

A golden resource for named entity recognition in Portuguese

Diana Santos* and Nuno Cardoso‡

Linguateca: *node of Oslo at SINTEF ICT, ‡node of XLDB at University of Lisbon

Abstract. This paper presents a collection of texts manually annotated with named entities in context, which was used for HAREM, the first evaluation contest for named entity recognizers for Portuguese. We discuss the options taken and the originality of our approach compared with previous evaluation initiatives in the area. We document the choice of categories, their quantitative weight in the overall collection and how we deal with vagueness and underspecification.

1 Introduction

HAREM [1] was the first evaluation contest for named entity recognition (NER) in Portuguese, gathering the community interested in named entity recognition, as well as developing a new methodology for the evaluation of NER systems. Following the well-known evaluation context paradigm [2, 3], participants cooperated in the design and population of the golden resource used in HAREM, the golden collection (GC), which we describe in detail here.

The golden collection is a set of texts with the correct identification and classification of the named entities according to the HAREM guidelines.¹ This resource – of which each participant had only seen a tiny extract – was then embedded in a much larger text collection, the HAREM collection, which was provided untagged to all participating systems, which in turn had two days to return it NE-annotated. Evaluation was performed by comparing results with the answers previously encoded in the golden collection.

Although this is a fairly standard model, already used for Portuguese in Morfolimpíadas [4, 5], there are some interesting ways in which this evaluation contest differs from MUC [6] and other predecessors in the NE domain:

Kind of text HAREM encompassed a broad range of texts, of which language variety, genre and (implicit) publication mode was provided to the participants. In addition to allow a finer-grained characterization of the problem, such information allows more advanced systems to change behaviour according to any of these parameters.²

Tailored evaluation We allowed participants to choose to participate in all categories, or just to be evaluated for a subset of their own choice. We also gathered separate information for specific scenarios, per main category (e.g. if the system’s task was only to find locations, or organizations in text) or comparing with MUC.

¹ Guidelines and collection are available from <http://www.linguateca.pt/HAREM/>.

² Although in HAREM’s first edition we are not aware of any system profiting from this feature, in the future it is to be expected that textual genre is used by informed systems to select different interpretations of a particular NE.

Bottom-up categories Instead of using previous NE categories often designed for other languages, HAREM concerns a set of Portuguese linguistically motivated categories (and corresponding types), arrived at by careful consideration and manual annotation of a wide amount of text of different genres.

Context annotation Emphasis was put on the context of use: HAREM was meant to evaluate the type of a named entity in the **context** it is used, not its general description type. For example, the name of a newspaper may be classified as an organization (company), a place (of publication) or a person (the interviewer) depending on the context. This clearly makes the task evaluated by HAREM more difficult,³ but we believe such behaviour to be more useful in real applications.

Separation of identification and classification We conceptually separated identification (e.g. the correct delimitation of the NE) from classification (which could be semantic or morphological).

Morphology HAREM included a morphological task, since gender and number are properties of most Portuguese named entities, and may sometimes help semantic disambiguation.⁴

New measures HAREM introduced several new measures to assess the performance of the systems, dealing with indeterminacy and with partially correctly identified NEs.

We start the paper by providing, in section 2, some background on the named entity task and our motivation for organizing HAREM. In section 3 we describe the golden collection in detail, touching upon identification guidelines, classification guidelines, dealing with indeterminacy, and problems we are aware of. In section 4, we provide some first analyses making use of this collection.

2 Context and motivation

Named entity recognition was first coined in MUC 6, when it became an independent track in a larger effort to evaluate information extraction. Since then, several other evaluation contests have been organized around this task, for English as well as for other languages [9, 10]. More specific subsets of the original NE set, time expressions, have also been evaluated separately [11].

