Tirando o chapéu à Wikipédia: A coleção do Pág ico e o Cartola

Alberto Simões
Instituto de Letras e Ciências Humanas
Universidade do Minho
ambs@ilch.uminho.pt

Luís Costa
Linguateca/FCCN
luis.f.kosta@gmail.com

Cristina Mota
Linguateca/FCCN
cmota@ist.utl.pt

Resumo

Este artigo apresenta a coleção do Pág ico, ou seja, a coleção subjacente ao Pág ico, e o pacote de recursos do Pág ico, o Cartola, que inclui a própria coleção.

O artigo dá particular destaque à construção da coleção do Pág ico, uma coleção de documentos da Wikipédia portuguesa. Esta coleção foi criada com o objetivo de garantir (i) igualdade no recurso usado por todos os participantes, (ii) homogeneidade nas respostas e (iii) semi-automatização na avaliação das respostas. Em primeiro lugar, será justificada a necessidade da criação deste recurso. Posteriormente, serão apresentadas as alternativas existentes para a sua criação, qual a escolhida, e quais os problemas encontrados.

Além disso, o artigo caracteriza, segundo várias vertentes, a coleção do Pág ico bem como uma subcoleção desta, correspondente ao monte do Pág ico. O monte do Pág ico, também incluído no Cartola, inclui todas as respostas e justificações distintas encontradas pelos criadores de tópicos e pelos participantes.

Palavras chave

Wikipédia, Pág ico, Coleção, XHTML, Wiki

1 Introdução

Uma das grandes vantagens de uma avaliação conjunta é produzir um conjunto de recursos que podem ser usados no futuro para avaliar outros sistemas, estabelecendo uma bitola e a respetiva bancada de teste.

No decurso do Pág ico foi criado o Cartola, um pacote de recursos público constituído pela coleção do Pág ico (a coleção de documentos da Wikipédia de onde as respostas e justificações deviam ser escolhidas, primeiro, pelos criadores de tópicos e, depois, pelos participantes), por exemplos de tópicos e correspondentes respostas associadas às suas justificações, pelos tópicos de avaliação e pelas respostas dos criadores de tópicos e dos participantes com a respetiva avaliação feita pela organização. O Cartola é disponibilizado pela Linguateca em `http://www.linguateca.pt/Cartola/` e inclui especificamente:

- a coleção de documentos (689 629) da Wikipédia portuguesa usada no Pág ico;
- 11 exemplos de tópicos com as respetivas respostas e justificações (85);
- os 150 tópicos usados na avaliação;
- as corridas dos sistemas e as respostas dos participantes humanos em formato de corridas[1];
- o monte do Pág ico, ou seja, a coleção de todas as respostas com as suas justificações encontradas no Pág ico (quer pelos criadores de tópicos quer pelos participantes) e a respetiva avaliação.
- listas das respostas distintas corretas com (2 250) e sem as justificações (1 871);
- lista das respostas distintas corretas quer tenham sido bem ou mal justificadas, sem justificações (1 979);
- lista das respostas consideradas duvidosas.

Os tópicos de avaliação do Pág ico encontram-se descritos em (Freitas, 2012), enquanto (Freitas et al., 2012) discute a avaliação das respostas, analisando entre outras coisas as respostas duvidosas. Este artigo, por outro lado, foca a coleção do Pág ico e uma subcoleção desta, correspondente ao monte das respostas do Pág ico e que inclui, portanto, todos os documentos que foram usados como resposta ou justificação no Pág ico.

A coleção do Pág ico é uma coleção de documentos criada a partir de uma versão estática da Wikipédia portuguesa. Começaremos por justificar, na secção 2, a necessidade de criar este recurso. Discutiremos, em seguida, nas secções 3.1 e 3.2, um conjunto de definições e convenções usadas na Wikipédia que teria de ser estudado pelos participantes a fim de conseguirem processar de forma satisfatória as cópias disponibilizadas da

[1] Como referido em (Mota, 2012), a partir das respostas dadas no SIGA pelos participantes humanos foram criadas as corridas equivalentes.

Wikipédia. A secção 4 explica como o formato da Wikipédia foi processado e convertido num conjunto de documentos XHTML (que sendo um formato muito mais simples e amplamente usado facilita o processamento por parte dos participantes), o qual constitui a coleção do Págico.

Posteriormente, na secção 5, faremos uma caracterização da coleção do Págico e da subcoleção do monte do Págico de diferentes perspetivas. Esta caraterização irá permitir ao leitor ter uma noção da abrangência dos tópicos propostos para avaliação em relação à coleção como um todo. Além disso, permitirá que potenciais interessados no uso de coleções semelhantes em futuras avaliações fiquem com uma imagem do conteúdo real da Wikipédia portuguesa.

Terminaremos com algumas conclusões e propostas de melhoramentos para futuras edições, seja do Págico, seja de outras avaliações que usem a Wikipédia como fonte de informação.

2 Criar uma nova coleção, sim ou não?

A Wikipédia é um recurso em constante mutação. Por um lado, é o conteúdo que muda a cada instante, por outro, são as regras e a sintaxe que vão evoluindo. Esta constante mudança faz com que não seja um recurso fácil de usar para uma avaliação de qualquer tipo de ferramenta.

No caso concreto do Págico (consulte-se os restantes artigos nesta edição para mais informação sobre outros aspetos desta avaliação conjunta), em que se pretende avaliar ferramentas de recolha de informação na Wikipédia, este facto é de grande importância. No Págico, os participantes têm de encontrar artigos da Wikipédia que respondam a um tópico. Ora, se não existir uma versão estável, onde os participantes devam encontrar as ditas respostas, é possível que em determinado dia:

- exista um artigo que sirva de resposta (ou justificação) a um dos tópicos do Págico e que, no dia seguinte, esse artigo tenha desaparecido;
- não exista o artigo que sirva de resposta (ou justificação), mas no dia seguinte já tenha sido criado;
- exista um artigo que no dia seguinte é alterado de tal forma que invalida que seja uma resposta correta (ou que justifique adequadamente uma resposta).

Teria sido possível usar a coleção desenvolvida para o GikiCLEF (Santos et al., 2010), no entanto optou-se por usar uma versão mais recente da Wikipédia. Além do facto de garantir mais proximidade com a Wikipédia atual, também permite que possamos analisar (neste artigo) o estado da Wikipédia portuguesa. Infelizmente a abordagem usada para a construção da coleção para o GikiCLEF não foi possível de ser repetida já que a ferramenta usada já não é mantida.

