Portuguese at CLEF 2005: Reflections and Challenges

Diana Santos[†] and Nuno Cardoso[*]
[†]Linguateca, Oslo node, SINTEF ICT, Norway
[*]Linguateca, Lisbon node, DI-FCUL, Portugal
Diana.Santos at sintef.no, ncardoso at xldb.di.fc.ul.pt

Abstract

In this paper we report the addition of Portuguese to three new tracks in CLEF 2005, namely WebCLEF, GeoCLEF and ImageCLEF, and discuss differences and new features in the adhoc IR and the QA tracks, presenting a new Brazilian collection. Some critical remarks are made concerning the new tracks and the degree of success in adding Portuguese to them, reflecting about meaning and translation issues, as well as familiarity with the culture and users. We document briefly the changes occurred in adhoc and QA, compared to last year's campaign, and end by suggesting some improvements to the QA setup and evaluation practices.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Management**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Evaluation, Information retrieval, Question answering, Questions beyond factoids, Web retrieval, Multilingual retrieval, Geographical information retrieval, Image retrieval, Translation, Portuguese

1 Introduction

To add one more language (and/or culture) to a system or evaluation framework is not just hire a translator and have the job done, although the quest for language-independent systems is still mainstream in natural language processing [1]. This is one of the reasons why Linguateca has taken the role of organizing evaluation contests for systems dealing with Portuguese [2]. In order to evaluate cross-language retrieval, however, the obvious venue is CLEF. To have Portuguese as one of the languages which the systems must process, query and/or retrieve is undoubtedly beneficial to the processing of Portuguese language in general [3, 4].

So, for this year's campaign, we once again participated in the organization of the adhoc and the question answering tracks by suggesting topics and questions, and by evaluating the results in the Portuguese collections. Differences from last year's campaign and some suggestions for improvement will be offered in Sections 5 and 6 below. In addition, we provided translation into Portuguese of both image captions for ImageCLEF and geographical topics for GeoCLEF, and

produced some Portuguese topics to be used in WebCLEF. This very slight participation and the reflections it triggered will be the subject of Sections 2 to 4.

This paper has two kinds of readers in mind: both the present and future participants dealing with Portuguese, for whom we attempt to document what was done and what in our opinion is missing, and the other organizers and participants in CLEF in general, to whom we wish to provide critical feedback.[1]

One thought can, however, be stated at once. Our experience this year at CLEF 2005 reinforced what will be a recurrent idea through the paper: 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.

2 Portuguese at WebCLEF

We start by the most striking illustration of the advantage of knowing well the material – in this case the Portuguese Web. No matter the fact that it is the first time the WebCLEF track is run [5], we believe it could have been significantly improved if people with a working knowledge of each language (and Web) had been involved.

The Portuguese collection included in the EuroGOV[2] collection [6] is very weak indeed: near 70% of the 147,445 documents of EuroGOV under the `.pt` domain, henceforth EuroGOV-PT, belong to a single site, `www.portaldocidadao.pt`. This site is basically a hub of links to .gov.pt pages, and we suspect that the crawler was 'trapped' inside it. The remaining 30% of the crawl suffers from what we can call the virtual host problem, featuring 8,744 duplicated pages, almost 6% of the whole (EuroGOV-PT) collection.

It seems like the crawler did not normalize the URLs[3] as it collected links from the crawled pages, given cases as the one displayed in the example below: the document shown was harvested 153 times, due to different URLs with relative paths included in the string. Here we display some of the duplicate URLs:

```
http://www.mdn.gov.pt/../../Comunicados/../primeira.asp
http://www.mdn.gov.pt/../../Defesa/Legislacao/../../primeira.asp
http://www.mdn.gov.pt/../../destaque/../Links/../primeira.asp
http://www.mdn.gov.pt/../../fundos_ecran/../Glossario/../primeira.asp
http://www.mdn.gov.pt/../../Publicacoes/../primeira.asp
http://www.mdn.gov.pt/../../defesa/Historia/../../primeira.asp
```

In fact, EugoGOV-PT reflects some of the symptoms of the Portuguese Web found and reported on a previous characterization by Gomes and Silva [7], who report that the Portuguese Web is weakly inter linked, has few pages rich in links (or 'hubs'), and has plenty of virtual hosts, which are responsible for a considerable number of duplicated documents. They also claim that only 73% of the Portuguese Web was written in Portuguese. Among the documents, one could find PGP keys, protein sequences, source code of programming languages, music tablatures and all sorts of different content types, making it very hard to filter them all, as Thelwall [8] pointed out for English webpages.

