CHAVE: topics and questions on the Portuguese participation in CLEF

Diana Santos[1], Paulo Rocha[2]

Abstract

In this paper we report the work done by Linguateca in order to add Portuguese to two tracks of CLEF, namely the adhoc IR task and the QA task. We start by a brief description of Linguateca's aims and the way we see CLEF from the standpoint of Portuguese language processing; we then document several interesting problems posed by the tasks we added to organize, and offer some suggestions for improvement, as well as raise some possibly controversial points for discussion. Finally, we describe briefly the creation of the Portuguese collection and the process of checking answers in it.

1 Linguateca's role

Linguateca [1] originated from the realization that there were too little resources for the processing of Portuguese, and that the large resource centres such as LDC or ELRA could not take a primary role in the deployment of such resources, given world-wide priorities. In addition, there was little sense of community among practitioners of Portuguese language processing ("PLP"): The PLP community had scarcely met. Groups were not only scattered around Portugal and Brazil, but they dwelled in different departments with different practices – linguistics, IR, AI, NLP… There was no tradition of sharing results and comparing approaches. Therefore, Linguateca's main aims (which we call the IRE model) are to inform, create and disseminate resources and promote evaluation contests (or campaigns) dealing with Portuguese. See our site for a comprehensive description of our activity since 1999.

Linguateca therefore concentrates on Portuguese. We believe that, to improve PLP, one must start by studying the Portuguese language and comparing the state of the art of tools developed for Portuguese, in tasks that deal with Portuguese, evaluated by native speakers of Portuguese. See Santos (1999) [2] for a specific defense of this position, and also the websites of the two workshops organized on Portuguese evaluation [3]. This is why we have created resources for Portuguese, such as the large annotated corpus CETEMPúblico [4], and the Floresta Sintá(c)tica treebank [5], and have organized evaluation contests dealing with Portuguese only [6-7].

From this point of view, it may be somewhat puzzling that we joined the most international of all evaluation contests (in what concerns at least the number of different languages and of participants from different countries), namely CLEF. There are, however, many explanations for this:

First, one primary aim of CLEF is fostering crosslingual information retrieval. Therefore, in order to evaluate querying of a multilingual collection in Portuguese, or querying of a Portuguese collection in any other language, CLEF is the place to go. In fact, instead of copying or adapting something for another language to deal with Portuguese, we are adding Portuguese so that people primarily concerned with other languages may turn to process Portuguese as well. Even though remote, this is included in Linguateca's primary goal: our users are developers of PLP tools. Although one expects that they are mainly Portuguese or Brazilian, international ones are obviously most welcome.

Second, we have publically voiced the opinion that an all-Portuguese organization would give more weight to Portuguese-specific matters and more influence to participants dealing with Portuguese, and therefore one should start by Portuguese-only evaluation contests (see [8]). Still, we must weight this against a huge relief in organizational matters, which are already taken by the central organization of CLEF (and Linguateca has very little human resources).

Third, for some monolingual tasks, such as Q&A, there would not be enough participants to organize one evaluation contest on its own (only two groups competed), while QA@CLEF added Portuguese without hopefully too much bother, and, in fact, other languages (e.g., German) had a similar participation rate. So, joining a circle of international experts in order to circumscribe and make clear a given task is a sensible way to start, even if we may disagree or might have done it differently on our own.

Finally, the only unbiased way to assess whether it was worthwhile for the Portuguese community to participate in CLEF was to try it out.

[1] Linguateca, Oslo node, SINTEF IC, Norway (Diana.Santos@sintef.no)
[2] Linguateca, Braga node, Universidade do Minho (Paulo.Rocha@alfa.di.uminho.pt)

Having now participated in the organization of the Portuguese part, we think it was worthwhile to do it: it led to building resources that otherwise would have taken us much longer to build, and we learned a lot from working together with the organizers for the other languages.

The only thing we are not yet happy with, are the details and schedule of releasing the whole resource, which – had it only depended on us –, would have been made public long ago. Although we have the publisher's (PÚBLICO) consent to make the collection available on our own at once, we would rather release it together with the questions and answers, and with the topics, and this has to wait (at least) until the CLEF worskhop takes place. We hope it will not take too long after that. In case it does, we intend to make available (at least) the Q&A data created by us with the text collection. As far as the topics are concerned, since all topics were discussed together, we simply cannot do this unilaterally.

