Language Resources and Evaluation manuscript No.
(will be inserted by the editor)

Evaluation contests in Portuguese

Linguateca's contribution

Diana Santos

Received: date / Accepted: date

Diana Santos
Linguateca
University of Oslo
d.s.m.santos@ilos.uio.no

Abstract This paper presents four initiatives for fostering and jointly evaluating the progress of Portuguese language processing, organized for a decade (2002-2012) by Linguateca, concerning a) morphological analysis, b) named entity recognition, c) (crosslingual) information retrieval and question-answering, and d) Wikipedia search. In addition to summarize, for an international audience, the most important data and results coming from these activities, I discuss some issues critically and reflect on what was learned, also about the challenges of organizing language-specific venues.

1 Introduction

Linguateca was launched by the Portuguese government to give momentum and visibility to natural language processing for the Portuguese language, as one of the outcomes of a public discussion of the scientific community for the White book on Science and Technology (1999) in Portugal. This project, defined as an R&D network for resources and evaluation, had as main goals a) improve the informationa and collaboration among the actors in the field; b) develop public resources in cooperation with the scientific community, and c) organize evaluation contests that could both advance the area and evalutate the progress. It is solely this third axis we will be concerned in the present paper.

Diana Santos
Linguateca and University of Oslo
E-mail: d.s.m.santos@ilos.uio.no

Funding for Linguateca was active from 2000 to 2012, after which period Linguateca has only been given infrastructural support by the Portuguese authorities. Some of its members have, however, continued to work for Linguateca as part of their research committments, so that most of the resources are still available on the Web and some smaller projects have been launched from time to time. In addition, several researchers collaborate with us and even work voluntarily to improve or augment specific resources.

For further information on both the process of devising Linguateca, its activities, and the several phases and turns of its history, the readers are directed to [33,36–38], in Portuguese.

One important feature of Linguateca was that it was devised as addressing the computational processing of Portuguese as an international language: no variety or dialect should be preferred. Also, very early on it was decided to concern itself only with written language. So, in the remainder of the paper I use "Portuguese" as a cover term concerning Brazilian and Portuguese groups, and all the shades of Portuguese (written) language.

2 Evaluation

From the beginning, one of Linguateca's intended activies was joint evaluation, involving the community of researchers working in the area, in order to provide a fair assessment, to gather consensual materials for evaluation, and to develop agreed upon criteria.

We had as models the DARPA and MUC evaluations, so it should not be surprising that Lynn Hirschman's papers on the history of MUC – first at LREC in 1998 [18], and then in a journal [19] – directly inspired the present text. However, there are two main differences between the history of MUC and the history described here:

- In Linguateca we "wandered" among very different tasks to evaluate, in contrary to MUC, whose progress could be described as evolution
- While MUC was superseded/followed by different evaluation efforts, probably because it was considered to have done its job, the evaluation thread at Linguateca stopped abruptly because of quite orthogonal (external) considerations, and we felt no sense of "duty done, no need to go on".

The paper will thus present separately the four areas (some of them with ramifications and/or subareas), nothwithstanding temporal overlap and some reuse of ideas and even tools.

Figure 1 presents grossly the timeframes of the different evaluation initiatives, which wil be presented in turn.

Before the first evaluation contest proper took place, there was a consolidation period where we attemped to make everyone ready for this kind of practice: this campaign was called AvalON (a blending of Aval, the first four letters of *avaliação* – evaluation in Portuguese – and the English ON), and has been partially reported in Portuguese in [51]. In that initiative, we tried

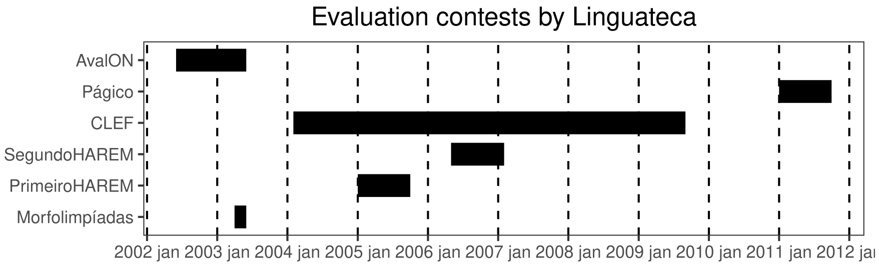


Fig. 1 Rough timeframe of the evaluation contests organized by Linguateca

to involve in evaluation literally everyone in the Brazilian and Portuguese NLP comunities, which meant that several proposals and/or ideas were presented in a sort of brainstorming conference, EPAV – obviously not all of them ended up being implemented. But this also resulted in a – (much) later – published book on the subject of this paradigm for Portuguese [34], whose first part also served as the official documentation of the first evaluation contest.[1] It is to this venue, *Morfolimpíadas*, that we turn now.

3 Morfolimpíadas

The first task we set up to study and evaluate was morphological analysis, because Portuguese has a rich inflection paradigm, and it seemed there was a fair number of groups that had some sort of morphological processing in their systems. Not only those who worked with parsing but also others who dealt with information retrieval or spellchecking.

And it was considered important that the first evaluation contest organized by Linguateca had as many participants as possible, to validate the model in the comunity and to develop trust.

The suggestion of Morfolimpíadas was presented for the first time in June 2002 at EPAV. After a dry run, to assess: a) the effort required from both participants and the organization, b) the golden list building process and c) the possibility og semiautomatically obtaining a silver list, the first (and only) edition of Morfolimpíadas was run from March to June 2003, with seven participants, 2 from Portugal, 3 from Brazil, 1 from a joint France/Portugal venue and 1 from Denmark.

3.1 Task definition

Very briefly, the participants were sent the test materials (613 texts) in three forms: running text (ts), one token per line (uul), and type list (uts). In these

[1] Other proposals deriving from this brainstorming activity were alternatively published in conference proceedings, like [2].

```
Amazonas÷PROP÷Amazonas÷.÷S÷.÷M÷.÷.÷.
Amazonas÷SUB÷amazona÷.÷P÷.÷F÷.÷.÷.
×
C÷NUM÷C÷.÷P÷.÷I÷.÷.÷CARD
C÷PROP÷C÷.÷S÷.÷M÷.÷.÷.
C÷PROP÷C÷.÷S÷.÷M÷.÷.÷QUIM
C÷SUB÷c÷.÷S÷.÷M÷.÷.÷LETRA
×
Capacite-se÷V+CL÷capacitar+se÷PR_C÷S+S÷3+3÷.+I÷.÷.÷.
×
Marinho÷ADJ÷marinho÷.÷S÷.÷M÷.÷.÷.
Marinho÷PROP÷Mário÷.÷S÷.÷M÷.÷DIM÷.
Marinho÷SUB÷marinho÷.÷S÷.÷M÷.÷.÷raro
×
Papa÷PROP÷Papa÷.÷S÷.÷M÷.÷.÷.
Papa÷SUB÷papa÷.÷S÷.÷F÷.÷.÷.
Papa÷SUB÷papa÷.÷S÷.÷M÷.÷.÷.
Papa÷V÷papar÷IMP÷S÷2÷.÷.÷.÷.
Papa÷V÷papar÷PR_I÷S÷3÷.÷.÷.÷.
×
descobri÷V÷descobrir÷IMP÷P÷2÷.÷.÷.÷.
descobri÷V÷cobrir÷IMP÷P÷2÷.÷.÷.÷deriv des
descobri÷V÷descobrir÷PSP_I÷S÷1÷.÷.÷.÷.
descobri÷V÷cobrir÷PSP_I÷S÷1÷.÷.÷.÷deriv des
```

Fig. 2 Examples of forms in the Morfolimpíadas golden list

texts or lists were included the items of our golden list, which included 200 forms which in turn corresponded to 345 different analyses. They should return all three test materials morphologically analised, without knowing which words would be looked into.

In Figure 2 we show six items in the golden list (corresponding to 19 analyses) as example. In [39] we discussed the options that underlie the choice of the golden list items. The categories and some of the choices came from the participants' morphological analysers. All participants were fully aware of the choices and the categories. Right from the beginning the classification of closed words had been removed from the realm of the comparison. It should perhaps be mentioned that, to minimize the work of participants, the output format of every system was converted by the organization into the golden list format for evaluation.

Given that only one participant dealt with multiword expressions, we were not able to jointly evaluate this issue.

3.2 Results and conclusions

Although morphological analysis seems to be a simple task, in order to produce a fair comparison of systems with very different goals, we had to devise results according to three axes:

- morphological analysis proper

- spellchecking
- stemming

Also, possibly the most interesting outcome of Morfolimpíadas was the realization that there remained – and to my knowledge remain to this day – considerable theoretical disagreement about the "proper way to do morphological analysis of Portuguese", as described in [45]. Briefly:

- Tokenization was far from consensual: there was disagreement for 15.9% of the tokens, and 9.5% of the types.
- There were linguistic disagreements on what lemmas should be, for example for noun/adjectives and adverbs
- There was disagreement on how to deal with verb derivation, compound nouns (hyphenated) and capitalization
- There was wide disagreement as far as past participles were concerned
- There was no agreement about the role and the limits of a morphological analyser: For example, should features like upper/mixed/lower case be assigned? Should morphological analysers deal separately with acronyms, or consider them as proper nouns? Should morphological analysers tag errors, and/or foreign words?
- Finally, should semi-formatted kinds of text (references to football matches, laws, bibliographic references, etc.) receive special treatment? And, in any case, which output should be expected in those cases?

There were also differences in the way the different morphological analysers dealt with contractions and clitics, but we had developed scripts to convert all formats to the form A+B, as shown in figure 2.

All results, data, and (Perl) programs used for organizing Morfolimpíadas were made public on Linguateca's site devoted to this contest[2] – where they can still be found, providing interesting data for NLP archaeology.

