A Golden Resource for Named Entity Recognition in Portuguese

By: Diana Santos
Nuno Cardoso

PROPOR '2006 presentation

15th May, 2006

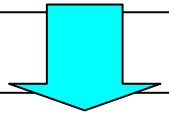
Linguateca

What is NER?

- **N**amed **E**ntity **R**ecognition: Roughly, identify and classify the entities (designated by proper names) in the text.

Example:

Eça de Queirós nasceu na Póvoa de Varzim em 1845, e faleceu 1900, em Paris. Estudou na Universidade de Coimbra.



Eça de Queirós nasceu na ***Póvoa de Varzim*** em ***1845***, e faleceu ***1900***, em ***Paris***. Estudou na ***Universidade de Coimbra***.

Semantic categories:

City, **Year**, **Person**, **University**

Same NE, different meanings

- Example: Portugal
  - **Portugal** venceu a Alemanha por 3-0.
  - **Portugal** votou 'não' na ONU.
  - O meu querido **Portugal** da infância...
  - **Portugal** tem muitas praias.
  - (João) **Portugal** canta hoje em Lisboa.
  - As acções da **Portugal** Telecom S.A....
- How many distinct meanings? How many more?

What is HAREM?

- **H**AREM - **A**valiação de **R**econhecimento de **E**ntidades **M**encionadas is the first evaluation contest on NER systems dealing with Portuguese

- See details of HAREM on http://www.linguateca.pt/HAREM

What is HAREM?

- The evaluation contest paradigm implies that together the participants agree on the task and the way to evaluate it.

everyone debates, everyone contributes, everyone participates, everybody wins!"

- Linguateca had previous experience with Morfolimpíadas and CLEF

Presentation Overview

- HAREM Overview
- The Golden Collection
- Discussion
- HAREM first results
- Outlook of HAREM

Motivation of HAREM

- Gather community around NER and measure the state-of-the-art of Portuguese NER
- Jointly discuss a NER evaluation methodology:
  - Define relevant tasks
  - Create directive sets
  - Develop new measures
  - Define procedure
  - Create new resources
- Make public HAREM resources (**Golden Collection**, evaluation platform) to help newcomers test and compare systems, ...

Background of HAREM

- There was no purely monolingual NER evaluation contest in a language different from English.
- Is different language relevant for NER?
  - Just change of modules (tokenization, spelling) and resources (gazetteers)? Minor adaptations...
  - ...Or a different language has different challenges? (Different things people talk about, different typographical conventions, different conceptualization of the world...)
- This is basically an empirical question...

HAREM in a Nutshell

- Several text **genres** and Portuguese **varieties**
- **Tailored** evaluation (suits systems with different objectives and purposes)
- Linguistically-motivated, bottom-up **categories**
- Tasks: identification, morphology, semantic
- Handles **vague** NEs (both in semantic and identification).

HAREM Task Description

- A collection of texts (the HAREM Collection) is delivered to the participants.
- Participants return it NE-annotated within 48 hours.
- One piece of the sent outputs is selected and compared against a **manually NE-annotated Golden Collection**
- Scores are computed, Global and Individual performance reports are generated and released.

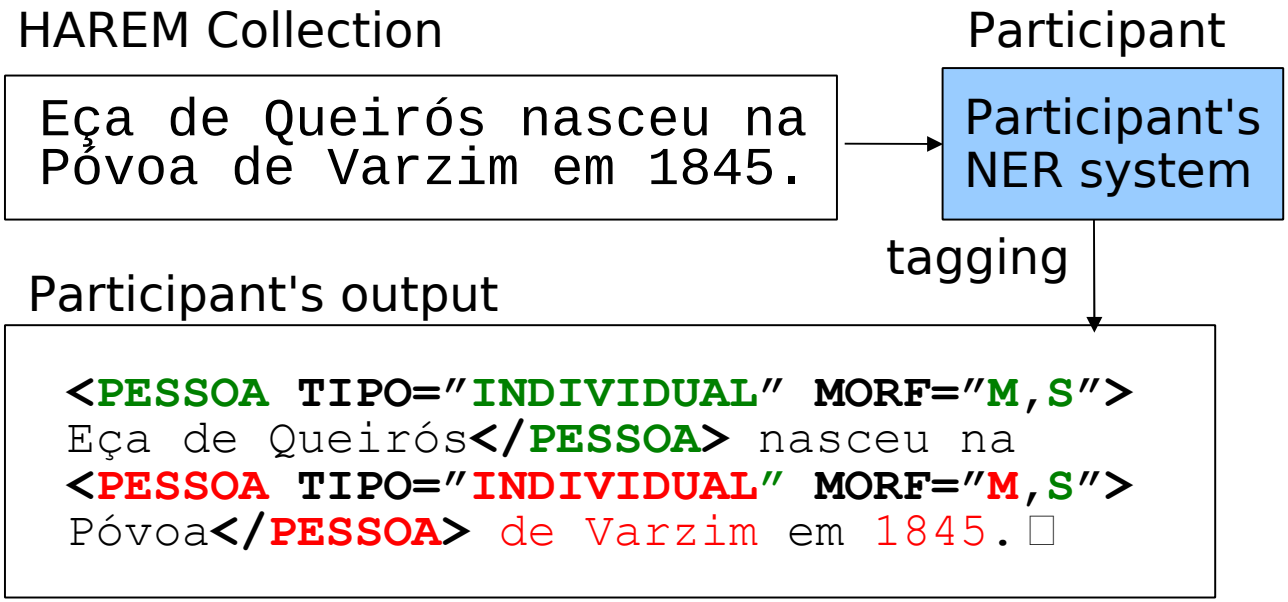
- **The core of HAREM is the Golden Collection.**

