The Key to the First CLEF with Portuguese: Topics, Questions and Answers in CHAVE

Diana Santos and Paulo Rocha

Linguateca, Oslo node, SINTEF ICT, Norway
Diana.Santos@sintef.no, Paulo.Rocha@di.uminho.pt

Abstract. In this paper we report the work done by Linguateca in order to add Portuguese to two tracks of CLEF, namely the *ad hoc* IR and the QA tracks. We start with a brief description of Linguateca's aims and the way we see CLEF from the standpoint of Portuguese language processing. We then comment on several interesting problems that emerged during our work and offer some suggestions for improvement, and finally raise some possibly controversial points for discussion.

1 The Role of Linguateca

The creation of Linguateca (http://www.linguateca.pt/) originated from the realization that there were too few resources for the processing of Portuguese, and that the large language resource centres such as LDC or ELRA could not take a primary role in the deployment of such resources, given their 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 were located 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.

Linguateca thus concentrates on Portuguese. To improve PLP, we believe that 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 – a language-specific bias as emphasized in [1]. We have, therefore, created resources for Portuguese, such as the large annotated corpus CETEMPúblico [2] and the Floresta Sintá(c)tica treebank [3], and organized evaluation contests dealing with Portuguese only [4].

There is no contradiction, however, in Linguateca joining CLEF, the most international of all evaluation contests (at least as far as the number of different languages and participants from different countries are concerned), given that the primary aim of CLEF is to foster **crosslingual** information retrieval. Thus, whether to evaluate querying a multilingual collection in Portuguese, or querying a Portuguese collection in another language, CLEF is the place to go. Instead

of copying or adapting something borrowed from another language to deal with Portuguese, we have added Portuguese, so that people primarily concerned with other languages may be encouraged to process Portuguese as well.

In any case, at present, for certain monolingual tasks, there would not be enough participants to organize one evaluation contest on its own: only two monolingual groups participated in the Portuguese QA task. Nevertheless, the QA@CLEF coordinators added Portuguese without too much work. This shows that joining a circle of international experts in order to define a particular task precisely is a sensible way to begin, even if one disagrees with some of the choices taken.

In fact, although we have publicly voiced the opinion that an all-Portuguese-speaking organization would give more weight to Portuguese-specific matters and more influence to participants dealing with Portuguese – and hence one should ideally start with Portuguese-only evaluation contests [5] –, this opinion must be weighed against the organizational relief of having general matters coordinated centrally.

Also, the only unbiased way to assess whether it was worthwhile for the Portuguese language processing community to participate in CLEF was to try it out, and we now believe it was worthwhile. This participation provided clear deadlines for building resources that otherwise would have taken us much longer to complete, and a lot was learned from working together with the teams for other languages.

As a result of our participation in CLEF, we have now released the CHAVE collection (www.linguateca.pt/CHAVE/), containing PÚBLICO newspaper's collection, the IR topics and relevance judgements, and the questions and answers created for the QA campaign.

2 Tasks

Portuguese was included in the Monolingual (non-English)/Bilingual/ Multilingual Information Retrieval (also called *ad hoc*) tasks and QA track. QA is, in fact, conceptually a more advanced IR task and the communities involved were different: not only the groups and systems that competed (at least for Portuguese) but the organizational apparatus and decisions.

The workload involved was also differently distributed: for IR, the topic creation and discussion was relatively light, but the evaluation of the results was demanding. On the contrary, the preparation and translation of the questions and answers, as well as finding justifications for them, represented the bulk of work for QA@CLEF, while evaluation was light and even intellectually rewarding.

In the following sections, we discuss in some detail our participation in each of the tracks. We avoid gory details and lengthy descriptions of issues which can only be fully apprehended by speakers of Portuguese (see [6] for this), trying instead to produce an interesting summary of our difficulties and remaining doubts, as well as provide some guidance to newcomers to the (CLEF) field.

2.1 IR Topic Preparation

The main issue in the ad-hoc topic preparation 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.

