Segundo HAREM, ReRelEM e LAMPADA 2.0

Cláudia Fre

Linguateca/PUC

PUC-Rio
20/04/2010

HAREM
Avaliação e **R**econhecimento de **E**ntidades **M**encionadas

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**</EM>, onde iniciou meteórica carreira...

A radiação de origem cósmica, prevista pelo <CATEG="**ABSTRACCAO**">**Big Bang** </EM>, seria descoberta em 1964...

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

... pelo qual tem início a expansão das galáxias que os cosmologistas descrevem como uma explosão ou seja um

«o problema do PSD é começar a ter só um <CATEG="**ACONTECIMENTO**">**Portugal**</EM> ou dois dentro de si»

<CATEG="**PESSOA**">**Portugal**</EM> 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**

Freitas et al. 2010

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**</EM>

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

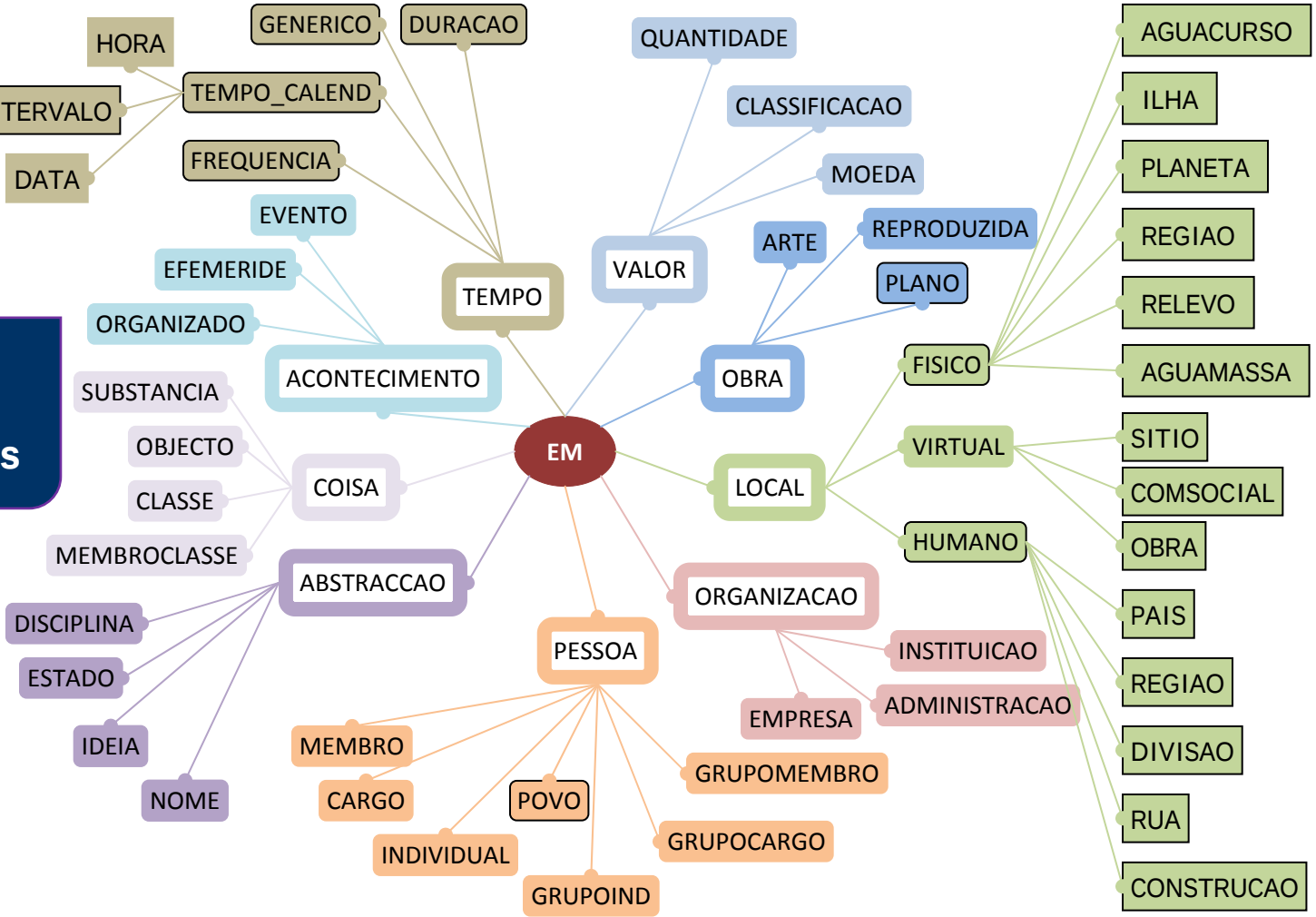
O carácter diferente da <CATEG="**ABSTRACCAO|ACONTECIMENTO** " TIPO="**IDEIA|EFEMERIDE**">**Reforma Inglesa** </EM> deve-se ao facto de ter sido promovida pelas necessidades políticas de Henrique VIII.