2 Tasks

The two tracks where Portuguese was included were the Monolingual (non-English)/Bilingual/Multilingual Information Retrieval (also called "adhoc") task and the QA task.

At first, it surprised us to see how little communication between the organization of these two tasks there was: different coordinators, different schedules, different sites, different basic decisions (e.g. how to identify each language), different encoding, even different collections... but then it became also clear that the communities involved were different: no matter the fact that Q&A is conceptually a more advanced IR task, the groups and systems that competed (at least for Portuguese) were disjoint.

Also the workload involved was different: while for IR the topic creation and discussion was relatively light, the evaluation of the results was demanding. Contrarywise, the preparation and translation of the questions and answers, as well as finding justifications for them, which we needed to do for QA@CLEF, was much work, while evaluation was light and even intellectually rewarding.

We detail what was our participation for each of the tasks in the following sections. We avoid gory details and lengthy description of issues whose main interest lies with Portuguese speakers (for this you may see [9]), but hope to give an interesting summing up of our difficulties and remaining puzzles, as well as some guidance to newcomers to the (CLEF) field.

2.1 IR topic preparation

The main issue was to come up with information needs that could be both representative of natural topics of interest for a Portuguese speaker, and relevant for an international (European) observer as well.

2.1.1 International vs. national

According to the IR coordinator's directives, a tripartite division should be aimed at: one third should cover international events (the world at large), another cover European news, and another third, language or country specific subjects. This was a rule of thumb for suggesting initial candidates; then, all topics were checked by all language groups and a final common decision was taken, based mainly on coverage in different collections. In fact, it is interesting to assess how the distribution of the final topics appears from each language standpoint.

In fact, as regards "internationality", it is not always clear whether some events are world-wide, European or just Portuguese (in fact, this does not depend on the event itself, but on its media coverage). It was quite interesting to check other groups' topics as well as learn about the relative importance of the Portuguese topics we expected to be reported elsewhere. There are studies on the relative impact of the Romance languages in the web as a whole (see e.g. [10]), and we suggest to do something related, in order to address, and measure, for each foreign collection, the degree of "Portugueseness" expectable. Unfortunately, we did not have access to the collections in the other languages at the time of topic preparation, so we can just hope to be able to do this later on.

2.1.2 Another classification of topics

Let us suggest here, however, another classification: One might also divide topics among: cyclic events; once-only events; states of broader events; impact measures; and atemporal subjects. Examples of the latter kinds[3] follow:

[3] For lack of space, we do not present the whole topics but just their titles. The readers must trust us that the topics we mention (and that were either CLEF 2004 topics or suggested topics) are of the kind we describe, although obviously most of these title names in isolation could describe totally different information needs.

As to states (or sub-events) of broader events, "East Timor guerilla" or "civil war in Rwanda" can be considered "states" of a larger war. The same concerns the subject of "Fight against AIDS in Africa" or "Russian-Finnish relations" (hopefully both spanning a much larger period than 1995 alone).

Impact measures can be illustrated by topics such as: "Tourism information on the internet"; "Music in digital form"; "Prevention of human rabies in France; "EU and the price of food". Basically, for this kind of topic, we are interested in how do these subjects fare in news coverage in 1995. In fact, in some of these cases the topic may have been raised by specific matters taken up in the (local?) press, but it is obviously possible for a user to want to know about these topics in collections that cover other years.

Atemporal subjects are exemplified by: "Dam building", "the deaf and society", "domestic fires" and "Iranian cinema" or "seal-fishing". One may argue that the latter can also be interpreted as states of a larger event (e.g. the whole history of Iranian cinema), or impact measures, i.e., one is looking for events concerning seal-fishing (like laws and debates) occurring in 1995. Still, we believe that one may be interested in knowing about seal-fishing or deaf people in society without a temporal grounding, while "EU and the price of food" seems to make sense only at a particular time.

In any case, the purpose of this classification is to consider carefully whether these different kinds of topics, which we argue reflect different users' needs, and possibly even different kinds of Q&A applications, should (or not) have different evaluation practices, and consequently also different forms of description (and narrative).

