Capítulo 12

Functional aspects on Portuguese NER

Eckhard Bick

This chapter is republished, with kind permission from Springer-Verlag, from Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira & Maria Carmelita Dias (eds.), *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 2006. Proceedings*, LNAI series, Vol. 3960, pp. 80-89. ISBN-10: 3-540-34045-9.

Therefore, we restrained from doing any changes to the original text, even notational conventions, adding instead editors' notes commenting on possible mismatches.

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM, *a primeira avaliação conjunta na área*, Capítulo 12, p. 145–155, 2007.

The PALAVRAS-NER parser is a progressive-level Constraint Grammar (CG) system, treating Named Entity Recognition (NER) as an integrated task of grammatical tagging. The original version, presented at the PROPOR 2003 (Bick, 2003) and also used for Linguateca's avalia-SREC task 2003, implemented a basic tag set of 6 NER categories (person, organisation, place, event, semantic products and objects) with about 20 subcategories, following the guidelines of a joint Scandinavian NER project (Nomen Nescio (Johannessen et al., 2005)). Category tag candidates were added at three levels, and subsequently disambiguated by CG-rules:

a) known lexical entries and gazeteer lists (about 17.000 entries)

b) pattern-based name type prediction (morphological module)

c) context-based name type inference for unknown words

Since PALAVRAS originally was conceived primarily as a syntactic parser (Bick, 2000), it fuses fixed expressions with non-compositional syntactic-semantic function into multi--word expressions (MWEs), creating complex tokens and in the process making life easier for the token-based syntactic CG-rules as well as avoiding arbitrary descriptive decisions as to the internal structure of such MWE[1]. Names, too, are treated as MWEs, and semantic NER-classes are assigned to the whole, not the parts.

12.1 Recognizing MWE name chains

Identification of names, as a sequence of atomic tokens, was a separate task in the HAREM joined NER evaluation (www.linguateca.pt), and the PALAVRAS-system performed best, with an F-Score of 80.61%, in both the selective and total measures. Single-token names, with the exception of sentence-initial position, are clearly marked by upper case - therefore, since multi-token names can't be identified without chaining them into MWEs first, and since very few other (non-NE) cases involve productive MWE-chaining, the NE identification task is to a large degree identical to an MWE-recognition task[2]. The 2003 PALAVRAS--NER system (in this text, PAL-1), taking a more static approach, tried to fix MWE names *before* running the system's grammars – either by simple lexicon-lookup or by pattern-recognition in the preprocessor – and the only allowed post-grammar token alteration was fusion of adjacent name chains. This technique was replaced by a more dynamic, grammar based tokenisation approach in the new, 2005 system (henceforth, PAL-2), used for HAREM. Here, preprocessor-generated name candidate MWEs that cannot be verified in

[1] For corpus-users with a blank-space based token definition, MWEs can be unfolded and assigned an internal analysis by an add-on filter-program.

[2] Strictly speaking, the HAREM annotation and metrics did not employ MWEs per se, but rather XML-tags marking the start end end of name expressions. These XML tags were automatically added to PALAVRAS output before evaluation, at the same time turning semantic category tags into XML attributes.

