

Segundo HAREM, ReReIEM e LAMPADA 2.0

Cláudia Freitas

Linguateca/PUC-Rio

HAREM

Avaliação e Reconhecimento de Entidades Mencionadas

avaliação conjunta:

“modelo de avaliação em que vários grupos comparam, com base num conjunto de tarefas consensuais, o progresso dos seus sistemas numa dada área, usando para isso um **conjunto de recursos** comum e uma métrica consensual.”

HAREM

Características principais (Santos, 2007b)

I. Modelo semântico

□ EM em contexto

Regressou então a **<CATEG="LOCAL">Portugal**, onde iniciou
A radiação de origem cósmica, prevista pelo **<CATEG="ABSTRACCAO">**
meteorica carreira...
**Big Bang **, seria descoberta em 1964...

O acordo político foi obtido durante a Presidência Alemã, tendo cabido a
<CATEG="ORGANIZACAO">Portugal concluir o processo de
revisão.

... pelo qual tem início a expansão das galáxias que os cosmologistas
«descrevem como uma explosão a que se dá um
<CATEG="ABSTRACCAO">Portugal ou dois dentro de si»

<CATEG="PESSOA">Portugal perdeu com a Suíça por 2-0

Características principais (Santos, 2007b)

I. Modelo semântico

→ NE classificadas em contexto

*A morte é reportada no **Diário de Notícias** do dia*

('The death is announced in Diário de Notícias')

→ LOCAL VIRTUAL COMSOC / place

*A diferença entre o 'Jornal de Notícias' e o '**Diário de Notícias**'*

('The difference between Jornal de Notícias and Diário de Notícias')

→ COISA CLASSE / thing

*O seu pai era funcionário público do Ministério da Justiça e crítico musical do '**Diário de Notícias**'*

('His father was an employee of the Ministry of the Justice and a music reviewer for Diário de Notícias')

→ ORGANIZACAO EMPRESA/ org

*... foi fotografado pelo **Diário de Notícias** (DN) a fumar uma cigarrilha...*

('had a picture taken by Diário de Notícias smoking a cigarette')

→ PESSOA GRUPOMEMBRO / person

Características principais (Santos, 2007b)

II. Vagueza

→ Uma EM pode receber simultaneamente mais de uma classificação

Mais de 32 mil pessoas poderiam morrer se uma pandemia de gripe aviária atingisse

<CATEG="PESSOA|LOCAL" TIPO="POVO|HUMANO">Portugal

Assim aceitam sacramentos do <CATEG="ABSTRACCAO|OBRA" TIPO="IDEIA|PLANO">Evangelho : o Santo Batismo, através do qual....