In fact, the organizers of evaluation campaigns like ACE [12] and dedicated workshops [13] have claimed that the time was ripe to investigate more challenging goals. However, although several groups had their in-house procedures and gazetteers, there was still no independent off-the-shelf NE component for Portuguese at the time work was started in charting interest and tools in the area. See [7] for a detailed historical overview, providing also background and further motivation to identify and classify Portuguese NEs.

³ Other evaluation contests might consider “newspaper name” or “organization” as a correct label in every such case.

⁴ [7] note that organizations named after male persons tend to retain the feminine gender, while football clubs are often described as masculine even if named after something originally feminine. [8] mention the acronym *EUA* as referring to the United States (masculine plural) or the European Universities Association (feminine singular).

There were, anyway, other scientific and methodological reasons that led us to organize HAREM, in addition to the obvious relevance to the Portuguese language processing community:

First of all, there was not, as far as we know, any independent-from-English investigation of which needs (in terms of other categories) other languages might have in this respect. For example, would person, organization, and place names be all, only and forever the right categories to identify, as [9] seems to assume uncritically? Although Sekine et al. [14] have proposed an ontology large enough to encompass both English and Japanese, and others have looked into the problem of multilingual NER [15, 16], there was no bottom-up **monolingual** NER evaluation contest we knew of.⁵ As shown in the remainder of this paper, we found a significant number of other relevant NEs to be catered for in Portuguese.

Second, would the NE categories used be equally relevant no matter the register or genre of texts, almost a decade later, if one wished to parse the Web, literary text and technical documents, too? We wanted to investigate whether the sort of NEs in a given text might be a discriminating feature for classifying genre [20]. Preliminary results on this subject have already been published in [1].

Finally, an issue in which we had a special interest is how much language-dependent NER is [21], and how much (if anything) can be gained by attacking Portuguese as opposed to language-general methods, which we expected international participants would try out. As to this last point, let us stress that HAREM is not comparing the same task for different languages, as [22, 23] have done. Rather, given a specific, particular, language (Portuguese), one of our goals was to measure, instead, how well language-dependent and language-independent methods fare.⁶

The reasons to organize HAREM were thus manifold, ranging from practical opportunity to theoretical relevance: As to the former, there was a relatively large number of interested parties, which is a requirement for a successful event. In addition, in contradistinction to morphological analysis [5] where it was hard to establish a common set of agreed categories, because many choices of the participating systems were only motivated by the different behaviours of the parsers using them, NER in itself boasted of a number of almost direct applications, although they differed significantly in their requirements.⁷

3 The golden collection

We defined the global task of HAREM as follows: systems had to classify all descriptions of entities involving capitalization in Portuguese, i.e. proper names, in addition to expressions involving digits. So, capitalized section titles, textual emphasis, or common words that are spelled with capitals just because they originate from (other language's)

⁵ Although we were obviously aware of a number of detailed proposals for a.o. German [17], Spanish [18] and Swedish [19].

⁶ This analysis must be based on the methodologies employed by the different participants.

⁷ The original task specification, as opposed to MUC, drastically differed among participants, interested e.g in deploying geographically-aware search engines, or searching and indexing oral documents or semi-automatically producing terminology.

initials, like *CD*, are **not** considered NEs. Expressions that should denote named entities but have been misspelled by not being capitalized are not considered NEs, either.

3.1 Purpose and rationale

We attempted to include in the golden resource of HAREM the ideal set of marking needed by any text in Portuguese to represent the distinctions we were interested in. The annotation of the GC does not represent the desiderata an automatic system is supposed to achieve nowadays, or ever, nor is it assumed that most of the NEs should appear in a gazetteer – on the balance of several methods, see [24].

One of our goals with HAREM was, in fact, to measure the degree of difficulty of the general NER task in Portuguese, and not particularly the performance of real systems today. To arrive at a particular classification may require a huge amount of reasoning and world knowledge – probably something no present system is ready to do. Still, we wanted to estimate how often this is the case.