HAREM task description

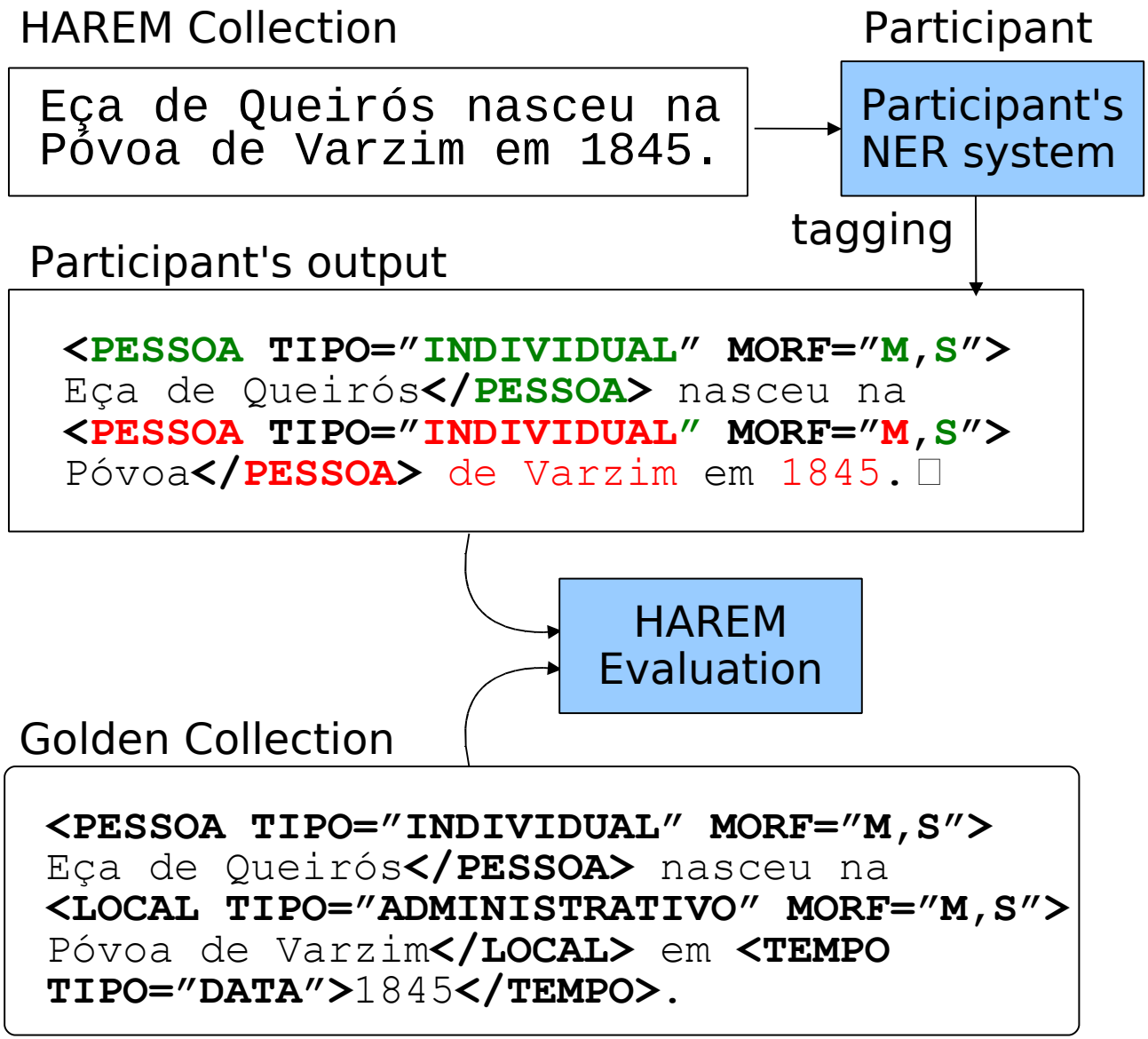
HAREM Collection

Eça de Queirós nasceu na Póvoa de Varzim em 1845.

