# Resultados págicos: participação, medidas e pontuação

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

**Resumo**

Este artigo descreve a participação no Págico, tanto a nível de sistemas, como a nível de participação humana. Além disso, caracteriza o processo de avaliação e apresenta as medidas de avaliação implementadas, introduzindo as novas medidas de pseudo-abrangência, pseudo-medida-F, originalidade e criatividade. Finalmente, mostra os resultados globais por participante em vários cenários de avaliação, bem como os resultados detalhados por temas e lugares dos tópicos no cenário completo do Págico, contrastando a participação humana e a automática.

**Palavras chave**

Recolha de informação, Resposta a perguntas, Avaliação, Cooperação pessoa-máquina, Wikipédia

## 1 Apresentação

O Págico foi uma avaliação conjunta em que sistemas e pessoas tiveram de fornecer respostas a 150 tópicos (consulte-se (Mota et al., 2012) para uma descrição da tarefa, e (Santos, 2012) para uma motivação da avaliação, e (Freitas, 2012) para uma apresentação e discussão do processo de criação dos tópicos). As respostas deveriam ser encontradas numa versão estática da Wikipédia portuguesa (veja-se (Simões, Costa e Mota, 2012) para uma descrição e avaliação deste recurso) e correspondiam aos títulos das páginas da Wikipédia. Nos casos em que o conteúdo da página não justificava só por si que a página (ou seja, o seu título) era a resposta correta, os participantes tinham de fornecer adicionalmente a(s) página(s) que permitia(m) chegar a essa conclusão - designaremos o conjunto das justificações simplesmente por justificação.

A título de exemplo, para um tópico como [Que cientistas ou avanços da ciência podem ser direta ou indiretamente relacionados com os jesuítas da escola de Coimbra?] esperava-se que os participantes identificassem *Nónio* (Wikipédia), entre outras respostas possíveis. Nesse caso, como a página não contém informação suficiente que justifique que o nónio foi um avanço relacionado com os jesuítas de Coimbra, os participantes tinham de associar a essa resposta a página *Pedro_Nunes_(matemático)* (Wikipédia), que era igualmente uma boa resposta, como justificação, pois é ela que contém essa informação.[1]

Sistemas e participantes humanos participaram no Págico de forma distintas. Enquanto os primeiros foram buscar os recursos da avaliação fornecidos pela organização (versão estática da Wikipédia portuguesa e lista de tópicos de avaliação) para os processarem e depois enviarem as respostas num ou mais ficheiros (até um máximo de três) que designaremos *corridas*, os participantes humanos forneceram as respostas através de uma interface desenvolvida para o efeito no SIGA (Costa, Mota e Santos, 2012) que permitia fazer pesquisas na versão estática da Wikipédia e adicionar as páginas como resposta ou justificação.

O calendário para participação também não foi o mesmo, uma vez que se deu mais tempo aos participantes humanos para responderem aos tópicos. O período para envio de respostas de participantes humanos e sistemas teve início ao mesmo tempo, a 4 de Novembro de 2011, e decorreu até 11 de Novembro para sistemas e 30 de Novembro para participantes humanos.

Após ter fechado o período de envio de corridas por sistemas[2], iniciou-se o processo de avaliação humana das respostas. Os primeiros resultados foram divulgados a 9 de Janeiro de 2012, quando mais de metade das respostas já tinham sido avaliadas por avaliadores humanos e a 21 de Janeiro concluiu-se a avaliação das respostas por pelo menos um avaliador humano. Os resultados finais, em que algumas das respostas foram revistas por mais de um avaliador e em

[1] Este exemplo foi retirado da página de apresentação e divulgação do Págico em `http://www.linguateca.pt/Pagico/`.

[2] Além das corridas oficiais, demos a possibilidade de serem enviadas corridas não oficiais pelos sistemas, após o prazo final. Essas corridas foram avaliadas automaticamente, mas neste artigo esses resultados não serão tidos em conta.

*This work is licensed under a*
*Creative Commons Attribution 3.0 License*

que foram revistas as justificações dadas pelos criadores dos tópicos, foram divulgados a 18 de Fevereiro.

Este artigo foca a participação no Págico, contrastando sistemas e participantes humanos, o processo geral de avaliação e as medidas utilizadas para avaliar as respostas dos participantes. Além de mostrar, por participante, os resultados globais da avaliação, mostra ainda esses resultados detalhados por tema, subtema e localização dos tópicos. Como, mais do que saber quem é o melhor participante, se pretende perceber em que difere a participação humana da automática, compara ainda os resultados entre ambas.

## 2 Participação no Págico

Inscreveram-se no Págico 21 equipas: 6 sistemas e 15 participantes humanos. No entanto, apenas um terço dos inscritos acabou por participar efetivamente: 2 sistemas e 5 participantes humanos. É notável também que pouco mais de um terço (4 sistemas e 4 participantes humanos) desistiram de participar sem sequer ver a coleção com os tópicos de avaliação `Págico`, embora dois sistemas e um dos participantes humanos ainda tenham visto a coleção com exemplos de tópicos, `PagicoEXEMPLOS`. Finalmente, pouco menos de um terço (os restantes 6 participantes humanos) desistiram depois de terem visto a coleção `Págico` e de terem feito algumas pesquisas na coleção, sendo que três deles ainda visualizaram documentos e chegando um deles a responder a dois tópicos. A tabela 1 sintetiza o perfil de envolvimento das equipas que se inscreveram no Págico.

Apresentamos sucintamente, em seguida, os sete participantes que forneceram respostas:

**GLNISTT** 23 estudantes organizados em 8 grupos, em que cada estudante respondeu em média a 7 perguntas. Os grupos responderam a um conjunto de tópicos disjuntos. Participaram no âmbito de um projecto para a cadeira de Língua Natural, de mestrado, sendo o principal objectivo perceber o que seria necessário fazer para construir um sistema capaz de participar no Págico. A participação dos 8 grupos foi avaliada como um todo, mas também individualmente por grupo, tendo sido facultado esse resultado à professora responsável pela cadeira. Recorreram a várias fontes, incluindo a Wikipédia atual. (Coheur e Ângela Costa, 2012)

**ludIT** Equipa de 6 pessoas que também se organizaram de modo a responderem a conjuntos de tópicos disjuntos. No entanto, colaboraram entre si em caso de dúvida. Usaram uma estratégia de pesquisa na Wikipédia atual e confirmação das respostas na versão da Wikipédia usada no Págico. (Veiga et al., 2012)

**João Miranda** Participou individualmente e usou uma estratégia de pesquisa com base nos termos do tópico ou tentando procurar pelo nome do artigo que contém a resposta no caso do tópico lhe ser familiar. (Miranda, 2012)

**Ângela Mota; Bruno Nascimento** Participaram individualmente, mas não enviaram relatos da participação.

**RAPPORTAGICO** Este sistema combina o reconhecimento de sintagmas frásicos com a identificação de sinónimos recorrendo a uma ontologia lexical. Enviou três corridas, em que uma delas serve de base de comparação (em inglês, *baseline*) às outras duas, que fazem expansão de sinónimos dos sintagmas verbais, cada uma com métodos de expansão diferentes: uma com *Bag of Words* e a outra com *Personalized Page Rank*. (Rodrigues, Gonçalo Oliveira e Gomes, 2012)

**RENOIR** Usou um sistema de recuperação geográfica que devolve os documentos mais relevantes para os tópicos, os quais numa das corridas não foram reformulados nem lematizados, numa outra foram lematizados mas sem que tivessem sido reformulados e na terceira foram lematizados e reformulados. (Cardoso, 2012)

A tabela 2 mostra informações sobre as respostas fornecidas pelos participantes. A distinção mais evidente é que os sistemas não forneceram justificações adicionais para as respostas, além de terem enviado, como se esperaria, um maior número de respostas. Repare-se também que os participantes humanos repetiram entre si menos respostas do que os participantes automáticos, embora se deva salientar desde já que apenas dois participantes humanos responderam a 148 ou mais tópicos, tendo um deles respondido a todos, e que os outros três responderam no máximo a um terço dos tópicos cada um. Sobre esta questão de a quantos tópicos os participantes responderam, bem como do número médio de respostas por tópico dadas pelos participantes, consulte-se a tabela 6, na secção 5.