International vs. National According to the ad-hoc track 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. It would be 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 an enlightening experience to check other groups' topics as well as to learn about the relative importance of the Portuguese topics that we expected to be reported elsewhere. There are studies on the relative impact of the Romance languages in the web as a whole [7], and we suggest doing something similar: to measure, for each foreign collection, the degree of "Portugueseness" to be expected. Unfortunately, we did not have access to the collections in the other languages at the time of topic preparation, so this must be postponed.

Another Classification of Topics We suggest a different classification of topics: cyclic events; once-only events; states of broader events; impact measures; and atemporal subjects. Examples of the latter kinds[1] follow.

As for *states* (or sub-events) *of broader events*, "East Timor guerilla" or "civil war in Rwanda" can be considered as "states" of a larger war. The same is true for "Fight against AIDS in Africa" or "Russian-Finnish relations" (both subjects concerning 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". For this kind of topic, we are interested in how these subjects fare in news coverage in 1995, although the topic may have been raised by specific events taken up in (local) press. Nevertheless, a user may 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, less straightforwardly, "Iranian cinema" or "Seal-fishing". One may argue that the latter can also be interpreted as states of

[1] For lack of space, we present only the topic titles, asking the readers to trust our judgement, although most of the title names, in isolation, could describe radically different information needs.

a larger event (e.g. the whole history of Iranian cinema), or impact measures, i.e., the user is looking for events concerning seal-fishing (like laws and debates) occurring in 1995. Still, we believe that searchers may be interested in knowing about seal-fishing or deaf people in society without a temporal grounding, while news covering "EU and the price of food" seems to make sense only at a particular time.

In any case, we suggest considering carefully whether these different kinds of topics, which we argue reflect different user needs, and consequently may even require different kinds of query applications, should have different evaluation practices (or not), and/or different forms of description (and narrative).

Different Answers in Different Collections We believe considerably more attention should be given to this issue. To us, topics with different answers in different collections are 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[2], but the (related) QA@CLEF campaign provided good examples: take the case of "Name some X", with X "person charged of paedophilia", or "what is the masonry?", in which different facets – and facts – about this organization in different countries might be uncovered in a multilingual collection.

This illustrates the strikingly fuzzy borderline between IR and QA. QA can be seen as a request for more precision about a topic, and some topics were even stated as questions. In fact, Magnini *et al.* [8] even report that the original set of questions used in QA@CLEF 2003 was inspired by the topics of the previous year's ad hoc competition. Having prepared the material for both, we cannot help stressing how both tracks are conceptually the same, despite testing different types of systems.

Topic Wording Although we have not received any specific instructions on this subject, we attempted to profusely word the topics, distributing paraphrases among title names, topic description and topic narrative in Portuguese.

Using as many synonyms and wording variants as possible, systems would get (almost) a synonym-expansion capability for free, if they used all material provided. For example, in topic C249 below, *dez mil metros* and *10.000 m* are alternative ways of stating "ten thousand meters" in Portuguese. And *campeã*

[2] "Sports women and doping", "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. Incidentally, and no matter their seemingly general character, atemporal subjects are not necessarily also alocational: "Seal-fishing", and "Avalanche disasters" are not often discussed in Portuguese media, for geographical reasons, and the same applies to topics on bowling or haunted buildings, suggested respectively by the Finnish and British teams. Apparently, these are, for cultural reasons, simply uninteresting subjects to Portuguese readers of newspapers.