I believe we can say that with Morfolimpíadas Linguateca showed that the evaluation contest model worked for Portuguese, and that much was learned about how to cooperatively evaluate a given application area, mainly through email contacts and two or three physical gatherings. Time was thus ripe to start braver endeavours.

4 HAREM

HAREM was by far the most charismatic evaluation contest organized by Linguateca, and had two independent editions, roughly in 2005 and 2007-2008. HAREM stands for "HAREM - Avaliação de Reconhecimento de Entidades Mencionadas", a recursive acronym roughly translatable as 'H - Evaluation of Named Entity Recognition'.

The task in itself was not known to have been attempted by any system for Portuguese so far, so this was a clear case of trying to foster work on an

[2] `https://www.linguateca.pt/Morfolimpiadas/`

unchartered area for Portuguese. As stated in the motivation page of the First HAREM, NER was a light task in terms of theoretical load (by this we meant that there were no different camps with unreducible positions, as could be said about e.g. syntactical analysis).

The idea of NER evaluation had already been advanced by Cristina Mota in EPAV, and an initial brainstorming had been written shortly after ([27], only published in 2007, unfortunately).

The number of participants (10 in both editions) – or intended participants (22 in the second) – was higher than in Morfolimpíadas, showing that either NER was a sexier topic, or that we as organizers had achieved the confidence and impact we lacked the first time, or both.

Also, we amassed a larger organizer group, from three of Morfolimpíadas to eight in First HAREM and (other) six in Second HAREM, and we published an (online) book right after each of the venues [43,26]. In addition, a subbranch of First HAREM (MiniHAREM) was the theme of a MSc Thesis [3], and several participants entered HAREM with systems developed in connection with their PhD work (which meant further dissemination). For Morfolimpíadas, there had been merely three papers, published by the organisers, and the first part of the book [34] on evaluation contest as a paradigm for progressing in the area of computational processing of Portuguese.[3]

It can be easily seen that we made NER – REM, in Portuguese – a considerably different task from MUC's, and its organization brought several different challenges, because we wanted, just like in Morfolimpíadas, to accomodate as many systems as possible. So, instead of a rigidly defined task, we gave the systems the possibility to compete, for example, only on some sets of categories.

As to the goal itself, we attempted a bottom-up analysis of the kinds of proper names occurring in Portuguese text, and accordingly come up with categories (and types, a subdivison of categories). No wonder that, by including all genres of texts in the material, we discovered much more than organizations, people, locations, weapons or joint ventures.

In terms of empirical semantic work, I still think it was the right way to proceed, but we should have thought that, by radically changing the task and the measures (a full new set of evaluation measures was developed for HAREM, see e.g. [55]) we would be compromising the comparability of the results to other languages, most prominently English. And, therefore, ever since we have heard this complaint, whenever systems that participated in HAREM report their results in English-speaking venues.

[3] This book was the only one published by a publisher, IST Press, and considerable care was put in its organization, in terms of design, language revision and cross-reviewing. However, the publishing time (four years after Morfolimpíadas took place!) and the fact that it had to be bought afterwards – that is, it was not available on the Web – made it quite a poor choice in retrospect, and this is also why the two books on HAREM, cross-reviewed, were published online at once.

4.1 The First HAREM

As already said, we created a set of categories and types for Portuguese through analysis of real texts and manual annotation, and wrote a lengthy set of directives to explain how the task should be performed, with plenty of examples, that was sent to the participants for comments and improvements. In fact, the golden collection was started as a cooperative endeavour, having different participants annotating different pieces, in order to have everyone on the same boat.

Those categories and types can be seen in Table 1, together with their translation into English.

At the same time of writing the annotation directives we created a golden collection in a pseudo-XML format, which is exemplified in figure 3.

```
O <PESSOA TIPO="CARGO">Presidente da ONU</PESSOA> foi abordado por um <PESSOA
TIPO="MEMBRO">GNR</PESSOA> à paisana quando ia no seu <COISA TIPO="MEMBROCLASSE">
Fiat Punto<COISA> para a cidade de <LOCAL TIPO="ADMINISTRATIVO">Viseu</LOCAL>.
```

Fig. 3 One (fictitious) example of how First HAREM's golden collection was encoded

One of the most important things we catered for was the possibility that a named entity in context represented more than one category. The participant systems were supposed to identify which categories a particular NE made reference to, and state the vagueness if it was vague in a particular context. For example, in a sentence like *Eu gosto de Portugal* (I like Portugal) one might be referring to the place, the people or even (although less likely) to the organization, state, or even abstraction – or to all of them at once. So we went much further than Markert and Nissim when they described some typical metonymic associations in NER in [24], or than ACE [7], which catered for geopolitical entities but not for all kinds of vagueness.

We also invested a large effort in the evaluation setup, which had some novel characteristics compared to other evaluation contests:

- We performed evaluation of three different tasks: NE identification, NE morphological analysis, and NE classification. By conceptually separating identification from classification, we could produce relative values as well, namely how good was a system in the classification task taking only in consideration the correctly identified NEs.
- We were careful to maintain vagueness/alternative interpretations whenever needed, using a codification using ALT for alternatives in the golden collection:

  ```
  <ALT><EM>Governo PSD de Cavaco Silva</EM>|<EM>Governo PSD</EM>
  de <EM>Cavaco Silva</EM>|Governo PSD de Cavaco Silva</ALT>
  ```

- We let participants compete in selective scenarios, which means that we had to compute evaluation measures for all sets of categories systems chose to participate in.

Category	Type	English gloss	Nr.
PESSOA	INDIVIDUAL	individual person	856
	CARGO	title	79
21.5%	MEMBRO	members	10
	GRUPOIND	group of people	10
BP: 58.75%	GRUPOCARGO	group of titles	19
BR: 72.72%	GRUPOMEMBRO	group of members	137
ORGANIZACAO	ADMINISTRACAO	administration	224
18.0%	INSTITUICAO	institution	462
BP: 51.01%	EMPRESA	company	230
BR: 62.72%	SUB	sub-organization	61
TEMPO	DATA	date	335
8.5%	HORA	time	39
BP: 77.68%	PERIODO	period	62
BR: 69.79%	CICLICO	cyclic	5
LOCAL	CORREIO	address	17
	ADMINISTRATIVO	administrative	906
25.0%	GEOGRAFICO	geographic	86
BP: 68.03%	VIRTUAL	virtual	126
BR: 73.91%	ALARGADO	extended	161
OBRA	PRODUTO	product	74
4.0%	REPRODUZIDA	reproducible work	89
BP: 20.58%	ARTE	unique work	10
BR: 18.85%	PUBLICACAO	publication	51
ACONTECIMENTO	EFEMERIDE	unique	23
2.5%	ORGANIZADO	large event	62
BP: 50.76%	EVENTO	atomic event	45
BR: 46.61%			
ABSTRACCAO	DISCIPLINA	subject	228
	MARCA	brandname	36
8.5%	ESTADO	condition	34
	ESCOLA	school	14
BP: 45.43%	IDEIA	ideal	45
	PLANO	plan	40
BR: 38.04%	OBRA	complete works	4
	NOME	name	76
COISA	OBJECTO	object	39
1.6%	SUBSTANCIA	substance	9
BP: 25.38%	CLASSE	class	37
BR: 40.74%			
VALOR	CLASSIFICACAO	classification	62
9.5%	QUANTIDADE	amount	370
BP: 84.82%	MOEDA	money	53
BR: 79.69%			
VARIADO		other	42
0.9%			

Table 1 First HAREM categories and types, their distribution in the golden collection, and best precision (BP) and recall (BR), from [53]

- For identification, we added two possible ouputs in addition to correct, spurious and missing, namely partially correct by excess, and partially correct by shortage.
- For morphological classification, we classified the outputs as correct, partially correct, overspecified, missing and spurious.
- For semantic classification, we provided for different measures, given that we had categories and types: only categories, the pair category and type, only types for correct categories, and a combined punctuation for category

and type:

$$\mathrm{P}_{CSC} = \begin{cases} 0 & \text{if the category is wrong.} \\ 1 & \text{if the category is right but the type is wrong.} \\ 1 + \left(1 - \dfrac{n_c}{n_t}\right) - \dfrac{n_e}{n_t} & \text{if the category is right and at least one type is correct.} \end{cases} \tag{1}$$

where n_c represents the number of correct types, n_e the number of spurious types, and n_t the possible number of types in the actual category.

All these different issues resulted in a fairly complicated evaluation architecture, described in [55]. All programs (of which half were written in Java, and the other half in Perl) were made public and downloadable from our site. Additionally, a huge amount of documentation of the programs, the architecture and the evaluation measures was also made available.[4]

4.2 Mini-HAREM

In order to study better the issue of statistical validation, it was necessary to gather some more runs, and we asked the participants of the First HAREM to compete once more at a later stage, with of course a new set of texts. The results of the comparison and validation of the system rankings were the subject of Nuno Cardoso's MSc thesis [3], who used approximate randomization, a non-parametric approach that had also been used in MUC, to assess whether the differences in ranking of the systems were statistically significant, and whether the size og the golden collection(s) was appropriate.

All in all, we considered Mini-HAREM to be part of First HAREM, and therefore First HAREM's golden collection contains the data of both venues. In addition to be available for download, we also made it available for search on the Web through the AC/DC project [37], so that linguists interested in the linguistic phenomenon of named entities (or onomastics, etc.) could look at language without having to develop a system or participate in HAREM.[5]

Although the corpus was later increased with the golden collection of the Second HAREM, created in 2008, the creation of the corpus was an idea and a result of the First HAREM. It comprises more than 225,000 words, including ca 16,000 named entities.