Another issue we believe more attention should be given to are topics with different answers in different collections. To us, they seem the cases where CLIR and MLIR make the most sense from an arbitrary user's point of view: Situations in which the addition of results provides genuinely more information. Apparently there were not many of these topics in this year's campaign[4], but the (related) QA@CLEF exercise did provide good examples, in the cases of "Name some X", with X "person charged of paedophilia", or questions about organizations like "what is the masonry?", in which different facets – and facts – about this organization in different countries might be uncovered.

In fact, and although, as mentioned above, the organization of the QA track and of the monolingual IR track were completely separate, we believe that many of the remarks above could apply to both. After all, QA can be seen as a request for more precision about a topic, and some topics were even stated as questions. Furthermore, Magnini *et al*. [11] report that the original set of questions used in QA@CLEF 2003 was inspired by the topics of that year's IR competition. Given that we prepared the material for both, we could not help comparing and thinking about both tasks as conceptually the same, despite testing different systems.

2.1.3 Distributionf of information among the topic

Another matter relevant to discuss is the way we worded, or attempted to word, the topics, distributing the text by names, description and narrative in Portuguese.

We tried hard to use as many synonyms and wording variants as possible, so that systems would get (almost) a synonym-expansion capability for free, if they used all content words as keywords. For example, in topic C249 below, *dez mil metros* and *10.000 m* are variant ways of stating "ten thousand meters" in Portuguese. And *campeã* (champion), *vencedora* (winner) and *venceu* (won) are closely related, but different ways of expressing the concept at stake.

```
<num> C249 </num>
<PT-title> Campeã dos 10.000 metros femininos </PT-title>
<PT-desc> Quem venceu os 10.000 metros femininos nos Mundiais de Atletismo
em Gotemburgo? </PT-desc>
<PT-narr> Documentos relevantes devem nomear a vencedora da final dos dez
mil metros nos Mundiais de Atletismo em Gotemburgo. </PT-narr>
```

Of course, we had no idea of which systems would compete and how they would use the information we provided. Since this was the first time Portuguese was included, however, we thought all (lawful) ways to

[4] "Sports women and doping", "the sales of the "Sophie's world" book" and "change of sex operations" are possible ones, but "multibillionaires", although apparently possible to find everywhere, are not evenly distributed (at least in the European countries participating in CLEF). Also, no matter their apparent "general" character, atemporal subjects are not necessarily also a-locational: "Seal-fishing", and "avalanche disasters" are not often discussed in Portuguese media, for geographical reasons, and the same applies to "bowling" or "haunted buildings", suggested respectively by the Finnish and British teams. These seem to be, by cultural reasons, simply uninteresting subjects for Portuguese readers of newspapers.

help participants should be used. At the time of writing the present paper, we do not (yet) know whether this capability was of any help for the actual participating systems.

2.2 Q&A preparation

Let us now report on the work and challenges of preparing the resources for the Q&A track, organized by Alessandro Vallin and Bernardo Magnini, see [12]. It should be noted that the opinions stated here are just our own, and do not intend to represent those of the QA organization as a whole.

Very briefly, for QA@CLEF we had to create 100 natural Portuguese questions with answers, indicate and associated document where the answers could be found; translate them into English; and then translate 600 other questions (with answers) from English (and/or from the original language) into Portuguese. Furthermore, for 100 of those we had to check the answer in our collection and provide it. This was hard work. Evaluating the results, however, was relatively straightforward (just in very few cases the whole document had to be read in order to decide whether it justified the answer or not).

The main challenges in our work were:

1) For our questions
   a) Coming up with a set of natural questions, not too difficult, with a straightforward answer
   b) Identifying clearly that answer: basically, all sets of plausible answers that we could find in our collection
   c) Providing a natural English translation with (if possible) the same presuppositions of the Portuguese one
2) For the questions coming from other groups (which we had in English and in the original language)
   a) Translating the question into Portuguese so that it made (some) sense to a Portuguese speaker
   b) Translating the answers as close as possible to the answers found in our collection, in case the question was answered, and adding other answers (either more correct in case a wrong answer had been supplied, or more Portuguese-like as regards measures or spelling).
   c) In case no answer could be found in our collection, trying to provide suitable translations of the answers as well as of the questions (as will be seen below, sometimes this was impossible)

Interestingly, neither of these subtasks was straightforward, as we try to illustrate in what follows.