Foi, então, necessário construir uma coleção estática que pudesse ser usada por todos os participantes, e que tornasse a avaliação mais simples (ou mesmo, possível). Para isso foi usada uma cópia estática da Wikipédia (a própria Fundação Wikimedia disponibiliza cópias regulares das várias versões da Wikipédia) de 25 de Abril de 2011[2].

Embora estas cópias estáticas da Wikipédia sejam disponibilizadas em vários formatos (como SQL, para introdução direta num gestor de bases de dados, ou num único documento XML com todos os artigos), esses formatos não são fáceis de processar, quer pelo seu tamanho, quer pelo próprio formato em que são disponibilizados, o que será discutido em seguida.

3 A Wikipédia

Todos conhecemos a Wikipédia, e já a consultámos pelo menos um par de vezes. No entanto, conhecemos a Wikipédia do ponto de vista de um utilizador comum, que consulta e lê um conjunto de artigos, e possivelmente não como um membro da comunidade da Wikipédia, tentando melhorar artigos, ou contribuir com novos artigos. Mesmo que já tenha editado um ou dois artigos da Wikipédia é natural que não tenha compreendido como a estrutura da Wikipédia é rica e, ao mesmo tempo, complexa.

A Wikipédia não é apenas um sistema *wiki* em que cada página corresponde a um artigo de uma enciclopédia. Existe uma estrutura de espaços de nomes (*namespaces*), entradas, entradas de desambiguação e de redireção, e macros e funções. Nesta secção apresentamos (de forma superficial) a estrutura e a sintaxe de macros e funções da Wikipédia relevantes à construção da coleção do Págico.

3.1 Estrutura da Wikipédia

A Wikipédia começou, como não podia deixar de ser, como um conjunto de páginas, em que cada uma correspondia a determinado artigo de uma enciclopédia virtual. Pouco tempo decorrido e

[2] Disponível no sítio da Wikipédia em `http://dumps.wikimedia.org/ptwiki/20110425/`.

surgiram espaços de nomes (*namespaces*) especiais, para guardar tipos de páginas que não correspondem a artigos. A secção 5.1.1 descreve um conjunto destes tipo de espaços. Enquanto que na navegação da Wikipédia é mais ou menos claro o que corresponde a um artigo da enciclopédia, e o que constitui um documento auxiliar de gestão, na cópia estática é necessário fazer essa divisão de forma manual, detetando em que espaço cada documento está.

Exemplos destes espaços de gestão são os *redireção* e *desambiguação*, que albergam páginas que servem de entradas preferenciais ou entradas de desambiguação para artigos (e que são descritos de seguida). Existe um outro espaço de gestão muito importante, denominado de *pré-definição*, que é explicado na secção 3.2.

3.1.1 Páginas de desambiguação

As páginas de desambiguação são usadas em situações em que uma palavra é polissémica. Nestes casos o utilizador é confrontado com um conjunto de resumos das páginas que representam cada um dos possíveis sentidos dessa palavra.

Por vezes a página de desambiguação não é logo apresentada. Por exemplo, ao procurar por *banco* o utilizador é redirecionado automaticamente para a página sobre a instituição financeira. Junto com o título da página aparece uma nota que permite ao utilizador saber que existem outros significados para a palavra, e deste modo aceder à página de desambiguação.

No entanto, se procurar por uma palavra ainda mais genérica, como *tipo*, a página de desambiguação é logo apresentada.

3.1.2 Redirecionamento

Durante a preparação da coleção do Págico foram encontrados dois tipos de redirecionamento, um dos quais está a cair em desuso.

O tipo de redirecionamento oficial serve para que um utilizador que procure um título que representa um tópico polimórfico (que pode ser descrito de diversas formas) o consiga encontrar. Exemplos são a pesquisa de um plural (*cavalos* em vez de *cavalo*) ou mesmo outro tipo de palavras relacionadas (*escravo* em vez de *escravidão*). Nestas situações a Wikipédia faz a ligação direta da pesquisa à página de destino, sem passar por uma página com o título procurado. No entanto, e junto do título (tal como no caso de palavras com página de desambiguação), é apresentada a forma original procurada pelo utilizador (*Escravidão (Redirecionado de Escravo)*).

O outro tipo de redirecionamento encontrado usa (ou usava) uma página intermédia, quase que como uma entrada remissiva num dicionário, que indicava ao utilizador que devia usar outra palavra para procurar a página desejada. Sendo apenas esta a informação que esta página continha não fazia sentido a sua existência, e talvez tenha sido essa a razão pela qual foram desaparecendo (durante a escrita deste artigo não se encontrou nenhum exemplo ilustrativo deste tipo de redirecionamento, no entanto foram encontrados vários casos na versão estática utilizada—que, note-se, tem cerca de um ano de idade).

3.2 A sintaxe MediaWiki

A sintaxe usada na Wikipédia é a sintaxe do sistema de Wiki MediaWiki. Não faz sentido nesta secção descrever toda a sintaxe suportada, já que corresponde a uma sintaxe Wiki comum, em que são usados carateres ASCII para a formatação do texto. A descrição oficial desta linguagem pode ser consultada, por exemplo, em `http://en.wikipedia.org/wiki/Wikipedia:Cheatsheet`.

Faz sentido, sim, referir o mecanismo de macros usado por esta linguagem, uma vez que se tornou uma pedra no processo de construção da coleção.

O mecanismo de macros permite que se definam abreviaturas, opcionalmente parametrizadas, que expandam em sintaxe Wiki ou diretamente em notação HTML.

Estas macros são definidas num espaço próprio (denominado *pré-definição* na Wikipédia portuguesa). Um exemplo de uma pré-definição é "`http://pt.wikipedia.org/wiki/Predefinição:POR`", que é uma macro para a inclusão da bandeira portuguesa juntamente com uma hiperligação para o artigo *Portugal*. Deste modo, basta usar `{{POR}}` numa página para que esta seja expandida na dita bandeira e hiperligação.

Existem macros bastante mais complexas. Um exemplo de uma macro parametrizada é a "`http://pt.wikipedia.org/wiki/Predefinição:Bandeira`," que permite a inclusão de bandeiras de qualquer país, com possibilidade de escolher uma variante (por exemplo, a da monarquia portuguesa), o tamanho da bandeira e o texto a ser apresentado. Um exemplo de uso desta macro será `{{Bandeira|Alemanha|império}}`.