Finally, it is relevant to note that we strove to get at linguistically motivated categories in the GC. That is, we did not start from an ideal ontology on how the real world should be conceived, but we observed what kinds of references were typically capitalized in Portuguese texts. After a first mental organization of these different cases, we attempted to posit their specific subcategorization (what we call types) which objectively showed different linguistic properties in Portuguese. For example, reproducible works of art are usually encoded within quotes, while unique ones do not; atomic events are often compound names as in *Benfica-Sporting*, while organized ones receive a more "ordinary" proper name; and references to publications, or laws, appear usually in a special and very specific format in running text.⁸

The release (and use) of the GC will help us assess whether we succeeded in the goal of proposing only linguistically different types, or whether still many iterations are required to obtain further consensus and more replicable guidelines.

3.2 Quantitative snapshot

HAREM's golden collection is a collection of texts from several genres (table 1) and origins (table 2), in which NEs have been identified, semantically classified and morphologically tagged in context, and revised independently, according to a large set of directives approved (and discussed) by all participants.

3.3 The classification grid

Our study of Portuguese texts uncovered that, in addition to locations (LOCAL), people (PESSOA) and organizations (ORGANIZACAO), as well as values (VALOR) and dates

⁸ They may pose interesting linguistic problems as well. Susana Afonso (p.c.) noted that a reference with multiple authors behaves differently as far as morphology – and therefore type of reference – is concerned. E.g., we may say *Santos et al. (2006) mostraram* (plural animated subject, "S et al showed") or *Cardoso and Santos mostraram em Cardoso & Santos (2006)* (singular inanimate location, "C&S showed in C&S").

Genre	Text extracts	Words	NEs
Web	40 (30.8%)	12 324	1 337
Newspaper	31 (24.6%)	11 614	1 129
Oral	16 (12.3%)	26 875	1 017
Expository	10 (7.7%)	6 647	471
E-mail	16 (12.3%)	4 405	438
Fiction	8 (6.2%)	9 956	327
Political	3 (2.3%)	5 470	312
Technical	5 (3.8%)	2 517	101

Table 1. Genre variety distribution

Origin	Text extracts	Words	NEs
Portugal	63	33 618	2 550
Brazil	60	42 073	2 274
Asia	3	2 864	233
Africa	3	1 253	75

Table 2. Language variety distribution

(TEMPO), there were enough data to justify classification in the following categories: OBRA: titles (works of art, man made things), COISA: things (objects which have names), ACONTECIMENTO: events, and a further bag of others (VARIADO).

Figure 1 shows how the different kinds of NEs are distributed in the GC, while in Table 3 we present the quantitative distribution of the categories as far as size in words is concerned, using Linguateca’s publicly available tokenizer.

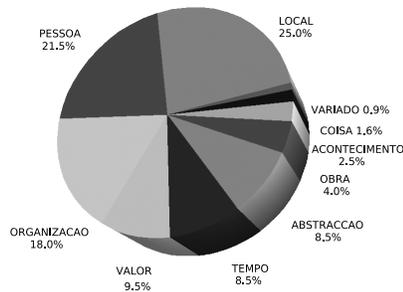


Fig. 1. Kinds of NEs in the GC

Category	Average	Median	Standard Deviation
OBRA	3.34	2	2.94
ACONTECIMENTO	3.25	3	2.52
ORGANIZACAO	2.25	1	1.94
VARIADO	2.24	1	2.60
PESSOA	1.97	2	1.16
ABSTRACCAO	1.92	1	1.25
TEMPO	1.86	1	1.33
VALOR	1.77	2	0.96
LOCAL	1.67	1	1.50
COISA	1.24	1	0.59
TOTAL	1.99	1	1.63

Table 3. Size in words of the NEs in HAREM’s golden collection

Let us present some of the distinctions encoded in the types for each category (for further details, see the guidelines in Portuguese):

PESSOA When NEs denote people, we make two kinds of orthogonal distinctions: One person vs. a group of people; and whether the person is described by its (own) name, by its job (title), or by its membership in a particular group (the metonymical use of companies or political movements). Observe that there are two kinds of groups: those who have a name as group, and those who are syntactically plural (pairs like *Romeu e Julieta* or plural CARGO (job/role) noun phrases, like *os Presidentes da Rússia e da China*, “Russia and China’s presidents”).