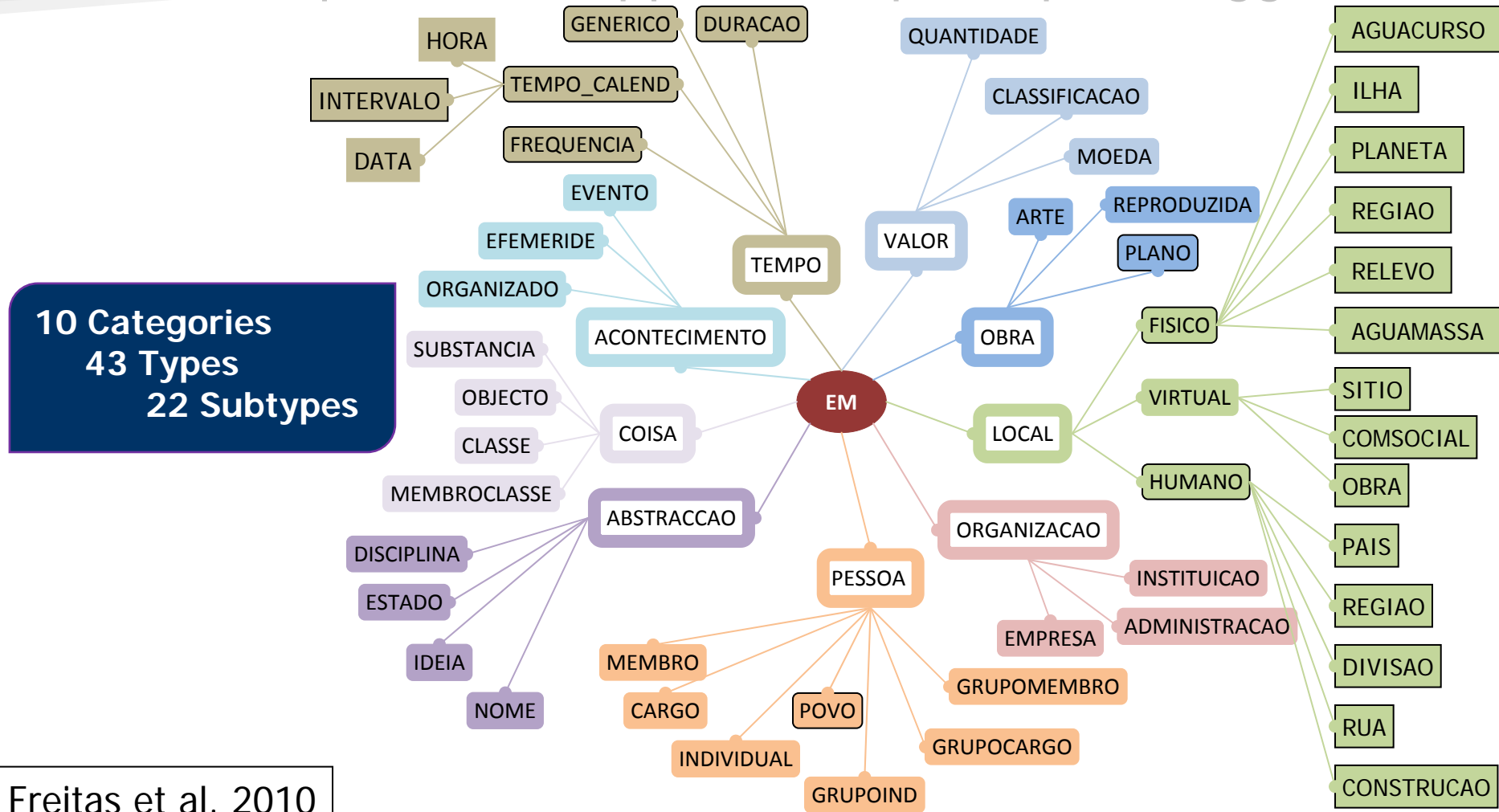
O carácter diferente da <CATEG="ABSTRACCAO/ACONTECIMENTO" TIPO="IDEIA/EFEMERIDE">Reforma Inglesa deve-se ao facto de ter sido promovida pelas necessidades políticas de Henrique VIII.

A <CATEG="PESSOA/ORGANIZACAO" TIPO="GRUPOIND/ADMINISTRACAO">Administração Bush identifica-se com a justiça divina

Características principais (Santos, 2007b)

III. Categorias

→ Initial corpus-based approach + participant suggestions

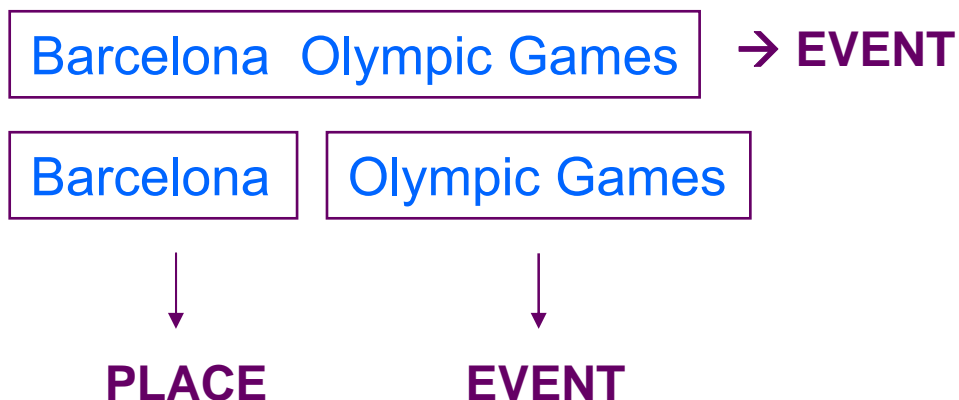


Características principais (Santos, 2007b)

IV. EM encaixadas

→ ALT mechanism

Quantos atletas participaram nos Jogos Olímpicos de Barcelona?



```
<ALT><Jogos Olímpicos de Barcelona |  
<Jogos Olímpicos> de <Barcelona>  
</ALT>
```