In 2004, the XLDB Group released a crawl of the entire Portuguese Web in the spring of 2003, called the WPT 03[4] [10, 9], in the scope of the *tumba!* project, a Web search engine for the

[1] At the time of writing the present paper, we only have full overview about participation and results of the QA track, due to different organization strategies, so comments concerning overall participation using in some way Portuguese will have to await another occasion.

[2] http://ilps.science.uva.nl/webclef/collection.html

[3] By this we mean taking into consideration that different navigation patterns in the same site can lead to the same absolute location. See e.g. http://www.ietf.org/rfc/rfc2396.txt on URI Normalization and Equivalence and Resolving Relative References to Absolute Form.

[4] http://poloxldb.linguateca.pt/index.php?l=WPT_03

Portuguese community [11]. This collection has approximately 3,775,000 documents.[5] *Tumba!*'s team has just finished another crawl, with over 10 million documents – 3 times more content than the WPT 03 in a two year's period – which we plan to make available to the R&D community as well. We will use these two collections to provide some comparison with EuroGOV-PT in what follows.

Even though considering that the EuroGOV collection was supposed to cover only governmental sites, we think the sample was too narrow: there were only 146 different hosts on EuroGOV-PT, 49 of which had less than 10 documents crawled. Table 1 displays EuroGOV-PT's distribution in terms of kind of URLs.

Kind of host (cluster)	Number of pages
.portaldocidadao.pt	102,303
.gov.pt	27,531
.min-*.pt	16,522
.parlamento.pt	784
.presidenciarepublica.pt	283
.moptc.pt	22

Table 1: Distribution of pages in EuroGOV-PT. These (regular expressions over) URLs are what we call EuroGOV-PT base.

We object to the preference given to the `www.portaldocidadao.pt` hub, which was probably the reason to leave behind many other relevant documents. Only 11 (5.7%) of EuroGOV hosts in the PT domain are not named `*gov.pt` or `min-*.pt`, from now on called "government pages". However, due to the aforementioned hub, government pages cover less than 30% of the pages.

A cursory investigation of WPT 03 displays 63,106 government pages, against 85,772 of the latest crawl, corresponding respectively to 189 and 344 hosts. So, we estimate that half of present-day government hosts are absent from EuroGOV-PT. Table 2 gives an overview of the coverage in the three collections.

	WPT03	Current crawl	EuroGOV-PT
Hosts	54,709	106,841	146
Government hosts	189	344	135
Pages	3,775,611	10,273,292	147,445
Government pages	63,106	85,772	44,053
Pages in EuroGOV-PT base	75,057	96,895	147,445

Table 2: Distribution of pages and hosts in different collections

Looking at Table 2, it may seem surprising that EuroGOV-PT has more pages than WPT 03 concerning the EuroGOV-PT base, but a more detailed analysis unveils that the reason is simple: the EuroGOV crawlers harvested almost 15 times more documents from `www.portaldocidadao.pt` than the other referred crawls, whose purpose was to retrieve all documents from the Portuguese web. This is due to *tumba!*'s crawler setup configuration, which limits the crawl up to 8,000 URLs per site, and a URL link depth of 5 levels, to avoid crawler traps, as detailed in [13]. We do not know which strategies were employed by the EuroGOV crawler, but the EuroGOV-PT collection provides clear evidence that crawlers should be aware of and avoid such issues, when harvesting the Web, otherwise the quality of a collection can be significantly degraded .