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

Características principais (Santos, 2007b)

III. Categorias

→ Initial corpus-based approach + participant suggestions



10 Categories
43 Types
22 Subtypes

Freitas et al. 2010

Características principais (Santos, 2007b)

IV. EM encaixadas

→ ALT mechanism

*Quantos atletas participaram nos **Jogos Olímpicos de Barcelona?***

Barcelona Olympic Games → **EVENT**

Barcelona | Olympic Games

↓ **PLACE** ↓ **EVENT**

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

Freitas et al. 2010

Nova tarefa - ReRelEM

Anaphora resolution
Mitkov, 2000; Collovoni 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 relacoes seriam encontradas nos textos

Criar tarefa piloto que comaparasse os sistemas que reconhecem essas relacoes

=

ReRelEM

Reconhecimento de Relações entre Entidades Mencionadas

Relation detection between named entities

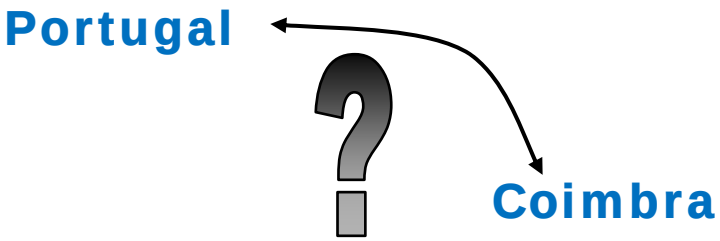
Freitas et al. 2010

ReRelEM ⟷ 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.

Freitas et al. 2008

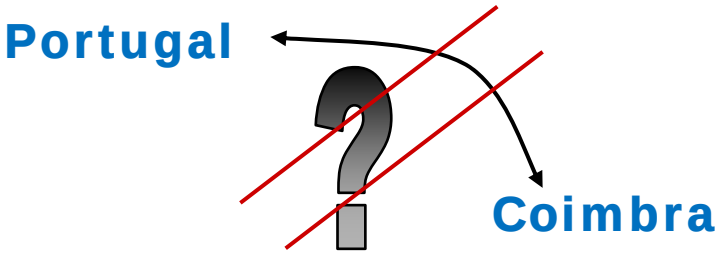
HAREM → ReRelEM



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

Freitas et al. 2008

HAREM → ReRelEM



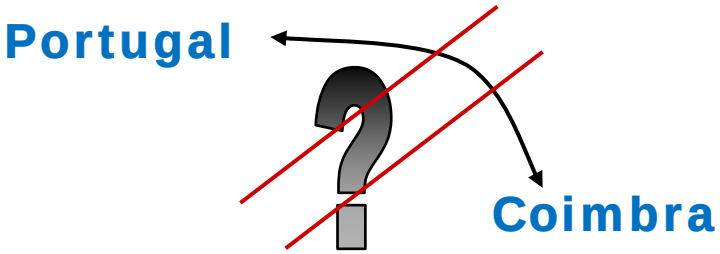
<EM ID="h-37" CATEG="**PESSOA**" TIPO="**GRUPOMEMBRO**">**Portugal**</EM>