4.3 The Second HAREM

The Second HAREM [46] was something that we proposed to do in the third funding phase of Linguateca, from December 2006 to December 2008, and therefore it was from the beginning time-limited. The organization team was

[4] See the site of the First HAREM, `https://www.linguateca.pt/primeiroHAREM/harem.html`. Everything is still available from there.

[5] `https://www.linguateca.pt/acesso/corpus.php?corpus=CDHAREM`

also almost totally changed (the author of this paper being the only link between the two teams).

The ostensibly more important difference in the Second HAREM relative to the First was that it had three tracks. This meant that, although related, there were three areas that were being separately evaluated. Moreover, the two new tracks, as opposed to the "classical" one, as it was called, were proposed by the community – and the temporal identification track [15,16] was organized by the group who proposed it, in what concerns the task definition, while annotation of the golden resource and the evaluation proper was done by Linguateca. ReRelEM, looking at relations among named entities, ended up by being organized by Linguateca.

Second HAREM, in its "classical" track, featured the following changes compared to the first:

- output codification was simplified in that all named entities were marked as EM, and both categories and types were features of the EM. See the following example:

Pela mão do <EM ID="a66435-10" CATEG="PESSOA" TIPO="INDIVIDUAL">**ministro Freitas do Amaral**</EM>, e sem necessidade alguma, <EM ID="a66435-10" CATEG="ORGANIZACAO—PESSOA" TIPO="ADMINISTRACAO—POVO">**Portugal**</EM> foi enxovalhado, coberto de vergonha e de cobardia, por um dos mais tristes textos políticos que já alguém escreveu.

- some categories underwen changes regarding the types allowed. The most radical change was the LOCAL category, which was changed into a tripartition: FISICO, HUMANO and VIRTUAL (new), each of which allowing for a finer-grained classification (subtypes). Figure 4 shows all changes.
- the delimitation of named entity suffered some changes: in particular, a list of allowed uncapitalized members of a named entity was made public, so that cases like *doença de Chagas* (Chagas' disease) or *rua das Amoreiras* (Amoreiras street) could be considered one named entity. Also person descriptions whose structure could be analysed as "honorific from a place", like *Marquês de Pombal* (Marquis of Pombal), were deemed to be only one named entity.
- the ALT mechanism was employed to accept embedded named enities, something which was not dealt with in the First HAREM, where systems were asked to identify the most encompassing one only.
- and we added another line of evaluation (dubbed "relaxed ALT") which gave total punctuation if the system came up with any of the alternative, as opposed to "rigid ALT", where the system was expected to provide all alternatives listed in the golden collection.

To improve the organization work, an annotation tool was created to speed up the creation of the golden resource. Etiquet(H)AREM [4] was a Java tool for XML documents to which one could add category, type, subtype, but also longer segments that should be ommited from evaluation, relations, and comments. Basically, it made it simple to guarantee the syntactic correcteness

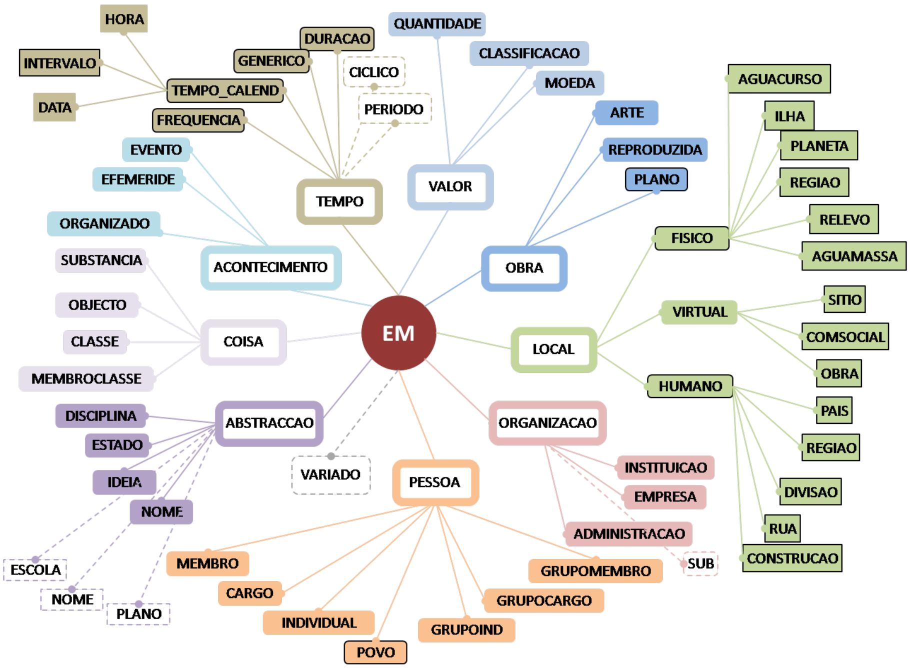


Fig. 4 The category tree in the Second HAREM: categories, types and subtypes with solid black contour belong to Second HAREM only; those with dashed contour existed only in the First HAREM

of a somehow complicated XML-syntax, as well as easier for linguist annotators to work with. (In the First HAREM the golden resource had been created manually in Emacs by the two main contributors.) This application, publicly available, was parametrizable through a text file where the allowed values of the categories and types where stored, and was to be installed in the annotators machine. One can say that it worked as a precursor of BRAT[6], but, to my knowledge, it was never used by others outside HAREM.

As to the evaluation, we used one single measure that weighted all the different kinds of information, shown in Figure 5.

4.4 ReRelEM

ReReLEM's goal was to identify a set of relationships among HAREM named entities, that chose – after close analysis of textual material – the following four relations: IDENTITY, INCLUSION, LOCALIZATION (or HAPPENED-IN), and OTHER.

One particular named entity can obviously be related to several others, as in the example

[6] `http://brat.nlplab.org/index.html`

$$
\begin{aligned}
&1+\sum_{i=1}^{N}((1-\frac{1}{n_{cats}})\cdot cat_{certa_i}\cdot\alpha+(1-\frac{1}{n_{tipos}})\cdot tipo_{certo_i}\cdot\beta+(1-\frac{1}{n_{sub}})\cdot sub_{certo_i}\cdot\gamma)\\
&\quad-\sum_{i=0}^{M}(\frac{1}{n_{cats}}\cdot cat_{esp_i}\cdot\alpha+cat_{certa_i}\cdot\frac{1}{n_{tipos}}\cdot tipo_{esp_i}\cdot\beta+tipo_{certo_i}\cdot\frac{1}{n_{sub}}\cdot sub_{esp_i}\cdot\gamma)
\end{aligned}
\tag{2}
$$

$$K_{certo_i}=\begin{cases}1 & \text{if the attribute } K_i \text{ is right,}\\ 0 & \text{if } K_i \text{ is wrong or missing}\end{cases}$$

$$K_{esp_i}=\begin{cases}1-K_{certo_i} & \text{if the attribute } K_i \text{ has a value}\\ 0 & \text{se } K_i \text{ is missing}\end{cases}$$

$K \in \{cat, tipo, sub\}$
N = number of different vague classifications in the golden collection, according to the selective scenario.
M = number of spurious classifications in the run, according to the selective scenario.
α, β, γ = parameters corresponding to the weights of the categories, types, and subtypes.

Fig. 5 Evaluation in the classical track of Second HAREM

> depois de partir em vantagem pontual no <EM ID="b13" CATEG="ACONTECIMENTO" TIPO="ORGANIZADO" COREL="b3 b5 b11" TIPOREL="ident ident ocorre_em">**Campeonato do Mundo**</EM>

This represents the information that the b13 NE is related to b3 and b5 by the identity relation, and to b11 by the happened-in relation.

Since it is exacly the same to say that A occurs-in B, or that B is the localization of A, and if A is identical to B and B includes C, then A also includes C, and so on, we developed a set of rules so that all possible logically valid relations were automatically filled in, inspired by [59] in MUC.

ReRelEM's golden collection (a subset of HAREM's golden collection containing 12 texts and 573 named entities) was annotated with the help of Etiquet(H)AREM. After automatic expansion, it contained 6,477 relations.

Only three systems participated, and in fact all in different configurations, which made it hard to provide a fair comparison: One used the full set, another competed only for places and for the inclusion relation, and the last one did not mark the OTHER relation, and only tried identification (not classification) in HAREM.

In any case, we were able to get a few results: we established that IDENTITY was far easier to identify than the other relations, and we found 21 different relations included in the OTHER portmanteau (like published in, born in, produced by, participant of and character of).

Although this was an interesting pilot, and we also published on it in an international venue [10], the definition of the relations was too tightly tied to the HAREM philosophy and options, so that ReReLEM was not possible to generalize or move outside HAREM.

4.5 Further comments on HAREM

Although named entity recognition seems a simple task, there have appeared in the literature a wealth of different approaches, definitions, and evaluation frameworks. I am not the first or the most vocal about it, but I discussed this in [35] and Nuno Seco had a chapter on the subject in English in the first HAREM book [54]. But one can see the discussion of what NE and NE evaluation means still coming in many different disguises and venues, see [20] in 2016.

5 CLEF

Already in 2003 we started collaboration with CLEF, the Cross-lingual evaluation forum, an European "spin-off" from TREC[7]which started in 1998 as the CLIT track and from 2002 on was financed by diverse European funding initiatives, such as the Fifth Framework Program, the DELOS Network of Excellence on Digital Libraries, etc. See [29,30] for an initial presentation of the global project.

From the point of view of Portuguese, it seemed like a good idea to participate in a setup with a large audience and much expertise behind, especially in tasks of which we had no experience whatsoever, such as information retrieval (IR) and automatic question answering (QA). In addition, the added bonus of getting non-lusophone participants for crosslingual tasks seemed a way to put Portuguese on the map.