| | | Tipo de participação | |
|---|---|---|---|
| | | Automática | Humana |
| Inscritos | | 6 | 15 |
| **- Participantes** | | **2** | **5** |
| - Desistentes | responderam a tópicos | - | 1 |
| | consultaram páginas | - | 3 |
| | fizeram pesquisas sem consultar | - | 2 |
| | viram apenas exemplos | 2 | 1 |
| | não viram coleções | 2 | 3 |

Tabela 1: Perfil de envolvimento no Págico.

| Tipo de Participação | Participante (Corrida) | # Respostas | # Com justificação |
|---|---|---|---|
| Humana | Ângela Mota | 157 | 8 (5%) |
| | GLNISTT | 1016 | 255 (25%) |
| | ludIT | 1387 | 489 (35%) |
| | João Miranda | 101 | 60 (50%) |
| | Bruno Nascimento | 34 | 1 (3%) |
| | **Total** | **2695** | |
| | **Distintas** | **2383** | |
| | **Total/Distintas** | **1.13** | |
| Automática | RENOIR (1) | 15000 | |
| | RENOIR (2) | 15000 | |
| | RENOIR (3) | 15000 | |
| | **Total** | **45000** | |
| | **Distintas** | **28626** | |
| | **Total/Distintas** | **1.57** | |
| | RAPPORTAGICO (1) | 1718 | |
| | RAPPORTAGICO (2) | 1736 | |
| | RAPPORTAGICO (3) | 1730 | |
| | **Total** | **5184** | |
| | **Distintas** | **2343** | |
| | **Total/Distintas** | **2.21** | |
| | **Total** | **50184** | |
| | **Distintas** | **30543** | |
| | **Total/Distintas** | **1.64** | |
| **Total** | | **52879** | |
| **Distintas** | | **32485** | |
| **Total/Distintas** | | **1.62** | |

Tabela 2: Participantes no Págico.

## 3 Procedimento de avaliação

Como referido anteriormente, na secção 1, sistemas e participantes humanos forneceram as respostas de modo distinto: os primeiros enviaram ficheiros com as respostas e os segundos utilizaram a interface do SIGA dedicada a esse fim.

No entanto, o procedimento de avaliação, adaptado do GikiCLEF (Santos et al., 2010), que não teve participação humana, não fez distinção entre os dois tipos de participação, a não ser durante a avaliação humana em que foram apresentados aos avaliadores, caso existissem, os comentários dados pelos participantes para complementar a justificação.

A avaliação processou-se então em cinco passos, que se descreverão em seguida: geração do monte (em inglês, *pool*) das respostas, avaliação automática das respostas, distribuição pelos avaliadores das respostas a avaliar, avaliação humana das respostas e cálculo das medidas de avaliação.

### 3.1 Geração do monte das respostas

O monte corresponde à união de todas as respostas, ou seja, ao conjunto de respostas distintas dadas por todos os participantes. É também criada uma lista com todas as respostas fornecidas pelos participantes. Antes da geração do monte, para cada participante humano é gerada uma corrida com as suas respostas, cujo formato é idêntico ao das corridas dos sistemas. Assim, as respostas dos sistemas e dos participantes humanos serão tratadas de forma

indistinguível. Em (Costa, Mota e Santos, 2012) justifica-se e descreve-se em maior detalhe esta opção.

Como já se viu na tabela 2, a partir das 52879 respostas enviadas, foi criado um monte com 32485 respostas, as quais foram então avaliadas. De notar que se não tivermos em conta as justificações, então o número de respostas distintas é 32086, o que quer dizer que 399 casos correspondem a respostas idênticas mas com justificações diferentes.

### 3.2 Avaliação automática das respostas

Após o monte ter sido gerado, as respostas nele contidas (32485) foram avaliadas automaticamente, comparando as respostas dos participantes e as fornecidas pelos criadores dos tópicos da seguinte forma:

- todas as respostas que correspondam a documentos inválidos (do ponto de vista de não poderem ser respostas no Págico), como sejam, páginas de categorias, de predefinição, de desambiguação, figuras, MediaWiki, ficheiros ou portal são avaliadas automaticamente como incorretas e não passarão para a fase de avaliação humana (4292 respostas);
- se o par (resposta, justificação) tiver sido fornecido pelos criadores dos tópicos, a resposta é considerada correta e justificada e será avaliada automaticamente como correta (420 respostas);
- se a resposta tiver sido fornecida pelos criadores dos tópicos, mas a justificação não, então a resposta é considerada correta, mas não justificada, e, como tal, a resposta será avaliada automaticamente como incorreta; este par passa para a fase de avaliação humana, de forma a validar se de facto a resposta é ou não justificada (235 respostas);
- se a resposta não tiver sido fornecida pelos criadores dos tópicos, então a resposta é avaliada automaticamente como incorreta; o par passa para a fase de avaliação humana, de forma a validar se a resposta é ou não correta e se está ou não justificada (27536 respostas).

A tabela 3 mostra o resultado da avaliação automática, contrastando a participação automática e humana. Salienta-se que a maioria das respostas avaliadas automaticamente como corretas e justificadas foram dadas exclusivamente por participantes humanos (58%), mas que uma parte significativa foi mesmo assim dada tanto por participantes humanos como sistemas (31%) e que, além disso, 11% foram dadas exclusivamente por sistemas. Também se pode ver que praticamente todas as respostas avaliadas automaticamente como corretas, mas cuja justificação não foi idêntica à dos criadores dos tópicos, foi dada por participantes humanos. No caso dos participantes humanos isso pode querer dizer que (i) a página dada pelo participante como justificação é diferente da usada pelo criador do tópico, (ii) o participante forneceu mais páginas de justificação além das dadas pelo criador do tópico, que pode até ter considerado a resposta como auto-justificada, e (iii) o participante não forneceu página de justificação quando o criador do tópico estabeleceu justificação adicional; nos casos as que as respostas dos sistemas foram consideradas corretas, mas não justificadas, são casos como em (iii), pois os sistemas não forneceram quaisquer justificações adicionais (cf. tabela 2).

As tabelas 11, 12 e 13, na secção 5, ilustram, respetivamente, os tópicos com mais respostas corretas e justificadas dadas exclusivamente por participantes humanos, exclusivamente por automáticos e por ambos, após concluída toda a avaliação (automática e humana).

Vale a pena referir que embora o número de respostas predeterminadas pelos criadores dos tópicos tenha acabado por ser inferior ao número de respostas corretas encontradas pelos participantes humanos, das 708 respostas definidas previamente pelos criadores dos tópicos, 288 não foram dadas por qualquer participante. No entanto, se não se tiver em conta as justificações a elas associadas, então o número de respostas não encontradas pelos participantes desce para 184. Isto quer dizer que 104 respostas dadas pelos criadores dos tópicos foram também usadas pelos participantes, mas estes não as justificaram da mesma forma.

### 3.3 Distribuição pelos avaliadores das respostas a avaliar

| Avaliador | # Todas | # Só humanas |
|---|---|---|
| Cláudia | 653 | 182 |
| Cristina | 279 | 265 |
| Diana | 570 | 121 |
| Luís | 818 | 299 |
| Paulo | 25896 | 1464 |

Tabela 4: Distribuição das avaliações humanas.

As respostas foram inicialmente distribuídas de forma disjunta pelos avaliadores e 266 respos-

| Avaliação | Humanas | Automáticas | Ambas | Total |
|---|---|---|---|---|
| Correta e justificada | 243 | 48 | 129 | 420 |
| Correta e não justificada | 221 | 8 | 6 | 235 |
| Incorreta (porque documento inválido) | 0 | 4292 | 0 | 4292 |
| Restantes casos | 1477 | 25753 | 306 | 27536 |
| Total | 1941 | 30101 | 441 | 32483 |

Tabela 3: Estatísticas da avaliação automática.

tas, todas avaliadas pelo mesmo avaliador, foram depois atribuídas a um segundo avaliador de um grupo de três avaliadores. A tabela 4 mostra que quase 92% das respostas foram avaliadas pelo mesmo avaliador, mas que se excluirmos as respostas fornecidas exclusivamente por sistemas, então a fatia avaliada por esse avaliador reduz para 62%.