Nova tarefa - ReReIEM

Anaphora resolution

Mitkov, 2000; Colloveni et al., 2007; de Souza et al. 2008

Focused on co-reference
Anaphoric chains in texts

+

Relation detection

Agichtein and Gravano, 2000; Zhao and Grishman, 2005; Culotta and Sorensen, 2004

Fact extraction
World knowledge

Investigar quais as relações
seriam encontradas nos textos

Criar tarefa piloto que
comparasse os sistemas que
reconhecem essas relações

=

ReReIEM

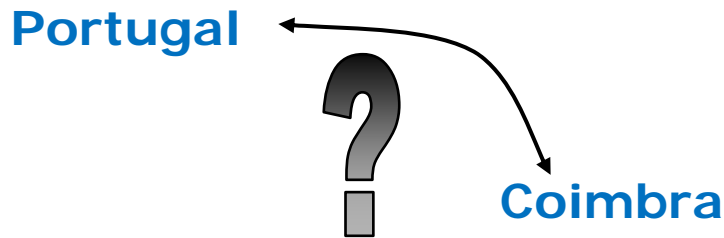
Reconhecimento de Relações entre Entidades Mencionadas

Relation detection between named entities

ReReIEM ↔ HAREM

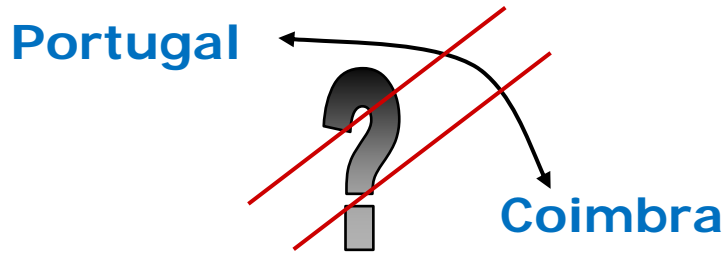
Portugal perdeu para a **Alemanha** nas quartas de final da **Eurocopa**. Vi o jogo na **Praça da República**, e mesmo com a derrota os bares de **Coimbra** continuaram cheios.

HAREM → ReReIEM



<EM ID="h-37" CATEG="PESSOA" TIPO="GRUPOMEMBRO">Portugal perdeu para a <EM ID="h-38" CATEG="PESSOA" TIPO="GRUPOMEMBRO">Alemanha na <EM ID="h-39" CATEG="ACONTECIMENTO" TIPO="ORGANIZADO">Eurocopa. Vi o jogo na <EM ID="h-40" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="RUA">Praça da República, e mesmo com a derrota os bares de <EM ID="h-41" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="DIVISAO">Coimbra estavam cheios.

HAREM → ReReIEM



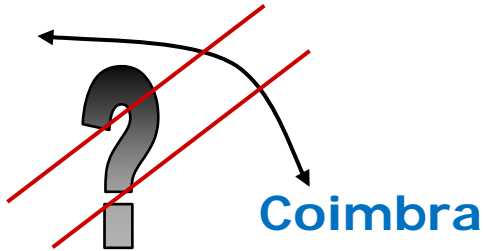
<EM ID="h-37" CATEG="PESSOA" TIPO="GRUPOMEMBRO">Portugal

<EM ID="h-41" CATEG="LOCAL" TIPO="HUMANO"

SUBTIPO="DIVISAO">Coimbra

HAREM → ReReIEM

Portugal



Coimbra

Portugal	→	Eurocopa
Alemanha	→	Eurocopa
Pr. da República	→	Coimbra

<EM ID="h-37" CATEG="PESSOA" TIPO="GRUPOMEMBRO">Portugal perdeu para a <EM ID="h-38" CATEG="PESSOA" TIPO="GRUPOMEMBRO">Alemanha na <EM ID="h-39" CATEG="ACONTECIMENTO" TIPO="ORGANIZADQ">Eurocopa. Vi o jogo na <EM ID="h-40" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="RUA">Praça da República, e mesmo com a derrota os bares de <EM ID="h-41" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="DIVISAO">Coimbra estavam cheios.

ReReIEM – o que anotar?

