciais	J. Almeida – 6 • José J. Almeida • José J. D. Almeida • José J. D. de Almeida • José J. Dias Almeida • José J. Dias de Almeida • José J. de Almeida	José Joaquim de Almeida -2 • José J. Almeida • José J. de Almeida José João Dias de Almeida – 6 • José J. Almeida • José J. D. Almeida • José J. D. de Almeida • José J. Dias Almeida • José J. Dias de Almeida • José J. de Almeida

Sequência com nomes pelo meio com expansão de iniciais	J. Almeida – 13 • José J. Almeida • José J. D. Almeida • José J. D. de Almeida • José J. Dias Almeida • José J. Dias de Almeida • José J. de Almeida • José Joaquim Almeida • José Joaquim de Almeida • José João Almeida • José João D. Almeida • José João D. de Almeida • José João Dias Almeida • José João Dias de Almeida	José Joaquim de Almeida -4 • José J. Almeida • José J. de Almeida • José Joaquim Almeida • José Joaquim de Almeida José João Dias de Almeida – 11 • José J. Almeida • José J. D. Almeida • José J. D. de Almeida • José J. Dias Almeida • José J. Dias de Almeida • José J. de Almeida • José João Almeida • José João D. Almeida • José João D. de Almeida • José João Dias Almeida • José João Dias de Almeida
Sequência aleatória	J Almeida – 6 • José J. Almeida • José J. D. Almeida • José J. D. de Almeida • José J. Dias Almeida • José J. Dias de Almeida • José J. de Almeida J Cura – 2 • J. José A. Almeida Cura • J. José Ançã Almeida Cura J Gonçalves – 4 • J. José Almeida S. Gonçalves • J. José Almeida Soares Gonçalves • J. José de Almeida S. Gonçalves • J. José de Almeida Soares Gonçalves	Joaquim José de Almeida Soares Gonçalves – 4 • J. José Almeida S. Gonçalves • J. José Almeida Soares Gonçalves • J. José de Almeida S. Gonçalves • J. José de Almeida Soares Gonçalves José Joaquim de Almeida – 2 • José J. Almeida • José J. de Almeida José João Dias de Almeida – 6 • José J. Almeida • José J. D. Almeida • José J. D. de Almeida • José J. Dias Almeida • José J. Dias de Almeida • José J. de Almeida João José Ançã Almeida Cura – 2 • J. José A. Almeida Cura • J. José Ançã Almeida Cura
Sequência aleatória com expansão de iniciais	35 resultados	57 resultados
Primeiro e últimos Nomes considerados	_	José Joaquim de Almeida (4) José João Dias de Almeida (11)
Expansão do 1º Nome e primeiro e último nome considerado	_	57 resultados

(J. Almeida)		

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 4.6.1 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 4.6.1. Exemplo do novo dicionário do Jspell contemplando nomes próprios, em que «%»corresponde a um espaço e «@» a um ponto.

ss ç

ç ss

ch x

x ch

Figura 4.6.2. 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 4.6.3 ilustramos alguns casos em que este Jspell adaptado 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 4.6.3. 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 de 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 4.7.1 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.»

Formatted: English (U.S.)
Formatted: English (U.S.)

Figura 4.7.1: Exemplos de falsos autores que foram encontrados por nós nos diários

Para estimar o esforço necessário neste tipo de deteção e correção, achámos contudo mais importante ver que tipo de problemas eram mais frequentes, e dedicámo-nos portanto primeiro ao estudo das sessões.

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 determinar 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.

Em primeiro lugar, vamos definir o que entendemos por sessão. Assim, uma sessão corresponde ao caso em que um utilizador faz mais de uma pesquisa consecutiva, e que entre duas pesquisas consecutivas não ultrapassa um determinado valor temporal. De forma a determinar este valor, observou-se o tempo que os utilizadores demoram entre duas pesquisas consecutivas. Duas pesquisas consecutivas de um mesmo utilizador podiam demorar menos de um segundo ou um mês ou mais, como se pode observar na figura 4.8.1.