Linguateca was therefore a member of the organization from 2003 to 2009 (while there was enough funding for Linguateca), and gradually we moved from just ad-hoc IR and QA to also organizing more innovative tasks such as GeoCLEF, GikiP and GikiCLEF, while also giving a hand to ImageCLEF, WebCLEF, ResPubliQAa and LogCLEF.

Contrary to what could be expected, there were always fewer Portuguese-processing participants in CLEF than in HAREM. In fact, often there was only one group participating in Portuguese IR. QA had two participants in 2004 [57], three participants in 2005 [58], five in 2006 [21], six in 2007 [13], and six in 2008 [8]. In 2009 the task changed into QA over European legislation, ResPubliQA, and only one Portuguese participant remained [31].

After some time we got more involved and familiar with the (amazingly productive and challenging) CLEF community, and started to also propose new challenges. After having embarked in GeoCLEF [12] in 2005 (and went on until 2008, so there were four GeoCLEF venues with Portuguese), we suggested GikiP in 2008 as a GeoCLEF pilot (and organized it for English, German, and Portuguese) and led a larger group who organized GikiCLEF in 2009.

[7] TREC, the Text REtrieval Conference, started in 1992 and yearly after that, played the same role for information retrieval as MUC played for information extraction, see [17].

5.1 Adhoc CLEF

Let us start by the most classic track of CLEF, a standard information retrieval task, in a crosslingual setting. And because not every NLP reader is aware of how things are done in the information retrieval world, I present here the task: given a collection of documents, and a set of topics, the competing systems should provide a ranked set of documents about the topic. Either monolingually (searching topics expressed in Portuguese in a collection of documents in Portuguese) or crosslingually (e.g. searching topics expressed in French over a Portuguese collection, or vice-versa), or multilingually across all collections.

Adding Portuguese meant in practice three things, discussed at some length in [52,32]:

- we had to make a text collection available – the CHAVE collection (*chave* means key in Portuguese, just like *clef* means key in French), featuring full texts from two main newspapers, one from Portugal and one from Brazil from 1994 and 1995.
- we had to come up with topics appropriate for the collection, as well as for multilingual querying. For example we would strive to get topics about lusophone countries which were also present in other language collections, and to find for example Finnish themes that occurred in the Portuguese collection as well as in Finland. (These topics had also to be translated into English, and we had to translate the other languages topics into Portuguese)
- we had to evaluate the results in the Portuguese collection, in that a number of documents for each topic had to be tagged as relevant or irrelevant. (To give an idea of the pool sizes, in 2004 22,311 documents were evaluated, with an average of 446 documents per topic.)

From the beginning, after the evaluation contest had taken place we made the topics and their binary judgements on the CHAVE collection public, so that systems could use them for training.[8]

5.2 Question answering

Question answering followed a similar evaluation setup over the same collection, just instead of topics one created questions (classified by the type of answer expected), and instead of relevant documents one evaluated specific answers. Before suggesting a question, we had to check that the answer could be found in the collection, preferably in more than one document, so that the question could be used in QA@CLEF.

Just like for the adhoc track, all answers to a particular question were pooled and made available together with the CHAVE collection. An example can be seen in Figure 6:

[8] `https://www.linguateca.pt/CHAVE/`

```
<pergunta ano="2004" id_org="0470" categoria="F" tipo="LOCATION" restrição="X"
ling_orig="IT" tarefa_pt="0002" tarefa_it="0020" tarefa_nl="0007">
<texto>
Onde era o campo de concentração de Auschwitz?
</texto>
<resposta n="2" docid="LING-940804-089">
Sul da Polónia
</resposta>
<resposta n="3" docid="LING-941120-083">
Polónia
</resposta>
<resposta n="4" docid="LING-950126-162">
Sudoeste da Polónia ocupada
</resposta>
</pergunta>
```

Fig. 6 Example of a question and its correct answers, as made available in CHAVE. This question was proposed by the Italian organizers in 2004.

Having participated in the organization of the QA track from 2003 to 2008 (corresponding to the tasks of 2004, 2005, 2006, 2007 and 2008), there are 4380 questions available in Portuguese, together with their answers in the CHAVE collection. (It should be noted that the questions were collectively gathered by all organizers since QA was both monolingual and crosslingual, so that some of the questions were suggested by Linguateca, while others were brought in by e.g. the Bulgarian, Dutch or French groups. This is the reason why not all questions have answers in CHAVE – they have in other collections.)

It is important to report that although this was the setup in the beginning, the QA task was improved and enriched from one edition to the other, and a reader interested in multilingual QA at CLEF should read all the overview papers, also because there were several subtracks. Here I just comment on a few things that are relevant for the Portuguese part.

For example, from 2006 on, in addition to providing an answer, a justification for it using the collection had to be provided, to prevent systems from knowing the answer from other sources, like the internet or their own knowledge bases.

This led to a fivefold classification of answers: correct (and justified), inexact, unsupported, wrong and missing, see [22].

Also, questions were created in clusters about a topic, accepting for example co-reference among them, instead of just independent questions, see [13].

Finally, list questions, that is, questions that required a closed or open list as answer were added, as well as temporally restricted questions.

From 2007 on, wikipedia collections were also considered as a valid answer source, and questions could have answers in both or just one of the two collections. For the QA tasks of 2007 and 2008, the Wikipedia collections from November 2006 were made available to the participants [9]. This was something that, while providing some extra work for the organization, also opened for novel tasks, namely GikiP and GikiCLEF.

But first we report on GeoCLEF.

5.3 GeoCLEF

GeoCLEF [23] was an endeavour started by Fred Gey and colleagues which began while Portuguese was already well established in CLEF, after a pilot in 2015 on German and English, and whose main goal was to develop and evaluate geographical information retrieval. Contrary to other tracks, it had a lower number of languages (and therefore organizers): in 2006, German, Portuguese, Spanish and English, in 2007 and 2008 Spanish dropped out. The idea was to create challenging topics which required geographical knowlegde and therefore reasoning.

Two examples are "Cities within 100km of Frankfurt" and "Malaria in the tropics", respectively illustrating both precisely delimited and vague regions. But there were a large number of interesting issues while organizing GeoCLEF that were beyond the scope of "subject in region" topics, as discussed in [12].

5.4 GikiP

GikiP was a pilot of GeoCLEF 2008. The name of the task includes G for geographic, iki for Wikipedia, and P for pilot.

Using the Wikipedia collections already available for the QA track, the task was to answer (list) questions/information requests that required geographical reasoning of some sort, providing Wikipedia pages as answer. This meant a sort of hybrid between QA and IR, because we did not think it made sense to sharply distinguish the two tasks/research domains.

One of the motivations for using Wikipedia in a multilingual/crosslingual context was that Wikipedia is an interesting mixture of comparable and translation corpora, given the language links between different languages, see [44].

In figure 7 one can see the kind of results for a topic like: "Which Swiss cantons border Germany?"

```
de/k/a/n/Kanton_Aargau.html
de/k/a/n/Kanton_Baasel-Landschaft.html
en/a/a/r/Aargau.html
en/b/a/s/Basel-Land.html
pt/a/r/g/Argóvia_(cantão).html
pt/b/a/s/Basileia-Campo.html
```

Fig. 7 A (reduced) example of the output of a GikiP topic.

Only three systems participated, but it laid the ground for a larger venue the next year, called more pompously GikiCLEF. As is Linguateca's practice, all 15 topics and answers are available from the GikiP site.[9]

[9] https://www.linguateca.pt/GikiP/

5.5 GikiCLEF

GikiCLEF ran in nine languages (ten Wikipedias, because Norwegian has two) and received 17 runs from 8 participants, of which one was Portuguese and another Brazilian. It had 22 members in the organization committee, plus 10 others for answer assessment.

There were 50 topics (prepared in 10 languages), and systems had to provide Wikipedia page ids, plus a justification. The scoring gave better results the more languages a system could find an answer in. Topics were on purpose chosen so that they reflected cultural aspects, so that we would not find a similar answer in all languages. But as we discuss in [40], that had the perverse consequence that for any topic whatsoever there were more answers in English, and that since the 50 topics covered very different cultures and themes, it was hard to make statistical generalizations. In any case, GikiCLEF produced 1,009 answers in a set of 10 wikipedias, made available in the GIRA package.[10]

Of all the evaluation contests we have organized, I believe this was the one who required most computation and work for dealing with relatively complicated and large-sized (for the time) materials and answers, see [41]. This led to the development of the SIGA system [6], which was made available to the community. However, and notwithstanding the heavy work and the highly international participation, this task only occurred once, and I have my doubts that it added to any substantial increase in the state-of-the art of information gathering. And as far as Portuguese is concerned, it hardly materialized in any progress. After all, and as reported in [40, page 219], there was only one topic that had a lusophone theme: "Brazilian coastal states"!

One can however trace its influence in the last evaluation contest organized by Linguateca, Págico, where we were back to Portuguese-only challenges, and added a human in the loop, or rather, developed a contest where humans could compete as well.

5.6 Other CLEF tasks

CLEF was a world on its own, and our participation in the organisation was in a way dependent on having participants in Portuguese. Although we had a minimal contribution in tracks like WebCLEF, pilot GeoCLEF and ImageCLEF, simply translating topics or analysing the collections, we already in 2005 [42, page 2] argued for the importance of language and culture:

> you have to know well a language and culture in order to organize meaningfully evaluation contests dealing with it. Just performing translation afterwards, no matter how good, is never enough.

[10] `https://www.linguateca.pt/GikiCLEF/GIRA/`

5.7 Final remarks on CLEF

Like any choice, to participate in the organization of CLEF had advantages and disadvantages. While it did create more opportunities to evaluate different tasks dealing with Portuguese, it also moved the responsibility and significantly reduce the feeling of participating in their own future that was achieved in Morfolimpíadas and especially HAREM. The Portuguese community entered a much larger community, but not everyone was thrilled by it – in fact, we lost a significant part of the community that did not feel this was where they wanted to head. We entered an European club – but mostly lost the lusophone feeling.