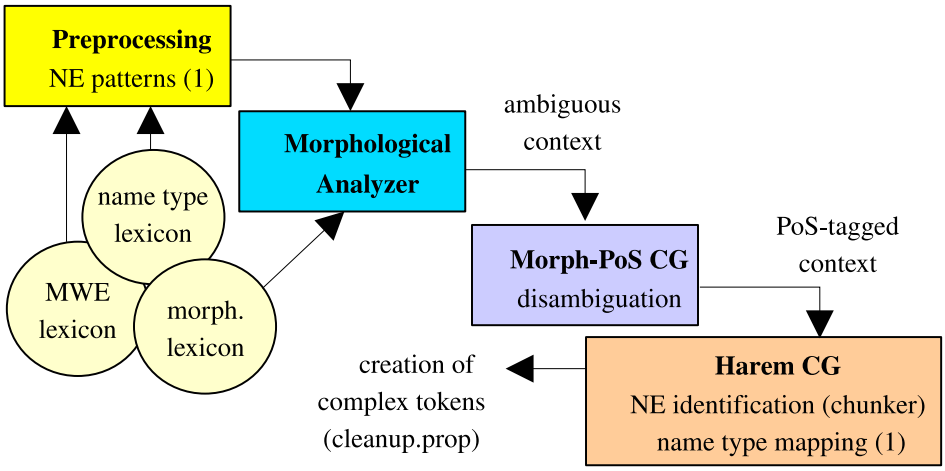

Figure 12.1: Name chain identification modules

the lexicon as either known names or non-name polylexicals, are fed to the morphological analyser not as a whole, but in individual token parts, with < and > tags indicating start and stop of name MWE candidates. Thus, parts of unknown name candidates will be individually tagged for word class, inflexion and - not least - semantic prototype class. In addition, each part is tagged either @prop1 (leftmost part) or @prop2 (middle and rightmost parts). This technique has two obvious advantages over the old approach:

1. It allows the morphological disambiguation grammar to establish the gender and number of names from their constituents, as well as internal morphological features, name-internal pp-constructions etc.

2. A specialized, newly-written name grammar can *change* the very composition of a name MWE, by removing, adding or replacing @prop1 start and @prop2 continuation tags.

For instance, the grammar can decide contextually whether sentence initial upper case is to be treated as a part of a name or not. Thus, certain word classes (prepositions, adverbs, conjunctions, finite verbs) can be recognized and tagged as no-name even with another upper case word to the right. Though a simple preprocessor might have sufficed to check for the closed classes, this is problematic due to ambiguity, and certainly not true of finite verbs, which are both open-class and often ambiguous with nouns, so the task has to be done after morphological analysis and disambiguation (illustration 12.1).

The name-chunker part of the Harem CG can progressively increase the length of a half-recognized chunk in a grammatically founded and context-sensitive way, for instance by adding conjuncts (e.g. the last two tokens in ... *Doenças Infecciosas e Parasitárias*, a1) or

PPs (e.g. the last part of *a Câmara Municipal de Leiria*, a2). Since the parts of name chains at this stage are "perspicuous" as nouns or other word classes, valency potential may be exploited directly (a3). In the rules below, the MAP operator adds information (tags) to a TARGET for a given context (1 meaning "one word to the right", -2 "two words to the left" etc.). BARRIER conditions can block a context if the barrier tag is found between the target and the context tag in question, while LINK conditions add secondary context conditions to an already instantiated context.

```
(a1)
MAP (@prop2) TARGET (KC) (-1 <prop2> LINK 0 ATTR) (1 <*> LINK 0 ATTR)
MAP (@prop2) TARGET <*> (0 ATTR) (-1 KC) (-2 <prop2> LINK 0 ATTR) ;
where <*> = upper case, KC = coordinator, ATTR = attribute

(a2)
MAP (@x @prop2) TARGET PRP-DE (*-1 N-INST BARRIER NON-ATTR LINK
0 <prop1>) (1PROP LINK 0 <civ> OR <top>)
MAP (@x @prop2) TARGET PROP (0 <civ> OR <top>) (-1 PRP-DE) (*-2 N-INST
BARRIER NON-ATTR LINK 0 <prop1>); where PROP = (atomic) proper noun, N-
INST = nouns with a semantic-prototype tag of institution, <civ> = known
civitas names, <top> = known place names, <prop1> = preprocessor-proposed
start of name chunk.

(a3)
MAP (@prop1) TARGET <*> (0 <+a>) (1 PRP-A) (NOT -1 >>>) ; where <+a> =
noun's or participle's binding potential for the preposition a, >>> =
sentence start
```