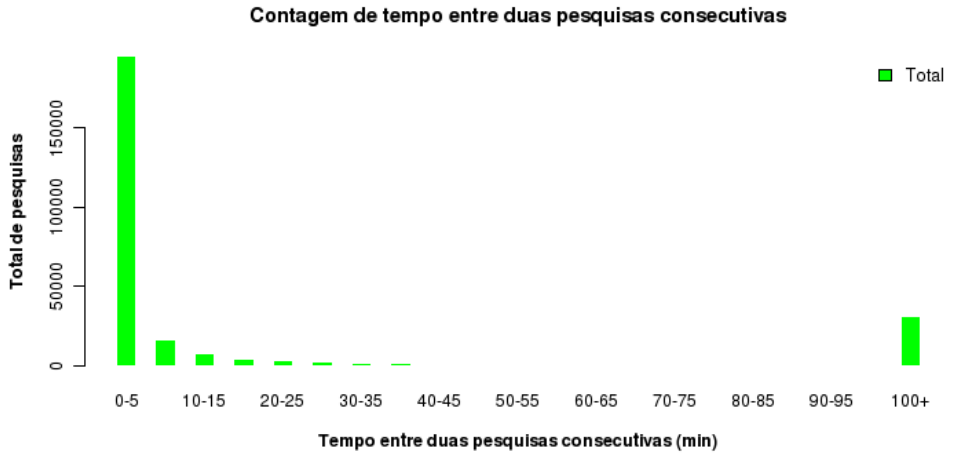

Figura 4.8.1 – Contagem do tempo entre duas pesquisas consecutivas (Até 31 de Maio de 2011).

No entanto, na sua maioria, o tempo gasto entre duas pesquisas consecutivas não é superior a cinco minutos. A partir deste valor, os valores são quase residuais. Com base na análise dos resultados e no gráfico, determinou-se que o tempo ótimo para operacionalizar uma sessão seria de 60 minutos, ou seja, definiu-se uma sessão como uma interação com o RCAAP que contém duas ou mais pesquisas, e em que o tempo entre duas pesquisas consecutivas não é superior a 60 minutos.

Depois de determinado este valor, fez-se a contagem de sessões e obtiveram-se os seguintes resultados:

Sessões	1	2	3	4	5	6	7	8	9	10+
Utilizadores	38830	4269	1135	487	249	136	96	45	36	297

Tabela 4.8.1. Número de sessões por utilizadores (dados até 31 de Maio de 2011).

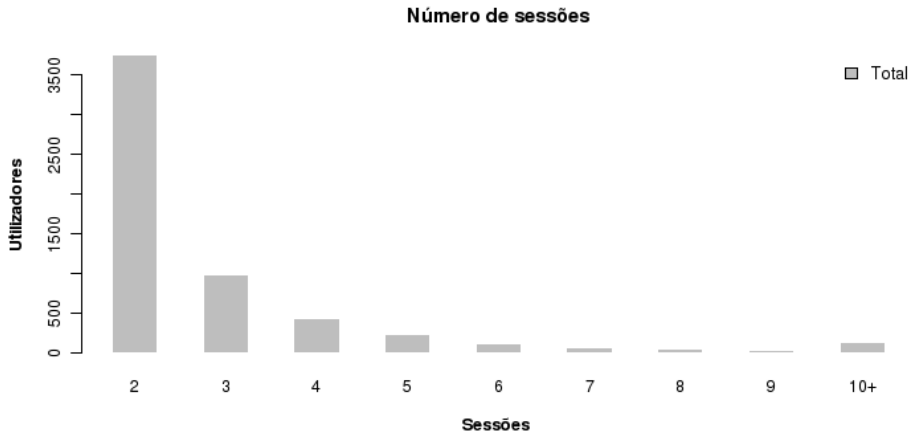

Figura 4.8.2 – Número de sessões por utilizador. Representação dos resultados com duas ou mais sessões (dados de 31 de Maio de 2011).

A tabela mostra que a maioria dos utilizadores apenas tem uma sessão, ou seja, embora faça mais de uma pesquisa fá-lo na mesma sessão. O gráfico 4.8.2 mostra-nos que, à medida que o número de sessões vai aumentando, diminui o número de utilizadores que voltam para fazer mais pesquisas. Este facto permite dar uma ideia de como funciona a interação com o sistema – um utilizador faz várias pesquisas até encontrar o que pretende e, aparentemente, na maioria das vezes isso é suficiente e ele não volta ao RCAAP. Contudo, é preciso

relembrar que o nosso conceito de «utilizador» é muito fraco – se uma mesma pessoa interagir com o RCAAP doutro computador, é considerado como novo utilizador.

Também analisámos o número de pesquisas por cada sessão e os resultados podem ser observados na tabela 4.8.2.