Estas macros podem conter código condicional, opções condicionais, opções com valores por omissão e mais uma panóplia de opções que as tornam muito poderosas. Por exemplo, as

célebres tabelas (denominadas por *infobox*) usadas em páginas como as de países, cidades ou animais, que sistematizam alguma informação numa barra vertical ao lado direito, são geradas usando macros.

4 Construção da coleção do Págico

Tendo sido decidido que o formato original da Wikipédia não seria o ideal para a coleção, por obrigar os participantes a compreender o funcionamento quer da sintaxe Wiki, quer das macros, foi decidido que a melhor opção seria converter os artigos em documentos XHTML. É certo que podíamos ter optado por soluções como a apresentada em (Junior et al., 2011), em que a Wikipédia é, de algum modo, simplificada ou sumariada, mas passaríamos a estar mais longe do que é a Wikipédia original.

Em todo o caso, a escolha da conversão da Wikipédia em XHTML faz sentido uma vez que grande parte da recolha de informação dos dias que correm é feita sobre a Rede, em que grande parte dos documentos estão codificados em HTML ou XML, ou sobre documentos estruturados, armazenados por ferramentas específicas e que, na sua grande maioria, também são armazenadas em XML.

O uso de HTML (ou XHTML) como formato de eleição para a coleção do Págico teve outras vantagens, nomeadamente o de possibilitar o uso de uma ferramenta já desenvolvida para a gestão de avaliações deste tipo (o SIGA(Costa, Mota e Santos, 2012), por exemplo).

Nesta secção faremos uma apresentação inicial das alternativas para o processamento da coleção e geração de documentos XHTML, seguindo-se uma breve explicação de quais as ferramentas escolhidas, e de como foram usadas. Terminaremos com alguns dos problemas encontrados, bem como a solução adotada.

4.1 Ferramentas disponíveis

Grande parte das ferramentas disponíveis para a conversão da Wikipédia para outros formatos não tem tido atualizações recentemente[3]. Além disso, o facto de serem ferramentas não desenvolvidas pelos programadores da ferramenta MediaWiki leva a que não suportem a totalidade da sintaxe usada na Wikipédia. Ora, não havendo atualizações para estas ferramentas, e estando a Wikipédia em constante evolução, este problema é acentuado. Foram testadas várias ferramentas, como o *FlexBisonParse*, *Wiki2XML mediawiki-parser*, entre outros. Alguns não se conseguiram instalar, outros não reconheciam o formato XML da Wikipédia, e outros ainda geravam documentos de forma não satisfatória.

A abordagem mais prometedora seria a instalação de um servidor HTTP e uma base de dados para onde se importasse toda a Wikipédia, e instalar uma versão recente do MediaWiki. Tendo esta configuração, muitas ferramentas estavam disponíveis, e mesmo que não estivessem, uma ferramenta de *crawling* conseguiria, de forma simples, obter uma cópia local em HTML. No entanto a meta-informação (como quais as páginas que são de redireção) seria perdida.

A primeira ferramenta que mostrou resultados aceitáveis foi a `mwlib`[4], um conjunto de conversores em Python. Dada a proximidade do evento optou-se por usar esta biblioteca mesmo com todos os problemas encontrados (e que serão descritos mais à frente).

Para auxiliar o processo, foi usado um módulo Perl, `MediaWiki::DumpFile`[5], que permite percorrer a cópia estática em XML e extrair meta-informação.

4.2 Abordagem adotada

O processo detalhado de conversão do formato XML em ficheiros XHTML está descrito na página do Págico, em `http://linguateca.pt/Pagico/`. Nesta secção limitar-nos-emos a enumerar os passos necessários.

O processamento foi feito com base na cópia estática da Wikipédia, nomeadamente na sua cópia em formato XML, de nome `pages-articles.xml.bz2`. Este documento inclui todos os artigos da Wikipédia num único documento XML. A anotação XML é usada para toda a meta-informação, e os artigos estão descritos de forma textual, na sintaxe wiki.

Infelizmente a ferramenta que escolhemos (`mwlib`) foi desenvolvida para a versão inglesa da Wikipédia, o que nos trouxe alguns problemas. Nomeadamente, foi necessário realizar alterações diretamente no código fonte da ferramenta para que esta considerasse o documento XML na língua portuguesa.

O módulo Perl `MediaWiki::DumpFile` percorre todo o ficheiro XML obtendo meta-informação sobre cada artigo e, dependendo do seu tipo, tomando diferentes ações. No caso de

[3] Existe uma lista de ferramentas de conversão disponíveis em `http://www.mediawiki.org/wiki/Alternative_parsers`.

[4] Disponível em `http://pediapress.com/code/`.

[5] Disponível em `http://search.cpan.org/~triddle/MediaWiki-DumpFile-0.2.1/`.

ser um artigo comum, a ferramenta da `mwlib` para conversão em XML era invocada. No caso de ser uma página de redireção oficial, era gerado um documento HTML apenas com a ligação para a página oficial. Finalmente, em casos especiais, como páginas de desambiguação e páginas referentes a imagens, foram simplesmente descartadas.

Os documentos produzidos em XHTML foram arrumados numa árvore de diretorias, organizados pelos três primeiros carateres do título do documento. Além disso, os documentos foram processados pela ferramenta `xmllint` para garantir a correção dos documentos gerados.

4.3 Problemas encontrados

Foram vários os problemas encontrados durante a criação da coleção, o que explica a disponibilização quase consecutiva de 7 versões da coleção. Muitos destes problemas deveram-se a comportamentos não esperados por parte das ferramentas utilizadas. Por exemplo, a primeira versão disponibilizada a 1 de Agosto de 2011 incluía algumas páginas de redireção não detetadas.

Outras versões foram criadas por pequenos erros incluídos na preparação das coleções anteriores, como a incorreta normalização de títulos (carateres não previstos) ou a correção das hiperligações internas à coleção.

No entanto, os principais problemas encontrados foram as páginas de redireção não oficiais e o processamento das macros.

Em relação às páginas de redireção não oficiais, a decisão foi ignorar. Felizmente, não foram detetadas muitas destas páginas. Em todo o caso, a decisão seria a mesma, já que não existe uma forma clara para distinguir a página de redireção (intermédia) de uma página comum.