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

SUBTIPO="**DIVISAO**">**Coimbra**</EM>

Freitas et al. 2008

HAREM → ReRelEM

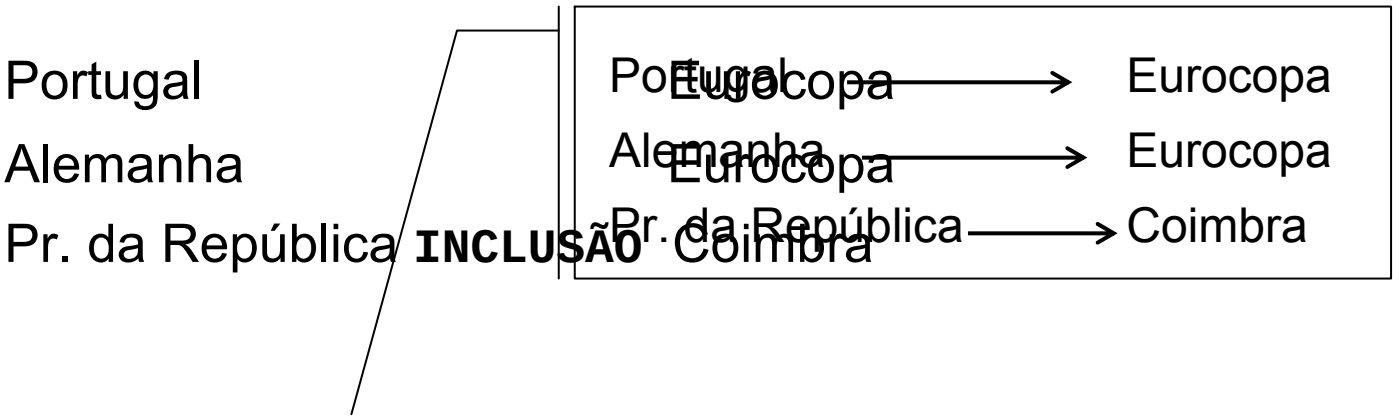


Portugal	→	Eurocopa
Alemanha	→	Eurocopa
Pr. da República	→	Coimbra

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

Freitas et al. 2008

ReRelEM – o que anotar?



Freitas et al. 2008

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

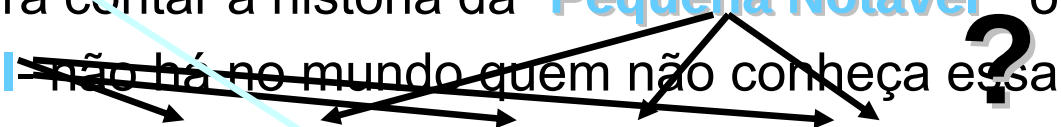
Freitas et al. 2008

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

Freitas et al. 2008

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

Freitas et al. 2008

ReRelEM: 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

Freitas et al. 2008

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 / incluido)