Also, very few other participants in CLEF tried their hand at Portuguese – which was, anyway, for them "just another language". So, ultimately there was not very much to be gained by having international competitions where people were not even able to do a superficial error analysis, because they did not understand the language in the first place.

Although we tried to create interesting topics and questions, which can be used for training new systems, this work was not taken up by many – in fact, we see now several papers about question answering in Portuguese that don't even cite these materials. So, apparently the internationalization did not help to make the materials known.

6 Págico

Págico was organized already under much weaker conditions from Linguateca's side, and it worked as a test of how much we could do without Linguateca's funding,[11] using small funds for small projects from the University of Oslo.

The idea was to make the evaluation contest also relevant for human participation – for example students of Portuguese as foreign language, or researchers of Portuguese and Brazilian culture – using questions that related to culture of lusophone countries [28].

It was marketed as having as goal the obtaining of non-trivial answers for complex information needs in Portuguese, and its name was a blend of the words *página* (page) and *mágico* (magic), directed at the Portuguese Wikipedia. It had a group of five organizers, four of which organizers of previous evaluation contests. The reports on the organization, the participants, and the results were published in a special volume of the *Linguamática* journal [48].

Págico had seven participants (only two with automatic systems). It was the only evaluation contest where no Brazilian team or individual participated (although co-organized by a Brazilian university, PUC-Rio, and ironically the majority of Págico's topics concerned Brazil). This clearly indicates that the task suffered from lack of publicity, for example not reaching people who might be interested in lusophone culture. Or that humnities scholars were not ready for these venues. In fact, most human participants had a technological (NLP or

[11] More precisely, we had a very restricted funding in 2011 and 2012, and none after.

Table 2 Examples of some topics/questions in Págico

Topic	English translation
Tribos indígenas que vivem na Amazónia	Indigenous Amazonian tribes
Locais mencionados nos Lusíadas	Places mentioned in Lusíadas
Museus em capitais de países lusófonos	Museums in lusophone capitals
Políticos da África lusófona que estudaram na União Soviética	Lusophone African politicians who studied in the Soviet Union
Frutos de Angola	Fruits from Angola

IR) background, and participated in order to design future automatic systems to compete. One can also claim that it was a too difficult task – and definitely not a priority for most people working in Portuguese NLP.

6.1 Task

More specifically, the goal of Págico was to answer 150 questions about lusophone culture in Portuguese Wikipedia (a version provided for Págico inside SIGA) which had non-trivial answers, in the sense that they would not be covered by a single page or hub. In other words, we were targeting aggregate answers: to obtain a justified list given a specific information need, rewarding variation/quantity of answers.

Some examples of questions/topics can be seen in Table 2. They tried to cover all places where Portuguese is spoken, and were distributed among humanities, arts, geography, culture, politics, sports, science and economy. Some were considered local, others global, that is, related to more than one country/region where Portuguese is spoken. One of Págico's challenges – far from perfectly solved – was to provide an additional environment for human participants to compete, in addition to all possibilities and features for assessing and evaluating already available in SIGA from GikiCLEF.

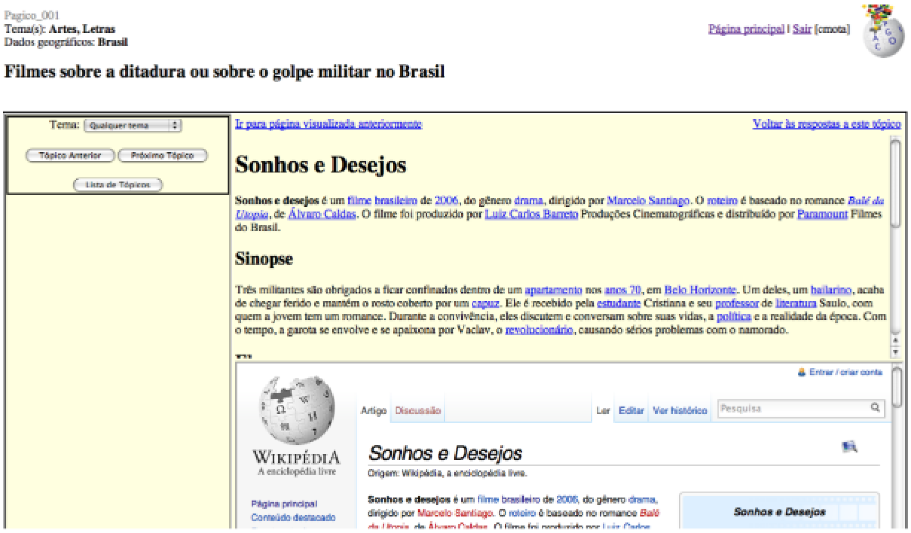
Figure 8 shows one view of SIGA's interface for human search, while Figure 9 illustrates the interface for answer assessment.

While creating the topics, the topic creators also gathered possible answers, in order to be able to later compute pseudo-recall (and corresponding pseudo F-measure). Answers were also marked as "self-justified" when it was clear from the title page that the page was an answer.

During the contest, we also measured the time people used searching for answers, and how many pages were visited, since one of the goals was to study human performance vs. machine performance. (It should be mentioned in passing that, while automatic systems only had one week to return their results, human participants had three).

6.2 Results

In Págico we proposed several different measures in order to evaluate the participation: In addition to the usual precision and (pseudo)-recall, using all



Pagico_001
Tema(s): **Artes, Letras**
Dados geográficos: **Brasil**

Filmes sobre a ditadura ou sobre o golpe militar no Brasil

Tema: Qualquer tema

Ir para página visualizada anteriormente

Voltar às respostas a este tópico

Sonhos e Desejos

Sonhos e desejos é um filme brasileiro de 2006, do gênero drama, dirigido por Marcelo Santiago. O roteiro é baseado no romance *Balé da Utopia*, de Álvaro Caldas. O filme foi produzido por Luiz Carlos Barreto Produções Cinematográficas e distribuído por Paramount Filmes do Brasil.

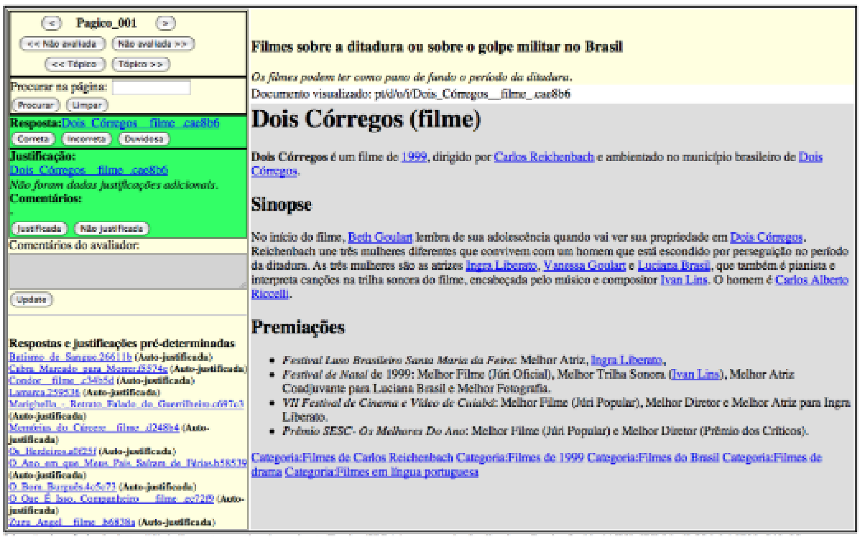
Sinopse

Três militantes são obrigados a ficar confinados dentro de um apartamento nos anos 70, em Belo Horizonte. Um deles, um bailarino, acaba de chegar ferido e mantém o rosto coberto por um capuz. Ele é recebido pela estudante Cristiana e seu professor de literatura Saulo, com quem a jovem tem um romance. Durante a convivência, eles discutem e conversam sobre suas vidas, a política e a realidade da época. Com o tempo, a garota se envolve e se apaixona por Vaclav, o revolucionário, causando sérios problemas com o namorado.

Sonhos e Desejos
Origem: Wikipédia, a enciclopédia livre.

Sonhos e desejos é um filme brasileiro de 2006, do gênero drama, dirigido por Marcelo Santiago. O roteiro é baseado no romance *Balé da Utopia*, de Álvaro Caldas. O filme foi produzido por Luiz Carlos

Fig. 8 Interface for a human participant, after selecting one page as answer, for the topic "Films about the Brazilian dictatorship or the military coup".



Pagico_001

Procurar na página:

Resposta: Dois_Córregos__filme_.cae8b6

Justificação:
Dois_Córregos__filme_.cae8b6
Não foram dadas justificações adicionais.
Comentários:
-

Comentários do avaliador:

Respostas e justificações pré-determinadas
Batismo_de_Sangue.266l1b (Auto-justificada)
Cabra_Marcado_para_Morrer.f5574c (Auto-justificada)
Condor__filme_.c34b5d (Auto-justificada)
Lamarca.259536 (Auto-justificada)
Marighella_-_Retrato_Falado_do_Guerrilheiro.c697c3 (Auto-justificada)
Memórias_do_Cárcere__filme_.d248b4 (Auto-justificada)
Os_Herdeiros.a0f25f (Auto-justificada)
O_Ano_em_que_Meus_Pais_Saíram_de_Férias.b58539 (Auto-justificada)
O_Bom_Burguês.4c5e73 (Auto-justificada)
O_Que_É_Isso,_Companheiro__filme_.cc72f9 (Auto-justificada)
Zuzu_Angel__filme_.b6838a (Auto-justificada)

Filmes sobre a ditadura ou sobre o golpe militar no Brasil

Os filmes podem ter como pano de fundo o período da ditadura.

