Functional Aspects in Portuguese NER

Eckhard Bick
Institute of Language and Communication
University of Southern Denmark
eckhard.bick@mail.dk

Outline

- Two versions of a rule based NER system:
  lexematic versus functional
- Identification of name chains: Name part mapping and disambiguation
- Semantic classification: Lexical vs. contextual
- Micromapping and macromapping
- Evaluation results
- Perspectives

Introduction

- State-of-the-art NER systems often use lexical and grammatical information, as well as extra-textual gazeteer knowledge
- BUT: Most do so in a framework of **data-driven statistical learning** (HMM, Maximum Entropy, Memory based or Transformation based learning)
- While this is fine where **language independence** is desired (e.g. CoNLL shared tasks 2002 &2003), **language-specific** systems or subsystems may well profit from explicit linguistic knowledge (i.e. Hand-written rules or lexica), e.g. Johannesen et al. 2005 (CG), Petasis 2004 (human rule-modification)
- The system presented here (PALAVRAS-NER) is an extreme case, since it is entirely based on hand-written rules, both locally and globally (sentence contex), not only in assigning grammatical tags for use by the NER system, but also within the latter itself

Previous work: Pal-1 NER

- Based on a syntactic CG parser (PALAVRAS)
- NER-module for PROPOR '03, Linguateca's avalia-SREC (03)
- 6 basic name categories (recommended by Nomen Nescio project)
- Names as MWEs (with categories assigned to the whole, not the parts)
- Category assignment (for later CG-disambiguation) at 3 levels
  - Known lexical entries and gazeteer lists (ca. 17.000)
  - Pattern-based name type prediction (morphological module)