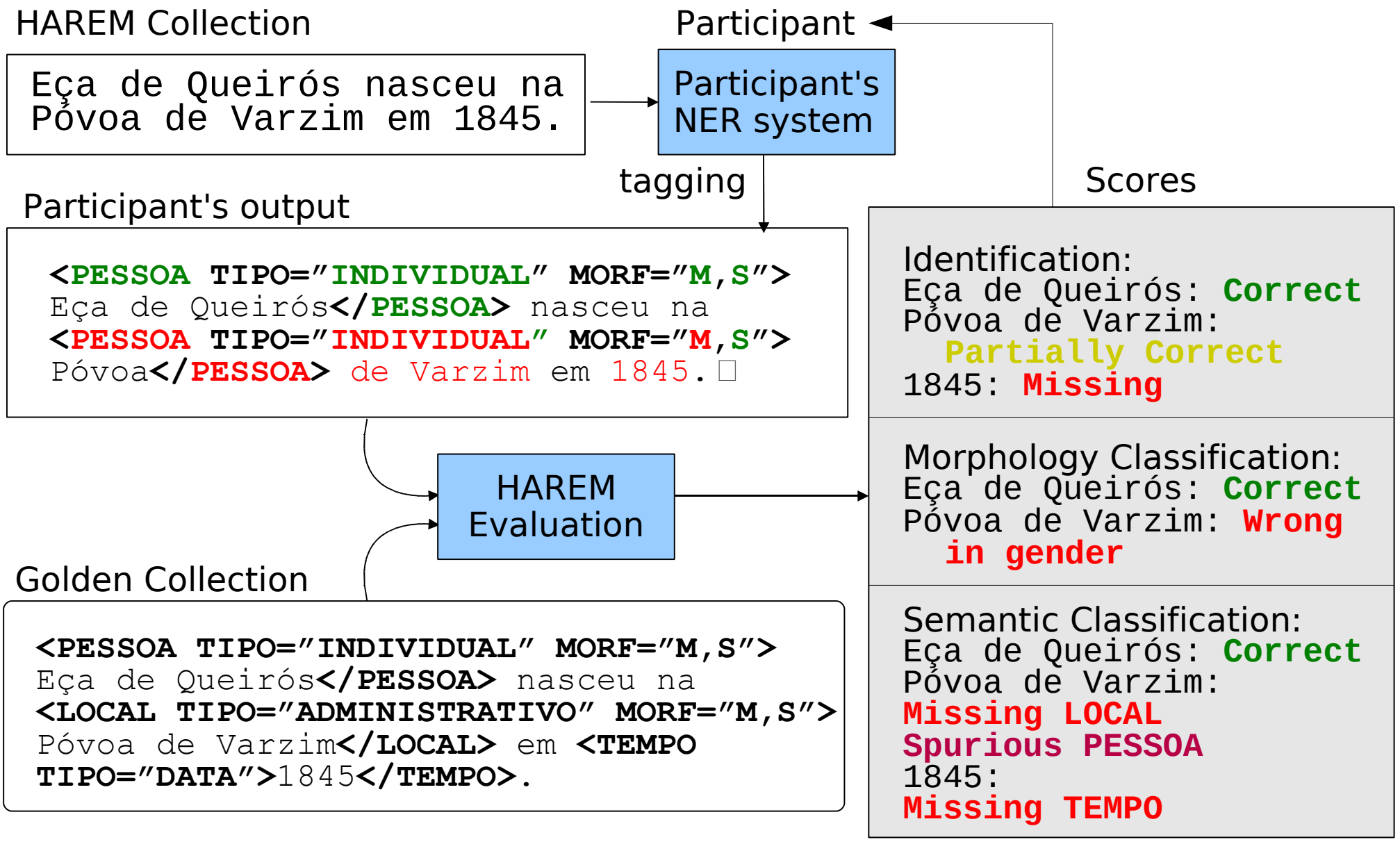
HAREM task description



HAREM task description



HAREM task description



HAREM Collections

- **HAREM Collection**: text collection of several genres and varieties, with no annotation of NEs.
- **Golden Collection (GC)**: a subset of HAREM Collection, manually NE-annotated ('2005 and '2006)

Collection sizes	HAREM Collection	GC 2005	GC 2006	Both GCs
Words	520 752	92 761	75 664	168 425
Documents	1 202	129	128	257
NEs	~ 40 000	5 270	3 858	9 128
Vague NEs (class.)	~ 1 000	133	142	275
Vague NEs (ident.)	~ 500	71	56	127

Desiderata of the Golden Collection

- Include all relevant NEs of Portuguese text, in an ideal set of marking
- NE categories empirically motivated from the text. We organized a two-level categorization (categories and types)
- It doesn't represent what NER systems are supposed to achieve now, instead it allows to:
  - measure the difficulty of the NER task
  - establish a ceiling
- There is much more to NER that just persons, organizations, places and numbers...

Golden Collection Creation

HAREM Organization

HAREM Collection (HC)

Subset of HC

Golden Collection (GC)

Pieces of the GC

NE-Annotated GC

- Selection
- Discussion
- Annotation
- Proofreading
- Improving directives

Final GC

Participants

Golden Collection Categories

- **PESSOA** (persons, groups, members)
- **ORGANIZAÇÃO** (associations, administration, firms)
- **ABSTRACÇÃO** (abstract concepts, tendencies)
- **COISA** (objects, class instances, classes)
- **LOCAL** (places, addresses, locations)
- **OBRA** (unique works of art, reproductions, publications)
- **ACONTECIMENTO** (historical events, organized events)
- **TEMPO** (time, date, periods, cyclic dates)
- **VALOR** (currencies, rankings, figures)

Golden Collection types, in 2005

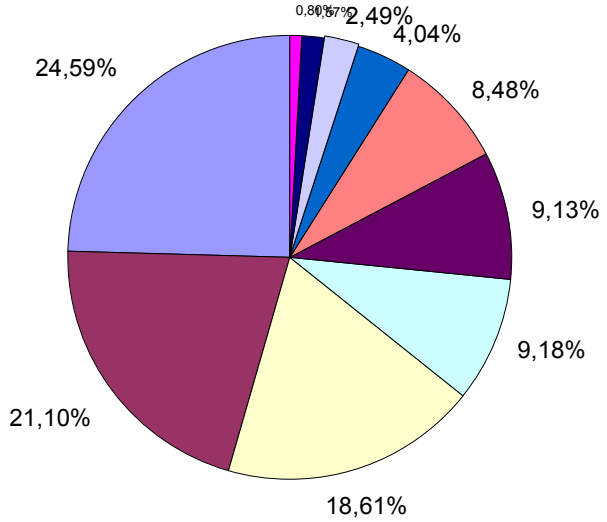
- **ABSTRACCAO**
  - DISCIPLINA
  - ESTADO
  - ESCOLA
  - OBRA
  - PLANO
  - IDEIA
  - NOME
- **OBRA**
  - ARTE
  - REPRODUZIDA
  - PRODUTO
  - PUBLICACAO
- **ACONTECIMENTO**
  - EFEMERIDE
  - ORGANIZADO
  - EVENTO
- **COISA**
  - OBJECTO
  - SUBSTANCIA
  - CLASSE
- **ORGANIZACAO**
  - INSTITUICAO
  - ADMINISTRACAO
  - EMPRESA
  - SUB
- **TEMPO**
  - DATA
  - HORA
  - PERIODO
  - CICLICO
- **PESSOA**
  - INDIVIDUAL
  - GRUPOIND
  - CARGO
  - GRUPOCARGO
  - MEMBRO
  - GRUPOMEMBRO
- **LOCAL**
  - GEOGRAFICO
  - ADMINISTRATIVO
  - VIRTUAL
  - ALARGADO
  - CORREIO
- **VARIADO**
  - OUTRO
- **VALOR**
  - MOEDA
  - CLASSIFICACAO
  - QUANTIDADE