Not all name-part mapping rules are unambiguous - (a2), for instance, includes @x, meaning "wrongly assumed name part", along with @prop2, meaning "second part of name". Ultimately, a set of REMOVE and SELECT rules decides for each name part candidate if it is valid in context and if it is a first or later part of the chain. For instance, definite articles or the preposition *de* cannot be part of a name chain, if the token immediately to the right is not a second part candidate, or has been stripped of its name tag by another, earlier, rule:

```
REMOVE (@prop2) (0 <artd> OR PRP-DE LINK 0 @y) (NOT 1 @prop2)
```

The result, an unambiguous tag (@prop1=first part, @prop2=later part, @x=ex-name, @y=confirmed no-name) is implemented by a filter program, *cleanup.prop*, such that later programs and grammars will see only ready-made complex name tokens.

12.2 Semantic typing of name tokens: Lexematic versus functional NE categories

The next task, after identifying the name chain tokens, was to assign them a semantic category and subtype. The original PAL-1 did subdivide the 6 *Nomen Nescio* supercategories into subcategories, but recognized only about 17 partly experimental categories, while the new PAL-2 had to accommodate for HAREM's 9 categories and 41 subcategories[3]. This meant more than doubling the category inventory, and category matching was in many cases complicated by the fact that matches were not one-to-many, but many-to--many. This difference was not, however, the most important one. Far more crucial, both linguistically (i.e. in terms of descriptive meaning) and application ally (i.e. in terms of parsing grammars), was the treatment of metonymy. For many name types, metonymy is a systematic, productive and frequent phenomenon – thus, author names may be used to represent their works, city names may denote soccer clubs and a country name may be substituted for its government. Here, PAL-1 subscribed to a lexeme based definition of name categories, while HAREM used a function-based category definition. In the former tradition, a given name would have one, unchanging lexematic category, while in the latter it would change category according to context. Thus, the name of a country would always be `<civ>` (civitas) in PAL-1, a hybrid category of place and organisation, allowing, for instance, both +HUM subject-hood, and BE-IN-LOC-adverbiality. According to the HAREM guidelines, however, hybrid categories were not allowed[4] , and simply turning `<civ>` into `<top>` (place) would result in a considerable error rate in those cases, where the country--name *functions* as an organisation or a humanoid group, i.e. where it announces, suffers or goes to war. Likewise, institutions `<inst>` can be seen as both places and organisations, while the erstwhile `<media>` category implies a function-split between a newspaper being read (semantic product), burned (object) or sued in court (company). On the other hand, HAREM also introduced some distinctions that *were* lexematic rather than functional, for instance the split between the (money-making) *company* subtype and the non-profit institution subtype of the organisation category.

In order to handle the lexeme-function difference, PAL-2 had not only to increase its category inventory, but treat lexicon-, morphology- and pattern-derived categories as "potentialities" to a much higher degree than PAL-1 had done. 5 levels can be distinguished for such lexicon-dependence or -independence of name tagging:

1. lexicon-entered names that have a reasonably unambiguous name category (e.g. Christian names, to a lesser degree surnames, which can denote styles or an artist's

[3] **Editors' note.** There are 10 categories in HAREM; the author is here ignoring the `VARIADO` category.

[4] **Editors' note.** A little precision is in order here: Since no system at the First HAREM reported that it would use the OR notation (in this case, `LOCAL | ORGANIZACAO`) in its output, "hybrid" categories were only used in the golden collection. In fact, the PALAVRAS-NER system could have used them, but then it would still not fare well in the cases where the golden resource had only `LOCAL` or `ORGANIZACAO`, which we believe to be Eckhard Bick's main message in this context.