2.2.1 What is a natural question?

What is a "natural" question is something that eludes a precise definition, and has been discussed often in the contest of QA systems. In general, the solution is to stick to a particular user's model.

We just mention here a few cases that we have not seen documented elsewhere. If a given role is occupied by a woman, should the natural question be in the feminine or in the masculine (neutral in Portuguese if you don't know the gender)?[5] We decided to use the easier kind in our set, as shown in question 337 (*ministra* is the feminine form).

```
337 F PERSON Quem é a ministra do Ambiente alemã?
   1            SEARCH[Angela Merkel]
```

Another cpmcern was the following: Should one use the informal way of posing questions in (European) Portuguese, or suppose that normal users of a Q&A system will not use it, given that it implies more typing? We tried to address this issue by using both ways. So, some questions were provided featuring the emphatic "é que", and others not.

```
612 F LOCATION   Onde é que nasceu Álvaro Cunhal?
   1 LING-941219-076 Coimbra
```

[5] In fact, the feminine form would only be natural if one knew the minister was a woman. This might have occurred if the word "minister" (in the feminine) had been mentioned before, and the user wanted to know who she was. A politically correct asker might use *Quem é o ministro ou ministra do Ambiente?* ("who is the male minister or the female ministre of the environment?") but we strongly doubt such users will ever amount to the majority of Portuguese speakers.

2.2.2 Presuppositions abound

How many presuppositions should be allowed in a natural (as opposed to a tricky) question? In looking for the answer to question 327, we found out that Deng Xiao Ping has not one, but two daughters. Apparently, therefore, this question was ill-posed.

```
327 F PERSON Como se chama a filha do líder chinês Deng Xiaoping?
   1            SEARCH[Deng Rong]
   2 940217-004 Deng Nan
```

In general, anyway, most questions presuppose that it is possible to share the referent with the reader. See 2.2.4 below for definitions.

2.2.3 Classification of questions

In addition to provide the questions, we had to classify them in a particular grid, according to the track instructions [12]. In our opinion, the kind of question required is not very meaningful. Although it should apparently be determined by the semantic category of the right answer (person, location, manner, object, measure…), this had little correlation with the linguistic properties of both question and answer.

In fact, the next questions are precisely after the same kind of information (What is X's nationality? and Which country is X from?), but have been classified differently.

```
558 F OTHER Qual a nacionalidade do tenista Sergi Bruguera?
   1            SEARCH[espanhol]
582 F LOCATION De que país é a escritora Taslima Nasreen?
   1            SEARCH[Bangladesh]
```

Also, one might argue that, although the next question asks for a postal address, classified as OTHER, an address is ultimately a linguistic specification of a LOCATION, and should thus be classified as such.

```
688 F OTHER   Qual o endereço da Livraria Barata?
   1 LING-940102-050 Av. de Roma, 11
```

Conversely, some questions are classified as MANNER when one is looking for causes. "How does cancer begin?" can be interpreted as what is the cause (or what precedes what). Likewise, the most frequent kind of MANNER questions had to do with cause of death ("How did X die?")[6]. To further prove our point, note that other "manner" questions such as "How is indocyanine angiography performed?" have been rightly translated as "what is…" in a number of different languages, showing that the kind of answer is not a semantic invariant.

Definition questions are, in our opinion, especially tricky. Consider question 693:

```
693 D PERSON Quem é Guilherme da Fonseca?
  1 LING-940127-152 juiz do Tribunal Constitucional
```

This is the same as asking what is his profession, which could then be classified as F OTHER… Incidentally, in order to provide a more accurate answer, one should have said "one (of several) judges". In other words, the undefinite article should not have been left out (see 2.2.5 below).

Therefore, we believe that a more objective way of classifying questions, such as the one presented by Costa in [13], is preferable. Alternatively, one could also classify questions according to the kind of linguistic entity expected as the answer, using thus categories like "proper name", "common name", "toponym", on a par with "definition" (that is a kind of answer, not an object).

2.2.4 More against "definitions"

Definition questions have always the lurking presupposition that there is no one (in the case of persons) or no other organization (in the case of organizations) bearing the same name. This is not always possible to ascertain, and, in fact, while asking about who is Fernando Gomes (the mayor of Porto in 1995) we found out that there was, in the very same collection, a reference to a football player of the same name. We leave for the reader to try out how many organizations called GIA (in all languages covered by QA@CLEF) exist.