Golden Collection types, in 2006

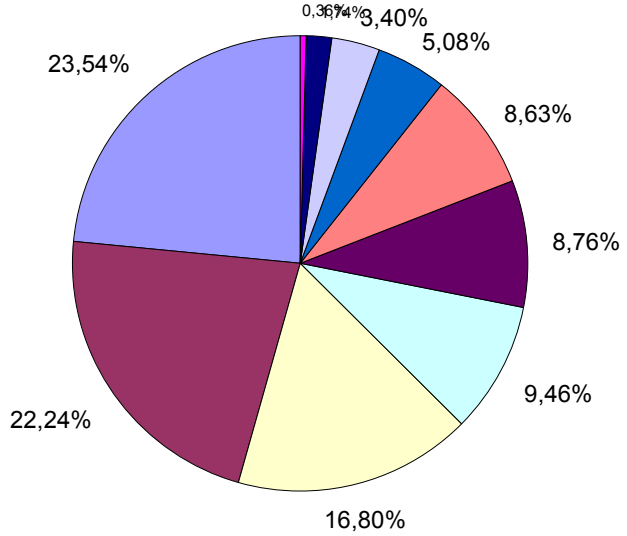
- **ABSTRACCAO**
  - DISCIPLINA
  - ESTADO
  - ESCOLA
  - OBRA
  - PLANO
  - IDEIA
  - NOME
- **OBRA**
  - ARTE
  - REPRODUZIDA
  - ~~**PRODUTO**~~
  - PUBLICACAO
- **ACONTECIMENTO**
  - EFEMERIDE
  - ORGANIZADO
  - EVENTO
- **COISA**
  - OBJECTO
  - SUBSTANCIA
  - CLASSE
  - **MEMBROCLASSE**
- **ORGANIZACAO**
  - INSTITUICAO
  - ADMINISTRACAO
  - EMPRESA
  - SUB
- **TEMPO**
  - DATA
  - HORA
  - PERIODO
  - CICLICO
- **VARIADO**
  - OUTRO
- **PESSOA**
  - INDIVIDUAL
  - GRUPOIND
  - CARGO
  - GRUPOCARGO
  - MEMBRO
  - GRUPOMEMBRO
- **LOCAL**
  - GEOGRAFICO
  - ADMINISTRATIVO
  - VIRTUAL
  - ALARGADO
  - CORREIO
- **VALOR**
  - MOEDA
  - CLASSIFICACAO
  - QUANTIDADE

GC Distribution by Categories

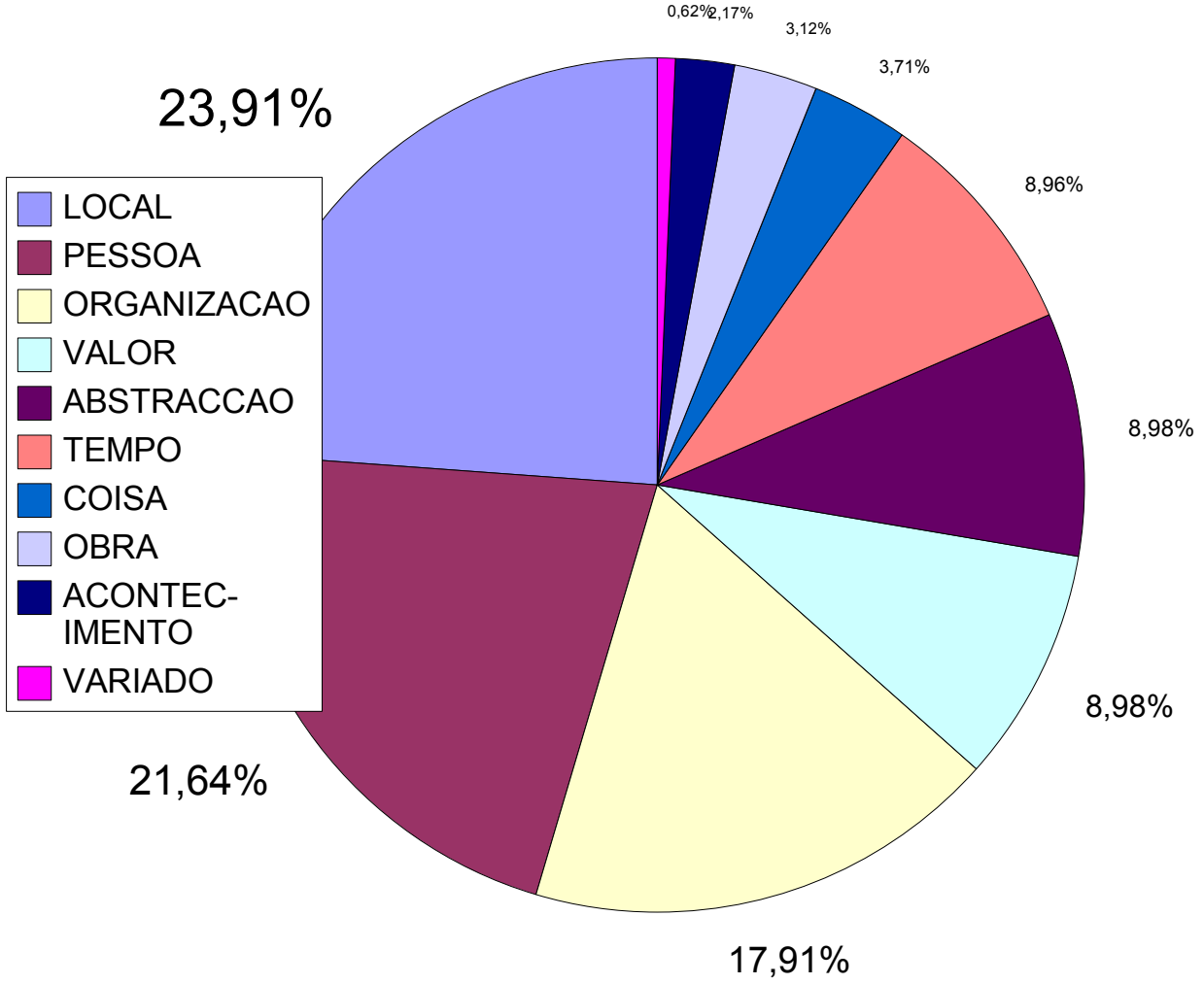
2005 Golden Collection (2005 rules)