Such an unbalanced collection made it furthermore quite difficult to come up with interesting topics, even if we assume that the information in the Portuguese government Web is a valid subset

[5] For comparison purposes, let us mention that a Brazilian web collection from 1999, WBR99 [12], had nearly 6 million documents. WBR99 is also publicly available from Linguateca, from http://www.linguateca.pt/Repositorio/WBR-99/.

for navigation of a large enough community of users. So far, preliminary studies of WPT 03 users' query logs[6] lead us to believe that, in Web navigation, user goals like research, on-line purchases and general info about travel, holidays and leisure, in addition to the usual sex and latest news, are much more common than information about public services, but more work has to be done on this subject.[7]

Given that a large crawl of the Portuguese Web was already available and freely distributed for R&D purposes, to restrict the material to official pages only becomes less defensible. In any case, previous acquaintance with the aforementioned collections and papers would have significantly improved the "debut" of WebCLEF.

3 Portuguese at GeoCLEF

Although we volunteered (acknowledgely quite late) to provide genuine geographical topics in the Portuguese collections, in GeoCLEF only the German and English collections were used, so our participation as "organisers" was limited to the translation of the topics (and geographical relations).

This is the first year such a contest is organized, so it could not be expected that all multilingual aspects of the problem would have been dealt with. Still, we feel that our attempt to add Portuguese to this track succedeed in pointing out a few serious weaknesses in it.

As explained elsewhere [15], geotopics were created by selecting "ordinary" location-related topics and adding specific geographical relations. We had either to translate these geotopics – if there was not yet a Portuguese version of them – or at least the added geographical relations.

Our first problem was to make sense of what these relations really meant, and whether they simply boiled down to (Germanic language) preposition usage. If the "relations" were supposed to convey meaning, this would have different implications for translation than if they were simply indicating preposition, in which case other prepositions could have been freely chosen in our rendering, for example when expressing geographical topics in Portuguese.

In fact, even the choice of English "relation" was sometimes problematic for translation. For example, we could not see any way to express the distinction between "in the south of" and "south of", in the sense of a subpart of a larger region versus adjacency or simply relative location. Worse still, the use of "and" and "or" inside geographical relations, instead of stating multiple relations in the definition of a same topic, obscured what exactly was meant. For example, what fine distinction hinges upon "in or around" versus "in and around"?

In a nutshell, a clear semantics for these geotopics was lacking and then, obviously, translation was hampered. We decided to do a literal translation in most of the cases, but were not happy with the resulting "Portuguese" topics.

Still, in some cases we just had to provide a different "semantic relation": while "near" or "nahe" is apparently an idiomatic enough way to locate shark attacks off California, we had to say "on the coast" or at most "near the coast", *nas costas da Austrália e da Califórnia*.

But our troubles did not end with the geographical relations. We actually often disagreed with the scope of what was being located as well. Let us present three different examples. While the original topic required documents about "Amnesty International reports on human rights in Latin America", it got converted into the following trio: `concept:` "Amnesty International Human Rights Reports", `spatial relation:` "in", `location:` "Latin America", which is altogether a different question. Of course, one may claim that the original topics were only a source of inspiration to create new geotopics, but the original user need (reports about human right violations that took place in Latin America) seems to make considerably more sense than the quest for arbitrary AI reports that happen to be (published? referred? criticized?) in Latin America.

[6] Performed by Nuno Seco.

[7] It is interesting to stress that the work reported in [14], based on an extensive study of logs in (Brazilian) Portuguese, has not even considered interaction with public institutions in one of the seven user needs' categories worth to distinguish.

A similar reinterpretation happens in the topic "Vegetable exporters **of Europe**", where we strongly doubt whether the purpose of the original topic refers to countries **inside Europe** or to the economic agents, i.e., the European countries as political objects.[8] Basically, our reading of the new geotopic asks for vegetable exporters in Europe into Europe, and not of Europe into whatever other location (which was the purpose of the original topic).

Even worse is the topic of "environmental concerns in and around the Scottish Trossachs", which refers to something which is not location restricted, but the subject of concerns (who can be had in London or Oslo, or wherever people have environmental concerns at all). Likewise, it is bold to state that all "factors influencing tourist industry in Scottish Highlands" are actually located in the Scottish Highlands themselves, as the new geotopic has it. When abstract concepts like concerns or influence are at stake, it is usually a bad idea to interpret them as happening at specific locations.