Portugal

Alemanha

Pr. da República

INCLUSÃO

Portugal

Alemanha

Pr. da República

Eurocopa → Eurocopa

Eurocopa → Eurocopa

Pr. da República → Coimbra

Depois de ser exibida no Rio, chega a São Paulo a mostra Carmen Miranda Para Sempre, que será inaugurada hoje para convidados no Memorial da América Latina. Fotos, roupas, objetos, são mais de 700 peças reunidas para contar a história da "Pequena Notável " ou a Brazilian Bombshell- não há no mundo quem não conheça essa genial estrela que conquistou o Brasil, a Broadway e Hollywood.

A mostra tem percurso cronológico e está dividida em núcleos. Inicia com o nascimento em Portugal e inclui imagens de sua família. Depois, vem a fase brasileira (...).Era uma "mulher art déco dos anos 30", que usava calças, ternos e vestidos belos - em particular, há uma sala especial com retratos da artista feitos em 1931, em Buenos Aires, pela alemã Annemarie Heinrich

Depois de ser exibida no Rio, chega a São Paulo a mostra Carmen Miranda Para Sempre, que será inaugurada hoje para convidados no **Memorial da América Latina**. Fotos, roupas, objetos, são mais de 700 peças reunidas para contar a história da "**Pequena Notável**" ou a **Brazilian Bombshell**- não há no mundo quem não conheça essa genial estrela que conquistou o **Brasil**, a **Broadway** e **Hollywood**.

A mostra tem percurso cronológico e está dividida em núcleos. Inicia com o nascimento em Portugal e inclui imagens de sua família. Depois, vem a fase brasileira (...).Era uma "mulher art déco dos anos 30", que usava calças, ternos e vestidos belos - em particular, há uma sala especial com retratos da artista feitos em 1931, em **Buenos Aires**, pela alemã Annemarie Heinrich

Depois de ser exibida no Rio, chega a São Paulo a mostra Carmen Miranda Para Sempre, que será inaugurada hoje para convidados no **Memorial da América Latina**. Fotos, roupas, objetos, são mais de 700 peças reunidas para contar a história da "**Pequena Notável**" ou a **Brazilian Bombshell** ~~não há no mundo quem não conheça essa~~ genial estrela que conquistou o **Brasil**, a **Broadway** e **Hollywood**. **?**

A mostra tem percurso cronológico e está dividida em núcleos. Inicia com o nascimento em Portugal e inclui imagens de sua família. Depois, vem a fase brasileira (...). Era uma "mulher art déco dos anos 30", que usava calças, ternos e vestidos belos - em particular, há uma sala especial com retratos da artista feitos em 1931, em **Buenos Aires**, pela alemã Annemarie Heinrich

ReReEM: o que anotar?

- Compatibilizar anotação linguisticamente motivada e interesses (e capacidades) dos sistemas

Visitei uma exposição de cavalos, no Peru, e vi raças que só conhecia de fotografia: ↑ ? Falabella, Hunter, Berbere, Andaluz e Paso

- Necessidades de informação imprevisíveis

Relações inicialmente consideradas

☐ Identidade (ident)

- ✓ foi fundada em 1131 por **D. Telo (São Teotónio)**
It was founded in 1132 by **D. Telo (São Teotónio)**
- ✗ Os adeptos do **Porto** invadiram a cidade do **Porto** em júbilo
The (FC) **Porto** fans invaded the (city of) **Porto**, very happy

☐ Inclusão (inclui / incluído)