In any case, we note that a definition is the most complex question one can ask, and that to give answers to "what is the masonry" or "what is indocyanine angiography?" one needs to be an expert in the field and still

[6] Not all answers to "How did X die?" had to do with cause. One was "In strange conditions". To us, this seems, again, rather difficult to accept as MANNER.

think a lot to come up with an appropriate rendering. Of course, if one is questioning a collection of expert writing (and, especially, didactic material), it may be possible to automatically extract definition-like passages. But, in a newspaper collection, we really doubt that more than `is-a` relations (which are not definitions) can be extracted.

We believe the "definitions" as described by the organization have in addition several other flaws:

1. they often overlap linguistically with factoid questions – F "Who is the pope?" D "Who is João Paulo II?" Often in free text it is difficult to know whether linguistic expressions are attributive or appositive, and in fact in most of the cases both questions (and corresponding answers) make sense
2. "definitions" of a person are in fact requirements for a specific kind information: questions about the most prominent role of a particular person, the one that allows the use of the definite article, or questions about his profession and nationality (in case of artists)
3. "definitions" of organizations are very often elicitations of the full name of something that is conveyed as an acronym, and should be called "expansions", or proper (anaphoric or cataphoric) antecedents

It seems to us, therefore, that considerable attention should be given to the user's (or system's) goal, and to what can sensibly be called a request for definition, as opposed to simply a question of the kind "Who occupies the <role>?" or "What <role, profession> X has?" (which idiomatically is expressed by *Quem é X?*).

2.2.5 Getting correct answers: articles, gender and redundancy

As mentioned above, it is not always obvious what the answer(s) to be claimed as the golden set should be, even if we are the question's authors. Should the answer be grammatically correct? In that case, prepositions are required in most cases, but they have consistently be left out. A more specifically Portuguese case follows:

```
647 F ORGANIZATION   Que empresa tem uma refinaria em Leça da Palmeira?
   1 LING-941223-116 Petrogal
```

In example 647, the proper name could have been preceded by the Portuguese article "a", meaning "the Petrogal". This would however probably confuse too much the other groups for the translation task, and we expected that most participating systems would throw articles away anyway. (That this should not be done lightly is illustrated by the two possible distinct questions "O que são os EUA?" "O que é a EUA"? The first having as right answer *os Estados Unidos da América* (USA) and the second the *European University Association.*)

In example 558 above (see 2.2.3), there are also two ways of answering the question: either *espanhola*, modifying the feminine noun *nacionalidade* (nationality), or *espanhol*, masculine (modifying Sergio Bruguera). We used the second, since this was the form present in our collection.

Another concern as to the proper specification of the golden answer is how much redundancy is acceptable. In the case of the first answer of 443, *volume de negócios* was repeated in order to translate the original answer (which specified "of the group"). In 588, on the other hand, the word *pessoas* ("persons") in the answer about how many employees is quite uninformative.

```
443 F MEASURE Que proporção do seu volume de negócios fez a HP na Europa?
   1            SEARCH[um terço do volume de negócios do grupo]
   2            SEARCH[35 por cento]

588 F MEASURE Quantos empregados tem o grupo Warburg?
   1        SEARCH[4.472 pessoas]
```

2.2.6 Translation is hard: Idiomaticity and presuppositions

Not surprising, not every question we came up with was equally easy to render in English.

In some cases, we simply made up what seemed to us the best translation, like "barefoot diva" for "diva dos pés descalços". We have, however, never seen Cesária Évora being thus described in English newspapers.

In addition, not all presupositions are easy to maintain: consider the possible question "Como se chamava a amante de Mussolini?" which could be appropriately rendered, in English, by "What was Mussolini's mistress called?". If one had used the expression "Mussolini's lover", however, the information that we were looking for a woman would be lost.