Processar as macros de forma satisfatória foi um problema mais complicado. Estas macros não podem ser ignoradas, já que levaria a que muita informação fosse perdida. Veja-se por exemplo a macro `{{POR}}` apresentada anteriormente, que se fosse ignorada levaria a que grande parte das ligações à página de Portugal fossem perdidas.

Embora os autores das `mwlib` digam que a ferramenta reconhece e trata corretamente as macros, não o conseguimos fazer para a versão portuguesa da Wikipédia (possivelmente pelo uso de *Predefinição* como prefixo, em vez do termo usado na Wikipédia inglesa, *Template*).

A solução foi implementada na casa: criou-se uma base de dados de macros, pré-processando o documento XML da Wikipédia, e para todas as páginas de pré-definição, foi introduzido um registo na base de dados, mapeamento do seu nome (nome da macro) e o conteúdo gerado pela macro (ignorando comentários usados para explicar como a macro se deve usar). Posteriormente, ao processar a Wikipédia, as macros seriam substituídas pela expansão respetiva.

Infelizmente esta abordagem não foi totalmente satisfatória, dado existir um conjunto de macros que geram etiquetas XHTML diretamente. Ora, ao interpolar as macros no XML com essas novas etiquetas, o documento XML deixava de ser bem formado, e a ferramenta `mwlib` não era capaz de o processar. Esta foi a principal razão pela qual se perderam as *Infoboxes* já mencionadas. Dada a necessidade de estabilizar rapidamente a coleção, e de estas caixas, embora contendo informação relevante, terem pouco que ver com língua natural (os dados são tabelados), a equipa do Págico decidiu ignorar este problema.

Existiu ainda um pequeno conjunto de macros que não foram expandidas corretamente dada a sua complexidade (número de argumentos, argumentos pré-definidos, aninhamento de macros, etc.).

5 Caracterização do Cartola

Esta secção faz uma caracterização preliminar do conteúdo do Cartola. Concretamente, apresenta estatísticas relativas à coleção do Págico, bem como diversas estatísticas relativas à subcoleção do monte do Págico. Esta subcoleção contém todos os documentos usados como resposta aos tópicos bem como os usados como justificações das respostas, não distinguindo se foram dados pelos criadores de tópicos ou pelos participantes.

O objetivo desta caraterização é permitir que o leitor consiga julgar a dificuldade (ou facilidade) da participação no Págico. Além disso, permite ter uma noção da abrangência dos tópicos em relação à coleção disponibilizada.

5.1 A coleção do Págico

Para que se tenha uma ideia do espaço de procura das páginas que podem ser respostas aos tópicos do Págico, apresentamos aqui várias quantificações em relação à coleção.

Começaremos por analisar o tamanho da coleção em número de documentos, e em número de documentos por tipo (ou *espaço de nomes*), o que indicará qual a percentagem de documentos da coleção que constituíam, realmente, espaço de procura das respostas.

Após a divisão de páginas pelo seu tipo, um sistema automático poderia tentar indexar os artigos pelas categorias que são usadas para os classificar. Deste modo, na secção 5.1.2 apresentamos algumas estatísticas que permitem analisar até que ponto as categorias usadas na Wikipédia podem ser úteis, ou não, na indexação dos artigos, e facilitação na pesquisa de respostas.

As secções que se lhe seguem tentam caracterizar a coleção de um ponto de vista mais concreto: qual é o tamanho da coleção? qual o número médio de palavras por artigo? Embora pouco relevante para a construção de um sistema ou para a indexação dos artigos, esta informação permite-nos saber o que constitui um artigo da coleção.

Finalmente, será apresentada uma análise temporal que permite caracterizar a coleção em termos de atualiadade. Possivelmente, esta análise é pouco relevante para o Págico, mas acaba por demonstrar que a maior parte dos artigos da Wikipédia portuguesa foram atualizados nos últimos 12 meses. Este facto só por si justifica a relevância em se ter criado uma nova coleção para o Págico (especialmente quando o Págico se propõe a sugerir temas ligados à cultura portuguesa), uma vez que a coleção do GikiCLEF foi criada a partir de uma versão de 2008 da Wikipédia.

5.1.1 Tipos de páginas

A coleção pode ser dividida em várias partições, de acordo com o tipo de conteúdo das páginas: páginas de pré-definições (com definições de funções, macros, etc.), páginas de desambiguação (que permitem ao utilizador escolher qual o artigo que realmente lhe interessa), páginas de redirecionamento (que funcionam como entradas remissivas), páginas relativas a conteúdo audiovisual (que descrevem imagens, sons, etc.) e as páginas de artigos propriamente ditos. A tabela 1 apresenta o número de páginas para cada um destes tipos. Destas, apenas as páginas relativas a conteúdo audiovisual não foram incluídas na coleção.

Tipo	Nº de documentos
Páginas de pré-definição	32 900
Páginas de desambiguação	5 006
Páginas de redireção	574 077
Páginas de audiovisuais	9 678
Artigos (e anexos)	856 005

Tabela 1: Distribuição de páginas da coleção por tipo.

Embora sejam 689 629 as páginas que fazem parte da coleção, e que não correspondem aos tipos descritas anteriormente, destes apenas 856 005 documentos correspondem a artigos propriamente ditos (e a anexos), onde, em princípio, se encontrarão as respostas aos tópicos do Págico.

Ou seja, uma quantidade razoável de documentos contidos na coleção não eram relevantes, nem constituíam o espaço de procura para as respostas aos tópicos do Págico. Uma nova versão da coleção poderia descartar essas páginas já que não traziam qualquer informação adicional, e acabam por gerar confusão, quer para os participantes, quer para os avaliadores.

5.1.2 Categorização das páginas

Um processo que pode ajudar na divisão do espaço de procura é o uso das categorias associadas a cada página da Wikipédia (colocadas em notação Wiki em cada página, na forma `[[Categoria:nome da categoria]]`)). Estas categorias são colocadas de forma *ad-hoc* por quem contribui com artigos e, embora existam algumas regras definidas, não podem ser consideradas parte de uma estrutura classificativa estruturada, mas antes de, no melhor dos casos, uma estrutura classificativa de dois níveis. Na verdade, as estruturas classificativas mais próximas deste tipo de classificação são as *Folksonomy* (Sinclair e Cardew-Hall, 2008).