ABSTRACCAO Abstract concepts turned out to constitute a sizeable proportion of capitalized entities in Portuguese, ranging from scientific areas to political programmes, to metonymical uses of an author name to refer to the collection of his works, etc. Also, the use of names as names (fairly common in all genres) was separately identified as a type of ABSTRACCAO called NOME.

ORGANIZACAO Organizations were divided into companies, administration (government) and the rest (called institutions). Still, we found the need to independently classify sub-organizations, such as committees or departments, which were a kind

of more flexible organization, but were nevertheless felt to be different from groups of people without a goal as group.

ACONTECIMENTO We distinguish three kinds of events: unique ones, which have a name identifying something that happened in a particular time, and other events which are often organized (and may be repeated) and encompass smaller (indivisible) events.

COISA Objects can also be named, although, contrarily to people, there are more metonymical references than unique objects named. Although there are unique objects (type OBJECTO) that do have a name, like boats: *a fragata D. Luís*, most often – at least in technical domains – things are referred by the classes they belong to. An interesting distinction we catered for in HAREM is instance of a given kind (type CLASSE) and metonymical reference through the use of the same class name to denote an individual object (type MEMBROCLASSE).⁹ Obviously, we only consider such cases as NEs when, because they are named after their discoverers, have to be spelled in capitals, like *aparelho de Golgi*, *constante de Planck*, *contador Geiger*. Avowedly an original decision, we expected systems to classify *de Golgi*, *de Planck* and *Geiger* alone (and not the full NPs) as COISAS of type CLASSE or MEMBROCLASSE, and not as PESSOA.

OBRA Another category found to account for a lot of proper names in Portuguese text concerns works of art and other products of intellectual activity. An interesting difference emerged in the way unique works as opposed to reproducible are mentioned in Portuguese: *Guernica*, *Gioconda* or *a Torre Eiffel* do not take quotes, contrary to talk about books, musical pieces or movies.

LOCAL Locations were divided into geographical landmarks and political or administrative locations.¹⁰ Other kinds of locations identified and accommodated by the HAREM scheme are virtual locations (such as newspapers or the Internet), addresses (with a specific syntax of their own), and any named physical body which in a particular context is used for naming a place (like *Teatro São João*).

3.4 Vagueness and alternative delimitation

It is not unusual that classification of linguistic objects is not mutually exclusive: in fact, we believe that vagueness is a constitutive property of natural language [25] and that any NLP system should take it into consideration.

In addition, sometimes human annotators do not have enough information to decide, even consulting other knowledge sources, such as the Web: How many (Portuguese speaking) readers know that *Miss Brown* is a Brazilian band or that *os pastéis de Belém* is often used as a place in Lisbon?

We therefore employed the disjunction operator (!) in both situations, abiding by two guiding principles:

⁹ The difference between “I love Ferraris” or “my Ferrari is here”.

¹⁰ Not to be confused with political actors represented by a place name, considered ORGANIZACAO in HAREM, as in *Os EUA não assinaram o tratado de Kyoto* (“USA did not sign the Kyoto treaty”). The distinction here is between human geography and nature, plausibly relevant for different applications.

1. We are dealing with real text and we are measuring systems' performances in real text. Therefore we should not exclude "too difficult" cases from the GR.
2. We do not want arbitrary decisions to be taken by annotators – if there is a ceiling, in the sense that humans cannot decide above it, it makes no sense to require that a machine should decide more fine-grainedly than humans [26].

The same principles are operational at the level of NE alternative segmentation as well, as in <ALT> <ORGANIZACAO>Governo</ORGANIZACAO> de <PESSOA>Mário Soares</PESSOA> | <ORGANIZACAO>Governo de Mário Soares</ORGANIZACAO> </ALT>.