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

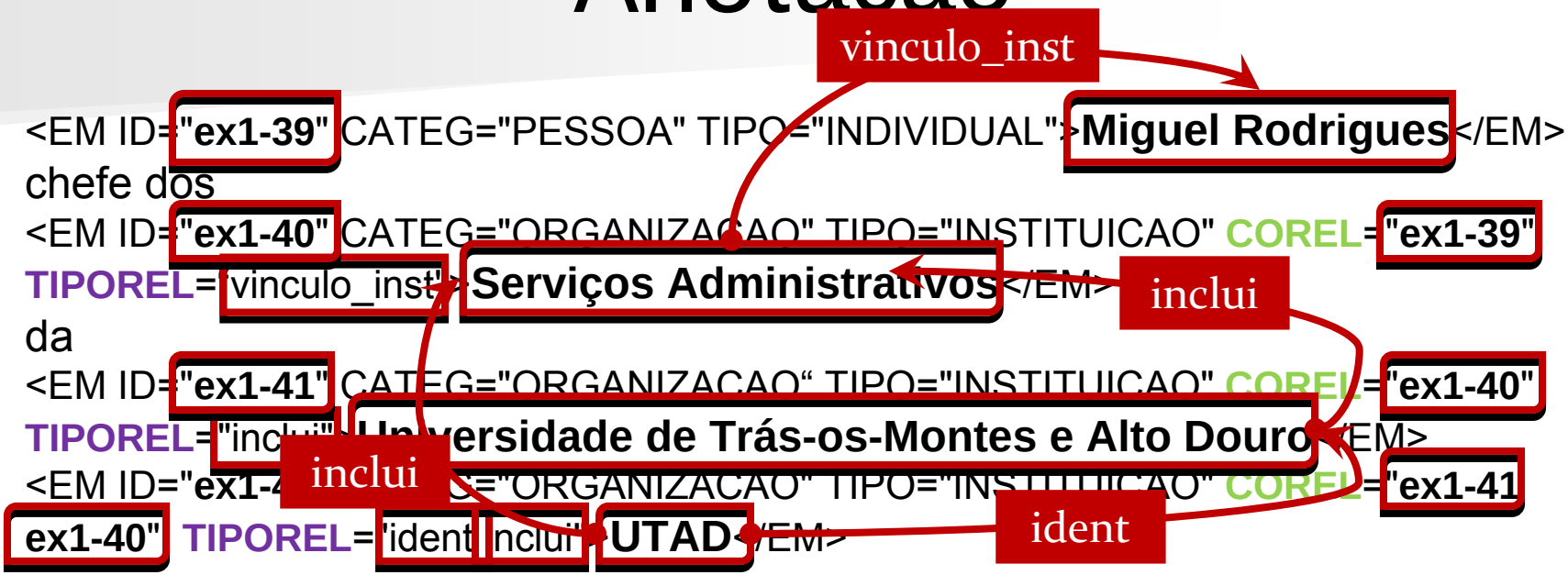
- Localizacao (ocorre-em / sede-de)

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

- Outra

Freitas et al. 2008

Anotacão

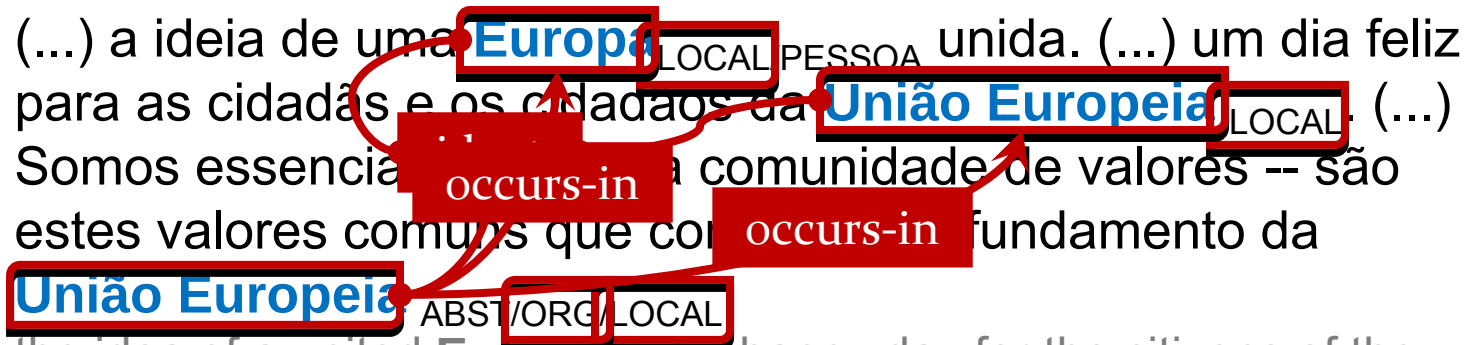


- No need to annotate all relations
- Evaluation program expands all possible relations