Documento visualizado: pt/d/o/i/Dois_Córregos__filme_.cae8b6

Dois Córregos (filme)

Dois Córregos é um filme de 1999, dirigido por Carlos Reichenbach e ambientado no município brasileiro de Dois Córregos.

Sinopse

No início do filme, Beth Goulart lembra de sua adolescência quando vai ver sua propriedade em Dois Córregos. Reichenbach une três mulheres diferentes que convivem com um homem que está escondido por perseguição no período da ditadura. As três mulheres são as atrizes Ingra Liberato, Vanessa Goulart e Luciana Brasil, que também é pianista e interpreta canções na trilha sonora do filme, encabeçada pelo músico e compositor Ivan Lins. O homem é Carlos Alberto Riccelli.

Premiações

- *Festival Luso Brasileiro Santa Maria da Feira*: Melhor Atriz, Ingra Liberato,
- *Festival de Natal* de 1999: Melhor Filme (Júri Oficial), Melhor Trilha Sonora (Ivan Lins), Melhor Atriz Coadjuvante para Luciana Brasil e Melhor Fotografia.
- *VII Festival de Cinema e Vídeo de Cuiabá*: Melhor Filme (Júri Popular), Melhor Diretor e Melhor Atriz para Ingra Liberato.
- *Prêmio SESC- Os Melhores Do Ano*: Melhor Filme (Júri Popular) e Melhor Diretor (Prêmio dos Críticos).

Categoria:Filmes de Carlos Reichenbach Categoria:Filmes de 1999 Categoria:Filmes do Brasil Categoria:Filmes de drama Categoria:Filmes em língua portuguesa

Fig. 9 Interface for the judges, after having accepted the justification as correct (and thus the green color)

pooled correct answers and the answers already gathered by the organisers, we added relaxed precision (without assessing justification), and two more original ones: originality, and creativity.

Originality, which is measured per run, and per participant (in case a participant submitted more than one run – and the two automatic systems both submitted three runs), rewards correct answers that were given only in that run/by that participant.

Creativity K, on the other hand, is a measure that punctuates (a run, or a participant) inversely proportionally to the number of different runs (or participants) which provided the same answer.

$$K_{p,c} = \sum_{i}^{T} \sum_{j}^{R_{p,c,i}} k(r_{p,c,i,j}) \tag{3}$$

$$k(r_{p,c,i,j}) = \begin{cases} \frac{1}{c(r_{p,c,i,j})} \times p(i) \; r_{p,c,i,j} \in C_{Pagico} \bigcup C_{aval} \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

$p(i)$ = number of participants who attempted to answer the topic$i$
$c(r_{p,c,i,j})$ = number of participants who gave the answer$r_{p,c,i,j}$
p = participant p
c = run (corrida) c
C = set of correct answers correcly justified
$\tilde{C}$ = set of correct answers incorrectly justified
R = set of answers
T = set of topics

It should not be a surprise that the human participants were more creative and original, but it was interesting to see that the automatic systems managed to come up with a few answers not found by the human participants. Given that most of the latter only answered a subset of the topics, it was also necessary to do selective evaluation (like in HAREM) in four different scenarios.

Although we made a tentative assessment of which topics were most difficult for humans and for machines in [25], the low participation in both categories did not allow us to generalize, and we concluded with a rather critical assessment of Págico in [49] despite the considerable workload involved.

As usual, all data (including the Wikipedia collection of 25 April 2011 in XML, comprising 681,058 documents), the answer pool and the results were made available in what we called the Cartola resource.[12]

7 Discussion

Hirschman makes several pertinent considerations about evaluation in [19], which deserve to be discussed or restated here: First, that the evaluation contest model does not exhaust the wide range of evaluation possibilities in NLP, something I absolutely subscribe.

Secondly, she calls our attention for whose stakeholders are concerned in a particular evaluation: the industry, the users, the researchers, or the funders? It is obvious that the evaluation contests organized by Linguateca were

[12] `https://www.linguateca.pt/Cartola`

researcher-based, or technology developer-based. It was quality and methods that were evaluated, not efficiency or price, and even less sustainability from a funding agency perspective.

The only thing that may distinguish Linguateca from other resource centers is that all resources created (and therefore also the evaluation resources) were equally free and unrestricted for industry use as for research use. There were never any strings attached. We can boast some industry participation in our evaluation contests: one Portuguese company participated both in Second HAREM and in three editions of QA@CLEF. Although it sounds very modest, this is more than for example most other QA@CLEF languages could boast of.

Another issue that may be relevant to discuss is the kind of task (and therefore evaluation). While morphological analysis is a user-transparent task according to Gaizauskas's bipartition [11], and NER also (though less), information retrieval and question-answering can be assessed by the man in the street, and finding justified answers in Wikipedia could even be done by humans (although I am not claiming that Págico's setup mimicked the way people use Wikipedia), thus user-visible. It is thus apparent that we consistently moved towards user-visibility in Linguateca's evaluation path.

This can be at least partly ascribed to CLEF's influence, not only by allowing us to offer the evaluation of IR tasks for Portuguese, but also because among the many tracks during Linguateca's presence there was one specifically devoted to non-technological issues, the interactive track iCLEF, see e.g. [14].

The other reason that made us not prioritize user-transparent evaluations, like POS-tagging, syntactic analysis or tokenization, is better explained by Gaizauskas [11, page 252]:

> Most user-transparent tasks rest on some theoretical assumptions about the modularization of language processing and about the content of intermediate representations. Since very little theory about language processing is universally shared, finding a community which shares these assumptions about modularization and, if so, shares assumptions about the informational content of the representations the intermediate module consumes or produces, is difficult.

A specific illustration of this problem using Portuguese syntactic analysis can be found in [47], and although Gaizaskas is careful not to conclude that user-transparent evaluations should not be organized, it did not seem to us a fruitful path for Portuguese.

8 Concluding remarks

Assessing the organization of evaluation contests by Linguateca in the decade from 2002 to 2012, one can see that we started very near the community – trying to deal with common tasks, and hearing almost everyone that could be interested – but progressively distanced us by looking at challenging problems to which we wanted people to address, as Págico illustrates so clearly.

During those years, Linguateca devoted most of its time to the organization of evaluation contests, while also allowing some of its junior members to participate in the scope of their PhD or other research tasks, provided they made their systems available for the community, as is the case of Esfinge [5], an open-source QA system. The two activities were, needless to say, mutually exclusive.

One aspect that was common to the four venues was the sheer amount of documentation produced, which, for the three concerning only Portuguese, was written in Portuguese. There exist several thousand pages about the tasks and their evaluation in Portuguese, together with a large set of still publicly available resources. Table 3 gives an overview of the main resources and documentation for quick reference.

Table 3 Overview of the evaluation contests organized by Linguateca: the references in bold are in English. All URLs start by `https://www.linguateca.pt/`

Resource	**Main documentation**
`Morfolimpiadas/`	**[45]**, **[39]**, chapters 2-11 of [34]
`primeiroHAREM/harem/ColeccaoDouradaHAREM.zip`, `primeiroHAREM/harem/software/ferramentas_HAREM_perl.tar.gz`, `primeiroHAREM/harem/software/ferramentas_HAREM_java.jar`	**[53]**, **[55]**,[43]
`HAREM/PacoteRecursosSegundoHAREM.zip`	[26], **[46]**
`CHAVE/`	**[52]**, **[42]**, chapter 13 of [34]
`GikiP/`	**[44]**
`GikiCLEF/GIRA`	**[41]**, **[40]**, **[6]**
`Cartola`	[48], **[28]**

These are two advantages that are hard to deny, especially when one acknowledges that this is not in general the case.

But, although we have always made our data, procedures, and programs (or systems) publicly available, and in that respect we have also contributed to increase the resources available for Portuguese, as far as I know there is not much research that uses those data. And, worse still, young students or researchers in NLP of Portuguese tend to go on citing English NLP and not Portuguese, even on areas where there would be a lot to start from, like question answering or NER.

Let me do an easy comparison: the annotated corpora CETENFolha and CETEMPúblico [50], which Linguateca makes available but requires register for download, feature five times more downloads than the CHAVE collection (the collection used in CLEF for Portuguese), which is comparable in terms of textual data[13], and in addition includes a wealth of other material, like questions and answers. And Floresta Sintáctica, the first public treebank for

[13] The source of the news is the same, it is also annotated, and incidentally for Brazilian Portuguese CHAVE includes twice as much text as CETENFolha.

Portuguese [1], has around 50 times more downloads and is probably the most cited and used resource created in collaboration with Linguateca.

So, comparing work in resource creation against the organization of evaluation contests, the time and effort spent in the first task was undeniably more productive in terms of direct impact and outreach.

Still, to our defence it can be said that one might expect different groups to create public resources, but one would need an organization like Linguateca to try and work as glue, as an independent, state-funded, institution to organize an evaluation contest, given that there is a lot – really a lot – of organisational work that does not qualify as research or even development. It would therefore be out of question to expect one R&D group to sacrifice itself for the sake of the others. And, in fact, this was one of the reasons why Linguateca was launched in the first place.