(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>
```

2.2 QA Preparation

Preparing the resources for the QA track presented another kind of challenge. Very briefly, our job was as follows: we had to create 100 natural Portuguese questions with answers, indicate an associated document where the answers could be found; translate them into English; and 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.

Each subtask was far from straightforward, the main challenges being: For our questions, (a) coming up with a set of not too difficult, natural questions with a straightforward answer; (b) identifying clearly the answer(s), finding all plausible answers in our collection; (c) providing a natural English translation with (if possible) the same presuppositions of the Portuguese one. For the questions coming from other groups (which we had both in English and in the original language), the main challenges were: (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 there were any), and adding other answers (either more correct in case a wrong answer had been supplied, or more Portuguese-like as regards measures or spelling); and (c) in case no answer could be found in our collection, trying to provide suitable translations of both answers and questions.

What is a Natural Question? A "natural" question is something that eludes a precise definition, and has often been discussed in the context 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)[3]? We decided to use the easier kind in our set, as shown in question 337

[3] 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 minister of the environment?) but we strongly doubt such users will ever amount to the majority of Portuguese speakers.

`F PERSON Quem é a ministra do Ambiente alemã?`, where *ministra* is the feminine form. Curiously, all other groups used the masculine form in their translation of this question.

Another concern was the following: Should one use the informal way of posing questions in (European) Portuguese, or suppose that normal users of a QA 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.

Question Classification In addition to coming up with questions, we had to classify them, according to the track instructions [12]. This turned out not to be as meaningful as expected. The classification was to be done according to the semantic category of the right answer (person, location, manner, object, measure...), but this in turn had little correlation with the linguistic properties of both question and answer.

In fact, questions `558 F OTHER Qual a nacionalidade do tenista Sergi Bruguera? SEARCH[espanhol]` and `582 F LOCATION De que país é a escritora Taslima Nasreen? SEARCH[Bangladesh]` are after precisely the same kind of information (What is X's nationality? and Which country is X from?), but have been classified differently.

Also, one might argue that, although question `688 F OTHER Qual o endereço da Livraria Barata? LING-940102-050 Av. de Roma, 11` 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.

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 were related to cause of death.[4]

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 – or that there were problems with the "semantic" classification.

Definition questions – which were introduced in the 2004 campaign – are, in our opinion, especially tricky. Consider question `693 D PERSON Quem é Guilherme da Fonseca? LING-940127-152 juiz do Tribunal Constitucional` (Who is X? with answer = supreme court judge). This is the same as asking what is X's profession, which should then be classified as F OTHER...[5]

Therefore, we believe that a more objective way of classifying questions, such as the one presented in [9], is preferable. Alternatively, one could classify questions according to the kind of linguistic entity expected as answer, using

[4] Not all answers to "How did X die?" had to do with cause. One was "In strange conditions"! Again, something rather hard to conceive as MANNER.

[5] Incidentally, in order to provide a more accurate answer, one should state "one (of several) judges". In other words, the indefinite article should not have been left out.

categories like "proper name", "common name", "toponym", on a par with "definition" (which is a kind of answer, not a real world object).

Presuppositions Abound How many presuppositions should be allowed in a natural (as opposed to a tricky) question? Looking for the answer to question `327 F PERSON Como se chama a filha do líder chinês Deng Xiaoping?`, what is Deng Xiaoping's daughter called?, we found out that he had not one, but two daughters ("Deng Rong" in the original Dutch collection, and "Deng Nan" in ours). Apparently, therefore, this question was ill-posed.

In general, anyway, most questions presuppose that it is possible to share the referent with the reader, as we point out in the next section, on definitions.

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 generally not possible to ascertain. In fact, asking who Fernando Gomes (the mayor of Porto in 1995) is, we found, in the very same collection, a reference to a football player of the same name. We leave the reader to try to find out how many organizations called GIA exist (in all languages covered in QA@CLEF).

A definition is the most complex question one can ask. To give answers to "what is the masonry?" or "what is indocyanine angiography?", one needs to be an expert in the field, and still consider carefully how to produce an appropriate rendering. Of course, if one is querying a collection of authoritative texts (and, especially, didactic material), it may be possible to automatically extract definition-like passages. But in a newspaper collection, it is doubtful whether more than `is-a` relations (which are not definitions) can be extracted.

The "definitions" as described by the CLEF organization have still other flaws:

1. They often overlap linguistically with factoid questions, cf. F(actoid) "Who is the pope?" D(efinition) "Who is João Paulo II?" In free text, it is often difficult to know whether linguistic expressions are attributive or appositive, and, in fact, in most 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.

Therefore, we propose giving more attention to the user's (or system's) goals in order to decide on what can sensibly be called a request for definition, as opposed to questions of the kind "Who occupies the *role* Y?"or "What *profession* does X have?" (which idiomatically is expressed by *Quem é X*?).

Getting Correct Answers: Articles, Gender and Redundancy Another interesting observation is that 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 been left out. A more specifically Portuguese case is example 647: the proper name could have been preceded by the Portuguese article *a*, meaning "the Petrogal". This would, however, probably confuse the other groups in the translation task too much, 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?* and *O que é a EUA?* – the first having as right answer *os Estados Unidos da América* (USA) and the second *a European University Association.*)

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

Finally, 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 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]`, *volume de negócios* was repeated in order to translate the original answer (which specified "of the group"). In `588 F MEASURE Quantos empregados tem o grupo Warburg? SEARCH[4.472 pessoas]`, on the other hand, the word *pessoas* (persons) in the answer about how many employees is quite uninformative.

Translation is Hard: Idiomaticness 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 "Party of National Solidarity" for *Partido da Solidariedade Nacional.*

In addition, not all presuppositions 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".[6]

Translation of Ungrounded Arbitrary Fragments The translation of other groups' questions, especially when there were no hits in our collection, or when the question seemed about unfamiliar subjects or contexts, also caused us problems. In question `293 F MANNER Como se garante a cobrança de sanções? SEARCH[pelo sistema de notificação de multas através de edictos]`, we had no idea of which

[6] Yet, one can easily conceive of questions which had to state gender: who was the first female president of Iceland? *Quem foi a primeira presidente da Islândia?*

kind of sanctions were mentioned, nor to whom the indeterminate *se* refers: government? tax authorities? sports club? Likewise, no clue was given as to who is supposed to pay them.

Translating the answers that came with the questions was even worse. In fact, it was in general a major headache, not only because of the reasons already discussed, but because it was not evident why some of the answers (paraphrases) had to be translated at all. And the shorter the units, the more difficult to translate them. Consider `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]
```