On the other hand, since "minister" is gender neutral in English, it would have been advisable, for most questions, not to add gender, thus rendering both *ministra* and *ministro* as "minister".[7]

2.2.7 Translation of ungrounded arbitrary fragments

We have also had problems with the translation of other groups' questions (2a), especially when there were no hits in our collection, or when the question seemed to concern unfamiliar subjects or contexts. In 293 below, we had no idea of which kind of sanctions was mentioned, nor to whom does the indeterminate "se" refer: government? tax authorities? sports club? Likewise, no clue was given as to who is supposed to pay them.

```
293 F MANNER Como se garante a cobrança de sanções?
  1 SEARCH[pelo sistema de notificação de multas através de edictos]
```

Translating the answers that came with the questions was, in fact, in general a major headache, not only because of the reasons already discussed, but because it was not evident why some of the translations (paraphrases) had to be translated at all.[8] Answer 3 in 172, for example, sounds very awkward, while we could give more precise and interesting definitions of AI (if one were after one gold standard with the "right" answers in Portuguese).

```
172 D ORGANIZATION O que é a Amnistia Internacional?
   1        SEARCH[grupo preocupado com os direitos humanos]
   2        SEARCH[organização de direitos humanos sediada em Londres]
   3        SEARCH[organização de direitos dos prisioneiros sediada em
Londres]
   4        SEARCH[um grupo privado de voluntários à escala mundial
dedicado a proteger prisioneiros políticos e outras vítimas de violações
dos direitos humanos]
```

What we tried to do in general, was to match the most similar answer form(s) to the answers in our collection, and put those as "translations", since we did not see the point of doing literal translations that sounded far-fetched.

Still, in many cases (especially in the cases of subjects not treated in the PÚBLICO collection), we had to engage in the translation of answers that did not really feel adequate, like in "Tell me a reason for teenage suicides" "Who are the Simpsons?" (some of the answers), "How can you save energy?", "How do they plan to carry out family planning in Peru?", "What does the company Victorinox produce?". Example 480 shows how little informative, and possibly even erroneously translated, can be the translations eventually provided.

```
480 F OBJECT Que produz a MCC?
   1            SEARCH[o automóvel Micro Compact Car]
   2            SEARCH[o "carro urbano do futuro" de dois lugares]
   3            SEARCH[o carro compacto Smart]
   4            SEARCH[veículos]
   5            SEARCH[Swatchmobile]
   6            SEARCH[carro urbano]
```

In fact, the vendors of MCC products may come up with different ways of describing them in Portugal. In addition, it seems totally arbitrary to have to keep in the translation the fact that in some cases the word "car" is used and in others not, just because it happened to occur that way in the original collection.

2.2.8 Irrelevant questions

Finally, not all questions selected by the other groups make sense for Portuguese speakers to ask, as examples 174 and 202 should make obvious. In [14], similar cases are mentioned for German.

```
174 F OTHER O que significa «Forza Italia!»?
   1            SEARCH[Força, Itália!]
   2            SEARCH[Força Itália]
```

[7] However, one might easily conceive of questions which had to state gender, like "who was the first female president of Iceland?" for *Quem foi a primeira presidente da Islândia?*.

[8] It is well known from the translation studies discipline that the shorter the units, the more difficult to translate them.

```
   3            NIL

202 F OTHER Qual o acrónimo da Amnistia Internacional?
   1            SEARCH[AI]
   2 940113-144 AI
```

In fact, one might want to ask about acronyms in another language, given that an international organization can have different acronyms (such as NATO and OTAN) in different languages. This raises, in any case, the question of whether one was supposed to translate the original "Amnesty International" into *Amnistia Internacional*, or not.

3 Preparing and using the collection to develop the test material

The Portuguese collection, which we tentatively[9] call CHAVE (Portuguese translation of French "clef") was created using the same texts (restricted to years 1994-1995) that were used to build the CETEMPúblico corpus (for a description of the building process, see [15]). In CETEMPúblico, for legal reasons, the documents were split in extracts of about two paragraphs and shuffled so that no reconstruction of the full articles were possible. For CLEF, however, PÚBLICO allowed us to distribute the full texts, so our task was solely to adapt the original programs to the new format, while solving also some of the problems reported in [4]. A few cases, mostly having to do with proper separation of documents, were impossible to solve automatically, so we had to perform a limited manual clean-up. We know that some minor imperfections still persist, though.

We ended up with a collection of 106,821 documents (348Mb). Ideally, each document contains a single article in the newspaper. However, some "articles", coming from sections like "Last news", gather several different short news, about quite different subjects, which may cause problems to some IR systems.