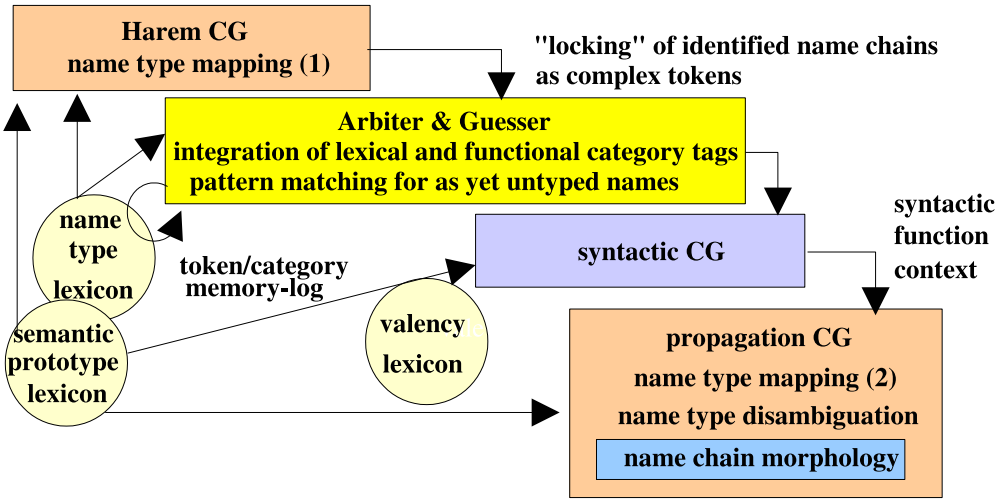

Figure 12.2: Name typing modules

collected work)

2. lexicon-entered names with semantically hybrid categories (<civ>, <media>, <inst>) or with systematic metaphoring (<brand> as <object>)
3. pattern/morphology-matched names of type (1)
4. pattern/morphology-matched names of type (2)
5. names recognized as such (upper case, name chaining), but without a lexicon entry or a category-specific pattern/morphology-match

Even in the PAL-1 evaluation (Bick, 2003), where hybrid categories did not have to be resolved and where only few, strong rules were allowed to override lexicon- or gazeteer--supported name-readings (1. and 2.), this group had an error rate of 5%, indicating that for many names, ambiguity is not merely functional, but already hard-wired in the lexicon (e.g. *Washington* as person or place name). In PAL-2, lexicon-derived categories were treated as contextual indications only, and the names carrying them were submitted to the same rule set as "unknown" names (3. - 5.), opening up for considerably more ambiguity and a correspondingly higher error risk.

Illustration 12.2 shows the distributed nature of PAL-2 and the interaction of its different name typing modules. An essential cut, the "locking"of identified name chains into complex tokens, is made between the (new) Harem CG on the one hand and the (modified) syntactic module and propagation CG on the other. While the former (*micromapping*) works on minimal tokens (name-part words) and can exploit their PoS, semantics and

morphology, this is not any longer possible for the latter, which is geared for syntactic clarity and therefore works on whole name chunks, and uses syntactic function and structure to "propagate" information from the rest of the sentence onto nouns (*macromapping*).

12.2.1 Micromapping: Name type rules based on name parts and patterns

Many of the micromapper's rules map chunking information at the same time as classifier tags, like in the following rule which types known towns or countries (<civ>) or typical *noun parts* (N-CIVITAS) of unknown towns or countries as "administrative", if they fill the subject slot of a human-agent or experiencer verb (V-HUM).

```
MAP (@admin @prop1) TARGET <*> (0 <civ> OR N-CIVITAS) (*1 V-NONAD
BARRIER CLB LINK 0 V-HUM) (NOT 0 <prop2>)
```

It is the first part of a complex name (@prop1) that will carry the classifier tag (@admin), and both tag types may be mapped ambiguously for later rule based disambiguation. Once output from the micromapper CG has been "frozen"into name chunks, the Arbiter module checks the result against lexical data and morphological patterns, adding pattern based classifier tags where no category has been mapped, or where tags are marked as unsafe (e.g. <hum?>) by the pre-CG inflexion and derivation analyzer. The Arbiter is the only part of the system that has a text-level memory - logging identified names and their types to resolve the classification of name abbreviations and the gender of person names. Thus, on a small scale, entity disambiguation is used for NE typing as suggested by Blume (2005).