By the way, note that even some of the original topics, if one required a clear separation of location from concept, could be claimed ambiguous or not well formulated enough. Take "Actions against the fur industry in Europe and the USA". This topic can, of course, be satisfied by documents referring to actions taking place in Europe or in the USA but, if the location prepositional phrase were attached to "the fur industry", actions in Japan against European or American industry would qualify as equally valid for the original topic.

After this exercise, we did not expect any Portuguese-aware participating system to fare well, since the topics were obviously geered to a Germanic view of locations, and most of our translations were probably not what Portuguese location-aware systems had been aiming for. As common sense would have it, a system interested in making sense of locations in Portuguese would have to start with a more Portuguese-like way to convey those. But to be able to question location-indexed news in Germanic languages is an interesting task in itself, also for Portuguese users. We therefore state our interest in further contributing to future editions of this track.

4 Portuguese at ImageCLEF

As should be expected, images are not very language dependent[9], but their descriptions may be. This is something that is not often stated, but which was clear from the simple task we were given at ImageCLEF [16]: translate the English captions into Portuguese, and/or provide a satisfactory description of the images in Portuguese.

First of all, most images are not self explanatory. And translation will not help if you do not know the subject. This was obvious for pictures like: "golfer putting on green" or "colour pictures of woodland scenes around St Andrews". Now, if one does not know the rules of golf, it is doubtful whether one is able to select the right images. Likewise, if one has never been to St Andrews (and admitting, for the sake of the argument, that woodland in St Andrews is not worldwide known for its unique and at once recognisable shapes), we can think of no intelligent system (or human who never saw similar pictures or been to St Andrews, for that matter) that, confronted with woodland pictures taken from many places in the world, could find the ones required. In fact, for this kind of pictures,[10] it is not image retrieval, but geographical retrieval (based on the captions) that one must be aiming at.

Then, other descriptions struck us as too artificial, although we are aware that often the craftmanship of expert librarians will have to be invoked before a collection of pictures is ready for perusal: compared to an art photographer who labelled a picture "woman in white dress", how many would label it instead "Clare during our honeymoon in Crete"? Instead of "dog in sitting position", what about "Timmy, summer holidays, 1990"? And would one prefer "people gathered at bandstand" to "tour to Bragança during high school"? Probably the next step for

[8] Incidentally, a distinction well known in the named entity and information extraction camps.

[9] Even though what people see – and consequently take pictures of – is extremely conditioned by culture.

[10] And there were many: "composite postcards of Northern Ireland", "postcards from Iona, Scotland", "Swiss mountain scenery", "royal visit to Scotland (not Fife)", etc.

a image retrieval system is to infer automatically, from pictures with these kind of more popular and personal labels, the more general classes of descriptions we were given in ImageCLEF.

In any case, and assuming that the English captions we had to translate were not already too much the result of an English conceptualization of the world, their translation into Portuguese – and we are only talking about 28 captions here – also provided us with plenty of interesting comments.

First of all, concrete descriptions of objects are vague, and it is well known that languages cut the semantic pie in different ways, something which is usually the first chapter in any book on introductory semantics. What is rarely discussed or quantized is how often this is the case and how often it brings consequences in real applications. In other words, empirical estimates of these phenomena for different language pairs are hard to find.

But it is remarkable that, in as much as seven cases, namely those mentioning boat, aircraft, ship, "cart or carriage", "church or cathedral", marketplace, and gateway, we had to either add a disjunction or be content with a more general or more specific term.

Then, there are also the well-know differences in attention given to manner between Germanic and Romance languages[11], that caused some relatively awkward – or unnatural – translations. Take "sitting dog" or "building with waving flag". We doubt that anyone would provide such specific descriptions in Portuguese, instead of a simpler "cão" (dog)[12] or "edifício com bandeira" (building with flag).