### 3.4 Avaliação humana das respostas

Após a avaliação automática, as respostas que foram avaliadas automaticamente como corretas mas não justificadas, e como incorretas (por essa resposta não ter sido dada pelos criadores de tópicos[3]), foram alvo de avaliação por avaliadores humanos. Esta avaliação engloba os seguintes passos:

- Avaliação das respostas por avaliadores humanos. A avaliação humana foi feita através do SIGA. Cada avaliador teve acesso à avaliação automática das respostas que lhe foram atribuídas. Além de poder considerar a resposta como correta ou incorreta, podia deixar a resposta por avaliar ou considerá-la duvidosa; caso a considerasse correta (ou a resposta já tivesse sido avaliada automaticamente com correta), tinha também de julgar se estava ou não justificada. Além de avaliarem as respostas e justificações, os avaliadores podiam associar comentários a cada avaliação que fizeram.
- Resolução de conflitos e revisão. Casos duvidosos foram sendo discutidos durante e após a avaliação.

Em (Freitas et al., 2012) descreve-se em maior detalhe a avaliação humana, apresentando-se os critérios para considerar ou não uma resposta como correta, e discutindo-se as várias dificuldades envolvidas nesta fase.

[3]Como referido antes, respostas que não foram dadas pelos criadores de tópicos, mas que correspondem a documentos inválidos - páginas de desambiguação, redireção, predefinição ou com conteúdo audiovisual, - são automaticamente consideradas incorretas, mas não passam para a avaliação humana.

### 3.5 Cálculo das medidas de avaliação

Este passo consiste no cálculo de cada uma das medidas de avaliação de acordo com a avaliação feita para cada resposta. Foram calculadas as seguintes medidas, descritas na secção 4: precisão, precisão tolerante, pseudo-abrangência, pseudo-medida-F, originalidade, criatividade e pontuação final.

Dado que aos participantes humanos foi dada a possibilidade de responderam a um subconjunto dos tópicos de avaliação por si escolhidos de entre os 150, seguimos a tradição das avaliações da Linguateca de fazer a avaliação dos participantes por cenários (cf. (Costa, Rocha e Santos, 2007; Gonçalo Oliveira et al., 2008; Oliveira et al., 2008)).

Em particular, adoptámos a avaliação por cenários tal como definida no Segundo HAREM, em que cada participante é avaliado no seu cenário, bem como em todos os outros, incluindo o do Págico que é constituído por todos os tópicos. Isso permite comparar participantes que responderam a um subconjunto dos tópicos de avaliação com os restantes participantes que responderam a todos ou a um outro subconjunto de tópicos. Para tal, no cálculo das medidas de avaliação ignora-se das corridas desses participantes as respostas a tópicos que não pertencem a esse subconjunto.

| Cenário | # Tópicos |
|---|---|
| Págico/ludIT | 150 |
| GLNISTT | 148 |
| Ângela Mota (AM) | 50 |
| João Miranda (JM) | 40 |
| Bruno Nascimento (BN) | 18 |

Tabela 5: Cenários do Págico.

No Págico, um cenário é então definido por um conjunto de tópicos. Além do cenário Págico, constituído pelos 150 tópicos, criámos um cenário por cada participante que respondeu a um subconjunto desse e que é constituído por esse subconjunto de tópicos. Uma vez que o participante ludIT respondeu a todos os tópicos, o seu cenário é igual ao cenário Págico. A tabela 5 mostra por quantos tópicos é constituído cada cenário.

## 4 Medidas de avaliação

Os participantes foram avaliados no Págico de acordo com as medidas de avaliação usadas no GikiCLEF (precisão e pontuação final), e também com as seguintes novas medidas: pseudo-abrangência, pseudo-medida-F, originalidade e criatividade.

Essas medidas foram calculadas para cada corrida, e as medidas de originalidade e criatividade foram também calculadas para cada participante. Neste último caso, as diferentes corridas de um mesmo participante foram vistas como uma única corrida.

Cada uma das medidas será descrita em seguida, usando a seguinte notação e terminologia:

$p$ = participante $p$

$c$ = corrida $c$

$C$ = conjunto das respostas correctas e justificadas corretamente

$\tilde{C}$ = conjunto das respostas correctas e justificadas incorrectamente

$R$ = conjunto das respostas fornecidas pelos participantes

$T$ = conjunto de tópicos

Designaremos simplesmente por *resposta* o par composto pela resposta e a sua justificação, uma vez que geralmente, em contexto, não é ambígua a sua interpretação. Assim, entende-se por *resposta correta* uma resposta que está correta e a sua justificação está correta também. Quando a resposta não foi justificada corretamente é designada por *resposta não justificada*.

### 4.1 Precisão

$$P_{p,c} = \frac{|C_{p,c}|}{|R_{p,c}|} \quad (1)$$

A precisão $P_{p,c}$ é uma medida que avalia a qualidade das respostas e respetivas justificações incluídas na corrida $c$ do participante $p$, e é dada pela fórmula 1, em que $R_{p,c}$ e $C_{p,c}$ são, respetivamente, o número de respostas dadas e de respostas corretas $c$ do participante $p$.

### 4.2 Precisão tolerante

$$\tilde{P}_{p,c} = \frac{|C_{p,c}| + |\tilde{C}_{p,c}|}{|R_{p,c}|} \quad (2)$$

A precisão tolerante $\tilde{P}_{p,c}$ é uma variante da medida de precisão $P_{p,c}$ que avalia a qualidade das respostas incluídas na corrida $c$ do participante $p$ sem ter em conta a correção das justificações. $\tilde{P}_{p,c}$ é então dada pela fórmula 2, em que $R_{p,c}$, $C_{p,c}$ e $\tilde{C}_{p,c}$ são, respetivamente, o número de respostas dadas, de respostas corretas e de respostas corretas e não justificadas da corrida $c$ do participante $p$.

### 4.3 Pseudo-abrangência

$$\alpha_{p,c} = \frac{|C_{p,c}|}{|C_{Pagico}| + |C_{aval}|} \quad (3)$$

Quando se conhece à partida todas as respostas corretas que um participante deve fornecer, é usual calcular uma medida de abrangência que avalia a quantidade de respostas que o participante forneceu relativamente ao que devia ter fornecido. Esse é o caso, por exemplo, em avaliações de reconhecimento de entidades mencionadas em que os textos anotados pelos participantes são comparados aos mesmos textos em que as entidades mencionadas a reconhecer foram exaustivamente anotadas pela organização da avaliação. O conjunto desses textos é designado por coleção dourada. Veja-se, por exemplo, (Gonçalo Oliveira et al., 2008) para uma descrição da abrangência usada no HAREM.

Em avaliações em que as respostas relevantes não se conhecem à partida, como acontece, por exemplo, em recolha de informação, em que não são conhecidos os documentos relevantes que os sistemas devem encontrar, a abrangência é calculada com base nos documentos relevantes conhecidos (**?**). Esses documentos são identificados por avaliadores humanos no monte dos documentos que é criado após todos os participantes terem enviado as suas corridas.

No Págico, calculámos uma variante da medida de abrangência a que chamámos *pseudo-abrangência* que tem em conta não só as respostas definidas à partida, mas também as respostas identificadas por avaliadores humanos como corretas no monte das respostas. Assim, a *pseudo-abrangência* $\alpha_{p,c}$ é dada pela fórmula 3 e calcula a quantidade de respostas corretas fornecidas pela corrida $c$ do participante $p$ relativamente ao total de respostas conhecidas no Págico, ou seja, relativamente ao total de respostas corretas fornecidas pelos criadores do tópicos, $C_{Pagico}$[4], juntamente com as respostas fornecidas por todos os participantes e que foram avaliadas como corretas, $C_{aval}$, e que não existem em $C_{Pagico}$.

### 4.4 Pseudo-medida-F

$$\phi_{p,c} = 2 \times \frac{P_{p,c} \times \alpha_{p,c}}{P_{p,c} + \alpha_{p,c}} \quad (4)$$

Em avaliações em que se calcula a precisão e a abrangência, também se costuma calcular

[4]Em alguns casos, os criadores dos tópicos forneceram respostas não justificadas que não são tidas em conta no cálculo da pseudo-abrangência.

a medida-F que combina as duas medidas anteriores num só valor.

No Págico, dado que temos precisão e pseudo-abrangência, calculámos a pseudo-medida-F, dada pela fórmula 4.