The Pal-1 based morphological analyzer only treats numbers as NE material if they are part of a larger NE, e.g. time and place units, not when occurring as mere quantifiers, as in the HAREM categories[5] of QUANTIDADE, CLASSIFICACAO and MOEDA. In PAL-2, it is the Arbiter's pattern-matching module, not the "character-blind" CG, who has to recognize such number expressions as names, as well as pre-classify them for later treatment in the CG macromapper.

12.2.2 Macromapping: Name type rules based on syntactic propagation

Macromapping is an adapted PAL-1 module that adds name type tags to already-identified name chains by using a number of syntactic "propagation"techniques (described in Bick (2003)), exploiting semantic information elsewhere in the sentence:

1. *Cross-nominal prototype transfer*: Postnominal or predicative names (NE @N<, PRP @N< + NE @P<, @SC, @OC) inherit the semantic type through of their noun-head

[5] **Editors' note.** We used the denomination "categories" for what the author refers as "major categories" elsewhere in this text, and "types" for "subcategories". So, in this case, the author is referring to HAREM types, and not categories.

2. Coordination based type inference: Types are propagated between conjuncts, if one has been determined, the other(s) inherit the same type.

3. Selection restrictions: Types are selected according to semantic argument restrictions, i.e. +HUM for (name) subjects of speech- and cognitive verbs, +TIME is selected after temporal prepositions etc.

In Constraint Grammar terms, macromapping is as much a mapping technique as a disambiguation technique, as becomes particularly clear from method (3), where many rules discard whole sets of name type categories by targeting an atomic semantic feature (+HUM or +TIME) shared by the whole group.

12.3 Evaluation

The complete HAREM evaluation computed a number of other metrics, such as text type dependent performance. PAL-2 came out on top for both European and Brazilian Portuguese, but in spite of its Brazilian-optimized lexicon and syntactic parser, it achieved a higher F-Score for the latter (60.3% vs. 54.7%), possibly reflecting sociolinguistic factors like the higher variation of person names in a traditional immigration country like Brazil, its Tupi-based place names etc. all of which hamper regular pattern/morphology-based name type recognition[6]. HAREM also had separate selective scores, where systems were allowed to compete only for certain categories and skip others. However, since PAL-2 competed globally in all areas, selective scores equaled total scores.

Another HAREM measure not presented in the overview table were relative performance, defined as category recognition measure separately for only those NEs that were correctly identified. Since this was not done by presenting systems with a ready-chunked ("gold-chunk-") corpus, but by measuring only against NEs correctly recognized by the system itself, PAL-2 had the relative disadvantage of being the best identifier and thus having to cope also with a larger proportion of difficult names than other systems, resulting in suboptimal rank performance.

For a direct performance comparison between PAL-1 and PAL-2, only the per-category scores are relevant, since even if subcategory scores had been available for PAL-1, score differences might simply reflect the difference in type set size. Even so, however, scores neither matched nor differed systematically. Of the major categories, *person* and *place* scored better in PAL-2/HAREM than what was published for the lexeme-based approach in PAL-1 (Bick 2003), while *organisation* and *event* had lower scores. Interestingly, the major categories (person, organisation, place) even *ranked* differently, with *person* higher (lowest in PAL-1) and organisation lowest (second in PAL-1). The reason for this may reside in the

[6] Alas, since all HAREM participants but the winner were anonymous, and different code names were used for the Brazilian and Lusitan evaluation, this pattern could not at the time of writing be verified as either general or system-specific.