Pesquisas	2	3	4	5	6	7	8	9	10+
Sessões	18267	11271	7772	5333	3883	2828	2045	1464	6007

Tabela 4.8.2. Número de utilizadores que têm sessões com 2, 3, … 10 ou mais pesquisas. (Dados de 31 de Maio de 2011)

A tabela 4.8.2 não apresenta valores para uma pesquisa só, uma vez que só se considera sessão quando existe mais de uma pesquisa. O gráfico 4.8.3 mostra a distribuição do número de pesquisas por sessão.

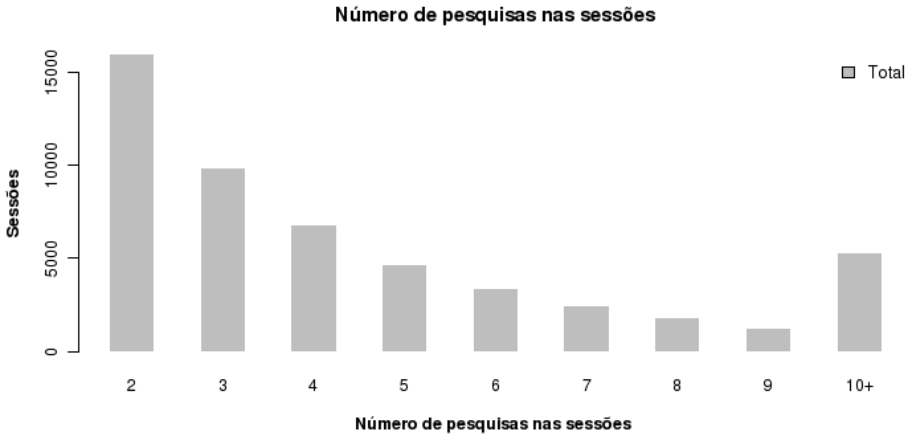

Figura 4.8.3 – Gráfico da distribuição do número de pesquisas nas sessões. (Dados de 31 de Maio de 2011)

Como se pode observar na figura 4.8.3., as sessões mais frequentes têm apenas duas pesquisas. Contudo, as sessões com três e quatro pesquisas também têm valores significativos. Este facto permite verificar que não é raro um utilizador realizar várias pesquisas no RCAAP numa mesma sessão.

Finalmente, também analisámos a duração de cada sessão. Ao escolher-se um tempo máximo entre duas pesquisas consecutivas de 60 minutos não implica que as sessões tenham exactamente esta duração. Pode acontecer que haja sessões de apenas alguns segundos e outras que demorem horas. O que tínhamos de garantir é que, entre duas pesquisas consecutivas, o tempo não poderia ultrapassar os 60 minutos. Se isso acontecesse

então não contava como uma sessão. O gráfico 4.8.4. mostra a distribuição dos tempos de sessão. Podemos verificar que as sessões são maioritariamente de curta duração, geralmente não ultrapassando os 5 minutos.

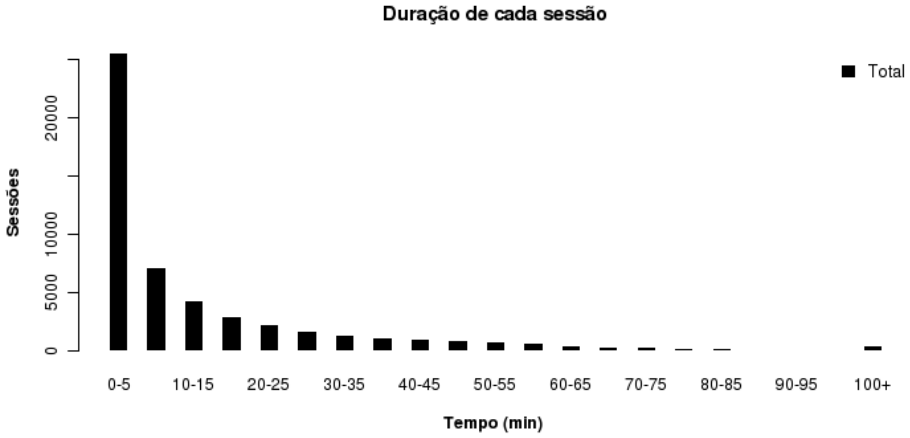

Figura 4.8.4 – Gráfico do tempo que dura cada sessão. (Dados de 31 de Maio de 2011)

Em segundo lugar, vamos definir o conceito de «visita». Este conceito vai ser importante para podermos analisar o comportamento dos utilizadores e também está relacionado com o conceito de sessão. Uma visita corresponde a um acesso do utilizador ao sistema. Uma visita pode ser:

- Uma pesquisa solitária, a que chamamos pois **visita unitária**;

- Um conjunto de pesquisas que constituem uma sessão – neste caso, sessão e visita correspondem à mesma realidade.