### 4.5 Originalidade

$$O_{p,c} = \sum_{i}^{T} \sum_{j}^{R_{p,c,i}} o(r_{p,c,i,j}) \tag{5}$$

$$o(r_{p,c,i,j}) = \begin{cases} p(i) & r_{p,c,i,j} \in C_{aval} \quad \wedge \\ & r_{p,c,i,j} \notin C_{Pagico} \quad \wedge \\ & r_{p,c,i,j} \notin \bigcup_{m \neq p, n \neq c} R_{m,n} \\ 0 & \text{c.c.} \end{cases} \tag{6}$$

No Págico definimos uma medida de originalidade por corrida, $O_{p,c}$, dada pela fórmula 5, que contabiliza o número de respostas corretas e originais da corrida $c$ do participante $p$, ou seja, o número de respostas corretas que existem exclusivamente nessa corrida e que também não pertencem ao conjunto de respostas fornecidas pelos criadores dos tópicos. Uma resposta é tão mais original quanto maior for o número de participantes $p(i)$ que tentaram responder ao tópico $i$, como se pode ver pela fórmula 6, que calcula a originalidade da resposta $j$ ao tópico $i$ da corrida $c$ do participante $p$, $o(r_{p,c,i,j})$.

$$O_{p} = \sum_{i}^{T} \sum_{j}^{R_{p,i}} o(r_{p,i,j}) \tag{7}$$

$$o(r_{p,i,j}) = \begin{cases} p(i) & r_{p,i,j} \in C_{aval} \quad \wedge \\ & r_{p,i,j} \notin C_{Pagico} \quad \wedge \\ & r_{p,i,j} \notin \bigcup_{m \neq p} R_{m} \\ 0 & \text{c.c.} \end{cases} \tag{8}$$

Nos casos em que o participante tem mais de uma corrida, a mesma resposta correta em corridas diferentes não é contabilizada como resposta original, mesmo que só tenha sido dada por esse participante. Por essa razão definimos ainda a originalidade por participante, $O_p$, dada pela fórmula 7, em que todas as corridas desse participante constituem uma só corrida.

Repare-se que tanto para se calcular $O_{p,c}$ como $O_p$ a originalidade de uma resposta é proporcional a $p(i)$, ou seja, ao número de participantes que tentaram responder ao tópico. No caso de $O_{p,c}$, se a originalidade fosse proporcional ao número de corridas que tentaram responder ao tópico, estar-se-ia a penalizar os participantes (automáticos) que enviaram mais do que uma corrida, e entre as quais é natural que haja respostas repetidas.

### 4.6 Criatividade

$$K_{p,c} = \sum_{i}^{T} \sum_{j}^{R_{p,c,i}} k(r_{p,c,i,j}) \tag{9}$$

$$k(r_{p,c,i,j}) = \begin{cases} \frac{1}{c(r_{p,c,i,j})} \times p(i) & r_{p,c,i,j} \\ & \in C_{Pagico} \bigcup C_{aval} \\ 0 & \text{c.c.} \end{cases} \tag{10}$$

$p(i)$ = número de participantes no tópico $i$
$c(r_{p,c,i,j})$ = número de participantes que deram a resposta $r_{p,c,i,j}$

Uma resposta correta de um participante pode não ser original, por existir no conjunto de respostas determinadas pelos criadores dos tópicos ou por ter sido dada por mais do que um participante. No entanto, pode ser mais ou menos criativa, no sentido de haver menos ou mais participantes a darem a mesma resposta.

Definimos então uma medida de criatividade por corrida $K_{p,c}$, dada pela fórmula 9, que contabiliza quão criativas são as respostas da corrida $c$ do participante $p$. A criatividade $k(r_{p,c,i,j})$ de uma resposta $i$ ao tópico $j$ da corrida $c$ do participante $p$ é inversamente proporcional ao número de participantes que deram a mesma resposta, $c(r_{p,c,i,j})$, e diretamente proporcional ao número de participantes que tentaram responder ao tópico, $p(i)$, tal como se pode ver na fórmula 10.

$$K_{p} = \sum_{i}^{T} \sum_{j}^{R_{p,i}} k(r_{p,i,j}) \tag{11}$$

$$k(r_{p,i,j}) = \begin{cases} \frac{1}{c(r_{p,i,j})} \times p(i) & r_{p,i,j} \\ & \in C_{Pagico} \bigcup C_{aval} \\ 0 & \text{c.c.} \end{cases} \tag{12}$$

À semelhança do que acontece na originalidade por corrida, em que respostas dadas unicamente por um único participante não contribuem para a originalidade da corrida se ocorrerem em mais do que uma corrida do mesmo participante, na criatividade por corrida, a criatividade de uma resposta é menor se tiver sido dada por corridas diferentes de um mesmo participante. Isso penaliza não só a criatividade

das corridas desse participante, mas também a das corridas de outros participantes que deram a mesma resposta.

Assim, definimos também a criatividade por participante, $K_p$, dada pela fórmula 11, que considera para cada participante a união de todas as suas corridas, e em que a criatividade $k(r_{p,c,j})$ é também proporcional ao número de participantes $p(i)$ que tentaram responder ao tópico $i$.

### 4.7 Pontuação final no Págico

$$M_{p,j} = |C_{p,c}| \times P_{c,j} \quad (13)$$

Embora tenhamos definido várias medidas para avaliar as corridas de perspectivas diferentes, os participantes foram classificados no Págico de acordo com a medida de classificação final por língua do GikiCLEF, que aqui designaremos $M_{p,c}$ e que é dada pela fórmula 13.

Esta medida, baseada na precisão $P_{p,c}$, permite distinguir participantes que tenham a mesma precisão com um número de respostas corretas diferentes. Mais especificamente, nessa situação, a medida atribui uma melhor pontuação final aos participantes que tenham mais respostas corretas.

## 5 Pontuação no Págico

Verificar se participantes humanos encontrariam mais respostas corretas na Wikipédia do que sistemas, ou se teriam um melhor desempenho do que sistemas nessa tarefa, não era um dos objectivos do Págico. Partiu-se do princípio de que encontrariam e de que seriam melhores. No entanto, comprova-se isso mesmo pelas tabelas 6 e 7: a primeira contém diversas estatísticas sobre as respostas ($|T|$ é número de tópicos respondidos, $|R|$ é o número de respostas dadas, $|R|/|T|$ é o número médio de respostas dadas por tópico, $|C|$ é o número de respostas corretas e bem justificadas e $\tilde{C}$ é o número de respostas corretas mas não justificadas corretamente; a segunda apresenta a avaliação dos participantes no Págico para cada uma das várias medidas de avaliação apresentadas nas secção 4. Ambas as tabelas mostram os valores calculados em cada um dos cenários do Págico.

Como seria de esperar, os participantes humanos tiveram uma precisão melhor do que os sistemas, acima dos 56% indo até quase 90% enquanto a dos sistemas não passou de 12%. O que talvez já não seja tão expectável é que os participantes humanos também acabaram por fazer uma melhor abrangência das respostas, isto se compararmos apenas os participantes que responderam a todos ou quase todos os tópicos (ludIT, GLNISTT, RAPPORTAGICO e RENOIR).

Como se vê claramente na tabela 6, e tal como já foi referido antes, os participantes humanos não responderam todos a todos os tópicos. A avaliação por cenários da mesma tabela evidencia que o número de tópicos em comum é baixo entre participantes, sobretudo entre os que responderam a um subconjunto dos 150 tópicos: os participantes Ângela Mota, que respondeu a 50 tópicos, e João Miranda, que respondeu a 40, partilham entre si apenas um décimo do total de tópicos, enquanto esses dois participantes com o participante Bruno Nascimento partilham somente 3 e 5 tópicos, respetivamente. Naturalmente, isso faz com que na avaliação por cenários (ver tabela 7), em termos de pontuação final, cada um desses participantes fique em terceiro lugar quando avaliado no seu cenário, não conseguindo mesmo assim superar os participantes ludIT e GLNISTT que deram em média um número maior de respostas por tópico do que esses participantes, e consequentemente um maior número de respostas corretas.