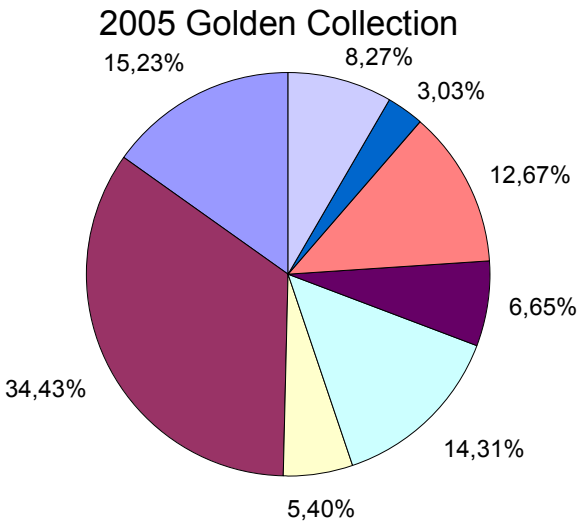
2006 Golden Collection (2006 rules)

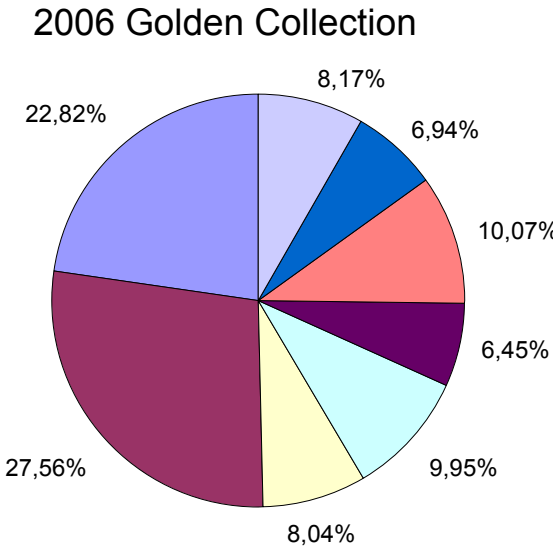


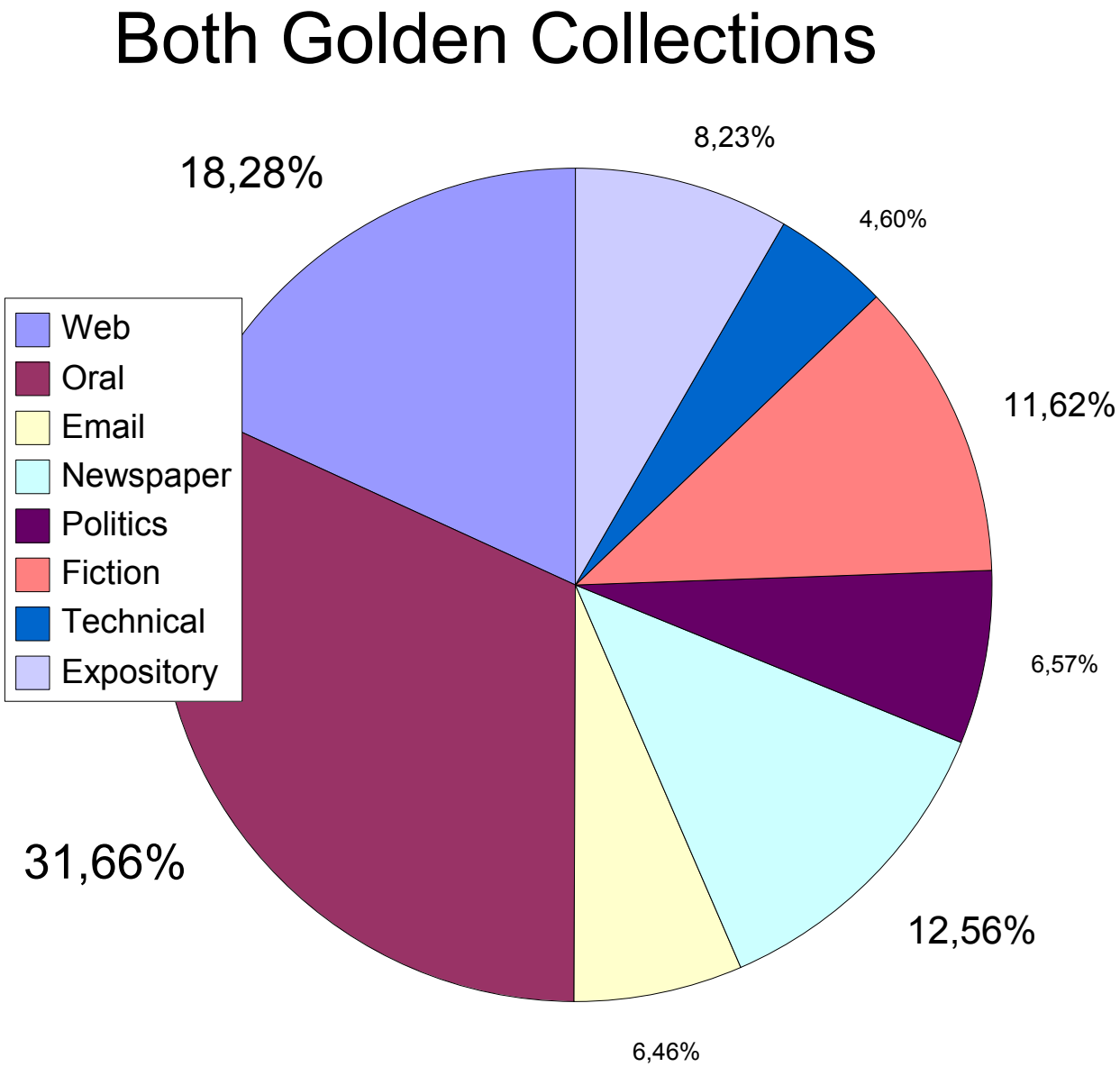
Both Golden Collections (2006 rules)