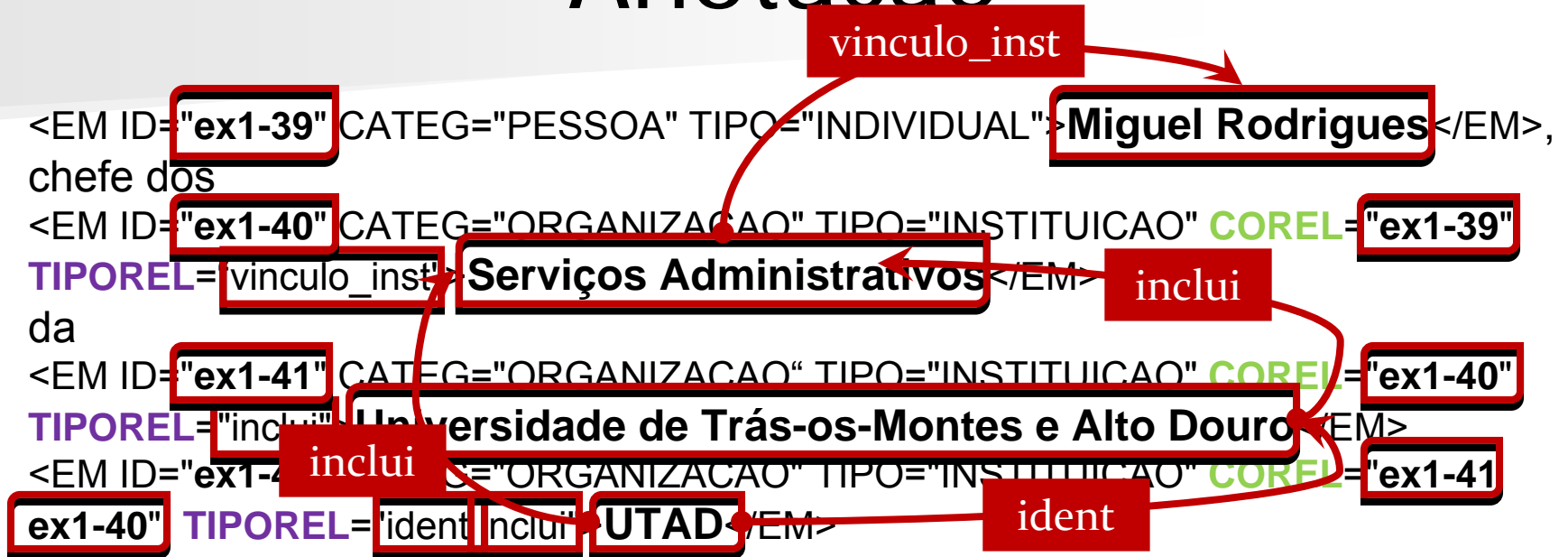
Hamilton, colega de Alonso na **McLaren**
Lewis Hamilton, Alonso's team-mate in **McLaren**

☐ Localizacao (ocorre-em / sede-de)

GP Brasil – Não faltou emoção em Interlagos no **Circuito José Carlos Pace** desde a primeira volta...

☐ Outra

Anotação



- ❑ No need to annotate all relations
- ❑ Evaluation program expands all possible relations

$A \text{ ident } B \wedge B \text{ ident } C \Rightarrow A \text{ ident } C$
 $A \text{ inclui } B \wedge B \text{ inclui } C \Rightarrow A \text{ inclui } C$
 $A \text{ inclui } B \wedge B \text{ sede_de } C \Rightarrow A \text{ sede_de } C$
 $A \text{ ident } B \wedge B \text{ any_rel } C \Rightarrow A \text{ any_rel } C$

Relations and vague categories

(...) a ideia de uma **Europa** LOCAL PESSOA unida. (...) um dia feliz para as cidadãs e os cidadãos da **União Europeia** LOCAL. (...) Somos essencialmente uma comunidade de valores -- são estes valores comuns que constituem o fundamento da

União Europeia ABST/ORG/LOCAL

the idea of a united **Europe** (...) a happy day for the citizens of the **European Union** (...) We are mainly a community of values and these common values constitute the foundation of the **European Union**