PALAVRAS Subtype	Category (incidence)	HAREM Subtype	F-Score (precision - recall) cat total	cat/types total	identification
hum official member grupoind groupofficial grouporg	**hum** PESSOA 20.5%	INDIVIDUAL CARGO MEMBRO GRUPOIND GRUPOCARGO GRUPOMEMBRO	**67.4** 61.1-75.2 rank 1	**65.6** 59.3-73.4 rank 1	**65.0** 58.6-72.7 rank 1
admin inst, party org suborg	**org** ORGANIZACAO 19.1%	ADMINISTR. INSTITUICAO EMPRESA SUB	**58.7** 53.3-65.4 rank 1	**50.0** 45.3-55.9 rank 1	**56.3** 51.0-62.7 rank 1
date hour period cyclic	TEMPO 8.6%	DATA HORA PERIODO CICLICO	**75.5** 79.8-71.7 rank 1	**72.2** 76.1-68.7 rank 1	**73.5** 77.7-69.8 rank 1
address admin top virtual site	**top** LOCAL 24.8%	CORREIO ADMINISTR. GEOGRAFICO VIRTUAL ALARGADO	**69.6** 75.1-64.8 rank 3	**64.3** 69.4-59.9 rank 4	**68.6** 74.1-63.9 rank 3
product, V copy, tit artwork pub	**tit** OBRA 4.3%	PRODUTO REPRODUZIDO ARTE PUBLICACAO	**21.3** 22.3-20.4 rank 1	**16.5** 17.3-15.8 rank 2	**19.7** 20.6-18.9 rank 1
history occ event	**event** ACONTECIMENTO 2.4%	EFEMERIDE ORGANIZADO EVENTO	**36.2** 28.9-48.6 rank 4	**30.8** 24.6-41.3 rank 4	**32.7** 26.0-43.8 rank 4
genre,brand, disease,idea, school,plan, author,abs-n	**brand** ABSTRACCAO 9.2%	DISCIPLINA,MARCA, ESTADO,IDEIA, ESCOLA,PLANO, OBRA,NOME	**43.1** 47.3-39.6 rank 1	**39.6** 43.3-36.4 rank 1	**41.4** 45.4-38.0 rank 1
object mat class,plant	**object** COISA 1.6%	OBJECTO SUBSTANCIA CLASSE	**31.3** 25.4-40.7 rank 1	**31.2** 25.5-40.3 rank 1	**31.3** 25.4-40.7 rank 1
prednum quantity currency	VALOR 9.5%	CLASSIFICADO QUANTIDADE MOEDA	**84.3** 87.0-81.7 rank 1	**82.5** 84.8-80.2 rank 1	**82.2** 84.8-79.7 rank 1

Table 12.1: Global HAREM results for PALAVRAS-NER, semantic classification absolute/total (i.e. all NE, identified or not) combined metric for 9 categories and 41 subcategories (types)

HAREM Category	combined		per category		PAL-1 F-Score
	Precision -recall	F-Score (rank)	Precision -recall	F-Score (rank)	
PESSOA	90.1-91.9	91.0 (3)	92.7-94.0	93.4 (3)	92.5
ORGANIZACAO	77.0-79.0	78.0 (5)	91.1-92.4	91.8 (7)	94.3
LOCAL	87.7-89.3	88.5 (7)	96.1-95.5	95.8 (5)	95.1
OBRA (tit, brand, V)	58.5-59.5	59.0 (3)	75.3-76.6	76.0 (3)	ABSTRACT 84.3 (tit, genre, ling) OBJECT: 57.1 (brand, V, mat)
ABSTRACCAO (genre, ling)	82.6-85.6	84.1 (1)	90.5-93.2	91.8 (1)	
COISA (brand, V, mat)	98.8-98.8	98.8 (1)	100-100	100 (1)	
ACONTECIMENTO	69.6-72.6	71.1 (5)	81.9-85.4	83.6 (5)	88.7
TEMPO	91.5-91.5	91.5 (4)	96.8-95.5	95.8 (5)	-
VALOR	94.2-95.8	95.0 (1)	96.6-97.6	97.1 (1)	-

Table 12.2: Relative HAREM performance of PAL-2.

fact that the function of human names is much more likely to stick to its lexeme category, while organisations frequently *function* as either human agents or place names[7]. The abstract and object categories of PAL-1 were not directly comparable to the ABSTRACCAO and COISA categories of HAREM, since the latter also had OBRA, drawing (book etc.) titles from PAL-1's *abstract* category and brands (unless *functioning* as objects) from the *object* category, with a number of minor subcategories and function distinctions further complicating this 2-to-3 category match.