As visitas podem ser únicas, em que o utilizador acede uma vez só ao sistema e nunca mais volta, ou repetidas, no caso em que o utilizador volta ao sistema. Do ponto de vista da constituição de cada visita, esta pode ser unitária, em que o utilizador apenas faz uma pergunta e portanto não cria uma sessão, ou múltipla, dando pois origem ao que chamamos sessão.

Para clarificar melhor estes conceitos, veja-se o seguinte exemplo:

100.100.100.001 : 26-Jan-2011 10.45
100.100.100.001 : 26-Jan-2011 10.55
200.200.200.002 : 26-Jan-2011 13:30
100.100.100.001 : 26-Jan-2011 13.55
200.200.200.002 : 26-Jan-2011 14:00
100.100.100.001 : 26-Jan-2011 17.55

O utilizador 100.100.100.001 fez três visitas, em que uma delas corresponde a uma sessão (a de 26 de Janeiro, com pesquisas às 10.45 e às 10.55). Destas três visitas, duas são unitárias, ou seja, só contêm uma pesquisa. Por outro lado, o utilizador 200.200.200.002 apenas fez uma visita (embora com duas pesquisas, constituindo portanto uma sessão), logo essa visita é única.

A tabela 4.8.4. apresenta o número de visitas ao RCAAP, considerando cada combinação «IP e agente» como um diferente utilizador.

Visitas	Utilizadores
1	55439
2	6705
3	1961
4+	2151

Tabela 4.8.4. Número de visitas por utilizador. (Dados de 31 de Maio de 2011)

Figura 4.8.5 – Número de visitas por utilizador. (Dados de 31 de Maio de 2011)

Como se pode verificar na tabela 4.8.4., os utilizadores na sua maioria fazem apenas uma visita ao RCAAP (relembramos que as visitas podem englobam várias pesquisas). Também se pode verificar pelo gráfico da figura 4.8.5 que, no caso de visitas múltiplas, a maioria só volta uma vez.

De qualquer maneira, é importante salientar que a heurística de identificação de um utilizador pelo seu IP, mesmo distinguindo entre agentes diferentes, não permite distinguir

entre vários utilizadores por trás da mesma *proxy*. Indicamos na próxima tabela, de qualquer maneira, a diferença entre o número de utilizadores usando apenas o IP e o número de utilizadores refinado pela identificação do agente. (Todos os resultados apresentados anteriormente, repetimos, são baseados no IP + agente.)

	nome - IP	nome - IP + agente	diferença
Utilizadores	43699	66256	+22557

Tabela 4.8.5 – Número de utilizadores sem considerar o agente e considerando o agente.(Dados de 31 de Maio de 2011)

Em relação às sessões, a diferença entre considerar o nome como o IP apenas ou considerar também o agente pode observar-se na tabela 4.8.6.

sessões	nome - IP	nome - IP + agente	diferença
1	26431	38830	+12399
2	2989	4269	+1280
3	852	1135	+283
4	373	487	+114
5	194	249	+55
6	118	136	+18
7	71	96	+25
8	64	45	-19
9	49	36	-13
10+	425	297	-128

Tabela 4.8.6 – Resultados das sessões, quando o nome é o IP e IP mais agente. Apresenta-se também a diferença entre estes valores. (Dados de 31 de Maio de 2011)

Como se pode verificar, usando o agente em conjunto com o IP obtêm-se resultados bastante diferentes. O número de utilizadores cresce de forma considerável, e apenas os utilizadores com muitas sessões decrescem ligeiramente.