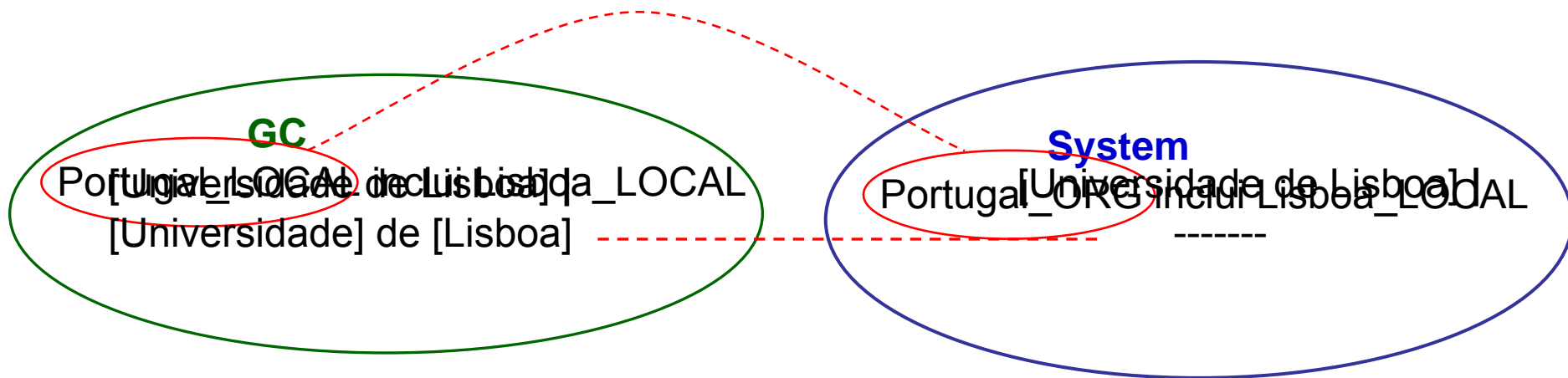
Avaliação

ReReIEM

□ Evaluate JUST the relations (not the NE)

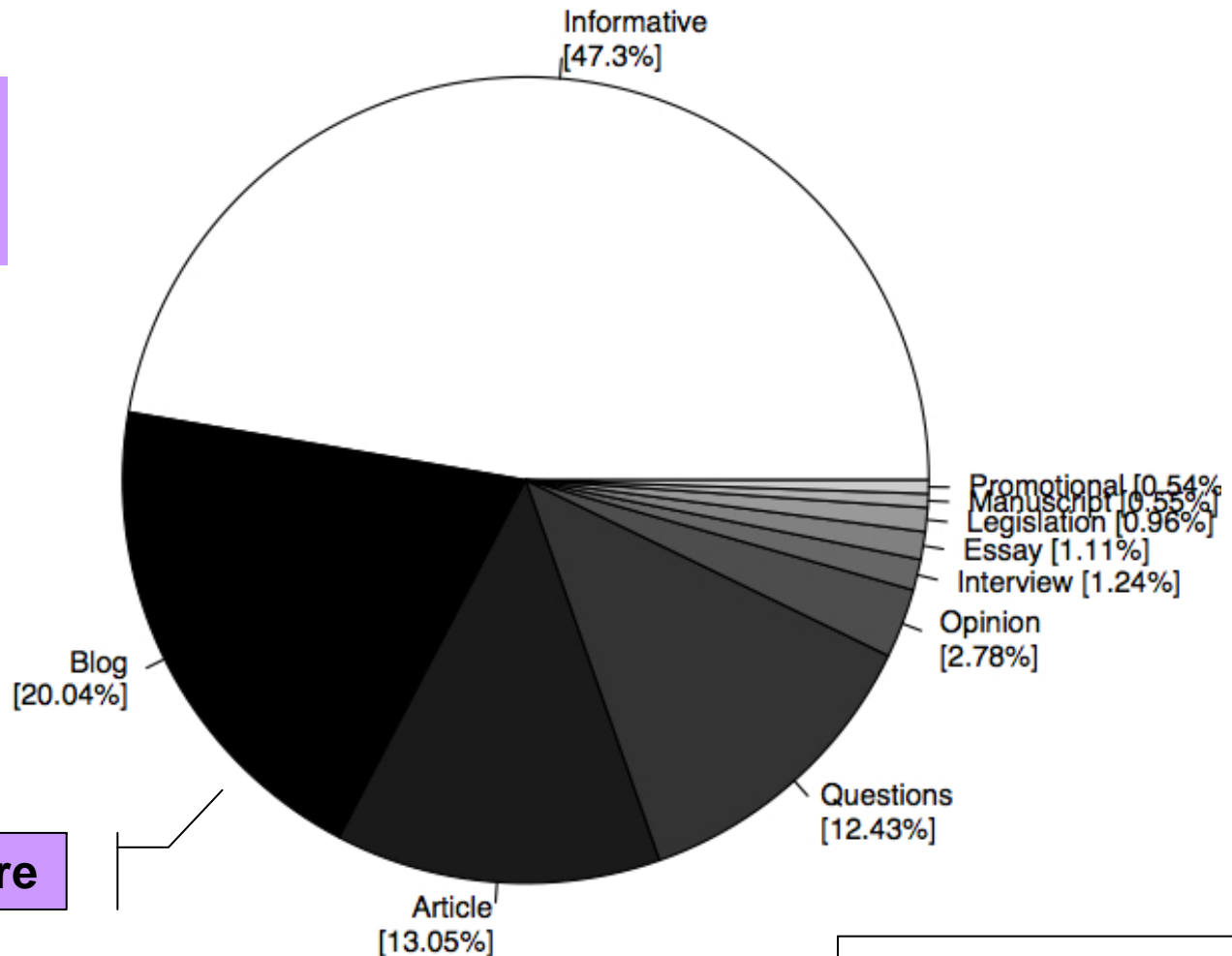
Relations with mismatched arguments were ignored