It is also easy to see in hindsight that imperceptibly we were drawn to the IR community due to our enthusiastic participation in CLEF. However, one might argue that the lusophone IR community, with very few exceptions, was not interested in CLEF, and in fact the highest participation we ever got in CLEF was on a more "traditional" NLP task like QA.[14]

Retrospectively, one may conclude that the work in CLEF did not pay off to foster NLP on Portuguese, although it allowed us to look at problems and challenges that were on the center of the European concerns, and include Portuguese there. But the European bias was also a problem, because it alienated many Brazilian groups we might have enrolled in a joint evaluation if instead of yearly CLEF workshops all around Europe we had had some evaluation workshops in Brazil.[15]

If one would start again from scratch, we would probably not have joined CLEF because, although it was an easy infrastructure to get things done, it alienated the "joint" feeling that the participants had in Morfolimpíadas and HAREM: to be together on the decision of what was the goal and the measures. CLEF, with a lot of languages and a broad international committee that soon embraced Asian countries as well, made it almost impossible for participants to have a say. And, as already remarked, it pulled too heavily in the direction of information retrieval, where there was not enough activity – at least in Portugal – to bring more participants on board.

There is no doubt that the greatest success of evaluation contests organized by Linguateca is HAREM, which is still routinely cited today.[16]

But we should have considered more seriously the possibility to do an (additional) MUC-style evaluation, to answer the need to compare the per-

[14] It is an interesting historical detail that the IR community (TREC) attempted the evaluation contest model for question answering rather than other probably more akin areas like parsing or database querying, something which has provoked a lot of debate and discussion at the time. But it is also the reason why it had a natural place in CLEF.

[15] For the record, there were three Brazilian NLP groups which participated in CLEF, and other Brazilian researchers working in European research institutions, but no pure IR groups.

[16] Just a random example: The first BERT word embeddings for Portuguese [56], published in 2020, use HAREM for evaluation.

formance with systems working on English. This would probably not only make the differences between the two evaluations much clearer, but might have helped HAREM to be used as the definitive measuring rod for Portuguese internationally. Because, let us be frank, only people from Brazil or Portugal cite HAREM. Non-lusophone researchers could not care less, although for all evaluation contests organized by Linguateca we duly published in English in conferences.

Anyway, and to conclude on a positive note, I believe that the number of systems involved in Linguateca's evaluation activities, and the number of people who participated and/or organized them, is significant: all in all, 37 different systems participated and 84 different people signed papers dealing specifically with these evaluation contests (in what concerns CLEF, only people and systems coming from lusophone countries were counted).

It is my belief that the work done, the discussions and papers, what was learned about Portuguese in the process, and especially the still available resources compiled, continue to be of value. This is the main reason why this paper was written, to allow newcomers to the field of Portuguese NLP to use the knowledge and data amassed, while also providing a more general view of Linguateca's role to an international audience.

Acknowledgements I am extremely grateful to my co-organizers and to the whole community who participated and contributed to the progress of the application areas in question. Specifically to Paulo Rocha, Luís Costa and Anabela Barreiro (Morfolimpíadas); to Nuno Cardoso, Nuno Seco, Rui Vilela, Anabela Barreiro, Paulo Rocha and Susana Afonso (First HAREM); to Cristina Mota, Hugo Gonçalo Oliveira, Cláudia Freitas and Paula Carvalho (Second HAREM); to Paulo Rocha (adhoc CLEF, QA@CLEF and GeoCLEF), to Luís Miguel Cabral (GikiP and GikiCLEF) and to Cristina Mota, Alberto Simões, Luís Costa and Cláudia Freitas (Págico).

I should also thank heartily the CLEF community for the amazing cooperation and intellectual stimulation in the years I belonged to it, and for the true sense of community. Although I may have reflected negatively on CLEF and Portuguese from a political point of view, CLEF was (and I suppose still is) a marvelous endeavour.

I acknowledge Linguateca's funding by the Portuguese Fundação para a Ciência e Tecnologia through project POSI/PLP/43931/2001, co-financed by POSI, and project POSC 339/1.3/C/NAC, and thank José Mariano Gago and Pedro Veiga for unfailing support during the active funding period of Linguateca.

Also Págico's support by the University of Oslo in 2011-2012 is gratefully acknowledged.

Finally, I gratefully acknowledge the current support of FCCN, Fundação para a Computação Científica Nacional, for hosting Linguateca's servers and the resources mentioned in this paper.

References

1. Afonso, S., Bick, E., Haber, R., Santos, D.: Floresta sintá(c)tica: a treebank for Portuguese. In: M.G. Rodrigues, C.P.S. Araujo (eds.) Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), pp. 1698–1703. ELRA, Paris (2002)
2. Aires, R., Aluísio, S., Quaresma, P., Santos, D., Silva, M.J.: An initial proposal for cooperative evaluation on information retrieval in Portuguese. In: J. Baptista, I. Trancoso,

M. das Graças Volpe Nunes, N.J. Mamede (eds.) Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003. Faro, Portugal, June 2003 (PROPOR 2003), pp. 227–234. Springer Verlag, Berlin/Heidelberg (2003)

3. Cardoso, N.: Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas. Master's thesis, Faculdade de Engenharia da Universidade do Porto (2006)
4. Carvalho, P., Oliveira, H.G.: Apêndice F: Manual de Utilização do Etiquet(H)AREM. In: C. Mota, D. Santos (eds.) Desafios na avaliação conjunta do reconhecimento de entidades mencionadas, pp. 339–346. Linguateca (2008)
5. Costa, L.: Esfinge - A Question Answering System in the Web using the Web. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), pp. 127–130 (2006)
6. Costa, L., Mota, C., Santos, D., Costa, L., Mota, C., Santos, D.: SIGA, a System to Manage Information Retrieval Evaluations. In: Computational processing of the Portuguese language (PROPOR2012), pp. 173–184 (2012)
7. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R.: The automatic content extraction (ace) program: Tasks, data, and evaluation. In: M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, R. Silva (eds.) Proceedings of LREC'2004, Fourth International Conference on Language resources and Evaluation (Lisboa, 26-28 May 2004), pp. 837–840 (2004)
8. Forner, P., Peñas, A., Alegria, I., Forascu, C., Moreau, N., Osenova, P., Prokopidis, P., Rocha, P., Sacaleanu, B., Sutcliffe, R., Sang, E.T.K.: Overview of the CLEF 2008 Multilingual Question Answering Track. In: F. Borri, A. Nardi, C. Peters (eds.) Cross Language Evaluation Forum: Working Notes for the CLEF 2008 Workshop (2008)
9. Forner, P., Peñas, A., Alegria, I., Forascu, C., Moreau, N., Osenova, P., Prokopidis, P., Rocha, P., Sacaleanu, B., Sutcliffe, R., Sang, E.T.K.: Overview of the CLEF 2008 Multilingual Question Answering Track. In: C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. J.F.Jones, M. Kurimo, T. Mandl, A. Peñas, V. Petras (eds.) Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers, pp. 262–295. Springer (2009)
10. Freitas, C., Santos, D., Mota, C., Oliveira, H.G., Carvalho, P.: Detection of relations between named entities: report of a shared task. In: Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, SEW-2009, pp. 129–137 (2009)
11. Gaizauskas, R.: Evaluation in language and speech technology. Computer Speech and Language **12**(4), 249–262 (1998)
12. Gey, F., Larson, R., Sanderson, M., Bischoff, K., Mandl, T., Womser-Hacker, C., Santos, D., Rocha, P., Nunzio, G.M.D., Ferro, N.: GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In: C. Peters, P. Clough, F.C. Gey, J. Karlgren, B. Magnini, D.W. Oard, M. de Rijke, M. Stempfhuber (eds.) Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers, *Lecture Notes in Computer Science*, vol. 4730, pp. 852–876. Springer, Berlin / Heidelberg (2007)
13. Giampiccolo, D., Forner, P., Peñas, A., Ayache, C., Cristea, D., Jijkoun, V., Osenova, P., Rocha, P., Sacaleanu, B., Sutcliffe, R.: Overview of the CLEF 2007 Multilingual Question Answering Track. In: C. Peters, V. Jijkoun, T. Mandl, H. Müller, D.W. Oard, A. Peñas, V. Petras, D. Santos (eds.) Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers, *Lecture Notes in Computer Science*, vol. 5152, pp. 200–236. Springer, Berlin (2008)
14. Gonzalo, J., Karlgren, J., Clough, P.D.: iclef 2006 overview: Searching the flickr WWW photo-sharing repository. In: A. Nardi, C. Peters, J.L.V. González, N. Ferro (eds.) Working Notes for CLEF 2006 Workshop co-located with the 10th European Conference on Digital Libraries (ECDL 2006), Alicante, Spain, September 20-22, 2006 (2006)
15. Hagège, C., Baptista, J., Mamede, N.: Identificação, classificação e normalização de expressões temporais do português: A experiência do Segundo HAREM e o futuro. In: C. Mota, D. Santos (eds.) Desafios na avaliação conjunta do reconhecimento de entidades mencionadas, pp. 33–54. Linguateca (2008)