Also, due to the different lexical meanings of the words employed, "people gathered at bandstand" could cover both political meetings or people just gathered for the picture taking occasion, while the even more general "pessoas junto a um coreto" (people near a bandstand), which was the translation chosen, allows for a number of people who were not really gathered but happened to be in the stand's vicinity. However, using "pessoas **reunidas** ao pé de um coreto" would imply a definite meeting purpose, which we judged too remote from the original intention, except if one were reporting revolutionary times.[13]

Another interesting remark concerns the form of the captions: it struck us that in Portuguese plural descriptions are more natural as a search topic, and so we turned "royal visit" into "royal visits", "tomb" into "tombs" and "viaduct" into "viaducts". However, we may have in some cases inadvertently discarded (or conveyed) a uniqueness presuposition that might be present (or not) in the original caption, such as referring to one particular royal visit, or more than one monument to Robert Burns.

Finally, three descriptions were particularly troublesome: "sun pictures, Scotland", "horse pulling cart or carriage" and "animal statue", because these short descriptions in English may cover more than one situation which would require structurally different translations into Portuguese. A translator would not be surprised, because the shorter/simpler the texts, the harder to translate, but this may be new information for IR researchers.[14]

4.1 The organiser's paradox?

The most interesting reflection, however, brought about by our participation in the ImageCLEF exercise – and which, incidently, applies equally well to GeoCLEF – is that, if one considers state of the art CLIR systems, which use machine translation and bag-of-words approaches, the more elaborate and idiomatic translation we provide, the more we are harming recall, since the more literal the translation, the easier for the systems to get at the right original.

So, we can state a kind of organiser's paradox: the more natural we render a translation into a new language, the more a human user is bound to understand the topic and phrase it that way, but the less a CLIR system (at least the ones existing nowadays) is able to get sensible answers.

[11] See the ample bibliography on this and other subjects concerning different language styles in [17].

[12] We have doubts about whether, in order to select dogs which were not moving, one would ask in Portuguese for "cães parados", or for "cães sentados" or "cães deitados" instead.

[13] This example is not intended to claim that there would not be a better translation, but simply to point out that many factors are generally at stake. In this case, "grupo de pessoas junto a um coreto" (a group of people near a bandstand) would be a better rendering, but we failed to come up with it during translation.

[14] [18] offers substantiation of this claim in the realm of children's books vs. adult fiction.

If we have a really multilingual collection, the well phrased topics or questions may turn out to be a real advantage, at least in their respective language subcollections, but when we attempt to search "language biased" collections, in which only a subset of languages is represented, one could argue that, the more literal the translations, the better, in order to help systems and even users to get at what they want.

In other words, one might predict that the future of CLIR lies in being able to word differently a same topic in the same language, when geered to different target language collections, that is, developing a kind of intentional translationese. Of course, in an ideal world, the (machine) translation system would accomodate the differences toward each target language separately, but in the real world even a human user with some knowledge of the target language would be tempted to help the system, if he realised that different renderings (just like different styles) would provide significantly better performance.

5 Portuguese at adhoc CLEF

Given the addition of new languages with newer collections, see [19], topics for this year's adhoc track had by necessity to be more restrictive, since they would have to feature hits both in 1994-1995 (for Portuguese, French and English) and in 2002 (for Hungarian and Bulgarian). Also, ideally one should provide some coverage of Brazilian news in addition to Portuguese ones, as was already the case with both American and British varieties for English.

In [3], in addition to (or as an alternative to) the central organization directives in 2004 (tripartition in international, European and national topics), we suggested a five-fold classification of topics, namely

1. cyclic events
2. once-only events
3. states of broader events
4. impact measures
5. atemporal subjects

Because of the different collections and dates involved, this year "national" topics were not really sought after, and once-only events had to be discarded, in favour of impact measures, cyclic events and states of broader events. Another category, possibly covered by impact measures but with a more cultural (and named entity flavour) could be identified, namely "same role of a given individual". Examples are topics like "James Bond films" and "Public performances of Liszt", as well as last year's "Films of Kieslowski".

The proliferation of cultures and temporal periods in the collections made paradoxically the set of topics more homogeneous and, in fact, we believe it also made them more realistic – or, at least, easier to interpret as a need for information not too time-specific. In our opinion, this is a good development for the adhoc track, in that the sister competition, QA, is advancing towards more specific and time constrained questions, and is therefore emptying the need to have general IR systems selecting specific news about a very specific event. These could – and should – be satisfied by short, albeit always justified, answers. Recall that, last year, one of the adhoc topics was "Women's ten thousand metres champion". We expect an ordinary user to prefer a name to several documents discussing the championship.