A demonstração desta anarquia é o número de categorias existente: 95 446 categorias para classificar 681 058 documentos (a diferença deste número de documentos para o número total de documentos — 689 829 — mostra a existência de mais de 8 500 artigos não categorizados), o que corresponde a uma média de 7 documentos por categoria. Também é relevante dizer que a página *Língua inglesa* (Wikipédia) é a que tem mais categorias associadas, num total de 62. Por sua vez, existem 32 652 categorias que contêm apenas uma página associada, e a categoria com mais páginas (32 645) corresponde aos *Asteroides da cintura principal*. As tabelas 2 e 3 resumem esta informação. Não são apresentados os respetivos histogramas na sua forma gráfica já que a discrepância de valores torna-os pouco legíveis.

Para facilitar a comparação com a caraterização da coleção composta apenas pelas páginas correspondentes a tópicos (secção 5.2), e dado que a maioria dos documentos tem entre 0 a 8 categorias associadas, a figura 1 apresenta uma estatística mais fina correspondente a este intervalo.

nº de documentos	total de cat.	percentual
$]0,1]$	32 652	34.21%
$]1,66]$	59 775	62.63%
$]66,130]$	1 789	1.87%
$]130,194]$	507	0.53%
$]194,260]$	231	0.24%
$]260,345]$	166	0.17%
$]345,442]$	108	0.11%
$]442,592]$	84	0.09%
$]592,862]$	68	0.07%
$]862,\infty[$	65	0.07%

Tabela 2: Número de documentos por quantidade de categorias (p.ex. existem 32 652 categorias que só classificam um documento; e existem 65 categorias que classificam mais de 862 documentos).

nº categorias	total docs.	percentual
0	8 771	1.271%
$]0,8]$	676 705	98.097%
$]8,15]$	4 008	0.581%
$]15,23]$	314	0.046%
$]23,33]$	25	0.004%
$]33,\infty[$	6	0.001%

Tabela 3: Número de categorias por quantidade de documentos (p.ex, existem 8 771 documentos sem categorias associadas, e existem 6 documentos com mais de 33 categorias associadas).

5.1.3 Tamanho das páginas

O tamanho médio (incluindo toda a anotação wiki) destes artigos é de 3 169 bytes, cerca de 968 formas[6] (os artigos mais pequenos estão (ou estavam) vazios; o maior artigo, com o título *Anexo: Lista de espécies da família Salticidae* (Wikipédia)[7] tem 334 083 bytes (106 140 formas)).

[6]De realçar que os valores de formas aqui apresentados não correspondem a palavras uma vez que devido à grande quantidade de anotação Wiki presente nos documentos, apenas uma percentagem corresponde, realmente, a palavras. Além do mais, esta percentagem não é mantida entre páginas já que algumas (como a que é referida, com 106 140 formas) são tabelas com uma grande quantidade de anotação, e outras páginas, de artigos convencionais, têm uma quantidade de anotação bastante menor.

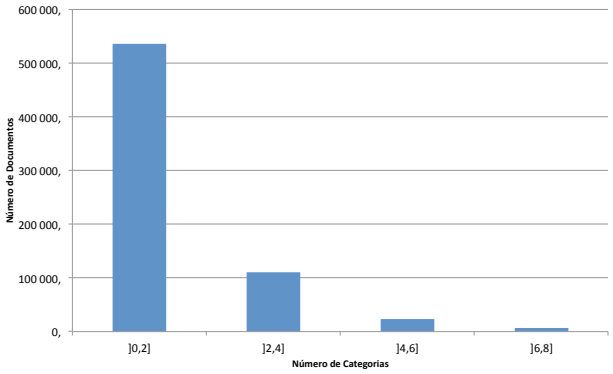

Figura 1: Número de categorias por quantidade de documentos, no intervalo de $]0, 8]$ categorias.

nº de formas	nº docs	percentual
$]0,5]$	1	0.00%
$]5,1042[$	541 628	78.54%
$]1042,2075[$	87 789	12.73%
$]2075,3108[$	26 527	3.85%
$]3108,4141[$	11 931	1.73%
$]4141,5176[$	6 501	0.94%
$]5176,6232[$	3 946	0.57%
$]6232,7378[$	2 711	0.39%
$]7378,8707[$	1 989	0.29%
$]8707,10256[$	1 691	0.25%
$]10256,12439[$	1 447	0.21%
$]12439,15585[$	1 256	0.18%
$]15585,21968[$	1 139	0.17%
$]21968,\infty]$	1 063	0.15%

Tabela 4: Número de documentos por classes de tamanhos (Por exemplo, a maioria dos documentos (78%) tem menos de 1042 formas).

5.1.4 Atualidade da coleção

O gráfico da figura 2, correspondente à tabela 5 mostra a evolução das páginas da coleção de acordo com a sua última edição.

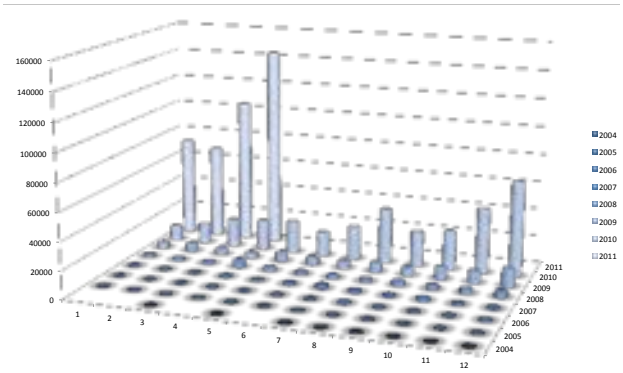

Figura 2: Número de artigos por ano/mês.

Embora o gráfico não permita ver as diferenças relativas aos primeiros anos torna mais visual a discrepância no número de artigos atualizados recentemente. Na verdade, esse valor aumenta à medida que nos aproximamos da atu-

[7]Note que este é o artigo maior em termos absolutos e não em termos de formas. Nesse caso, o artigo *Torneio de Wimbledon* (Wikipédia) estaria no topo, com 158 128 formas.