12.4 Conclusion: Comparison with other systems

Though state-of-the-art NER systems make often make use of lexical and grammatical information, as well as extra-textual gazetteer knowledge, most do so in a framework of data-driven statistical learning, using techniques such as HMM, Maximum Entropy, Memory or Transformation-based Learning. The statistical learning approach has obvious advantages where language independence is desired, as in the CoNLL2002 and CoNLL2003 shared tasks (Sang, 2002; Sang e Meulder, 2003), but language-specific systems or subsystems may profit from explicit linguistic knowledge (hand-written rules or lexica), as e.g. in a number of Scandinavian NER systems (Bick (2004) and Johannessen et al. (2005)). Petasis et al. (2004) describes a 4-language NERC system with hybrid methodology, where the French section relies on human modification of rules machine-learned from an human-annotated corpus. PALAVRAS-NER stands out by being entirely based on hand-written rules, both locally (morphological pattern recognition) and globally (sentence context) - not only in assigning the grammatical tags used as context by the NER-system, but also within the latter itself. However, though PAL-2's rule based method worked best in the Portuguese HAREM context, with overall F-Scores of 80.6 for identification and 63.0/68.3 for abso-

[7] the *commercial* vs. *administrative* distinction also increases PAL-2's error risk

lute/relative category classification, it is difficult to compare results to those achieved for other languages, due to differences in metrics and category set size. In the CoNLL shared tasks on newspaper-text, the best absolute F-scores were 88.8 (English), 81.4 (Spanish), 77.1 (Dutch) and 72.4 (German) for a 3-way category distinction: *person*, *organisation*, *place* (plus *miscellaneous*), and given PALAVRAS-NER's high *relative* scores for these categories (93.4, 91.8 and 95.8), its lower total scores may well be due to suboptimal identification, reflecting either shortcomings of the PAL-2 rule system in this respect or linguistic-descriptive differences between the gold-standard CD and PALAVRAS-NER[8]. However, it is not at all clear how the CoNLL systems would have performed on a large (41) subcategory set and HAREM style mixed-genre data[9]. On the other hand, HAREM's category-specific and relative rank scores clearly show that there is much room for improvement in Pal-2, especially for the place and event categories, where it didn't rank highest (Table 12.1). Also, Pal-2 appears to be *relatively* better at name chunk identification than at classification, since it ranked lower in the relative scores (on correct chunks only) than in the absolute scores (identification task included). However, improvements do not necessarily have to be Pal-2-internal: Given an integrated research environment and a modular perspective (for instance, a cgi-integrated web-interface), a joined Portuguese HAREM system could act on these findings by delegating the identification and classification tasks to different systems and by applying weighted votings to exploit the individual strengths of specific systems, thus seamlessly integrating rule based and statistical systems.

Acknowledgments

The authors would like to thank the Linguateca team for planning, preparing, organising and documenting HAREM, and for making available a multitude of evaluation metrics in a clear and accessible format.

[8] Such differences are particularly relevant for a system built by hand, not from training data. Thus, PAL-1 made far fewer chunking errors when evaluated internally (Bick, 2003).

[9] The MUC-7 MENE-system (Borthwick et al., 1998), for instance, experienced an F-Score drop from 92.2 to 84.2 even within the same (newspaper) genre, when measured not on the training topic domain, but in a cross-topic test.