As tabelas 8 e 9 mostram, respetivamente, quantos participantes responderam ao mesmo tópico e quantos responderam corretamente ao mesmo tópico. Salienta-se que:

- não houve nenhum tópico que tenha sido respondido pelos sete participantes, apenas 16 foram respondidos por seis e houve pelo menos três participantes a responder a cada tópico, sendo que 18 foram respondidos apenas por três participantes (ver tabela 8);
- dos 16 tópicos respondidos por 6 participantes, apenas 3 foram respondidos corretamente também por 6 participantes, pouco mais de 20% dos tópicos foram respondidos corretamente por apenas 1 ou 2 participantes, e um dos tópicos não foi respondido corretamente por nenhum participante (ver tabela 9);
- existe apenas uma resposta dada pelos 6 participantes (que responderam ao mesmo tópico) e essa resposta está correta; a única resposta que foi dada por 5 participantes também está correta e mais de metade das respostas corretas foram dadas por um único participante, o que não quer dizer que tenha sido sempre o mesmo (ver tabela 10).

| Cenário | Participante (Corrida) | $\|T\|$ | $\|R\|$ | $\|R\|/\|T\|$ | $\|C\|$ | $\|\tilde{C}\|$ |
|---|---|---|---|---|---|---|
| Págico | ludIT | 150 | 1387 | 9,25 | 1065 | 34 |
| | GLNISTT | 148 | 1016 | 6,86 | 661 | 52 |
| | João Miranda | 40 | 101 | 2,52 | 80 | 3 |
| | Ângela Mota | 50 | 157 | 3,14 | 88 | 3 |
| | RAPPORTAGICO (3) | 114 | 1730 | 15,18 | 208 | 13 |
| | RAPPORTAGICO (2) | 115 | 1736 | 15,1 | 203 | 13 |
| | RAPPORTAGICO (1) | 116 | 1718 | 14,81 | 181 | 11 |
| | Bruno Nascimento | 18 | 34 | 1,89 | 23 | 1 |
| | RENOIR (1) | 150 | 15000 | 100 | 436 | 38 |
| | RENOIR (3) | 150 | 15000 | 100 | 398 | 29 |
| | RENOIR (2) | 150 | 15000 | 100 | 329 | 25 |
| GLNISTT | ludIT | 148 | 1384 | 9,35 | 1063 | 34 |
| | GLNISTT | 148 | 1016 | 6,86 | 661 | 52 |
| | João Miranda | 39 | 100 | 2,56 | 79 | 3 |
| | Ângela Mota | 48 | 152 | 3,17 | 85 | 3 |
| | RAPPORTAGICO (3) | 112 | 1702 | 15,2 | 206 | 13 |
| | RAPPORTAGICO (2) | 113 | 1708 | 15,12 | 201 | 13 |
| | RAPPORTAGICO (1) | 114 | 1692 | 14,84 | 179 | 11 |
| | Bruno Nascimento | 18 | 34 | 1,89 | 23 | 1 |
| | RENOIR (1) | 148 | 14800 | 100 | 433 | 38 |
| | RENOIR (3) | 148 | 14800 | 100 | 395 | 29 |
| | RENOIR (2) | 148 | 14800 | 100 | 327 | 25 |
| AM | ludIT | 50 | 585 | 11,7 | 490 | 9 |
| | GLNISTT | 48 | 430 | 8,96 | 289 | 19 |
| | Ângela Mota | 50 | 157 | 3,14 | 88 | 3 |
| | João Miranda | 15 | 39 | 2,6 | 31 | - |
| | RAPPORTAGICO (2) | 44 | 743 | 16,89 | 105 | 8 |
| | RAPPORTAGICO (3) | 44 | 732 | 16,64 | 104 | 7 |
| | RAPPORTAGICO (1) | 44 | 722 | 16,41 | 85 | 7 |
| | RENOIR (1) | 50 | 4999 | 99,98 | 223 | 17 |
| | RENOIR (3) | 50 | 5000 | 100 | 194 | 13 |
| | RENOIR (2) | 50 | 5000 | 100 | 160 | 9 |
| | Bruno Nascimento | 3 | 5 | 1,67 | 3 | - |
| JM | ludIT | 40 | 430 | 10,75 | 344 | 10 |
| | GLNISTT | 39 | 342 | 8,77 | 224 | 18 |
| | João Miranda | 40 | 101 | 2,52 | 80 | 3 |
| | RAPPORTAGICO (3) | 30 | 488 | 16,27 | 59 | 5 |
| | RAPPORTAGICO (2) | 30 | 487 | 16,23 | 56 | 4 |
| | Bruno Nascimento | 5 | 12 | 2,4 | 8 | - |
| | Ângela Mota | 15 | 25 | 1,67 | 11 | - |
| | RENOIR (1) | 40 | 4002 | 100,05 | 128 | 16 |
| | RAPPORTAGICO (1) | 29 | 465 | 16,03 | 42 | 5 |
| | RENOIR (3) | 40 | 4000 | 100 | 122 | 13 |
| | RENOIR (2) | 40 | 4000 | 100 | 110 | 8 |
| BN | ludIT | 18 | 177 | 9,83 | 135 | 4 |
| | GLNISTT | 18 | 64 | 3,56 | 47 | 3 |
| | Bruno Nascimento | 18 | 34 | 1,89 | 23 | 1 |
| | Ângela Mota | 3 | 18 | 6 | 14 | - |
| | João Miranda | 5 | 15 | 3 | 11 | - |
| | RAPPORTAGICO (3) | 12 | 220 | 18,33 | 35 | 1 |
| | RAPPORTAGICO (1) | 12 | 179 | 14,92 | 28 | 1 |
| | RAPPORTAGICO (2) | 12 | 183 | 15,25 | 28 | 1 |
| | RENOIR (1) | 18 | 1800 | 100 | 60 | 1 |
| | RENOIR (3) | 18 | 1800 | 100 | 46 | 1 |
| | RENOIR (2) | 18 | 1800 | 100 | 29 | 1 |

Tabela 6: Estatísticas sobre as respostas.