As in last year's campaign, we attempted to have some topics phrased in the Brazilian variety as well as in the one from Portugal, in order to create a competition as much variety-neutral as possible and attract broader participation [3]. We selected the topics to be conveyed in each variety randomly, without no apriori correlation among the variety of the topic and the variety of the document(s) that answer it. Table 3 shows the distribution of answers in the two collections. One can appreciate that both varieties contributed fairly for the Portuguese document pool and for the final results.

Topics	Candidates in Folha	Relevant in Folha	Candidates in Público	Relevant in Público
50	8213	1,035	12,326	1,869

Table 3: Distribution of relevant documents according to Portuguese collection

6 Portuguese at QA@CLEF

Compared with last year's track, the changes in QA@CLEF were few [20], which may either denote that a stable setup has been found and this is a mature track, or that the large number of languages involved (nine) actually brings some inertia and prevents changes.

There were, in any case, two modifications of this track on which we would like to cast a critical look, since we do not believe they are justified: (i) the increase in the number of definition questions (whose exclusion we had advocated in [3]); and (ii) the introduction of temporally restricted questions. Finally, we suggest some improvements for the future of the QA track.

Notwithstanding, we would like to stress that this year not only the participation of systems dealing with Portuguese increased, but their overall performance considerably improved, advocating for some continuity in this track, which proved to be a real incentive for participants.

Definitions Let us start by the "definitions" issue. Although it seemed consensual that no objective way to evaluate answers to this sort of question was available, and therefore they were further restricted to either ask for a profession/role/position of specific named persons, or to require the expansion of organizations' acronyms,[15] still their number in the present campaign was increased from 30 to 50.

This year's "definitions" were much simpler, but still no general evaluation rules were provided. We had to define our own, and in this process some interesting questions arose. In what concerns "definition" questions about people, we assigned a number of information pieces, and evaluated answers as incomplete ("X") if they included any of these pieces (but not all). For example, if the expected correct answer was "prime minister of Finland", both "prime minister" and "Finn" (or "Finland") alone would grant the system an "X", and the same for the three pieces relative to "minister of Education of Nigeria" (minister, Education, Nigeria(n)). The justification for this behaviour was that there could be contexts where just one of the pieces would satisfy the user.

However, a consequence of this course of action was that it was no longer possible to guarantee perfect overlap (or perfect correctness, given the collections) with the golden resource, since the right answers (pieces) could be scattered among different documents. This led to some cases where systems got answers classified with "X", although there stood NIL in the golden collection (since there was no document that provided the full answer).

Another remaining problem with "definition" questions was that there still crept questions whose only correct answer would be the full document (or an extended abstract of it) as in the case of a short biography of Iqbal Masih (in the question "Who is Iqbal Masih?").

Temporal restrictions The problem of the temporally restricted questions (T questions) was that, again, there was no formal definition of what the introduction of this kind of questions was supposed to assess or represent. First of all, there was no distinction between meta temporal restriction (like "temporal location" analogous to GeoCLEF) and factual temporal restriction (inside the text). In fact, the way questions were formulated, in natural language, allowed for the simple strategy that the system searched for questions whose answer included also the temporal restriction, and therefore not too different from more complex (longer) questions in the first place.

On the other hand, a truly temporally restricted question, about past ("Which was the largest Italian party?", meaning "was but no longer is"), was not classified as "T", although it was critically temporally dependent and was, in fact, a genuine cause for complaint from the participants

[15] A case where multiple answers should be provided in a real world setting, by the way.

[21] because of its lack of unambiguous temporal context.[16]

Justifications We believe that for the QA track to develop into something that really evaluates useful systems for real users, justification passages need be required of a QA system, in addition to the short answer, instead of just providing the whole document id.