16. Hagège, C., Baptista, J., Mamede, N.: Portuguese Temporal Expressions Recognition: from TE characterization to an effective TER module implementation. In: The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009) (2009)
17. Harman, D.: The text retrieval conferences (trecs): Providing a test-bed for information retrieval systems. Bulletin of the American Society for Information Science **24**(4), 11–13 (1998)
18. Hirschman, L.: Language Understanding Evaluations: Lessons Learned from MUC and ATIS. In: Proceedings of The First International Conference on Language Resources and Evaluation, LREC'98, vol.1, pp. 117–122 (1998)
19. Hirschman, L.: The evolution of Evaluation: Lessons from the Message Understanding Conferences. Computer Speech and Language **12**(4), 281–305 (1998)
20. Jiang, R., Banchs, R.E., Li, H.: Evaluating and combining named entity recognition systems. In: Proceedings of the Sixth Named Entity Workshop, joint with 54th ACL, Berlin, Germany, August 12, 2016, pp. 21–27 (2016)
21. Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Jijkoun, V., Osenova, P., Peñas, A., Rocha, P., Sacaleanu, B., Sutcliffe, R.: Overview of the CLEF 2006 Multilingual Question Answering Track. In: A. Nardi, C. Peters, J.L. Vicedo (eds.) Cross Language Evaluation Forum: Working Notes for the CLEF 2006 Workshop (CLEF 2006), p. s/pp (2006)
22. Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Jijkoun, V., Osenova, P., Peñas, A., Rocha, P., Sacaleanu, B., Sutcliffe, R.: Overview of the CLEF 2006 Multilingual Question Answering Track. In: C. Peters, P. Clough, F.C. Gey, J. Karlgren, B. Magnini, D.W. Oard, M. de Rijke, M. Stempfhuber (eds.) Evaluation of Multilingual and Multimodal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers, *Lecture Notes in Computer Science*, vol. 4730, pp. 223–256. Springer, Berlin / Heidelberg (2007)
23. Mandl, T., Gey, F., di Nunzio, G., Ferro, N., Sanderson, M., Santos, D., Womser-Hacker, C.: An evaluation resource for Geographical Information Retrieval. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), p. s/pp. European Language Resources Association (ELRA) (2008)
24. Markert, K., Nissim, M.: Towards a corpus annotated for metonymies: the case of location names. In: M.G. Rodríguez, C.P.S. Araujo (eds.) Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation (Las Palmas de Gran Canaria, Spain, 29-31 May 2002), pp. 1385–1392. ELRA (2002)
25. Mota, C.: Resultados págicos: participação, medidas e pontuação. Linguamática **4**(1), 77–91 (2012)
26. Mota, C., Santos, D. (eds.): Desafios na avaliação conjunta do reconhecimento de entidades mencionadas. Linguateca (2008)
27. Mota, C., Santos, D., Ranchhod, E.: Avaliação de reconhecimento de entidades mencionadas: princípio de AREM. In: D. Santos (ed.) Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa, pp. 161–176. IST Press, Lisboa, Portugal (2007)
28. Mota, C., Simões, A., Freitas, C., Costa, L., Santos, D.: Págico: Evaluating Wikipedia-based information retrieval in Portuguese. In: N. Calzolari, K. Choukri, T. Declerck, M.U. Do?an, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (eds.) Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12), pp. 2015–2022 (2012)
29. Peters, C.: The contribution of evaluation. In: F. Gey, C. Peters, N. Kando (eds.) Cross-Language Information Retrieval: A Research Roadmap, Workshop at SIGIR-2002, Tampere, Finland August 15, 2002 (2004)
30. Peters, C., Braschler, M., Choukri, K., Gonzalo, J., Kluck, M.: The future of evaluation for cross-language information retrieval systems. In: M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, R. Silva (eds.) Proceedings of LREC'2004, Fourth International Conference on Language resources and Evaluation, (Lisboa, 26-28 May 2004), pp. 841–844 (2004)
31. Peñas, A., Forner, P., Álvaro Rodrigo, Sutcliffe, R., Forascu, C., Mota, C.: Overview of ResPubliQA 2010: Question Answering Evaluation over European Legislation. In: ResPubliQA - Multilingual Question Answering at CLEF 2010 (QA@CLEF 2010 - ResPubliQA) (2010)

32. Rocha, P., Santos, D.: CLEF: Abrindo a porta à participação internacional em avaliação de RI do português. In: D. Santos (ed.) Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa, pp. 143–158. IST Press, Lisboa, Portugal (2007)
33. Santos, D.: O projecto Processamento Computacional do Português: Balanço e perspectivas. In: M. das Graças Volpe Nunes (ed.) V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000), pp. 105–113. ICMC/USP, São Paulo (2000)
34. Santos, D. (ed.): Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. IST Press, Lisboa, Portugal (2007)
35. Santos, D.: Evaluation in natural language processing (2007). URL `http://www.linguateca.pt/Diana/download/EvaluationESSLLI07.pdf`
36. Santos, D.: Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. Linguamática **1**(1), 25–59 (2009)
37. Santos, D.: Corpora at Linguateca: Vision and Roads Taken. In: T.B. Sardinha, T. de Lurdes São Bento Ferreira (eds.) Working with Portuguese Corpora, pp. 219–236. Bloomsbury (2014)
38. Santos, D.: Para documentar o "ministro da língua". In: C.C. Alves (ed.) Caminhos do Conhecimento: O Legado de José Mariano Gago. Dia Nacional dos Cientistas, pp. 145–161 (2018)
39. Santos, D., Barreiro, A.: On the problems of creating a consensual golden standard of inflected forms in Portuguese. In: M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, R. Silva (eds.) Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004), pp. 483–486 (2004)
40. Santos, D., Cabral, L.M.: GikiCLEF : Expectations and lessons learned. In: C. Peters, G.D. Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Peñas, G. Roda (eds.) Multilingual Information Access Evaluation, VOL I, 1, pp. 212–222. Springer (2010)
41. Santos, D., Cabral, L.M., Forascu, C., Forner, P., Gey, F., Lamm, K., Mandl, T., Osenova, P., Peñas, A., Rodrigo, A., Schulz, J., Skalban, Y., Sang, E.T.K.: GikiCLEF: Cross-cultural issues in multilingual information access. In: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (eds.) Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010), pp. 2346–2353. European Language Resources Association (2010)
42. Santos, D., Cardoso, N.: Portuguese at CLEF 2005: Reflections and Challenges. In: C. Peters (ed.) Cross Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop (CLEF 2005). Centromedia, Wien, Österreich (2005)
43. Santos, D., Cardoso, N. (eds.): Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca, Linguateca (2007)
44. Santos, D., Cardoso, N., Carvalho, P., Dornescu, I., Hartrumpf, S., Leveling, J., Skalban, Y.: GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. In: C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. J.F.Jones, M. Kurimo, T. Mandl, A. Peñas, V. Petras (eds.) Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers, pp. 894–905. Springer (2009)
45. Santos, D., Costa, L., Rocha, P.: Cooperatively evaluating Portuguese morphology. In: J. Baptista, I. Trancoso, M. das Graças Volpe Nunes, N.J. Mamede (eds.) Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003. Faro, Portugal, June 2003 (PROPOR 2003), pp. 259–266. Springer Verlag, Berlin/Heidelberg (2003). (c) Springer-Verlag
46. Santos, D., Freitas, C., Oliveira, H.G., Carvalho, P.: Second HAREM: new challenges and old wisdom. In: A. Teixeira, V.L.S. de Lima, L.C. de Oliveira, P. Quaresma (eds.) Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008), vol. Vol. 5190, pp. 212–215. Springer Verlag (2008)
47. Santos, D., Gasperin, C.: Evaluation of parsed corpora: experiments in user-transparent and user-visible evaluation. In: M.G. Rodrigues, C.P.S. Araujo (eds.) Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), pp. 597–604. ELRA, Paris (2002)

48. Santos, D., Mota, C., Freitas, C., Costa, L. (eds.): Edição especial Págico - português mágico, vol. 4.1. Linguamática (2012)
49. Santos, D., Mota, C., Simões, A., Costa, L., Freitas, C.: Balanço do Págico e perspetivas de futuro. Linguamática **4**(1), 93–99 (2012)
50. Santos, D., Rocha, P.: Evaluating CETEMPúblico, a free resource for Portuguese. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, pp. 442–449 (2001)
51. Santos, D., Rocha, P.: AvalON: uma iniciativa de avaliação conjunta para o português. In: A. Mendes, T. Freitas (eds.) Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística (APL 2002), pp. 693–704. APL, Lisboa (2003)
52. Santos, D., Rocha, P.: The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. In: C. Peters, P. Clough, J. Gonzalo, G.J.F. Jones, M. Kluck, B. Magnini (eds.) Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers, *Lecture Notes in Computer Science*, vol. 3491, pp. 821–832. Springer, Berlin/Heidelberg (2005)
53. Santos, D., Seco, N., Cardoso, N., Vilela, R.: HAREM: An Advanced NER Evaluation Contest for Portuguese. In: N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odjik, D. Tapias (eds.) Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), pp. 1986–1991 (2006)
54. Seco, N.: MUC vs HAREM: a contrastive perspective. In: D. Santos, N. Cardoso (eds.) Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área, pp. 35–41. Linguateca (2007)
55. Seco, N., Santos, D., Vilela, R., Cardoso, N.: A Complex Evaluation Architecture for HAREM. In: R. Vieira, P. Quaresma, M. da Graça Volpes Nunes, N.J. Mamede, C. Oliveira, M.C. Dias (eds.) Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006 (PROPOR'2006), vol. LNAI 3960, pp. 260–263. Springer Verlag, Berlin/Heidelberg (2006)
56. Souza, F., Nogueira, R., Lotufo, R.: Bertimbau: Pretrained bert models for brazilian portuguese. In: R. Cerri, R.C. Prati (eds.) Intelligent Systems. BRACIS 2020, pp. 403–417 (2020)
57. Vallin, A., Magnini, B., Ayache, C., Erbach, G., Peñas, A., de Rijke, M., Rocha, P., Simov, K., Sutcliffe, R.: Overview of the CLEF 2004 Multilingual Question answering track. In: C. Peters, F. Borri (eds.) Cross Language Evaluation Forum: Working Notes for the CLEF 2004 Workshop (CLEF 2004), pp. 281–294. IST-CNR, Pisa, Italy (2004)
58. Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., Peñas, A., de Rijke, M., Sacaleanu, B., Santos, D., Sutcliffe, R.: Overview of the CLEF 2005 Multilingual Question Answering Track. In: C. Peters (ed.) Cross Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop (CLEF 2005). Centromedia, Wien, Österreich (2005)
59. Vilain, M., Burger, J., Aberdeen, J., Connolly, D., Hirschman, L.: A model-theoretic coreference scoring scheme. In: Proceedings of the 6th Message Understanding Conference (MUC-6), pp. 45–52. Morgan Kaufmann, Los Altos, CA, EUA (1995)