| Cenário | Participante (Corrida) | $M$ | $P$ | $\alpha$ | $\phi$ | $\tilde{P}$ | $O$ | $K$ |
|---|---|---|---|---|---|---|---|---|
| Págico | ludIT | 817,754 | 0,768 | 0,474 | 0,586 | 0,792 | 3442 | 3995,21 |
| | GLNISTT | 430,04 | 0,651 | 0,294 | 0,405 | 0,702 | 1767 | 2211,826 |
| | João Miranda | 63,366 | 0,792 | 0,036 | 0,068 | 0,822 | 202 | 287,139 |
| | Ângela Mota | 49,325 | 0,56 | 0,039 | 0,073 | 0,58 | 146 | 251,395 |
| | RAPPORTAGICO (3) | 25,008 | 0,12 | 0,092 | 0,104 | 0,128 | 29 | 297,003 |
| | RAPPORTAGICO (2) | 23,738 | 0,117 | 0,09 | 0,102 | 0,124 | 5 | 265,219 |
| | RAPPORTAGICO (1) | 19,069 | 0,105 | 0,08 | 0,091 | 0,112 | 22 | 224,72 |
| | Bruno Nascimento | 15,559 | 0,676 | 0,01 | 0,02 | 0,706 | 37 | 65,667 |
| | RENOIR (1) | 12,673 | 0,029 | 0,194 | 0,051 | 0,032 | 126 | 745,087 |
| | RENOIR (3) | 10,56 | 0,026 | 0,177 | 0,046 | 0,028 | 54 | 618,504 |
| | RENOIR (2) | 7,216 | 0,022 | 0,146 | 0,038 | 0,024 | 220 | 609,232 |
| GLNISTT | ludIT | 816,452 | 0,768 | 0,474 | 0,586 | 0,793 | 3438 | 3990,654 |
| | GLNISTT | 430,04 | 0,651 | 0,295 | 0,406 | 0,702 | 1767 | 2211,826 |
| | João Miranda | 62,41 | 0,79 | 0,035 | 0,067 | 0,82 | 202 | 286,583 |
| | Ângela Mota | 47,533 | 0,559 | 0,038 | 0,071 | 0,579 | 141 | 244,173 |
| | RAPPORTAGICO (3) | 24,933 | 0,121 | 0,092 | 0,104 | 0,129 | 29 | 295,614 |
| | RAPPORTAGICO (2) | 23,654 | 0,118 | 0,09 | 0,102 | 0,125 | 5 | 263,831 |
| | RAPPORTAGICO (1) | 18,937 | 0,106 | 0,08 | 0,091 | 0,112 | 22 | 223,331 |
| | Bruno Nascimento | 15,559 | 0,676 | 0,01 | 0,02 | 0,706 | 37 | 65,667 |
| | RENOIR (1) | 12,668 | 0,029 | 0,193 | 0,051 | 0,032 | 126 | 742,032 |
| | RENOIR (3) | 10,542 | 0,027 | 0,176 | 0,046 | 0,029 | 54 | 615,448 |
| | RENOIR (2) | 7,225 | 0,022 | 0,146 | 0,038 | 0,024 | 220 | 607,843 |
| AM | ludIT | 410,427 | 0,838 | 0,474 | 0,605 | 0,853 | 1868 | 2126,441 |
| | GLNISTT | 194,235 | 0,672 | 0,28 | 0,395 | 0,716 | 897 | 1091,941 |
| | Ângela Mota | 49,325 | 0,56 | 0,085 | 0,148 | 0,58 | 146 | 251,395 |
| | João Miranda | 24,641 | 0,795 | 0,03 | 0,058 | 0,795 | 90 | 125,472 |
| | RAPPORTAGICO (2) | 14,838 | 0,141 | 0,102 | 0,118 | 0,152 | 0 | 163,944 |
| | RAPPORTAGICO (3) | 14,776 | 0,142 | 0,101 | 0,118 | 0,152 | 11 | 174,111 |
| | RAPPORTAGICO (1) | 10,007 | 0,118 | 0,082 | 0,097 | 0,127 | 22 | 129,361 |
| | RENOIR (1) | 9,948 | 0,045 | 0,216 | 0,074 | 0,048 | 55 | 441,078 |
| | RENOIR (3) | 7,527 | 0,039 | 0,188 | 0,064 | 0,041 | 6 | 349,945 |
| | RENOIR (2) | 5,12 | 0,032 | 0,155 | 0,053 | 0,034 | 115 | 343,651 |
| | Bruno Nascimento | 1,8 | 0,6 | 0,003 | 0,006 | 0,6 | 6 | 9,667 |
| JM | ludIT | 275,2 | 0,8 | 0,436 | 0,564 | 0,823 | 1471 | 1616,256 |
| | GLNISTT | 146,714 | 0,655 | 0,284 | 0,396 | 0,708 | 725 | 882,214 |
| | João Miranda | 63,366 | 0,792 | 0,101 | 0,18 | 0,822 | 202 | 287,139 |
| | RAPPORTAGICO (3) | 7,133 | 0,121 | 0,075 | 0,092 | 0,131 | 0 | 111,136 |
| | RAPPORTAGICO (2) | 6,439 | 0,115 | 0,071 | 0,088 | 0,123 | 0 | 105,136 |
| | Bruno Nascimento | 5,333 | 0,667 | 0,01 | 0,02 | 0,667 | 12 | 23,25 |
| | Ângela Mota | 4,84 | 0,44 | 0,014 | 0,027 | 0,44 | 28 | 35,889 |
| | RENOIR (1) | 4,094 | 0,032 | 0,162 | 0,053 | 0,036 | 25 | 262,52 |
| | RAPPORTAGICO (1) | 3,793 | 0,09 | 0,053 | 0,067 | 0,101 | 12 | 69,886 |
| | RENOIR (3) | 3,721 | 0,03 | 0,155 | 0,051 | 0,034 | 15 | 242,387 |
| | RENOIR (2) | 3,025 | 0,028 | 0,139 | 0,046 | 0,03 | 112 | 261,187 |
| BM | ludIT | 102,966 | 0,763 | 0,5 | 0,604 | 0,785 | 410 | 504,441 |
| | GLNISTT | 34,516 | 0,734 | 0,174 | 0,281 | 0,781 | 110 | 152,762 |
| | Bruno Nascimento | 15,559 | 0,676 | 0,085 | 0,151 | 0,706 | 37 | 65,667 |
| | Ângela Mota | 10,889 | 0,778 | 0,052 | 0,097 | 0,778 | 48 | 62,667 |
| | João Miranda | 8,067 | 0,733 | 0,041 | 0,077 | 0,733 | 43 | 46,5 |
| | RAPPORTAGICO (3) | 5,568 | 0,159 | 0,13 | 0,143 | 0,164 | 16 | 54,438 |
| | RAPPORTAGICO (1) | 4,38 | 0,156 | 0,104 | 0,125 | 0,162 | 0 | 32,188 |
| | RAPPORTAGICO (2) | 4,284 | 0,153 | 0,104 | 0,124 | 0,158 | 5 | 35,522 |
| | RENOIR (1) | 2 | 0,033 | 0,222 | 0,058 | 0,034 | 47 | 126,855 |
| | RENOIR (3) | 1,176 | 0,026 | 0,17 | 0,044 | 0,026 | 0 | 69,105 |
| | RENOIR (2) | 0,467 | 0,016 | 0,107 | 0,028 | 0,017 | 36 | 69,857 |

Tabela 7: Avaliação dos participantes no Págico.

| # Participantes | # Tópicos |
|---|---|
| 6 | 16 |
| 5 | 59 |
| 4 | 57 |
| 3 | 18 |

Tabela 8: Participantes que responderam ao mesmo tópico.

| # Participantes | # Tópicos |
|---|---|
| 6 | 3 |
| 5 | 28 |
| 4 | 41 |
| 3 | 45 |
| 2 | 20 |
| 1 | 12 |

Tabela 9: Participantes que responderam corretamente ao mesmo tópico.

## 6 Comparação pessoa vs. máquina

Ao contrário do que seria desejável, como se realça no balanço do Págico (Santos et al., 2012), a participação no Págico não foi suficientemente grande para se poderem tirar conclusões comparativas fiáveis entre humanos e sistemas, ou mesmo entre sistemas.

Ainda assim, exploramos nesta seção alguns pontos de partida para uma análise futura mais profunda.

### 6.1 Há tópicos mais difíceis?

O facto de nem todos os participantes terem respondido aos mesmos tópicos dificulta a análise sobre se haveria tópicos mais difíceis do que outros, e se essa dificuldade é sensível ao tipo de participação. Uma primeira tentativa no sentido de aferir essa dificuldade é observando os tópicos com mais respostas corretas para cada um dos tipos de participação.

A tabela 11 mostra os cinco tópicos onde se verifica o maior número de respostas corretas dadas exclusivamente pelos participantes humanos, e em que tanto para as respostas enviadas como para as respostas corretas dos participantes foram dados: o total de respostas ($T$), o número de respostas dadas exclusivamente pelos participantes humanos ($H$), exclusivamente pelos sistemas ($S$) e dadas por ambos os tipos de participantes ($HS$). Esses, aliás, são também os tópicos com mais respostas dadas exclusivamente pelos participantes humanos, mas não pela mesma ordem: o tópico com mais respostas corretas exclusivamente humanas (**tópico 106** [Vice-reis da Índia Portuguesa]) é o terceiro tópico com mais respostas exclusivamente humanas.

| # Participantes | # Respostas | # Corretas |
|---|---|---|
| 6 | 1 | 1 |
| 5 | 1 | 1 |
| 4 | 42 | 24 |
| 3 | 126 | 59 |
| 2 | 792 | 115 |
| 1 | 31523 | 220 |

Tabela 10: Total de participantes que deu a mesma resposta.

Tal como mostra a tabela 12, dos tópicos que reúnem o maior número de respostas corretas exclusivamente dadas por participantes humanos, apenas um se encontra também entre os cinco que obtiveram mais respostas corretas dadas exclusivamente por sistemas (**tópico 19** [Tribos indígenas que vivem na Amazónia]). Ao contrário do que acontece com as respostas dadas exclusivamente por humanos, os cinco tópicos onde existe um maior número de respostas corretas dadas exclusivamente por sistemas não são os tópicos com maior número de respostas enviadas pelos sistemas.

O tópico 19 é também o único que está entre os tópicos que reuniram maior número de respostas corretas dadas por ambos os tipos de participante, sendo o tópico com mais respostas enviadas e também corretas nesse caso (ver tabela 13). Os tópicos com maior número de respostas corretas de ambos os tipos de participação são os mesmos que têm o maior de número respostas dadas por ambos os tipos de participante.