A ident B $\wedge$ B ident C $\Rightarrow$ A ident C
A inclui B $\wedge$ B inclui C $\Rightarrow$ A inclui C
A inclui B $\wedge$ B sede_de C $\Rightarrow$ A sede_de C
A ident B $\wedge$ B any_rel C $\Rightarrow$ A any_rel C

Freitas et al. 2009

Relations and vague categories



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**

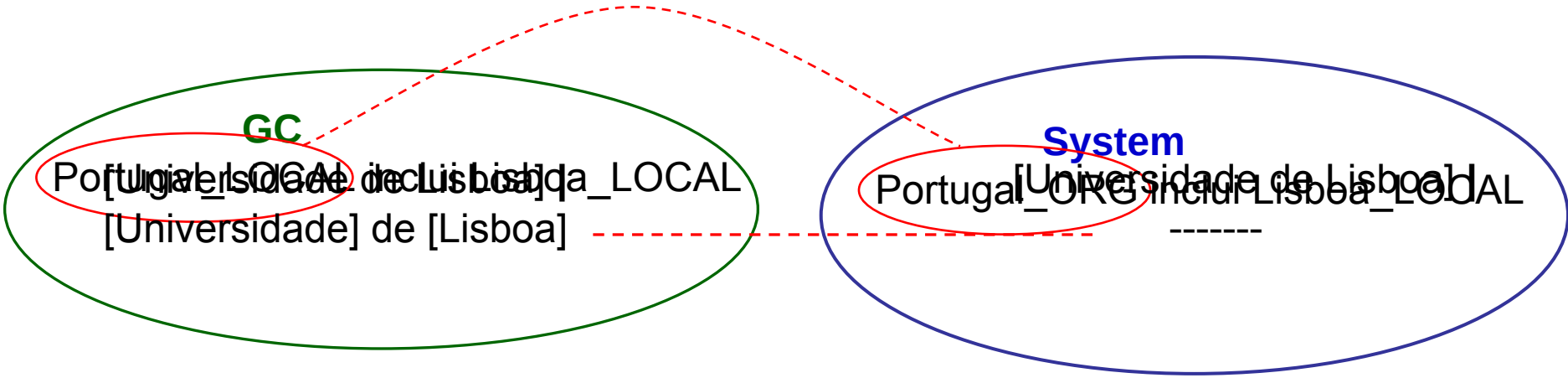
Freitas et al. 2009

Avaliação
ReRelEM

- Evaluate JUST the relations (not the NE)