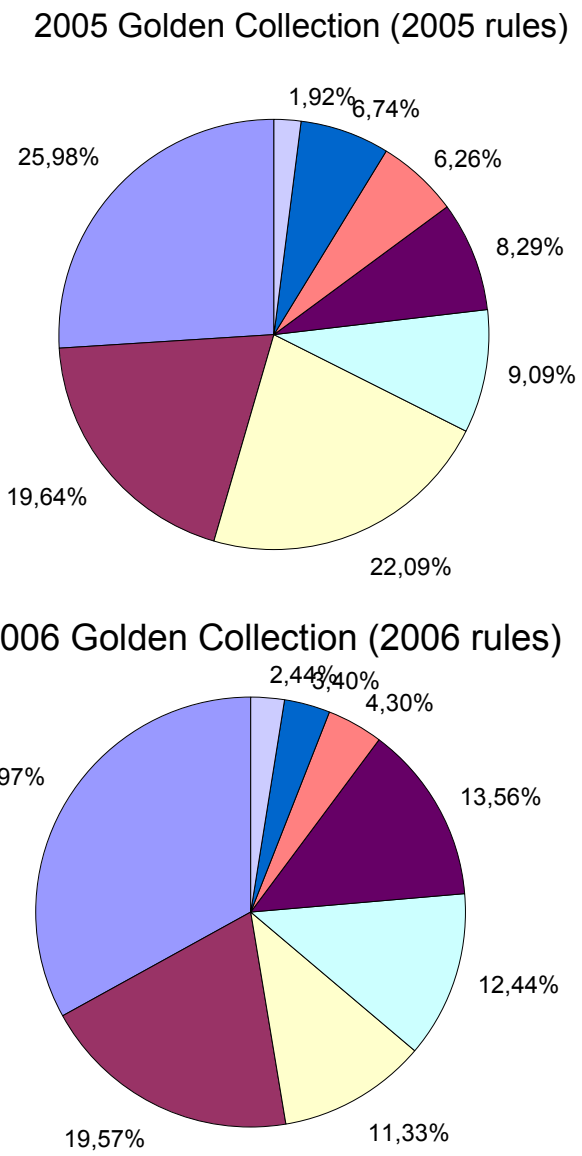
GC Distribution by Genre (Word Count)