Answer 3, for example, sounds awkward, while we could concoct more precise and interesting definitions of AI (if one were after one gold standard with the "right" answers in Portuguese).

Generally, we tried 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 mentioned 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", some of the answers to "Who are the Simpsons?", "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 result of this process. `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, MCC salespeople may come up with different ways of describing the products in Portugal. In addition, it seems totally arbitrary 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.

Irrelevant Questions Finally, not all questions selected by the other groups make sense for Portuguese speakers to ask, as examples `174 F OTHER O que significa Forza Italia!? 1 SEARCH[Força, Itália!] 2 SEARCH[Força Itália]` and `202 F OTHER Qual o acrónimo da Amnistia Internacional? SEARCH[AI]` should make obvious. In [10], similar cases are mentioned for German.

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" as *Amnistia Internacional*, or not, in question 202.

3 Preparing and Using the Collection

The Portuguese collection, which we called CHAVE (the 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 [11]). In CETEMPúblico, for legal reasons, the documents were split into extracts of about two paragraphs each 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 the proper separation of documents, were impossible to solve automatically, and 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", from sections like "Last news", gather several different short news about quite different subjects, which may harm the performance of some IR systems. The documents are only 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, but, to help systems that rely on titles (and would thus filter authors), we also provided a list of probable authors at our website.

We had no IR system or QA engine available. We therefore encoded the collection in the IMS Corpus Workbench [12], a powerful suite of programs designed to deal with large corpora, efficiently handling several kinds of annotations. For each document we encoded an unique ID, composed from its date and section), and used the corpus query processor (CQP) to retrieve concordances showing the ID of the document they occurred in.

So, checking whether the topics proposed by the other groups existed in our collection was considerably simplified: For example, to find whether we had any document referring to Sosnovyj Bor, we would look for `"Sosn.+" "Bor"`, allowing for variations in orthography.[7] We could also check which documents referred to a minor earthquake in Nice in the dates provided.