The documents are simply marked with date and kind of section (as provided by the newspaper). Neither titles nor authors have been marked as such, so they appear as free text. In order to help systems who might want to identify titles and authors, we provided a list of probable authors in our Portuguese@CLEF webpage [16].

Given that we had no IR system or Q&A engine available, we should report on the tools we used to find and check the topics and answers in the collection.

We compiled the collection with the IMS Corpus Workbench [17], a powerful suite of programs to deal with large corpora and efficiently encode several kinds of annotations. For each document we encoded its unique ID (composed by the date and section). Then, the corpus query processor (CQP) allowed us to retrieve concordances showing the ID of the document they occurred in. This eased substantially our workload, both while preparing the topics for the IR tasks (mainly when checking whether the topics proposed by the other groups existed in our collection), and preparing the resources for the Q&A task (allowing us to easily identify the documents which provided the answer to a particular question).

In fact, CQP proved useful for Q&A in no less than four stages:

- while searching for possible questions and their answers,
- while translating the other groups questions and their answers (checking the more usual Portuguese forms),
- while selecting the 100 additional questions among those, through searching for the translation of the answers
- and while evaluating the correctness of the answers provided by the participant groups.

For IR, CQP allowed us for example to check easily whether we had any document referering to Sosnovyj Bor (or to be exact, "Sosn.+ Bor", allowing for variations in orthography)[10], as well as to check which documents referred to a minor earthquake in Nice in the dates provided, since all documents were dated.

---

[9] As said in section 1, we are still unsure about the distribution of CLEF resources, and therefore also about the possibility of being able to baptize independently some of them.

[10] As anyone dealing with real text is aware of, there are often several spelling variants, even within a single language, especially if the texts have not been proofread. This is a problem particularly with less used foreign names: the Icelandic capital, Reykjavík, appears in six different forms in CETEMPúblico, and Antwerp is often written as *Anvers* in texts translated from French, despite having a name universally used in Portuguese, *Antuérpia*. Other common cases in Portuguese are the unstable use of the dash: "prime minister" can equally frequently appear as *primeiro-ministro* or *primeiro ministro*, and the use of capitalization, e.g. "in Northern China" is rendered both as *no Norte da China* or *no norte da China*.

As to the future, we suggest the possibility of creating a multilingual corpus contaning the texts of the several collections in order to speed up the process of checking the existence of references to any topics in all the collections, for the IR adhoc track.

4 Conclusions

One aim of this paper was to describe several difficulties in creating the topics and the questions for the CLEF campaign of 2004, with a view to help future groups when adding a new language, but also in order to suggest some improvements for future editions. In fact, many of the ideas stated here, especially in what concerns Q&A are probably relatively controversial, but we used this opportunity to stimulate discussion on the subject in the CLEF community.

Our main conclusion is that, in general, more reflection and study should be given to the process of selecting topics and questions, in order to maximize the utility of the collection for those topics and questions.

Having access to all collections, one might (collaboratively) study the collections and find out a) in which (subject) areas the information is conveyed by all languages, b) which areas exist where local information can be relevant for people of other languages, and c) areas (maybe the most interesting) where there is complementariness in the collections.

As regards Q&A categorization, we argued that the present classification does not seem very useful, especially because there may be different ways to look for the same information.

Moreover, as far as definition questions are concerned, not only there is no difference linguistically, but their specification stretches considerably what a "concise and informative answer" may be. In some cases one gets as golden answers "a sect" or "a movement" (obviously too little for a definition), in other cases one gets a lenghty description of a scientific procedure, in several sentences. In fact, it is a truism to say that the difference between IR and Q&A lies mainly in the form of the answer. Definition questions seem to require a passage and are not good examples of Q&A with unique and consensual answers.

We also suggest to integrate the work done in IR and Q&A more closely: We feel it would have been interesting to submit all questions as IR topics and see whether IR systems could provide the documents where the answers could be found. Conversely, it would be interesting to create a set of questions from the topic description and/or narrative and look for them in the QA exercise. More integration between both tasks might be able to shed light on the current state of the art of both kinds of systems. And, at least for the organisers, it would either save a lot of effort or produce twice as many results with the same amount of work.