Ano	Jan.	Fev.	Mar.	Abr.	Mai.	Jun.	Jul.	Ago.	Set.	Out.	Nov.	Dez.	Total
2004			4		9		5	5	4	5	7	8	47
2005	9	3	17	16	74	61	33	30	64	16	39	25	387
2006	120	96	101	316	125	228	268	1329	271	528	638	726	4746
2007	681	590	487	1023	834	1461	2933	1760	1007	2199	970	1058	15003
2008	1977	1654	1554	5385	2812	2125	2123	2328	3570	3148	3574	4883	35133
2009	4330	5876	4665	4024	6559	5558	5369	6364	5804	8866	8768	13098	79281
2010	10131	13988	19879	21241	22941	17257	23927	39281	24860	27785	46672	68136	336098
2011	71369	67126	103464	143351									385310

Tabela 5: Número de artigos por ano/mês.

alidade, o que sugere uma atualização contínua dos conteúdos.

5.2 A subcoleção do monte do Págico

Nesta subsecção, vamos debruçar-nos sobre a subcoleção do monte do Págico, ou seja, o subconjunto da coleção constituído pelos documentos usados como resposta ou justificação pelos criadores dos tópicos, no processo de criação dos mesmos, e por todos os participantes no Págico (tanto sistemas automáticos como participações humanas). Por simplificação, usaremos o termo *documento de resposta*, independentemente desse documento ter sido usado como resposta ou justificação.

Primeiro faremos uma análise sem ter em conta se as respostas do monte estavam ou não corretas, e sem seguida teremos apenas em consideração os documentos de resposta que correspondem a respostas e justificações corretas.

5.2.1 Visão sobre todas as respostas

A figura 3 apresenta uma panorâmica sobre a distribuição do número de documentos de resposta determinados pelos criadores dos tópicos e encontrados pelos participantes no Págico. Como se pode constatar, para a maior parte dos tópicos, o número de documentos associados varia entre 175 e 250 documentos. Se nos restringirmos aos documentos que existem apenas na Wikipédia portuguesa, portanto sem equivalentes noutras línguas, então obtemos o gráfico da figura 4, onde se pode ver que, para a maior parte dos tópicos, entre 20% e 50% dos documentos de resposta existem unicamente na Wikipédia em português.

Os tópicos mais especificamente lusófonos, se assim considerarmos aqueles para os quais uma maior percentagem dos documentos de resposta existe apenas na Wikipédia em português, são sobre samba (**tópico 36** [Escolas de samba fundadas ou sediadas em morros cariocas.], **tópico 51** [Além do samba, que outros gêneros musicais são populares no carnaval brasileiro] e **tópico 86** [Compositoras brasileiras de samba]) e São

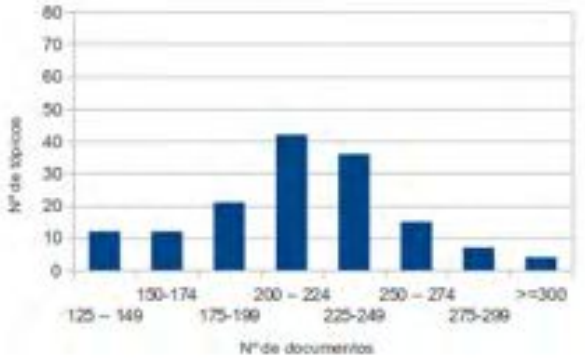

Figura 3: Número de tópicos agrupados por número de documentos de resposta.

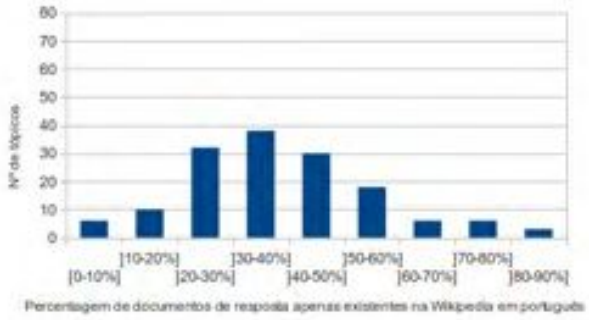

Figura 4: Número de tópicos agrupados pela percentagem de documentos de resposta apenas existentes na Wikipédia em português.

Tomé e Príncipe (**tópico 131** [Quem descobriu São Tomé e Príncipe?] e **tópico 95** [Partidos políticos de São Tomé e Príncipe]).

No pólo oposto, os tópicos para os quais uma menor percentagem dos documentos de resposta existe apenas na Wikipédia em português, os tópicos sobre desporto estão bem representados (**tópico 58** [Países que venceram a Copa do Mundo em uma disputa de pênaltis], **tópico 137** [Eventos onde Maria de Lurdes Mutola foi medalha de ouro] e **tópico 39** [Modalidades esportivas em que países lusófonos já ganharam medalha de ouro nos Jogos Olímpicos.]).

A figura 5 mostra o número total de palavras dos documentos de resposta. Este número varia bastante de tópico para tópico, havendo tópicos

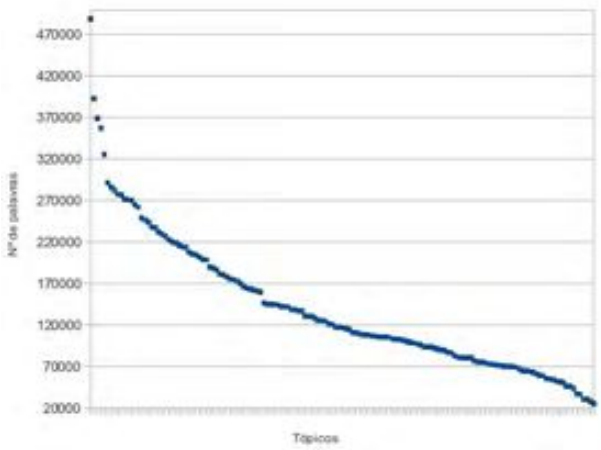

Figura 5: Número de palavras dos documentos de resposta.

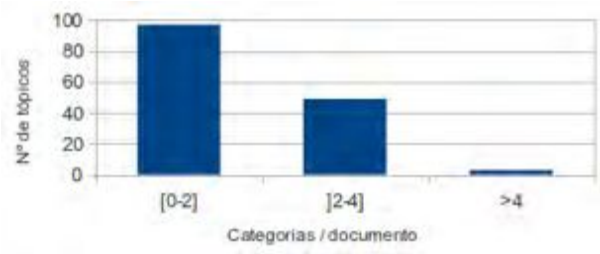

Figura 6: Número de tópicos agrupados pelo número de categorias em que estão classificados os documentos de resposta.

com menos de 50000 palavras, enquanto outros têm mais de 300000 palavras.