Uma vez que entre os tópicos com maior número de respostas corretas exclusivamente humanas, exclusivamente automáticas e de ambos os tipos de participação existe apenas um que é comum aos três casos, esse facto parece sugerir que existem tópicos para os quais participantes humanos encontram mais facilmente as respostas corretas, outros em que os sistemas serão mais bem sucedidos a encontrar as respostas corretas e, finalmente, ainda outros para os quais é indiferente se são humanos ou máquinas a tentar encontrar as respostas para eles. No futuro, analisar as diferenças entre estes tópicos, as suas respostas e as suas justificações, poderia ser um bom ponto de partida para identificar os tipo de tópicos onde é mais essencial melhorar os sistemas de modo a que estes possam auxiliaram mais os humanos no que precisam.

Além dos tópicos onde houve mais respostas corretas, observámos os tópicos sem respostas corretas para cada uma dos três casos acima referidos, uma vez que isso demonstra alguma

| ID | Tópico | Enviadas | | | | Corretas | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *T* | *H* | *A* | *HA* | *T* | *H* | *A* | *HA* |
| 106 | Vice-reis da Índia Portuguesa | 262 | 83 | 170 | 9 | 88 | **78** | 1 | 9 |
| 147 | Museus em capitais de países lusófonos | 285 | 86 | 198 | 1 | 65 | **65** | 0 | 0 |
| 144 | Locais referidos n Os Lusíadas | 351 | 85 | 265 | 1 | 62 | **60** | 1 | 1 |
| 19 | Tribos indígenas que vivem na Amazônia. | 250 | 59 | 160 | 31 | 115 | **56** | 35 | 24 |
| 16 | Membros da igreja associados à Teologia da Libertação. | 211 | 51 | 153 | 7 | 48 | **37** | 6 | 5 |

Tabela 11: Tópicos com mais respostas corretas exclusivamente de participações humanas

| ID | Tópico | Enviadas | | | | Corretas | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *T* | *H* | *S* | *HS* | *T* | *H* | *S* | *HS* |
| 135 | Aves de Angola | 154 | 12 | 141 | 1 | 54 | 10 | **44** | 0 |
| 19 | Tribos indígenas que vivem na Amazônia. | 250 | 59 | 160 | 31 | 115 | 56 | **35** | 24 |
| 90 | Filmes brasileiros premiados na categoria Montagem. | 211 | 14 | 190 | 7 | 34 | 8 | **19** | 7 |
| 13 | Dinossauros carnívoros que habitaram o Brasil. | 182 | 11 | 166 | 5 | 23 | 6 | **12** | 5 |
| 104 | Pesquisadores do folclore brasileiro | 203 | 14 | 179 | 10 | 33 | 13 | **11** | 9 |

Tabela 12: Tópicos com mais respostas corretas exclusivamente de sistemas

dificuldade por parte dos participantes em encontrar as respostas.

Existem dez tópicos sem respostas corretas dadas exclusivamente por participantes humanos, sendo que para um deles também não existem respostas corretas dadas por sistemas (**tópico 53** [Toureiros a cavalo de países lusófonos com carreira internacional]). Esses tópicos encontram-se na tabela 14. Como se pode ver na tabela, em três desses tópicos (**tópico 4** [Mulheres violoncelistas de língua portuguesa], **tópico 5** [Flautistas que se naturalizaram brasileiros ou portugueses] e **tópico 107** [Dioceses católicas de Moçambique]) não existem também respostas dadas apenas por participantes humanos; em quatro desses tópicos também não houve respostas corretas exclusivas de sistemas.

Cerca de um quarto dos tópicos (36) não tem respostas corretas dadas por sistemas (seja exclusivamente ou não) e 38% dos tópicos (57) tem respostas corretas dadas por um ou outro tipo de participação, mas não pelos dois.

De alguma forma, a ausência de respostas exclusivas de um dos tipos de participação demostra que esses tópicos são mais difíceis para esse tipo de participante, ou que pelo menos a dificuldade poderá ser semelhante ao do outro tipo de participação se as respostas forem comum aos dois.

### 6.2 Comparação por temas

No Págico os tópicos foram classificados em temas e grandes temas. Assim, mostramos nas tabelas 15 e 16 o desempenho comparativo entre sistemas e participantes humanos, em termos de pontuação final ($M$) e de precisão ($P$), discriminado por grande tema e tema, respetivamente.

Como é facilmente constatável, tanto participantes humanos como sistemas tiveram a melhor precisão no tópicos de geografia, se bem que a pior precisão dos primeiros foi em ciência e as dos segundos em política. No entanto, ao nível da pontuação final, os participantes humanos sairam-se significativamente melhor nos tópicos de letras, enquanto sistemas continuaram a ser melhores em geografia. Dado que o RENOIR é um sistema vocacionado para a recolha de informação geográfica, talvez não seja de espantar esse resultado.

| Tema | *M* | | *P* | |
|---|---|---|---|---|
| | Hum. | Auto. | Hum. | Auto. |
| Letras | 590.72 | 5.24 | 71.52 | 1.90 |
| Artes | 324.80 | 4.48 | 71.07 | 2.46 |
| Geografia | 268.88 | 8.86 | 71.70 | 3.62 |
| Cultura | 205.34 | 2.19 | 67.11 | 2.05 |
| Política | 107.58 | 0.77 | 65.60 | 1.39 |
| Desporto | 104.31 | 1.14 | 63.22 | 1.75 |
| Ciência | 59.08 | 1.88 | 61.54 | 2.57 |
| Economia | 45.10 | 0.32 | 71.59 | 1.61 |

Tabela 15: Pontuação final ($M$) e precisão ($P$) por tema e tipo de participação.

### 6.3 Comparação por localização

A comparação entre sistemas e participantes relativamente à classificação geográfica dos tópicos mostra que a melhor pontuação final para

| ID | Tópico | Enviadas | | | | Corretas | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *T* | *H* | *S* | *HS* | *T* | *H* | *S* | *HS* |
| 19 | Tribos indígenas que vivem na Amazônia. | 250 | 59 | 160 | 31 | 115 | 56 | 35 | **24** |
| 62 | Praias de Portugal boas para a prática de surf | 161 | 7 | 134 | 20 | 30 | 5 | 6 | **19** |
| 7 | Guitarristas portugueses que também foram compositores. | 242 | 26 | 197 | 19 | 34 | 17 | 0 | **17** |
| 11 | Filmes sobre o cangaço. | 223 | 21 | 185 | 17 | 41 | 20 | 4 | **17** |
| 79 | Povos indígenas brasileiros considerados extintos. | 199 | 29 | 153 | 17 | 49 | 27 | 6 | **16** |

Tabela 13: Tópicos com mais respostas corretas de ambos os tipos de participação

| ID | Tópico | Enviadas | | | | Corretas | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *T* | *H* | *S* | *HS* | *T* | *H* | *S* | *HS* |
| 4 | Mulheres violoncelistas de língua portuguesa | 242 | 0 | 240 | 2 | 3 | **0** | 1 | 2 |
| 5 | Flautistas que se naturalizaram brasileiros ou portugueses. | 203 | 0 | 201 | 2 | 3 | **0** | 1 | 2 |
| 71 | Doenças presentes no Brasil no século XVII | 197 | 5 | 189 | 3 | 2 | **0** | 1 | 1 |
| 94 | Parques nacionais de Moçambique | 172 | 1 | 167 | 4 | 4 | **0** | 0 | 4 |
| 107 | Dioceses católicas de Moçambique | 185 | 0 | 178 | 7 | 7 | **0** | 1 | 6 |
| 108 | Jogadores de futebol nascidos em Cabo Verde que representaram a seleção portuguesa | 141 | 2 | 136 | 3 | 2 | **0** | 0 | 2 |
| 111 | Padres católicos que estão ou estiveram ativos em Timor | 164 | 1 | 158 | 5 | 4 | **0** | 1 | 3 |
| 121 | Frutos de Angola | 125 | 2 | 117 | 6 | 4 | **0** | 1 | 3 |
| 132 | Deputados da FRELIMO | 188 | 1 | 185 | 2 | 1 | **0** | 0 | 1 |
| 153 | Toureiros a cavalo de países lusófonos com carreira internacional | 229 | 4 | 225 | 0 | 0 | **0** | 0 | 0 |