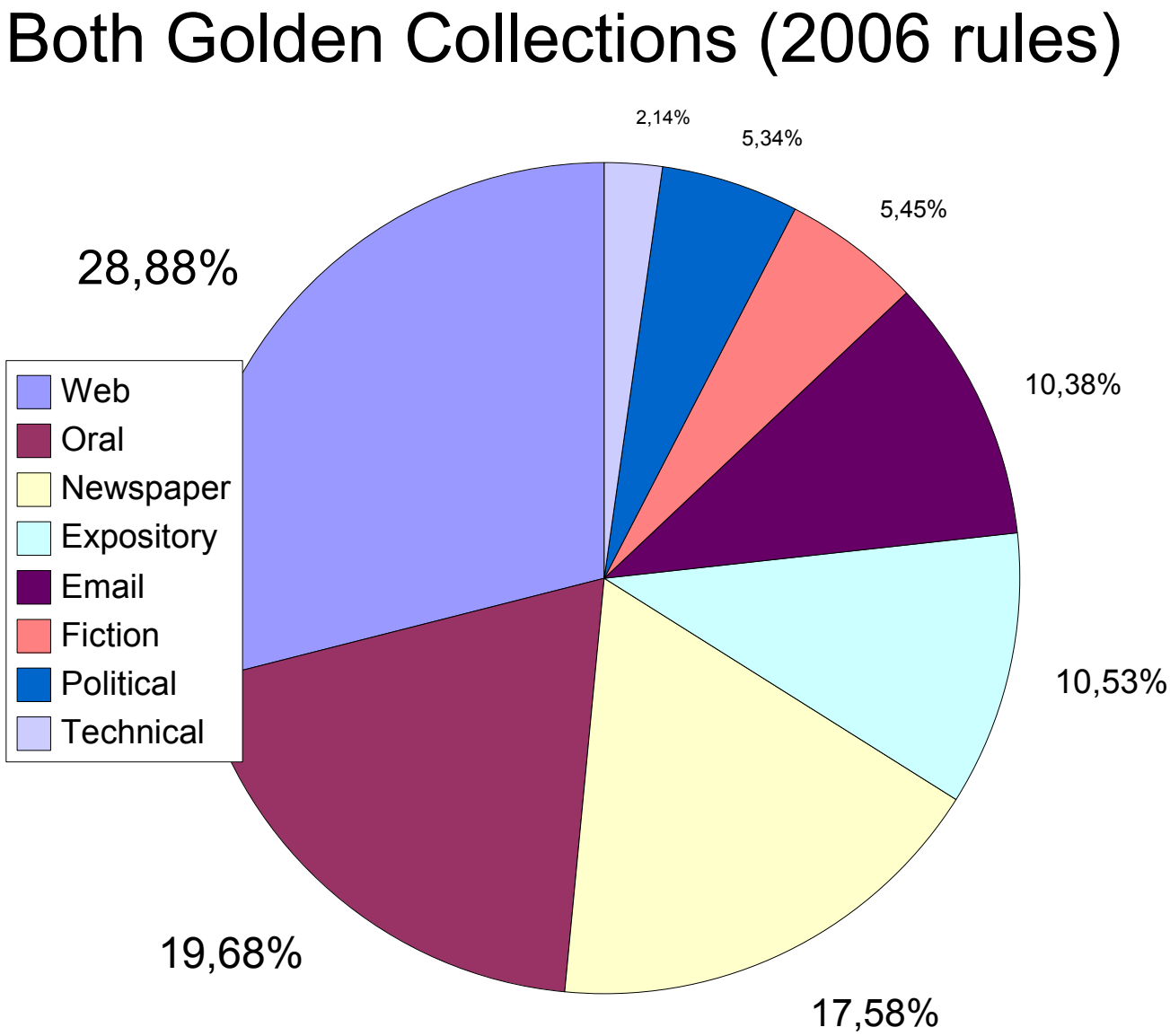




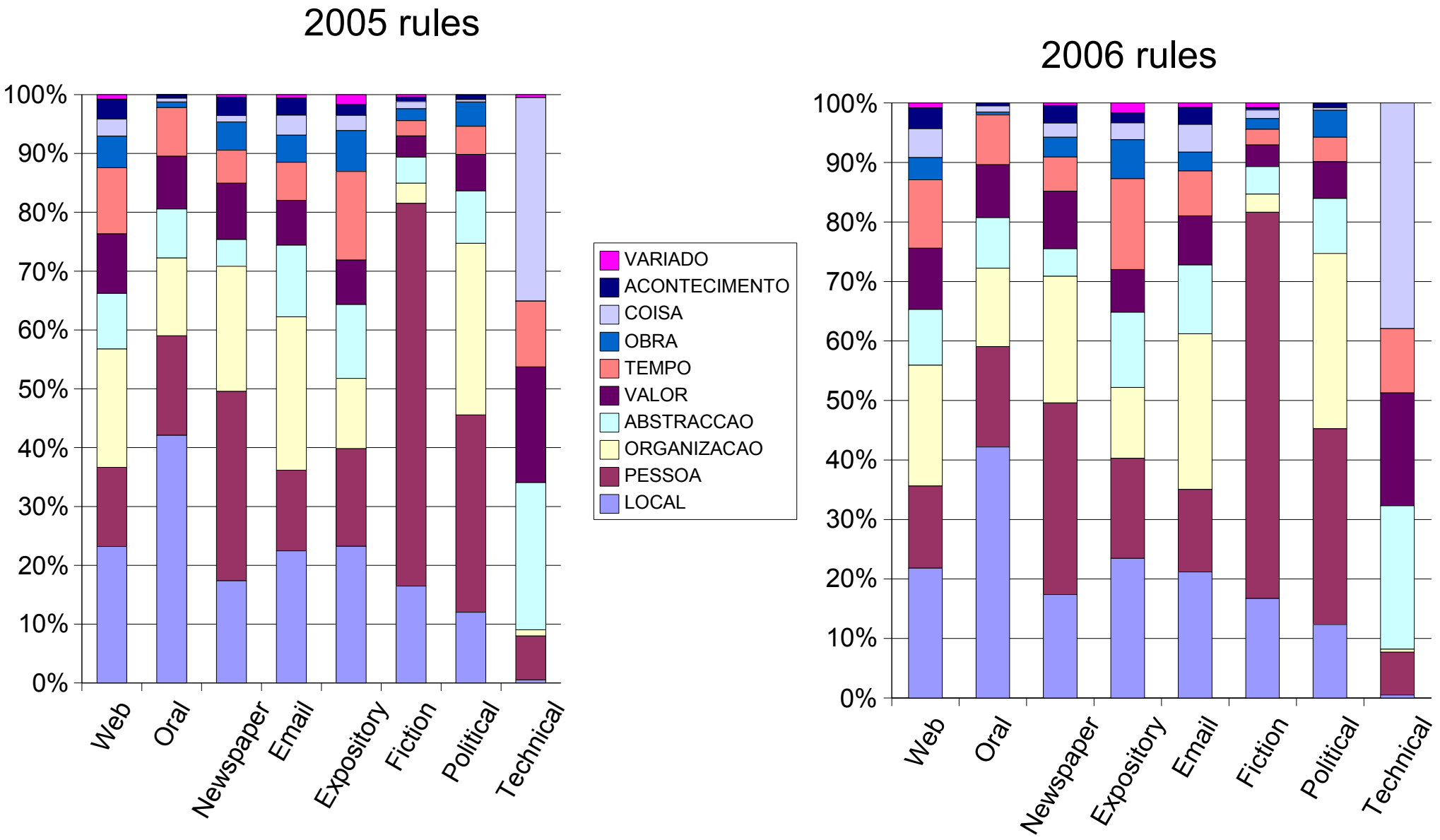


GC Distribution by Genre (NE Count)