Anyway, to avoid artificial proliferation of alternative NEs, we tried to separate the above cases where human readers had no clues to decide between different interpretations from those cases which could be easily formalized, and arbitrarily but consistently encoded. Examples of this latter are as following:

- encode only largest NE: if a title is followed by the name, mark only one person, as in *o Presidente do Parlamento Europeu Stefano Prodi*;
- do not practise recursive NE-labelling: place names in organization descriptions (as in *Câmara Municipal de Lisboa*), or organization names in title jobs (as in *Secretário-geral do PC do B*), are not tagged separately;¹¹
- do not allow an arbitrary number of uncapitalized NE beginnings: we allow only titles and address forms (such as in *major Otelio* and *senhor Alves*) and disease hyperonyms (such as in *doença de Alzheimer*, and *síndrome de Down*);¹²

4 Discussion and first results

No matter how simple the task to build a golden resource for NER may appear, many fundamental NLP issues had to be tackled. "Named entity" (or better, our Portuguese appropriation of it, *entidade mencionada*, literally "mentioned entity") did not have a consensual description, so we had to provide an operational definition for HAREM's purposes, which is obviously open to criticism.

One of these decisions, motivated by the need to compare fairly different systems was our definition of avowedly "non-standard" NEs like *de Planck*. However, this decision would not penalize systems that detected or classified whole terminological units, provided the systems included an "added-to-participate-in-HAREM"-rule, which would propagate their original semantic classification to the new NE including only the capitalized sub-expression, but including preceding prepositions.

The most important decision taken, in any case, was our choice of disambiguation in context. There are two approaches to NE recognition, which are actually not specific

¹¹ This is essentially linked to our wish to annotate in context, but one may later re-annotate the GC with some embedded NEs whose usefulness turns out to be consensual.

¹² This apparently unmotivated rule has a good explanation: if one was to mark **the entity** that included a named entity, too much dependence on systems' assumptions and complex parsing would occur. One might end up having to accept *a dona da barraquinha das bolas de Berlim* ("the owner of the candy shop selling Berlin balls", a kind of cake) or *o braço direito do rio Guadiana* ("the northern bank of the river Guadiana") as the right NEs.

of NER but of corpus annotation as a discipline, and have to do with how far one disambiguates or chooses a specific annotation value: Some researchers have opted to simplify NER by considering vague categories as final leaves in the classification, much in the same vein as early PoS ventures used so-called “portmanteau” categories, therefore failing to fully disambiguate the text [27]. We took the opposite path: We wanted to evaluate the task of selecting the (meaning of the) named entity in context. Accordingly, HAREM turned out to be more difficult than many of the tasks originally aimed by the participating systems.

Take the much discussed case of country names: to tag *Portugal* as a country requires a simple dictionary/gazetteer lookup, but to find places or political actions are two different goals, which we expect to be performed by different users with different user needs (and therefore possibly even implemented in different applications). That this is not an irrelevant theoretical concern, is evidenced by the real distributions for *Portugal*, *França* and *Brasil* in context in the GC (table 4):

NE	LOCAL (administrative place)	ORGANIZACAO (administration)	PESSOA (group of people)	ABSTRACCAO (ideal)
Portugal	32 (0)	16 (1)	0 (0)	0 (1)
Brasil	52 (1)	5 (2)	2 (0)	2 (1)
França	7 (0)	2 (0)	1 (0)	12 (0)

Table 4. NE counts for single categories (normal) and vague categories (enclosed in parenthesis)

Incidentally, the NEs that had more different contextual interpretations in HAREM’s GC were newspaper’s names, employed in news text in many ways: as a person, representing the journalist that asks questions in an interview; as a place (of publication); as a group of reporters; as a company; or as a product (that for example sells well). Our GC – if representative enough – will also allow researchers to estimate relative probabilities of these different uses, in addition to provide them with contexts to discover discrimination patterns. In fact, one important advantage of evaluation contests, is the creation of consensual available resources, and we agree that “reusable test collections – which allow researchers to run their own experiments and receive rapid feedback as to the quality of alternative methods – are key to advancing the state-of-the-art.”[28]