Tempo de sessão (min)	nome - IP	nome - IP + agente	diferença
0-5	23956	29241	+5285
5-10	6862	8128	+1266

10-15	4168	4877	+709
15-20	2864	3328	+464
20-25	2196	2495	+299
25-30	1638	1871	+233
30-35	1420	1584	+164
35-40	1147	1307	+160
40-45	981	1084	+103
45-50	880	960	+80
50-55	775	840	+65
55-60	637	692	+55
60-65	479	505	+26
65-70	338	354	+16
70-75	286	295	+9
75-80	219	218	-1
80-85	191	187	-4
85-90	137	142	+5
90-95	148	132	-16
95-100	83	88	+5
100+	576	512	-64

Tabela 4.8.7 – Resultados dos tempos médios que cada utilizador gasta por sessão, para nomes só com IP ou nomes com IP mais agente e a diferença entres os valores obtidos. (Dados de 31 de Maio de 2011)

Em relação ao tempo de sessão, há uma subida considerável nos casos que vão até aos cinco minutos. Isto explica-se porque no caso de um IP por utilizador estava-se a amalgamar casos de vários utilizadores a interagirem ao mesmo tempo, dando valores humanamente impossíveis (como 0) entre duas pesquisas, o que com IP-agente foi um pouco reduzido.

pesquisas	nome - IP	nome - IP + agente	diferença
2	14702	18267	+3565
3	9345	11271	+1926
4	6581	7772	+1191

5	4587	5333	+746
6	3443	3883	+440
7	2452	2828	+376
8	1822	2045	+223
9	1324	1464	+140
10+	5725	6007	+282

Tabela 4.8.8 – Resultados obtidos para as pesquisas por sessão, para nomes só com IP ou nomes com IP mais agente e a diferença entres os valores obtidos. (Dados de 31 de Maio de 2011)

visitas	nome - IP	nome - IP + agente	diferença
1	35898	55439	+19541
2	4623	6705	+2082
3	1360	1961	+601
4	563	786	+223
5	297	440	+143
6	183	245	+62
7	141	156	+15
8	86	109	+23
9	59	71	+12
10	62	66	+4
11	38	51	+13
12	28	30	+2
13	26	25	-1
14	18	16	-2
15	24	22	-2
16	16	17	+1
17	13	9	-4
18	21	10	-11
19	13	5	-8
20+	230	93	-137

Tabela 4.8.9 – Resultados obtidos para as visitas, para nomes só com IP ou nomes com IP mais agente e a diferença entres os valores obtidos. (Dados de 31 de Maio de 2011)

Os gráficos e tabelas que foram apresentados neste relatório estão em constante desenvolvimento. Assim, a actualização será feita directamente na *internet*, nos URL indicados):

Figura 4.3.2 –http://www.linguateca.pt/colabRCAAP/images/variedadeGruposMaximos.png

Figura 4.4.2 –http://www.linguateca.pt/colabRCAAP/images/empProcuraGruposMaximos.png

Figura 4.5.1 –http://www.linguateca.pt/colabRCAAP/images/ambMetadadosGruposUnicos.png

Figura 4.5.2 –http://www.linguateca.pt/colabRCAAP/images/variedadeGruposUnicos.png

Figura 4.5.3 –http://www.linguateca.pt/colabRCAAP/images/empProcuraGruposUnicos.png

Figura 4.8.1 –http://www.linguateca.pt/colabRCAAP/images/presessao.png

Figura 4.8.2 –http://www.linguateca.pt/colabRCAAP/images/sessoes.png

Figura 4.8.3 –http://www.linguateca.pt/colabRCAAP/images/procuraSessao.png

Figura 4.8.4 –http://www.linguateca.pt/colabRCAAP/images/tempoPorSessao.png

Figura 4.8.5 –http://www.linguateca.pt/colabRCAAP/images/visitas.png

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 fornecendo 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.

O nosso trabalho no futuro imediato é o de estudar as correções/repetições de procuras feitas pelos utilizadores, para tomarmos o pulso das intenções e problemas na interação, através do estudo das sessões encontradas.

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 será fazermos o mesmo processo para a procura livre e para a procura por título. Isso sem prejuízo de contemplarmos, quer no nosso estudo, quer nas funções que iremos desenvolver, todos os casos que consigamos 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

1 RESUMO 1

2 INTENÇÕES E PRESSUPOSTOS 2

3 OBJETIVOS DE UMA PRIMEIRA COLABORAÇÃO 3

4 TRABALHO REALIZADO 4

4.1 Estudo de nomes de pessoas (autores) 4
4.2 Estudo dos diários de acesso («logs») 7
4.3 Agrupamento dos nomes dos autores 9
4.4 Medição do emparelhamento das procuras 12
4.5 Agrupamento por autores únicos (mínimos) 14
4.6 Semelhança de autores 21
4.7 Filtragem dos diários 22
4.8 Identificação de sessões 23
4.9 Identificação do que o utilizador realmente quer 31

5 PRÓXIMOS PASSOS 32

6 REFERÊNCIAS 33