Relations with mismatched arguments were ignored

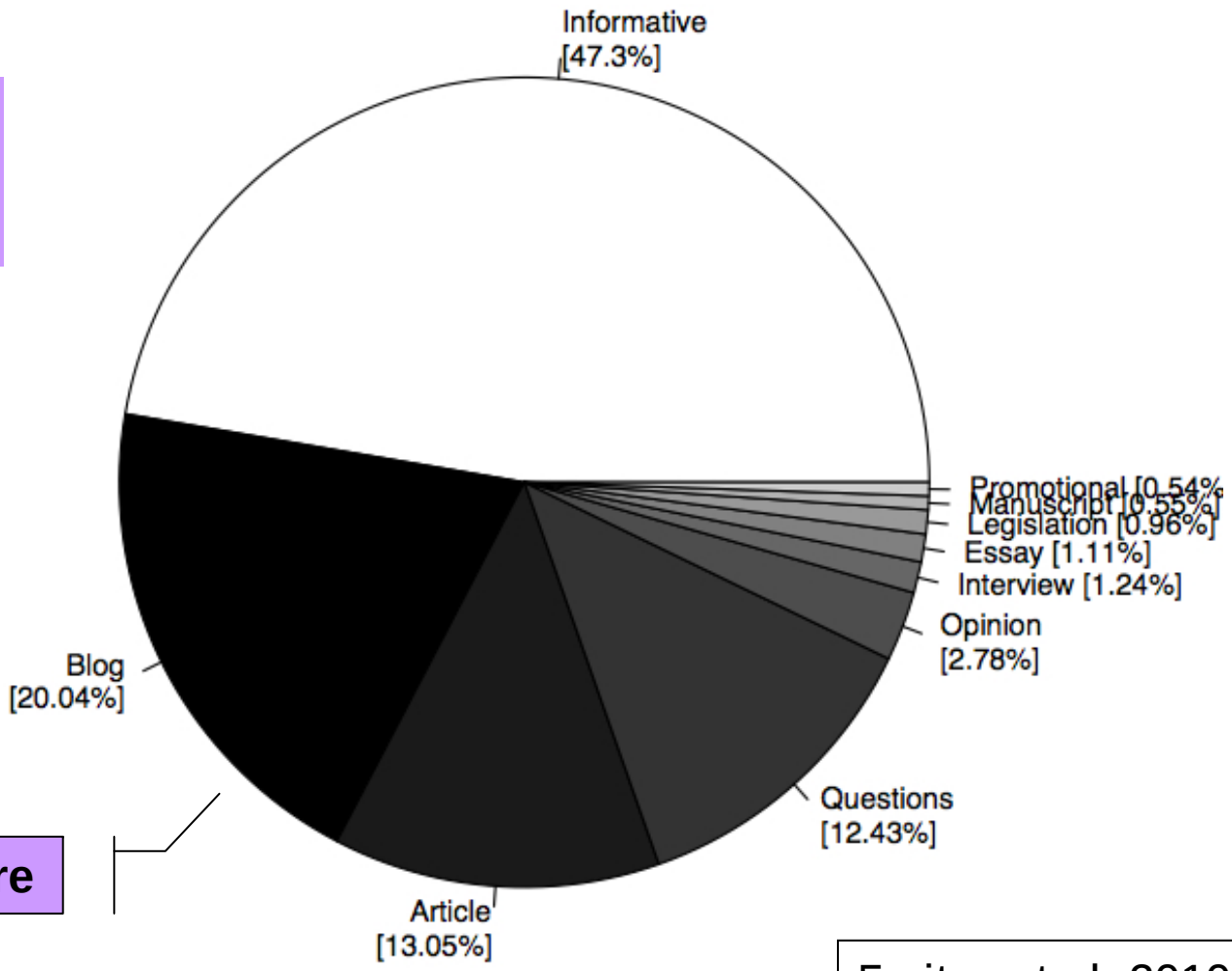
Alternative segmentations were ignored



Freitas et al. 2009

Second HAREM Collection

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

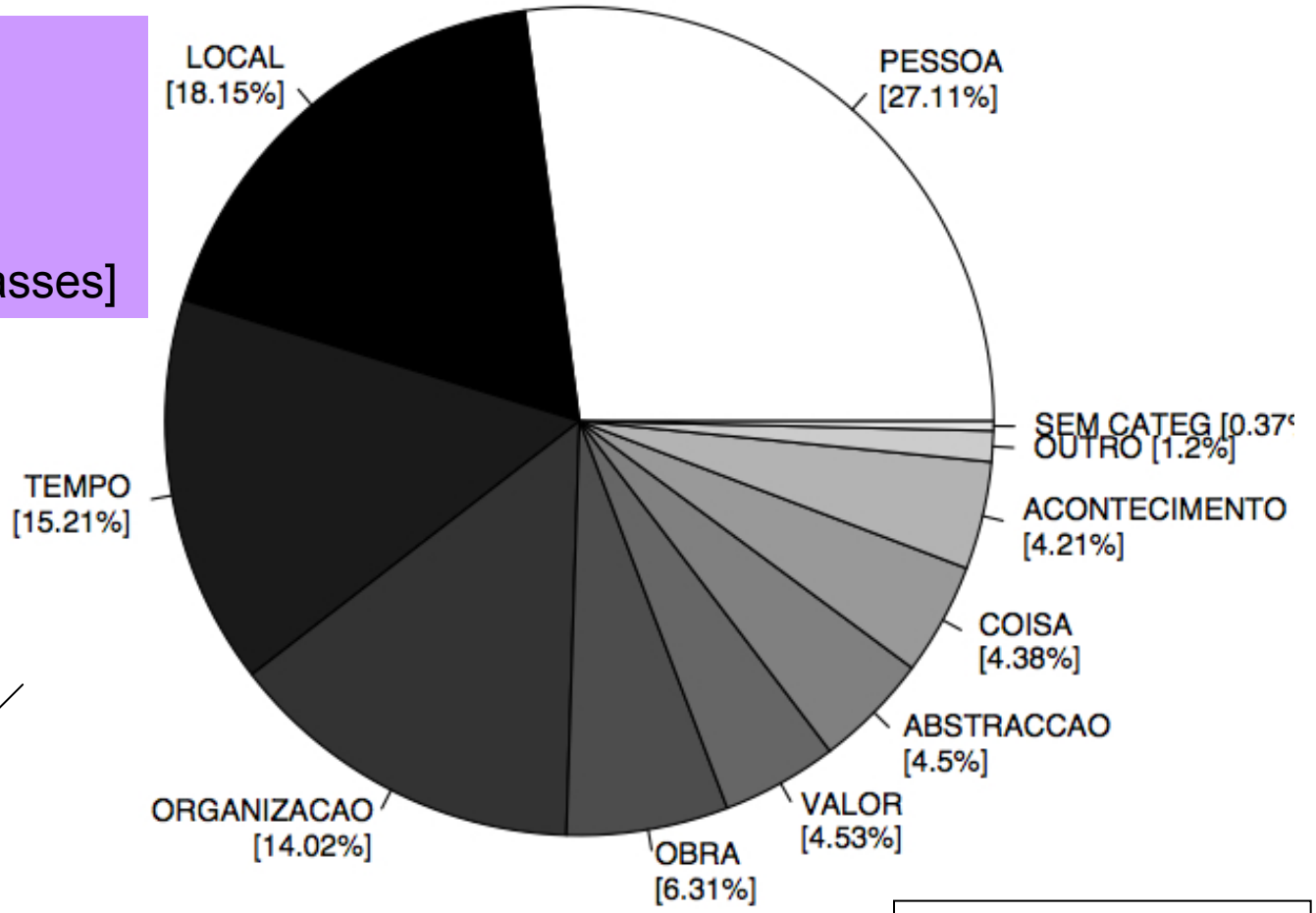


Distribution by text genre

Freitas et al. 2010

Second HAREM Golden Collection

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



NE distribution

Freitas et al. 2010

Freitas et al. 2010

ReRelEM Golden Collection – full version

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

ReRelEM relation types

ReRelEM 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
ident (identity)	**2234**
inclui/incluido (inclusion)	**869**
local_nascimento_de/natural_de (birth place)	142
localizado_em/localizacao_de (place of)	24
nome_de/nomeado_por (name-of)	57
ocorre_em/sede_de / (location)	**360**
outra_edicao (other edition)	3
outrarel (other relation)	94
participante_em/ter_participacao_de (participation-in)	155
periodo_vida (lifetime)	5
personagem_de (character of)	14
praticado_em/pratica_se/praticante_de/praticado_por (practicing)	99
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

Relations per category	#
ABSTRACCAO/ abstraction	258
ACONTECIMENTO/event	170
COISA / thing	175
LOCAL / place	963
OBRA / title	274
ORGANIZACAO / org	794
OUTRO / other	25
PESSOA / person	1289
TEMPO / time	193
VALOR / value	19

Second HAREM Resources

Second HAREM Collection

+

Second HAREM Golden Collection (GC)

+

TEMPO GC

+

ReRelEM GC

+

Evaluation programs

+

System runs

+

Documentation

=

LÂMPADA – Second HAREM Resource Package

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

Freitas et al. 2010

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. "ReRelEM: 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).