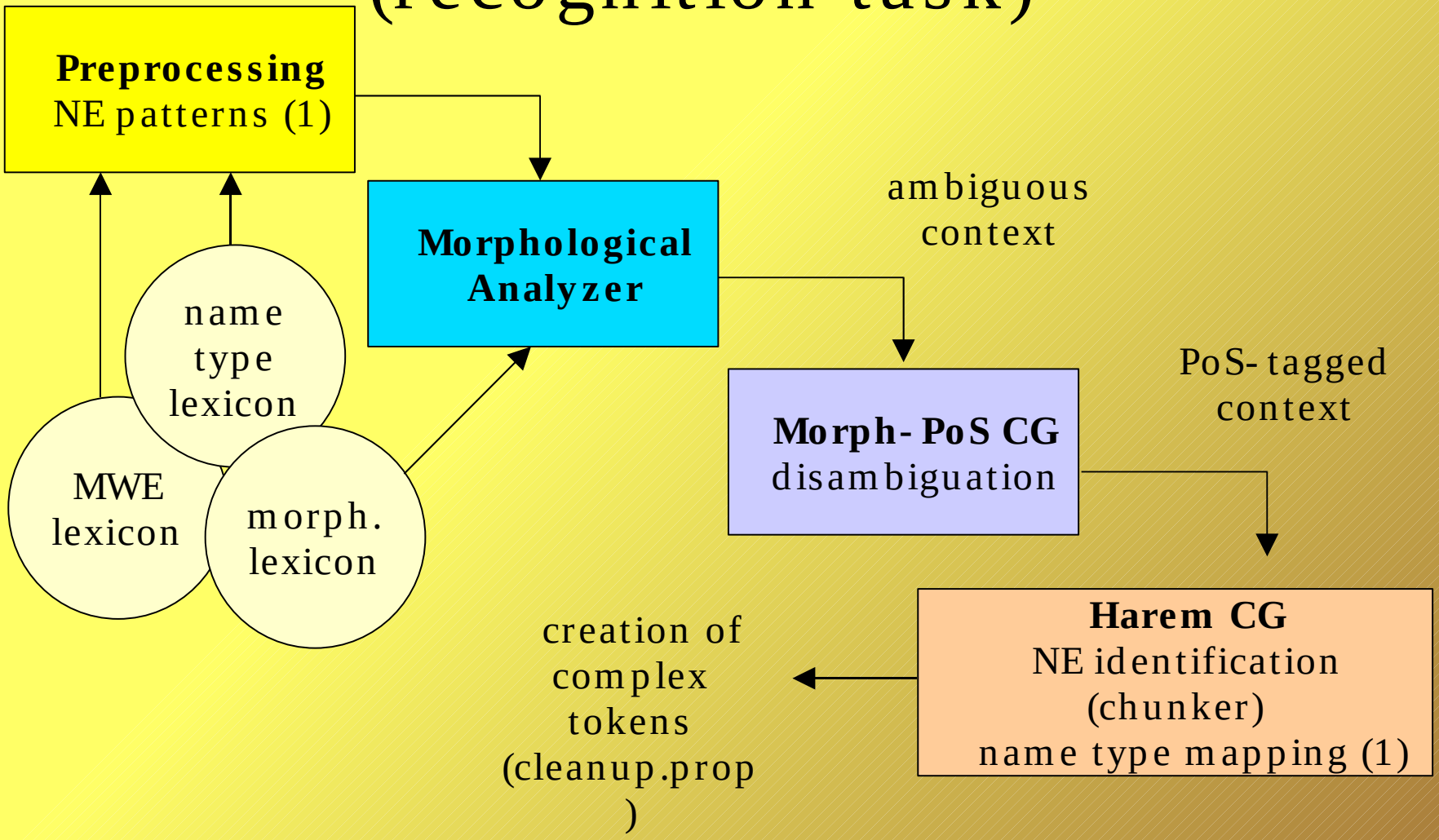
Core changes in Pal-2 NER

- Extension to over 40 NER categories (vs. 6/20)
- Change from lexeme-based to a token-based description: functional and context-dependent categories rather than stable lexematic categories
- As a consequence – substantial changes in the rule body, as well as remapping of also lexically known material
- Pattern-based name chain recognition now enhanced by rule-based name-

NE recognition as MWE recognition

- With the exception of sentence-initial position (PoS disambiguation problem), NE identification means MWE recognition
- Pal-1: preprocessor tokenisation,
  Pal-2: dynamic, grammar-based tokenisation:
  --> 1. pattern guess
  --> 2. lexicon check
  --> 3. MWE candidate parts are analysed individually (PoS, morphology, semantic prototype), allowing contextual chaining with BEGIN (@prop1) and CONTINUE (@prop2) tags
- Advantages:
  1. Analyzer can “conclude” gender and number from parts
  2. a special grammar can *change* the very composition of a name MWE, by removing, adding or replacing @prop1 and @prop2 continuation tags

Name chain identification modules (recognition task)



Name part mapping rules

Pal-2 can progressively increase the length of a half-recognized NE chunk in a grammatically founded and context-sensitive way by

- Adding conjuncts: *Doencas Infecciosas e Parasitárias*
  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)
- Adding pp's: *a Câmara Municipal de Leiria*
  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>)
- Exploiting valency:
  MAP (@prop1) TARGET <*> (0 <+a>) (1 PRP-A) (NOT

Name part disambiguation rules

REMOVE and SELECT rules decide for each name part candidate if it is valid in context and if it is a first (@prop1) or later (@prop2) part of the chain, a "misassumed" (i.e. ex-) name part (@x) or a confirmed no-name (@y)

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

HAREM- results:

F-Score of 80.61% in both the selective and total measures

Semantic typing (classification task)

- 6 super- and 17 partly experimental subcategories (Pal-1) had to be turned into 9 super- and 41 subcategories (HAREM, Pal-2)
- Many-to-many relation between categories, new areas (e.g. numbers as names)
- Descriptive and methodological problem: Metonymy
  - Lexematic view: <civ> = place *and* organisation, allowing both +HUM subjecthood *and* BE-IN-LOC adverbiality.
  - Functional view: <civ> unmappable, since mapping it into <top> would result in errors where a country *acts* like a humanoid group.

5 levels of lexicon (in)dependence

(1) Lexicon-entered names with a reasonably **unambiguous name category** (e.g. Christian names, but not surnames -> styles, work of art)

(2) Lexicon-entered names with semantically **hybrid categories** (<civ>, <medi>, <inst>) or with systematic mataphoring (<brand> as <object>)

(3) pattern/morphology-matched names of type (1)