For QA, CQP proved useful 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

[7] As anyone dealing with real text is aware, 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; similarly, Antwerp is often written as *Anvers* in texts whose original was published in French, despite having a name universally used in Portuguese, *Antuérpia*. Also common is the unstable use of the dash: prime minister can equally frequently appear as *primeiro-ministro* or *primeiro ministro*, and variable capitalization, e.g. "in Northern China" is rendered both as *no Norte da China* or *no norte da China*.

of the answers; and while evaluating the correctness of the answers provided by the participant groups.

4 Concluding Remarks

One aim of this paper was to describe some of the difficulties in creating the topics and the questions for the CLEF campaign of 2004, with a view to helping future groups when adding a new language, but also in order to suggest improvements for future editions. In fact, some of the ideas stated here, especially for what concerns QA may be relatively controversial, but we use 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. We feel it is extremely important to look at topics and questions really posed by actual users, also to ascertain how difficult and how frequent are the test data we have created, to eventually evaluate our work (and that of the CLEF organizers as a whole).

Having access to all collections, one might (collaboratively) study them 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 QA 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, and we also suggested removing definition questions, which seem to require a passage and hence are not good examples of QA with unique and consensual answers. We furthermore suspect that quite different subjects are asked by people looking at newspaper text, and some missing question types may be quite relevant. A case in point are confirmation questions[8] – people often want to confirm what they think they know, instead of asking about something they know nothing of.

We also suggest to integrating more closely the work for IR and QA: On the one hand, it would be 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 shed light on the current state of the art of both kinds of systems.

Acknowledgements We are grateful to José Vítor Malheiros and PÚBLICO for their material, and to Luís Costa and Nuno Cardoso for valuable comments on previous versions. We acknowledge grant POSI/PLP/43931/2001 from the Portuguese Fundação para a Ciência e Tecnologia, co-financed by POSI.

[8] Such as "Is Oslo the capital of Norway?", "Is Athens the first city where the modern Olympic games took place?", "Did James Joyce write *Finnegans Wake*?"

References

1. Santos, Diana: Toward Language-specific Applications. Machine Translation Vol.14 (1999) 83–112.
2. Santos, Diana, Rocha, Paulo: Evaluating CETEMPúblico, a free resource for Portuguese. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. Toulouse, July 2001, 442–449.
3. Afonso, Susana, Bick, Eckhard, Haber, Renato, Santos, Diana: "Floresta sintá(c)tica": a treebank for Portuguese. In Rodríguez & Araujo (eds.): Proceedings of LREC 2002, Las Palmas, May 2002, ELRA, 1698–1703.
4. Santos, Diana, Costa, Luís, Rocha, Paulo: Cooperatively evaluating Portuguese morphology. In: Mamede et al. (eds.). Computational Processing of the Portuguese Language, 6th PROPOR, Springer (2003) 259–266.
5. Santos, Diana, Rocha, Paulo: AvalON: uma iniciativa de avaliação conjunta para o português. In: A. Mendes & T. Freitas (orgs.), Actas do XVIII Encontro da Associação Portuguesa de Linguística, Porto, October 2002, APL (2003) 693–704.
6. Rocha, Paulo, Santos, Diana: 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.
7. Latin Union: L4: The fourth study on Languages and the Internet. http://www.funredes.org/LC/english/L4.html.
8. Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo, F., de Rijke, M.: Creating the DISEQuA Corpus: a Test Set for Multilingual Question Answering. In: C. Peters (ed.), Working Notes for the CLEF 2003 Workshop, August 2003, Trondheim. http://clef.isti.cnr.it/publications.html.
9. Costa, Luís: First evaluation of Esfinge, a question-answering system for Portuguese. This volume.
10. Magnini et al.: Overview of the CLEF 2004 Multilingual Question Answering Track. This volume.
11. Rocha, Paulo, Santos, Diana: CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In: M.G.V. Nunes (ed.). Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada. PROPOR2000, Atibaia-SP (November 2000), 131–140.
12. Christ, Oliver, Schulze, Bruno M, Hofmann, Anja, Koenig, Esther: 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).