Category	One category (<A>)	Vague class. (< A ₁ A ₂ >)	Vague ident. (<ALT>)
LOCAL	1227	44	10
PESSOA	972	59	30
ORGANIZACAO	895	64	13
VALOR	480	1	1
ABSTRACCAO	437	23	9
TEMPO	430	1	5
OBRA	194	22	4
ACONTECIMENTO	105	3	21
COISA	75	6	2
VARIADO	37	0	2

Table 5. Number of vague NEs, and number of alternative identifications, per category.

Some idea of the difficulty of human annotation of the different categories in context can be obtained from Table 5.

We conclude this paper by providing an overview of HAREM results, using the present GC: Figures 2 and 3 depict systems’ performance for semantic classification, according to two different scenarios: absolute (counting all NEs in the GC even if they

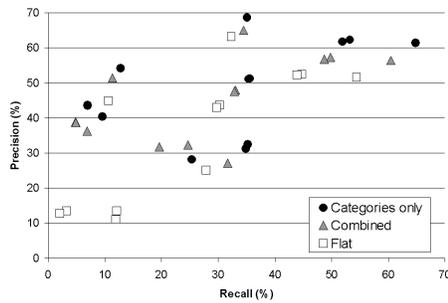


Fig. 2. Prec. vs Recall for Semantic Classification (Absolute scenario)

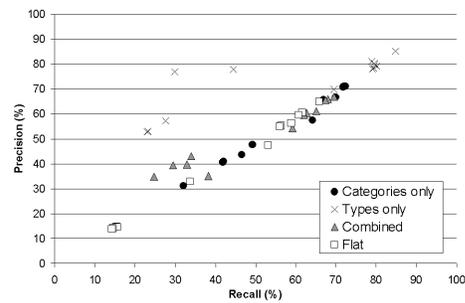


Fig. 3. Prec. vs Recall for Semantic Classification (Relative scenario)

had not been identified by the system), and relative (that is, relative only to correctly identified NEs). See [29, 30] for further details. We look forward to further studies in NER of Portuguese that employ this resource.

Acknowledgements

We are grateful to HAREM's organizing committee's other members and to all participants for feedback and enthusiasm, and especially to Cristina Mota for initiating the whole activity back in 2003. This work was supported by grant POSI/PLP/43931/2001 from Fundação para a Ciência e Tecnologia, co-financed by POSI.

References

1. Santos, D., Seco, N., Cardoso, N., Vilela, R.: HAREM: an Advanced NER Evaluation Contest for Portuguese. In: Proceedings of LREC'2006, Genoa, Italy (2006)
2. Hirschman, L.: The evolution of Evaluation: Lessons from the Message Understanding Conferences. *Computer Speech and Language* **12** (1998) 281–305
3. Santos, D.: Avaliação conjunta. In Santos, D., ed.: *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. (In press)
4. Santos, D., Barreiro, A.: On the problems of creating a consensual golden standard of inflected forms in Portuguese. In Lino et al, ed.: Proceedings of LREC'2004. (2004) 483–486
5. Santos, D., Costa, L., Rocha, P.: Cooperatively evaluating Portuguese morphology. In Nuno Mamede et al, ed.: *Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR*, Springer Verlag (2003) 259–266
6. Grisham, R., Sundheim, B.: Message Understanding Conference - 6: A Brief History. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), Copenhagen (1996) 466–471
7. Mota, C., Santos, D., Ranchhod, E.: Avaliação de reconhecimento de entidades mencionadas: princípio de AREM. In Santos, D., ed.: *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. (In press)
8. Rocha, P., Santos, D.: CLEF: Abrindo a porta à participação internacional em avaliação de RI do português. In Santos, D., ed.: *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. (In press)