Tabela 14: Tópicos sem respostas corretas exclusivamente humanas

| Subtema | *M* | | *P* | |
|---|---|---|---|---|
| | Hum. | Auto. | Hum. | Auto. |
| história | 407,42 | 3,89 | 72,37 | 1,93 |
| geografia | 197,15 | 8,56 | 70,92 | 4,05 |
| cinema | 137,93 | 5,04 | 84,10 | 4,89 |
| demografia | 135,19 | 11,15 | 85,03 | 12,12 |
| literatura | 123,60 | 0,57 | 67,91 | 1,30 |
| política | 107,58 | 0,77 | 65,60 | 1,39 |
| desporto | 104,31 | 1,14 | 63,22 | 1,75 |
| música | 96,66 | 1,10 | 57,53 | 1,69 |
| antro./folc. | 76,33 | 1,32 | 62,56 | 2,36 |
| religião | 70,92 | 1,08 | 70,92 | 3,08 |
| cultura | 50,24 | 0,02 | 67,89 | 0,62 |
| televisão | 49,42 | 0,09 | 91,53 | 0,98 |
| artes | 47,87 | 0,00 | 72,53 | 0,00 |
| economia | 45,10 | 0,32 | 71,59 | 1,61 |
| filosofia | 34,52 | 0,33 | 73,44 | 2,99 |
| linguística | 33,78 | 0,46 | 66,23 | 1,92 |
| culinária | 30,22 | 0,46 | 67,16 | 2,20 |
| arquit./urb. | 25,78 | 0,37 | 66,10 | 1,60 |
| zoologia | 19,70 | 7,50 | 72,97 | 12,30 |
| jornalismo | 18,23 | 0,04 | 67,50 | 0,78 |
| ciência | 13,33 | 0,00 | 66,67 | 0,00 |
| saúde | 10,62 | 0,04 | 55,88 | 0,70 |
| geologia | 8,76 | 0,00 | 48,65 | 0,00 |
| ensino | 6,05 | 0,00 | 55,00 | 0,26 |
| botânica | 4,92 | 0,07 | 61,54 | 1,32 |
| artes plásticas | 3,85 | 0,35 | 38,46 | 2,93 |
| matemática | 3,20 | 0,00 | 80,00 | 0,43 |

Tabela 16: Pontuação final ($M$) e precisão ($P$) por subtema e tipo de participação.

ambos os tipos de participação foi nos tópicos sobre o Brasil (veja-se a tabela 17). Isso talvez se deva ao facto de a maioria dos tópicos estar classificado com esse local, e que, como tal, terá à partida um maior número de respostas associado.

Ao nível da precisão, os participantes humanos obtiveram o melhor desempenho nos tópicos sobre a Guiné-Bissau, enquanto os sistemas obtiveram um melhor resultado nos temas de Angola.

## 7 Comentários finais

Com o Págico foram dados os primeiros passos no sentido de comparar o desempenho de humanos e sistemas numa tarefa de pesquisa de informação na Wikipédia. Embora o objectivo não tenha sido criar uma competição entre humanos e sistemas, mas sim uma colaboração entre ambos a fim de no futuro criar melhores sistemas que possam ajudar os humanos nessa tarefa, apresentámos neste artigo resultados detalhados, mas ainda superficiais, sobre a participação no Págico.

Além de ser necessário no futuro olhar mais aprofundadamente para estes resultados e de analisar as participações de outras perspetivas, realçamos aqui alguns pontos que talvez valha a

| Lugar | *M* | | *P* | |
|---|---|---|---|---|
| | Hum. | Auto. | Hum. | Auto. |
| Brasil | 462.28 | 9.73 | 72.69 | 3.08 |
| Lusofonia | 275.89 | 1.47 | 61.86 | 1.22 |
| Portugal | 202.75 | 2.50 | 73.73 | 2.75 |
| Geral | 64.46 | 0.10 | 65.77 | 0.87 |
| Moçambique | 36.91 | 0.29 | 68.35 | 1.22 |
| Angola | 36.05 | 3.87 | 69.33 | 5.23 |
| Macau | 23.44 | 0.42 | 75.61 | 2.44 |
| Cabo Verde | 19.88 | 0.19 | 76.47 | 1.38 |
| Timor | 13.83 | 0.83 | 62.86 | 4.17 |
| Guiné Bissau | 5.44 | 0.00 | 77.78 | 0.39 |
| São Tomé e Príncipe | 4.45 | 0.03 | 63.64 | 1.14 |

Tabela 17: Pontuação final ($M$) e precisão ($P$) por localização e tipo de participação.

pena explorar:

- caracterizar os tópicos com mais e menos respostas corretas para cada tipo de participação - em (Simões, Costa e Mota, 2012) é feita uma caracterzição pelo número de palavras e de documentos sem ter em conta o tipo de participação;
- apresentar estatísticas de participação humana: quanto tempo e qual a ordem pela qual os participantes humanos tentaram responder, se alteraram a ordem pré-estabelecida, se tentaram responder primeiro a tópicos de um determinado tema e só depois passar a outro, etc.. Este trabalho, em parte foi iniciado em (Costa, Mota e Santos, 2012);
- avaliar as medidas de avaliação para ver até que ponto são realmente úteis para julgar a qualidade das respostas dos participantes.

Para tal, os interessados em estudar estas questões poderão consultar resultados adicionais disponibilizados no sítio do Págico, bem como usar o pacote do Págico, o Cartola, descrito em (Simões, Costa e Mota, 2012) e o SIGA para obter ainda mais resultados.

## 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).

Agradeço à restante organização do Págico pelas várias sugestões de medidas de avaliação, bem como pelas discussões sobre as mesmas e os demais aspetos relacionados com a avaliação dos participantes.

Estou também agradecida aos revisores convidados, Luísa Coheur e Paulo Gomes, pelas suas críticas construtivas que ajudaram a melhor significativamente o artigo.

## Referências

Cardoso, Nuno. 2012. Medindo o precipício semântico. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Coheur, Luísa e Ângela Costa. 2012. Do tópico às respostas: do processo humano à sua simulação. *Linguamática*, 4(1), Abril, 2012. Neste volume.

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.

Costa, Luís, Paulo Rocha, e Diana Santos. 2007. Organização e resultados morfolímpicos. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal, capítulo 2, pp. 15–33.

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

Gonçalo Oliveira, Hugo, Cristina Mota, Cláudia Freitas, Diana Santos, e Paula Carvalho. 2008. Avaliação à medida no Segundo HAREM. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*. Linguateca, pp. 97–129, 31 de Dezembro, 2008.

Miranda, João. 2012. Desafios na recolha de informação baseada na Wikipédia portuguesa com o Págico. *Linguamática*, 4(1), Abril, 2012. "Neste volume".

Mota, Cristina, Alberto Simões, Cláudia Freitas, Luís Costa, e Diana Santos. 2012. Págico: Evaluating Wikipedia-based information retrieval in Portuguese. Em *Language Resources and Evaluation Conference*, Maio, 2012.

Oliveira, Hugo Gonçalo, Cristina Mota, Cláudia Freitas, Diana Santos, e Paula Carvalho. 2008. Avaliação à medida no Segundo HAREM. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*. Linguateca, pp. 97–129, 31 de Dezembro, 2008.

Rodrigues, Ricardo, Hugo Gonçalo Oliveira, e Paulo Gomes. 2012. Uma abordagem ao Págico baseada no processamento e análise de sintagmas dos tópicos. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Santos, Diana. 2012. Porquê o Págico? Razões para uma avaliação conjunta. *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, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, e Daniel Tapias, editores, *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, Maio, 2010. European Language Resources Association (ELRA).

Santos, Diana, Cristina Mota, Alberto Simões, Luís Costa, e Cláudia Freitas. 2012. Balanço do Págico e perspetivas de futuro. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Simões, Alberto, Luís Costa, e Cristina Mota. 2012. Tirando o chapéu à Wikipédia: A coleção do Págico e o Cartola. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Veiga, Arlindo, Carla Lopes, Dirce Celorico, Jorge Proença, Fernando Perdigão, e Sara Candeias. 2012. O desafio da participação humana do IT-Coimbra no Págico. *Linguamática*, 4(1), Abril, 2012. Neste volume.