Alternative segmentations were ignored



Second HAREM Collection

DOCS: 1,040
Paragraphs: 15,737
Words: 670,610

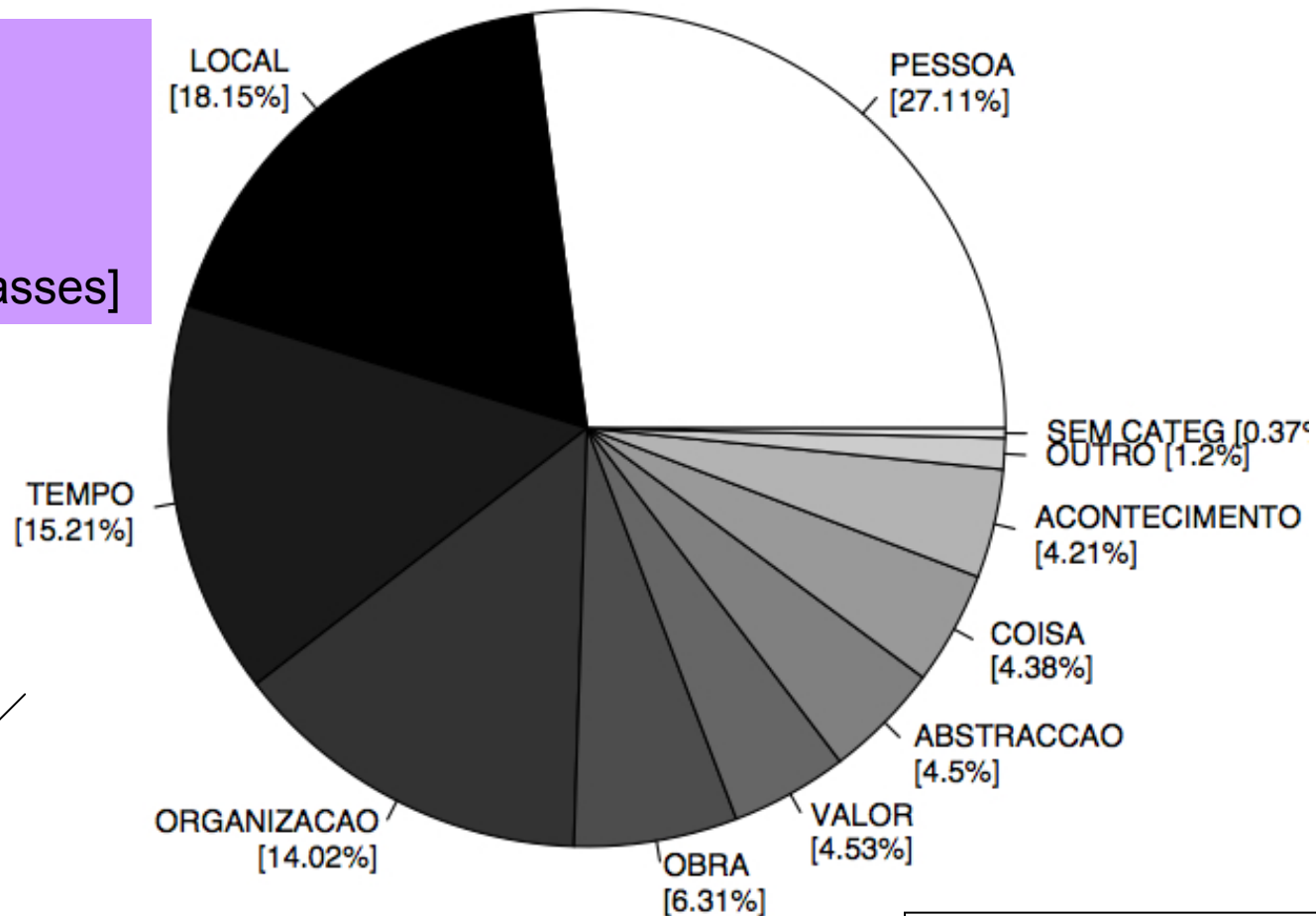


Distribution by text genre

Second HAREM Golden Collection

DOCS: 129
Paragraphs: 2,274
Words: 147,991
NEs: 7,847
Vague NEs: 633 [52 classes]