We conclude this paper by noting that it is extremely important to look at kinds of topics and questions really posed by real users in order to ascertain how difficult and how frequent are the test data we have created (in other words, to evaluate the work done by the organization).

For all we know, it is possible that quite different matters are asked by people looking at newspaper text, and that other types of questions are more relevant than we suppose. Confirmation questions[11], at least, seem to be worth while considering, since people often want to confirm what they already know or suspect, instead of asking about something they are completely ignorant about.

Acknowledgements

The authors would like to thank Carol Peters and Alessandro Vallin for answering all our doubts at once, as well as José Vítor Malheiros and PÚBLICO for allowing us to use their material. Thanks also to Luís Costa and Nuno Cardoso for valuable comments on a previous version. This work is financed by the Portuguese Fundação para a Ciência e Tecnologia through grant POSI/PLP/43931/2001, co-financed by POSI.

References

[1] Linguateca. www.linguateca.pt

[2] Santos, Diana. "Toward Language-specific Applications". *Machine Translation* **14** (2), June 1999, 83-112.

[3] www.linguateca.pt/aval_conjunta/.

[11] Such as "Is Oslo the capital of Norway?", "Is Athens the first city where the modern Olympic games took place?", "Was *Finnegan's Wake* written by James Joyce?".

[4] Santos, Diana & Paulo Rocha. “Evaluating CETEMPúblico, a free resource for Portuguese”, *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (Toulouse, 9-11 July 2001), pp. 442-449.

[5] Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos. “"Floresta sintá(c)tica": a treebank for Portuguese”, in Manuel González Rodríguez & Carmen Paz Suárez 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), ELRA, 2002, pp. 1698-1703.

[6] Santos, Diana, Luís Costa & Paulo Rocha. “Cooperatively evaluating Portuguese morphology”, in Nuno J. Mamede, Jorge Baptista, Isabel Trancoso & Maria das Graças Volpe Nunes (eds.), *Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003, Faro, 26-27 June 2003, Proceedings*, Springer Verlag, 2003, pp. 259-66.

[7] Santos, Diana & Anabela Barreiro. “On the problems of creating a consensual golden standard of inflected forms in Portuguese”, in Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of LREC'2004, Fourth International Conference on Language resources and Evaluation*, (Lisboa, 26-28 May 2004), pp. 483-6.

[8] Santos, Diana & Paulo Rocha. “AvalON: uma iniciativa de avaliação conjunta para o português”, in Amália Mendes & Tiago Freitas (orgs.), *Actas do XVIII Encontro da Associação Portuguesa de Linguística* (Porto, 2-4 de Outubro de 2002), APL, 2003, pp. 693-704.

[9] Paulo Rocha & Diana Santos. “CLEF: Abrindo a porta à participação internacional em RI do português”, in Santos, Diana (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*, in print.

[10] “A presença das línguas e das culturas latinas na Internet”, União Latina, 28 de setembro de 1998, http://www.unilat.org/dtil/lenguainternet/pt/l_latinas_pt.asp. English version as “L4: The fourth study on Languages and the Internet”, http://www.funredes.org/LC/english/L4.html.

[11] Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo, F. and de Rijke, M.: “Creating the DISEQuA Corpus: a Test Set for Multilingual Question Answering”, in Carol Peters, editor, *Working Notes for the CLEF 2003 Workshop*, 21-22 August, Trondheim, Norway, 2003.

[12] QA@CLEF-2004 Guidelines. http://clef-qa.itc.it/2004/guidelines.html

[13] Costa, Luis. “First evaluation of Esfinge – a question-answering system for Portuguese”. This volume.

[14] Magnini et al. “Overview of the CLEF 2004 Multilingual Question Answering Track” This volume.

[15] Rocha, Paulo Alexandre & Diana Santos. “CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa”, in Maria das Graças Volpe Nunes (ed.), *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR’2000)*, Atibaia, São Paulo, Brasil (19 a 22 de Novembro de 2000), pp.131-140.

[16] CLEF 2004 - Presença do português. http://acdc.linguateca.pt/aval_conjunta/Merlin/prCLEF.html.

[17] Christ, Oliver, Bruno M. Schulze, Anja Hofmann & Esther Koenig. “The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual”, Institute for Natural Language Processing, University of Stuttgart, March 8, 1999 (CQP V2.2), accessed 28 May 1999, http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/.