A figura 6 ilustra a distribuição do número de categorias por documento em que estão classificados os documentos de resposta. Como se pode constatar para a maior parte dos tópicos, este número não ultrapassa as duas categorias por documento.

A tabela 6 apresenta os cinco tópicos com maior e menor número de documentos de resposta. É curioso verificar que os cinco tópicos para os quais foram encontrados menos documentos de resposta são todos sobre temas africanos o que parece indicar que a Wikipédia conterá menos informação sobre esses temas.

5.2.2 Visão sobre as respostas corretas do Págico

A figura 7 apresenta uma panorâmica sobre a distribuição do número de documentos de resposta corretos, ou seja, relativos às respostas e justificações determinadas pelos criadores dos tópicos e encontradas pelos participantes no Págico que foram consideradas corretas. Como se pode constatar, para a maior parte dos tópicos este número situou-se abaixo dos dez documentos. Se nos res-

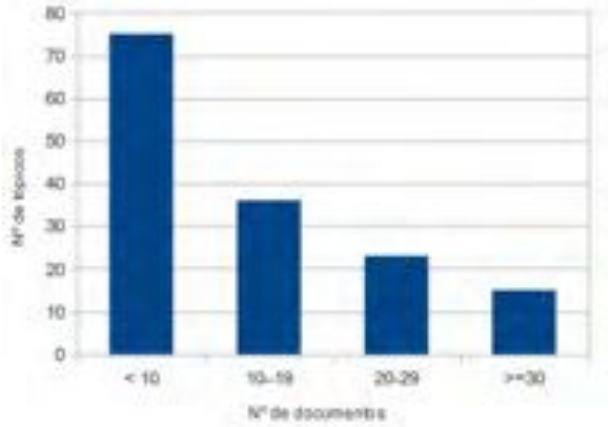

Figura 7: Número de tópicos agrupados por número de documentos de resposta corretos.

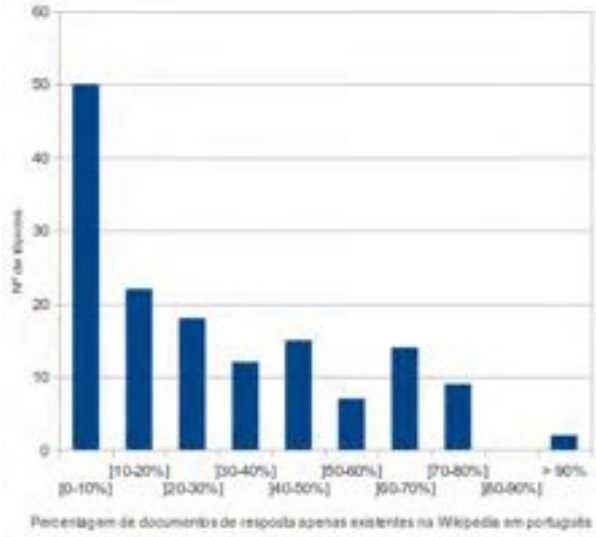

Figura 8: Número de tópicos agrupados pela percentagem de documentos de resposta corretos apenas existentes na Wikipédia em português.

tringirmos aos documentos que existem apenas na Wikipédia portuguesa, portanto sem equivalentes noutras línguas, então obtemos o gráfico da figura 8. Estes valores diferem bastante dos encontrados para todos os documentos de resposta (cf. figura 4). Neste caso para um terço dos tópicos, a percentagem de documentos de resposta apenas na Wikipédia em português situa-se entre os 0% e 10%, existindo apenas dois tópicos onde este valor é superior a 90% (**tópico 41** [Congressos ou conferências que têm por tema as relações culturais e/ou sociais entre África e demais países lusófonos] e **tópico 54** [Igrejas do Rio de Janeiro construídas por irmandades ou confrarias de negros]).

A figura 9 mostra o número total de palavras dos documentos de resposta correspondentes a respostas e justificações corretas. Este número

ID	Tópico	# Documentos
83	Que equipes da primeira divisão do futebol brasileiro desceram para a segunda divisão e nunca mais conseguiram voltar?	330
142	Locais referidos n' "Os Lusíadas"	327
17	Documentários sobre políticos brasileiros.	325
29	Escritores lusófonos que se filiaram a partidos políticos	315
35	Que autores não lusófonos escreveram sobre o Brasil nos séculos XVIII e XIX?	294
	(...)	
109	Candidatos a alguma das eleições presidenciais na Guiné-Bissau	129
95	Partidos políticos de São Tomé e Príncipe	128
129	Antigos alunos da Universidade Eduardo Mondlane e da sua antecessora, a Universidade de Lourenço Marques	128
100	Ilhas de Moçambique	125
121	Frutos de Angola	125

Tabela 6: Tópicos com maior e menor número de documentos de resposta.

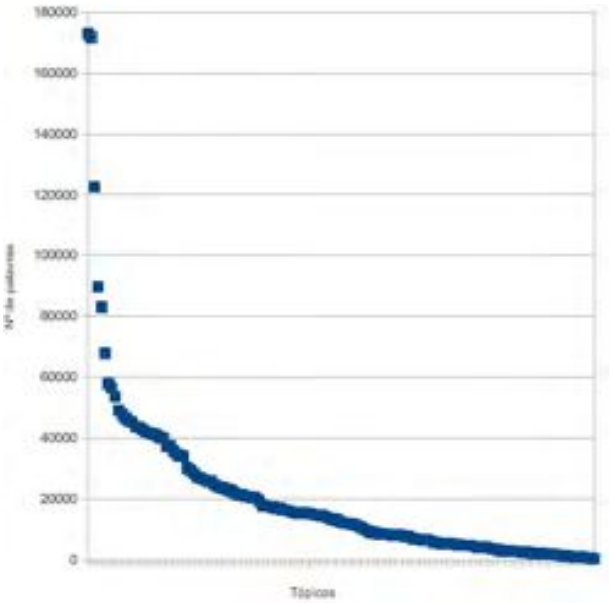

Figura 9: Número de palavras dos documentos de resposta corretos.