9. Sang, E.F.T.K.: Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of CoNLL-2002, Taipei (2002) 155–158
10. Sang, E.F.T.K., Meulder, F.D.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proc. of CoNLL-2003, Edmonton (2003) 142–147
11. Lisa Ferro et al: TIDES 2003 Standard for the Annotation of Temporal Expressions. Technical report, MITRE (2004)
12. George Doddington et al: The Automatic Content Extraction (ACE) Program. Tasks, Data and Evaluation. In Lino et al, ed.: Proc. LREC 2004, Lisbon (2004) 837–840
13. Guthrie, L., Basili, R., Hajicova, E., Jelinek, F.: Beyond Entity Recognition - Semantic Labelling for NLP Tasks. In: Workshop proceedings, ELRA, Lisboa (2004)
14. Sekine, S., Sudo, K., Nobata, C.: Extended Named Entity Hierarchy. In González Rodríguez, M., Araujo, C.P.S., eds.: Proceedings LREC'2002, Las Palmas (2002) 1818–1824
15. Christian Bering et al: Corpora and evaluation tools for multilingual named entity grammar development. In Newman, S., Schirra, S.H., eds.: Proceedings of Multilingual Corpora Workshop at Corpus Linguistics 2003, Lancaster (2003) 43–52
16. Merchant, R., Okurowski, M.E., Chinchor, N.: The Multilingual Entity Task (met) overview. In: Proceedings of TIPSTER Text Program (Phase II), Tysons Corner, Virginia (1996)
17. Callmeier et al: COLLATE-Annotationsschema. Technical report, DFKI (2003) <http://www.coli.uni-sb.de/~erbach/pub/collate/AnnotationScheme.pdf>.
18. Arévalo, M., Carreras, X., Márquez, L., Martí, M.A., Padró, L., Simón, M.J.: A Proposal for Wide-Coverage Spanish Named Entity Recognition. *Revista da SEPLN* **1** (2002) 1–15
19. Kokkinakis, D.: Reducing the effect of name explosion. In Guthrie, L., Roberto Basili, Eva Hajicova, Frederick Jelinek, eds.: Beyond Named Entity Recognition - Semantic Labelling for NLP Tasks. Pre-conference Workshop at LREC'2004, Lisboa, Portugal (2004) 1–6
20. Karlgren, J., Cutting, D.: Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In: Proceedings of COLING 94, Kyoto, Japan (1994) 1071–1075
21. Santos, D.: Towards language-specific applications. *Machine Translation* **14** (1999) 83–112
22. Palmer, D.D., Day, D.S.: A Statistical Profile of the Named Entity Task. In: Proceedings of ANLP-97, Washington D.C. (1997) 190–193
23. Bick, E.: Multi-level NER for Portuguese in a CG framework. In Nuno Mamede et al., ed.: Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003, Springer Verlag (2003) 118–125
24. Mikheev, A., Moens, M., Grover, C.: Named Entity recognition without Gazetteers. In: Proceedings of EACL'99, Bergen (1999) 1–8
25. Santos, D.: The importance of vagueness in translation: Examples from English to Portuguese. *Romansk Forum* **5** (1997) 43–69
26. Calzolari, N., Corazzari, O.: Senseval/Romanseva: The Framework for Italian. *Computers and the Humanities* **34** (2000) 61–78
27. Macklovitch, E.: Where the Tagger Falts. In: Proc. of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation, Montréal (1992) 113–126
28. Voorhees, E.M., Tice, D.M.: Building a Question Answering Test Collection. In Nicholas Belkin et al, ed.: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens (2000) 200–207
29. Cardoso, N.: Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas. Master's thesis, FEUP, Porto, Portugal (2006) In preparation.
30. Seco, N., Santos, D., Cardoso, N., Vilela, R.: A complex evaluation architecture for HAREM. (This volume)