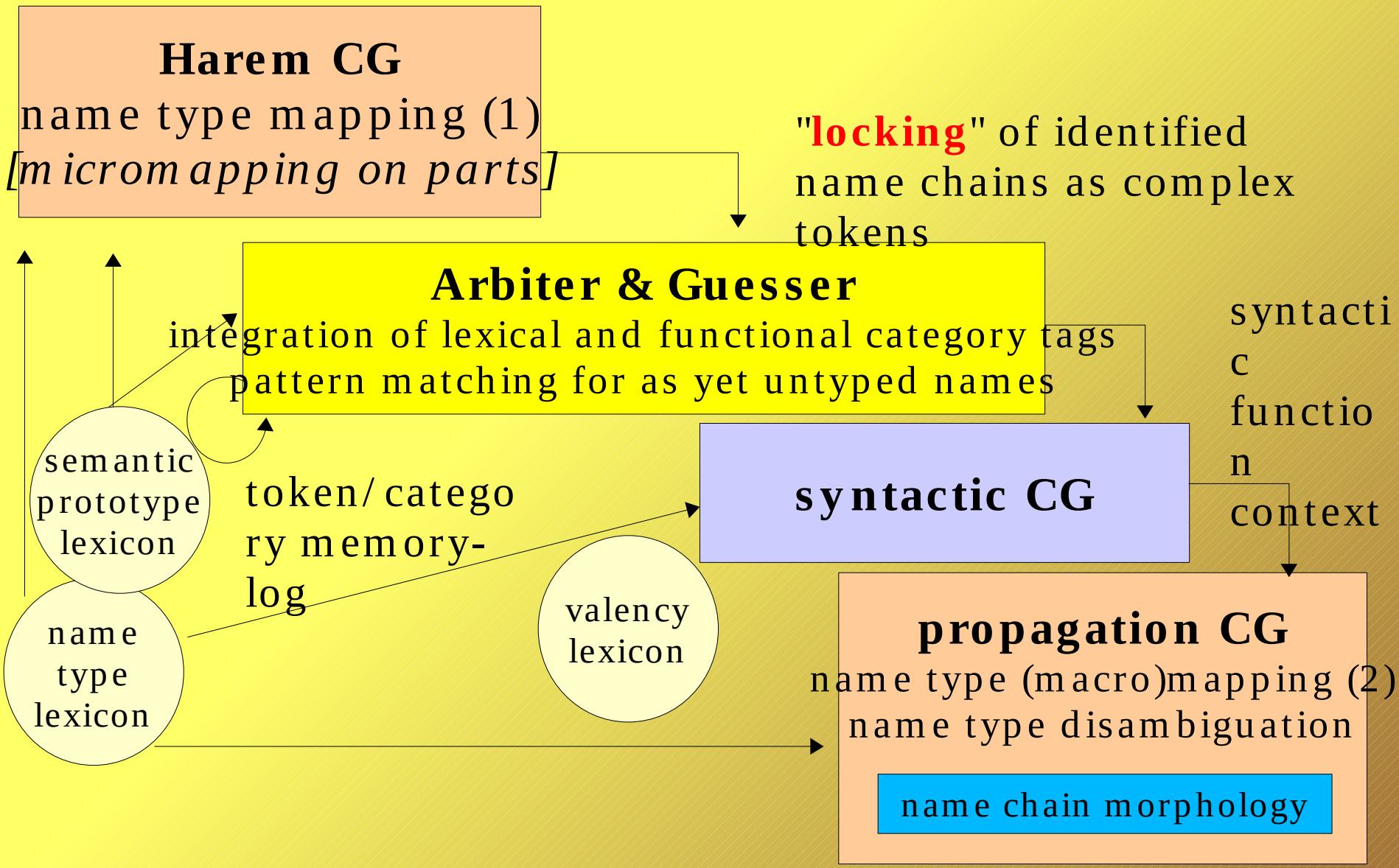
(4) pattern/morphology-matched names of type (2)

(5) Names recognized as such (upper case, name chaining), but **without a lexicon entry** or a category-specific pattern/morphology match

Pal-1: "hard-wired" ambiguities, only few override rules, 5% errors for lexicon-derived material

Pal-2: lexicon-derived categories are weighted as heuristic indications only, "known" (1-2) and "unknown" (3-5) names are submitted to the same rules --> higher ambiguity and correspondingly higher error risk

Name typing modules (identification task)



Micromapping

(name type rules based on name parts and patterns)

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

- first NE part carries type tag, **type information** and **chunking information** can be mapped at the same time
- After "freezing" NE chunks, the *Arbiter* checks unsafe (e.g. <hum?>) or nil-readings against lexicon data and morphological patterns
- The *Arbiter* logs names and types to help resolve e.g. Abbreviations and gender for person names
- Number expressions (= names in HAREM, even when quantifiers) need to be micromapped, because CG – so far – is "character-blind"

Macromapping

(Name type rules based on syntactic propagation)