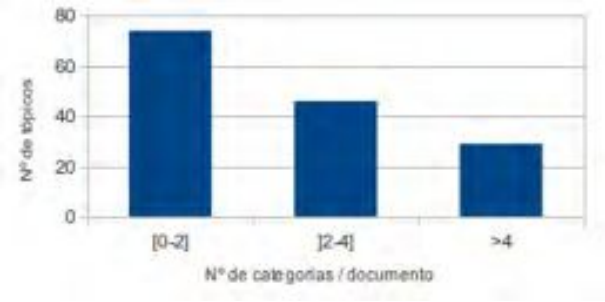

Figura 10: Número de tópicos agrupados pelo número de categorias por documento em que estão classificados os documentos de resposta corretos.

varia bastante de tópico para tópico, havendo tópicos com menos de um milhar de palavras, enquanto outros têm mais de cem mil palavras.

A figura 10 ilustra a distribuição do número de categorias por documento em que estão classificados os documentos de resposta corretos. Como se pode constatar para a maior parte dos tópicos, o número de categorias por documento situa-se entre as zero e as quatro categorias.

A tabela 7 apresenta os cinco tópicos para os quais foram determinados o maior e menor número de documentos de resposta corretos. Em relação aos tópicos com menos documentos de resposta corretos, a maior parte deles são sobre temas africanos, tal como se verificou considerando o conjunto total de respostas (corretas e incorretas). Relativamente aos tópicos com mais respostas corretas, parecem ser tópicos que de facto têm naturalmente um número elevado de respostas tais como **tópico 19** [Tribos indígenas que vivem na Amazônia] e **tópico 147** [Museus em capitais de países lusófonos].

6 Comentários finais

É certo que a coleção desta edição do Págico tem muitos problemas. O principal problema é depender de uma ferramenta externa para a produção dos documentos num formato menos complicado. Poder-se-ia ter disponibilizado aos participantes a versão original em XML disponibilizada pela própria Wikipédia, mas isso obrigaria os participantes a processar a marcação Wiki, processamento este que iria influenciar os resultados da participação, mas que nada têm a ver com a tarefa do Págico de encontrar as respostas aos tópicos.

ID	Tópico	# Documentos
19	Tribos indígenas que vivem na Amazônia.	95
147	Museus em capitais de países lusófonos	62
144	Locais referidos n' "Os Lusíadas"	51
79	Povos indígenas brasileiros considerados extintos.	50
106	Vice-reis da Índia Portuguesa	48
	(...)	
110	Políticos da África lusófona que estudaram na União Soviética	2
54	Igrejas do Rio de Janeiro construídas por irmandades ou confrarias de negros.	1
132	Deputados da FRELIMO	1
116	Escritores moçambicanos que receberam o Prémio Camões	1
55	Escritores estrangeiros que visitaram Portugal no século XIX e que publicaram descrições das suas viagens	1

Tabela 7: Tópicos com maior e menor número de documentos de resposta corretos.

Numa próxima edição a solução deverá passar por usar uma versão do motor da Wikipédia em modo local, e pela extração dos documentos HTML através de *crawling*. Esta abordagem irá desencadear um conjunto de outros problemas mas que, esperamos, serão menos graves que os encontrados com a coleção atual.

Com a compilação do Cartola, o recurso público criado no decurso do Págico, pretendemos que o trabalho e a experiência no Págico possa ser o mais proveitosa possível, mesmo após o término do mesmo. Ou seja assumindo naturalmente que nem sempre tomámos as melhores opções no decorrer da organização do Págico, disponibilizamos todos os resultados obtidos, para que possam ser usados e eventualmente melhorados por quem estiver interessado nas áreas abordadas pelo Págico.

Ideias para trabalho futuro seriam, por exemplo:

- o estudo da evolução da Wikipédia ao longo dos últimos anos, usando para isso quer as coleções desenvolvidas no contexto do GikiCLEF e no contexto do Págico, ou diretamente usando as cópias estáticas disponibilizadas pela Wikipédia.
- aferir a lusofonia da Wikipédia portuguesa, por um lado, a nível de conteúdo, por exemplo, analisando os topónimos e gentílicos usados nas categorias das páginas, e, pelo outro, em termos de quem a escreve, por exemplo, analisando a grafia e o vocabulário.

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 em 2011 pela Fundação da Ciência e da Tecnologia (FCT) e pela Fundação para a Computação Científica Nacional (FCCN). O trabalho de Alberto Simões foi parcialmente suportado pela bolsa da Fundação para a Ciência e a Tecnologia SFRH/BPD/73011/2010.

Agradecemos a Cláudia Freitas e Alice Gonçalves pela paciência de nos irem relatando os vários erros encontrados na coleção do Págico enquanto utilizadoras da mesa no SIGA, o que ajudou a melhorar a qualidade do recurso criado.

Estamos também gratos a Sandra Aluísio, Diana Santos e António Teixeira pelos comentários e sugestões que recebemos durante a preparação do artigo e que enriqueceram o mesmo, tornando-o também mais claro.

Referências

Costa, Luís, Cristina Mota, e Diana Santos. 2012. SIGA, a Management System to Support the Organization of Information Retrieval Evaluations. Em Helena Caseli, Aline Villavicêncio, António Teixeira, e Fernando Perdigão, editores, *Computational Processing of the Portuguese Language, PROPOR'2012*, pp. 284–290, Berlim/Heidelberg. Springer.

Freitas, Cláudia. 2012. A lusofonia na wikipédia em 150 tópicos. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Freitas, Cláudia, Paulo Rocha, Cristina Mota, Luís Costa, e Diana Santos. 2012. O que é uma resposta? Notas de uns avaliadores esta-

fados. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Junior, Arnaldo Candido, Ann Copestake, Lucia Specia, e Sandra Maria Aluísio. 2011. Towards an on-demand simple portuguese wikipedia. Em *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, SLPAT '11, pp. 137–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mota, Cristina. 2012. Resultados págicos: participação, medidas e pontuação. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Santos, Diana, Luís Miguel Cabral, Corina Forascu, Pamela Forner, Fredric Gey, Katrin Lamm, Thomas Mandl, Petya Osenova, Anselmo Peñas, Álvaro Rodrigo, Julia Schulz, Yvonne Skalban, e Erik Tjong Kim Sang. 2010. Gikiclef: Crosscultural issues in multilingual information access. Em Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, e Daniel Tapias, editores, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may, 2010. European Language Resources Association (ELRA).

Sinclair, James e Michael Cardew-Hall. 2008. The folksonomy tag cloud: when is it useful? *Journal of Information Science*, 34(1):15–29, February, 2008.