NE distribution



ReReIEM Golden Collection – full version

DOCS: 129
Paragraphs: 2,274
Words: 147,991
NE: 7,847
Relations: 4,852
NE with relations: 3,784

ReReIEM relation types

ReReIEM relations per category

Relations that the systems had to explicitly name

Relations under OUTRA/OTHER

Relation type	#
autor_de/obra_de (authorship)	145
causador_de (agent)	22
consequencia_de (result_of)	1
data_de	105
data_morte (death date)	10
data_nascimento (birth date)	6
ABSTRACCAO/ abstraction	258
ident (identity)	170
acontecimento/event	2234
incli/incluido (inclusion)	869
COISA / thing	175
local_nascimento_de/natural_de (birth place)	142
localizado_em/localizacao_de (place of)	24
LOCAL / place	963
nome_de/nomeado_por (name-of)	57
OBRA / title	274
ocorre em/sede de / (location)	360
ORGANIZACAO / org	794
outra_edicao (other edition)	3
outrarel (other relation)	94
OUTRO / other	25
participante_em/ter_participacao_de (participation-in)	155
PESSOA / person	1289
periodo_vida (lifetime)	5
personagem (character of)	14
TEMPO / time	193
praticado_em/pratica_se/praticante_de/praticado_por (practicing)	99
VALOR / value	19
produtor_de/produzido_por (manufacturing)	53
proprietario_de/propriedade_de (ownership)	39
relacao_familiar (kinship relation)	88
relacao_profissional (professional relation)	17
residente_de/residencia_de (place of residence)	19
vinculo_inst (affiliation)	282
TOTAL	4852

Second HAREM Resources

Second HAREM Collection

+

Second HAREM Golden Collection (GC)

+

TEMPO GC

+

ReReIEM GC

+

Evaluation programs

+

System runs

+

Documentation

=



LÂMPADA – Second HAREM Resource Package

<http://www.linguateca.pt/HAREM/PacoteRecursosSegundoHAREM.zip>

SAHARA and AC/DC: further access to HAREM and ReReLEM resources

- Sahara web service (Gonçalo Oliveira & Cardoso, 2009), <http://www.linguateca.pt/SAHARA/>
 - Submit new runs and...
 - select different options for scoring against the GC(s);
 - use several scenarios;
 - check the relative performance against the official runs.
- AC/DC, interaction with the parsed GC (Rocha & Santos, 2007) <http://www.linguateca.pt/ACDC/>

Acknowledgements

- Linguateca and HAREM were funded by the Portuguese government and the European Union with contract number 339/1.3/C/NAC, UMIC and FCCN



Slides referidos nesta apresentação:

[Carvalho et al. 2008]

Paula Carvalho, Hugo Gonçalo Oliveira, Cristina Mota, Diana Santos & Cláudia Freitas. "Segundo HAREM: Modelo geral, novidades e avaliação". *Encontro do Segundo HAREM* (Universidade de Aveiro, Portugal, 7 de Setembro de 2008).

[Freitas et al. 2008]

Cláudia Freitas, Diana Santos, Hugo Gonçalo Oliveira & Paula Carvalho. "ReReIEM: Relações Semânticas no Segundo HAREM". Encontro do Segundo HAREM (Universidade de Aveiro, Portugal, 7 de Setembro de 2008).

[Freitas et al. 2009]

Cláudia Freitas, Diana Santos, Cristina Mota, Hugo Gonçalo Oliveira & Paula Carvalho. "Detection of relations between named entities: report of a shared task". In Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, SEW-2009 (Boulder, Colorado, USA, June 4, 2009).

[Freitas et al. 2010]

Cláudia Freitas, Paula Carvalho, Hugo Gonçalo Oliveira, Cristina Mota & Diana Santos. "Second HAREM: advancing the state of the art of named entity recognition in Portuguese". In The seventh international conference on Language Resources and Evaluation (LREC 2010) (Malta, 10-21 de Maio de 2010).