Pragmatic assessment Furthermore, instead of classifying answers as correct or incorrect according to a predefined correct answer set, we should aim at a more pragmatically valid evaluation, trying to assess things like the following: Is the answer nonsensical, so that any user can discover this at once by consulting the alleged justifying passage? Is the answer incomplete but useful? Is the answer complete and right but not supported? Is the answer wrong but (at least apparently) supported?[17] Is the answer informative enough to lead to follow-up or reformulation questions from an interested user? In [20], we provide a first attempt towards this goal by dividing questions into rubbish, uninformative (empty), and dangerous questions. We suggest here that this way of reporting results be further developed in future editions.

Question difficulty Finally, we believe that a real assessment of the difficulty of the questions (given the collections) should be attempted. Although the decision of not to provide NIL questions that were trivial to uncover was a real improvement of this year's track, we were still forced to re-assess our golden answer set for three different questions, which had been assumed not to have answers in the collection, and which different systems, with different strategies, were able to counterproof and actually find a satisfactory answer.

Some criteria for ranking QA pairs according to difficulty could be: (a) literal answers, (b) answers in the same sentence (or clause) but with a wording different from the question, (c) answers in separate sentences, (d) answers requiring some reasoning from a human (although not necessarily from a system).

7 Brazilian Portuguese at CLEF

As already mentioned above, this year a new Portuguese collection was added, containing all editions of the Brazilian newspaper Folha de São Paulo in 1994-1995. We present here the size of the present collection for documentation purposes. More information can be found at the Portuguese CLEF site.[18]

Origin	Público	Folha de São Paulo
Documents	106,821	103,913
Size (kB)	348,078	226,690
Tokens	64,573,983	42,317,112
Types	605,092	530,382
Word tokens	55,538,483	35,907,591
Word types	392,999	497,798

Table 4: The Portuguese collections at CLEF 2005. "Tokens" include as well punctuation marks, numbers, email addresses, etc, while "Word tokens" only counts words. "Types", as usual, counts one token/word once, no matter how many times it occurrs.

[16] In Portuguese, the use of the Imperfeito tense pressuposes that there is an understood temporal period in mind; if Perfeito had been used, it would convey either a strong pressuposition that there had been only one largest Italian party in the past, or that a list of all parties having had that role once was being asked for.

[17] An interesting example, due to Luís Costa (p.c.), is last year's Esfinge answer to the question "What country is the world football champion?", where the indoor soccer ("futebol de salão") winner was named.

[18] http://www.linguateca.pt/CLEF/

Although it is disappointing to acknowledge that still no Brazilian participants turned up in CLEF, the existence of this new collection is obviously an asset and some of its consequences for CLEF 2005 can already be discussed.

For example, the same QA participants of last year came back, together with a newcomer, as can be verified in the QA track overview [20]. Precisely due to the way their systems are designed, the addition of a Brazilian collection allowed us to predict an improvement of the performance of one system while it would make things harder for the other. In fact, while Costa [22] mentioned that the absence of Brazilian Portuguese in the collection caused problems for the Esfinge system to justify the answers found on the Web, Quaresma and his colleagues [23] stated that one difficulty for their system was to obtain a general enough thesaurus to find answers in unrestricted text, which is arguably more difficult if equal treatment of the two main varieties of Portuguese is required. An interesting subject is the distribution of answers in the two collections, on which we plan to report later on.

Acknowledgements The organization and translation work in CLEF 2005 described in this paper was performed mainly by Paulo Rocha, supervised by the first author. Raquel Marchi helped with the Brazilian version of the topics. Evaluation was in addition performed by Diana Santos, Paulo Rocha, Luís Costa, Débora Oliveira, Rui Vilela and Alberto Simões.

We are grateful to Daniel Gomes for providing the figures about the new crawl of the Portuguese Web and to Luís Costa for valuable comments on previous versions.

We thank Público and Folha de São Paulo for allowing us to use their material, and respectively José Vítor Malheiros and Carlos Henrique Kauffmann for making this practically possible.

We acknowledge grant POSI/PLP/43931/2001 from the Portuguese Fundação para a Ciência e Tecnologia, co-financed by POSI.

References

[1] Santos, Diana: Toward Language-specific Applications. Machine Translation Vol.14 (1999) 83–112.