Adds name type tags to already-identified name chains by using a number os syntactic "propagation" techniques (adapted Pal-1), exploiting semantic information elsewhere in the sentence, plus the fact that besides functioning as subjects and objects like other np's, names can fill certain more specific syntactic slots:

- @N< (valency governed nominal dependents): *o presidente americano* ***George Bush***
- @APP (identifying appositions): *uma moradora do palácio,* ***Júlia Duarte****, ...*
- @N< PRED (predicating appositions)
- Os marchadores Susana Feitor ***(CN Rio Maior)*** e José Magalhães ***(Alfenense)***

Macromapping 2

***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

- MAP (<top>) TARGET (PROP @N<) (-1 N-TOP) ;
- MAP (<top>) TARGET (PROP @P<) (-1("de" PRP @N<) (-2 N-TOP) ;
- SELECT (<top>) (0 @SUBJ>) (*1 @<SC BARRIER @SUBJ LINK 0 N-TOP) ;

More detailed rules can match such information between main and relative clauses.

Macromapping 3

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

- the syntactic module supplies a secondary tag for "close/safe coordinators" (&KC-CLOSE), with one rule for each matched syntactic function, then uses it for disambiguation:
- REMOVE %non-h (0 %hum-all) (*-1 &KC-CLOSE BARRIER @NON->N LINK *-1C %hum OR N-HUM BARRIER @NON-N<);
- SELECT (<top>) (1 &KC-CLOSE) (*2C <top> BARRIER @NON->N) ;

Macromapping 4

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.

- *REMOVE %non-hum*
  (0 @SUBJ> LINK 0 %hum-all)
  *(*1 @MV BARRIER ser/estar/ficar LINK 0 V-HUM);*
  [@MV = main verb, @SUBJ = subject]

- *REMOVE %non-org*
  (0 @<ACC LINK 0 %org/inst) (-1 @MV LINK 0 V-ADMIN);*
  [@<ACC = accusative/direct object]

CG: macromapping is both mapping and disambiguation, cf. (3), where many rules discard whole sets of name type categories by targeting an *atomic semantic feature* (+ HUME or + TIME) shared by the whole group.

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)

PALAVRAS Subtype	**Category** (incidence)	**F-Score** (precision - recall)		
		cat total	**cat/types total**	**identification**
hum official member groupind groupoffici grouporg	**hum** PESSOA 20.5 %	**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 %	**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 %	**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	**top** LOCAL 24.8 %	**69.6** 75.1-64.8 rank 3	**64.3** 69.4-59.9 rank 4	**68.6** 74.1-63.9 rank 3

PALAVRAS Subtype	Category (incidence)	F-Score (precision - recall)		
		cat total	**cat/types total**	**identification**
product, V copy, tit artwork pub	**tit** OBRA 4.3 %	**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 %	**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 %	**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 %	**31.3** 25.4-40.7 rank 1	**31.2** 25.5-40.3 rank 1	**31.3** 25.4-40.7 rank 1

Other metrics

- **Selective** = total (Pal-2 participated for all categories)
- **European > Brazilian** (F 60.3 vs. 54.7 %): general or system-specific?

  pattern and rule problems with immigrant names, TUPI-place names?
- **Relative performance** (typing accuracy measured for correctly recognized

  names only (possible disadvantage for a good recognizer, because it will

  get a larger proportion of difficult names than other systems)

Pal-1 versus Pal-2 performance

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)
ABSTR. (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)	-

Conclusion

- It was possible succesfully to change a rule-based NER system from lexematic categories to functional categories
- The system had the overall best F-scores in the HAREM evaluation: 80.6 for identification and 63.0 and 68.3 for total types and category classification, BUT: performance is uneven (event and place score lower than the rest)
- Performance was lower than the best CoNLL-results (F 88.8 English, 81.4 Spanish, 77.1 Dutch, 72.4 German), BUT: CoNLL used a different metric and did a 3-way distinction only (hum, org, top + misc.), not 41 (!) like HAREM, and not as genre-mixed
- Since Pal-2 has high *relative* scores (over 90) for the 3 CoNLL categories, its identification module is a crucial candidate for improvement

Improvement strategies

- Identify strengths and weaknesses of subsystems
- If necessary, delegate the identification and classification tasks to different (sub)systems (possibly across research-groups)
- Integrate rule-based and statistical systems through a weighting scheme

Acknowledgments

I 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.