[2] Santos, Diana (ed.): Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa, in print.

[3] Santos, Diana; Rocha, Paulo: The Key to the First CLEF with Portuguese: Topics, Questions and Answers in CHAVE. In: Carol Peters and Francesca Borri (eds.), Cross Language Evaluation Forum: Working Notes for the CLEF 2004 Workshop (CLEF 2004), Bath, UK, 15-17 September 2004. 2004. IST-CNR, pp. 639–648.

[4] Rocha, Paulo; Santos, Diana: CLEF: Abrindo a porta à participação internacional em RI do português. In: [2].

[5] Sigurbjornsson, Borkur; Kamps, Jaap; de Rijke, Maarten: Overview of WebCLEF2005. This volume.

[6] Sigurbjornsson, Borkur; Kamps, Jaap; de Rijke, Maarten: EuroGOV: Engineering a Multilingual Web Corpus. This volume.

[7] Gomes, Daniel; Silva, Mário J.: Collecting Statistics about the Portuguese Web FCUL Technical Report DI/FCUL TR 03-10. June 2003

[8] Thelwall, Mike: Text Characteristics of English Language University Web Sites. Journal of the American Society for Information Science and Technology. May 2004,

[9] Martins, Bruno; Silva, Mário J.: A Statistical Study of the Tumba! Corpus. DI/FCUL TR 4-4, May 2004.

[10] Cardoso, Nuno; Martins, Bruno; Gomes, Daniel; Silva, Mário J. : WPT 03 - recolha da Web portuguesa. In: [2].

[11] Silva, Mário J.: The Case for a Portuguese Web Search Engine. Proceedings of the IADIS International Conference WWW/Internet 2003, ICWI 2003. IADIS, Algarve, Portugal, 5-8 Novembro 2003, pp. 411–418.

[12] Calado, Pável: The WBR-99 Collection: Description of the WBR-99 Web collection data-structures and file format. LATIN - Laboratório para o Tratamento de Informação, Departamento de Computação, Universidade Federal de Minas Gerais, Brasil.

[13] Gomes, Daniel; Silva, Mário J.: Characterizing a National Community Web. ACM Transactions on Internet Technology. In press.

[14] Aires, Rachel; Aluísio, Sandra; Santos, Diana: User-aware page classification in a search engine. In: Proceedings of Stylistic Analysis Of Text For Information Access, SIGIR 2005 Workshop (Salvador, Bahia, Brazil, August 19, 2005).

[15] Gey et al. Overview of GeoCLEF 2005. This volume.

[16] Clough et al. Overview of ImageCLEF 2005. This volume.

[17] Santos, Diana: Translation-based corpus studies: Contrasting Portuguese and English tense and aspect systems. 2004. Amsterdam/New York, NY: Rodopi.

[18] Santos, Diana: O tradutês na literatura infantil traduzida em Portugal. Actas do XIII Encontro da Associação Portuguesa de Linguística (Lisboa, 1-3 de Outubro de 1997), pp. 259-74.

[19] Peters, Carol: Overview of the CLEF 2005 AdHoc Track. This volume.

[20] Vallin, Alessandro et al.: Overview of the CLEF 2005 Multilingual Question Answering Track. This volume.

[21] Costa, Luís: 20th century Esfinge (Sphinx) solving the riddles in CLEF 2005. This volume.

[22] Costa, Luís: First evaluation of Esfinge, a question-answering system for Portuguese. In: Carol Peters and Francesca Borri (eds.) Cross Language Evaluation Forum: Working Notes for the CLEF 2004 Workshop (CLEF 2004), Bath, UK, 15-17 September 2004, 2004. IST-CNR, pp. 393-402.

[23] Quaresma, Paulo; Quintano, Luís; Rodrigues, Irene; Saias, José; Salgueiro, Pedro: The UE approach to QA@CLEF-2004. In: Carol Peters and Francesca Borri (eds.). Cross Language Evaluation Forum: Working Notes for the CLEF 2004 Workshop (CLEF 2004), Bath, UK, 15-17 September 2004. 2004. IST-CNR, pp. 403-411.