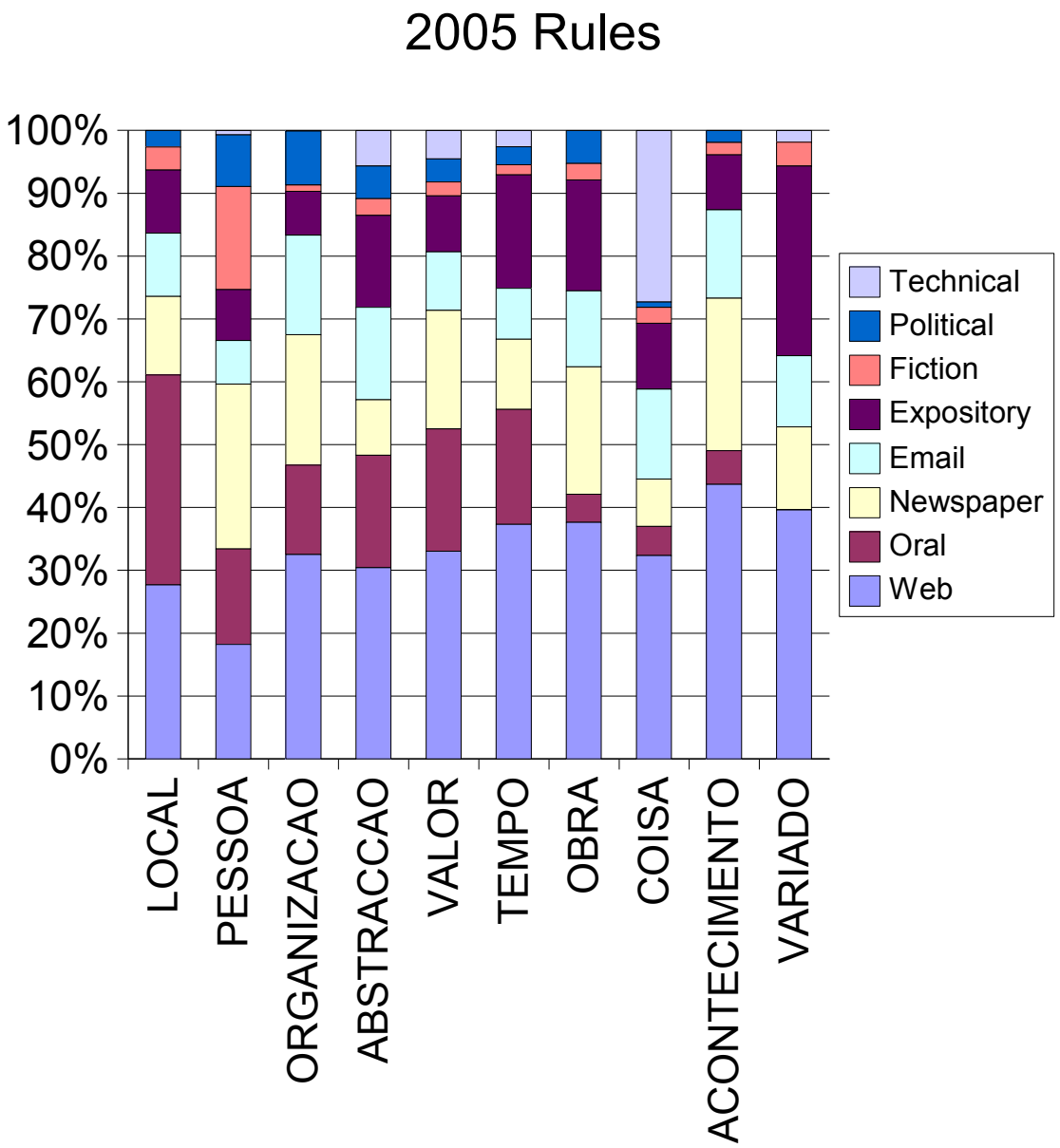


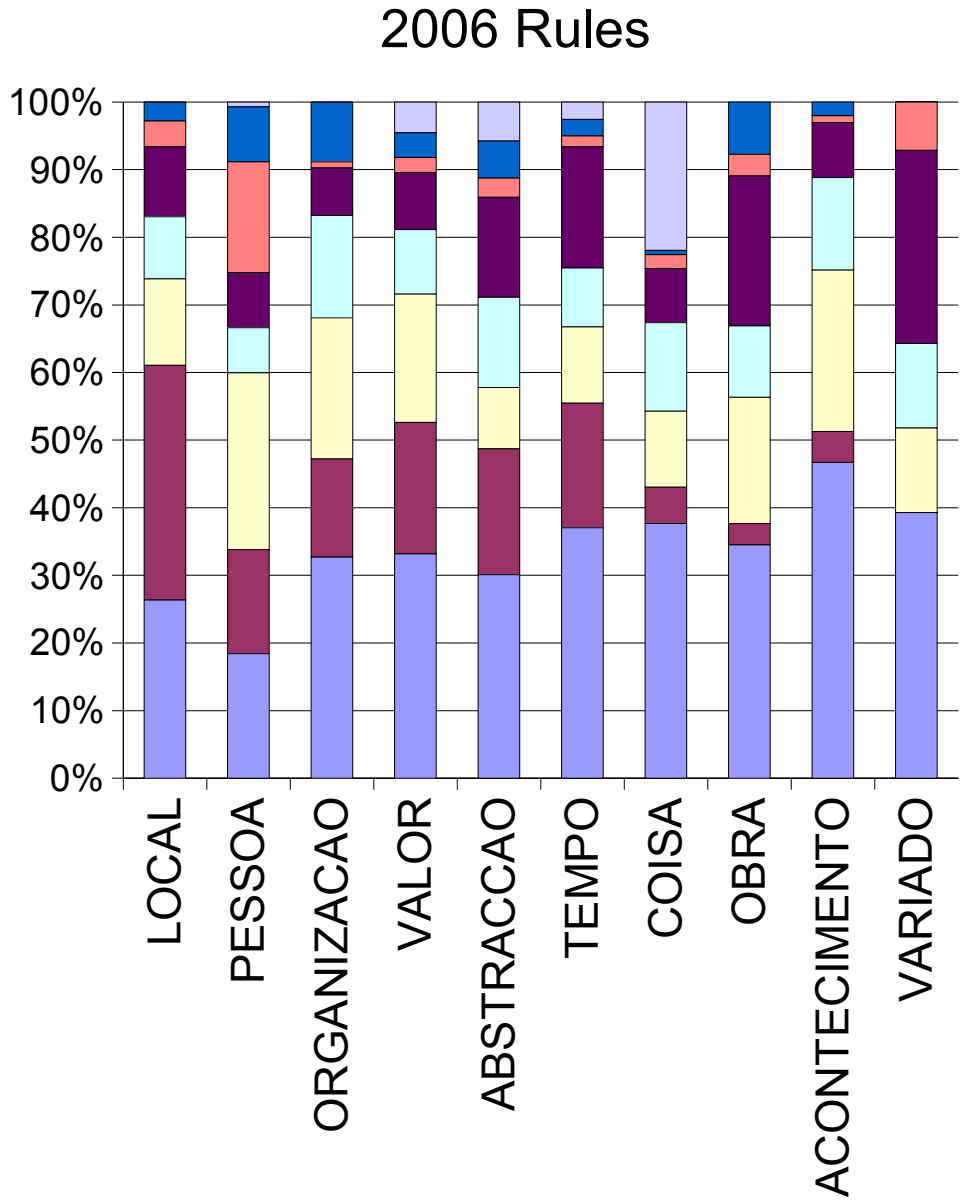


Category Distribution per Genre



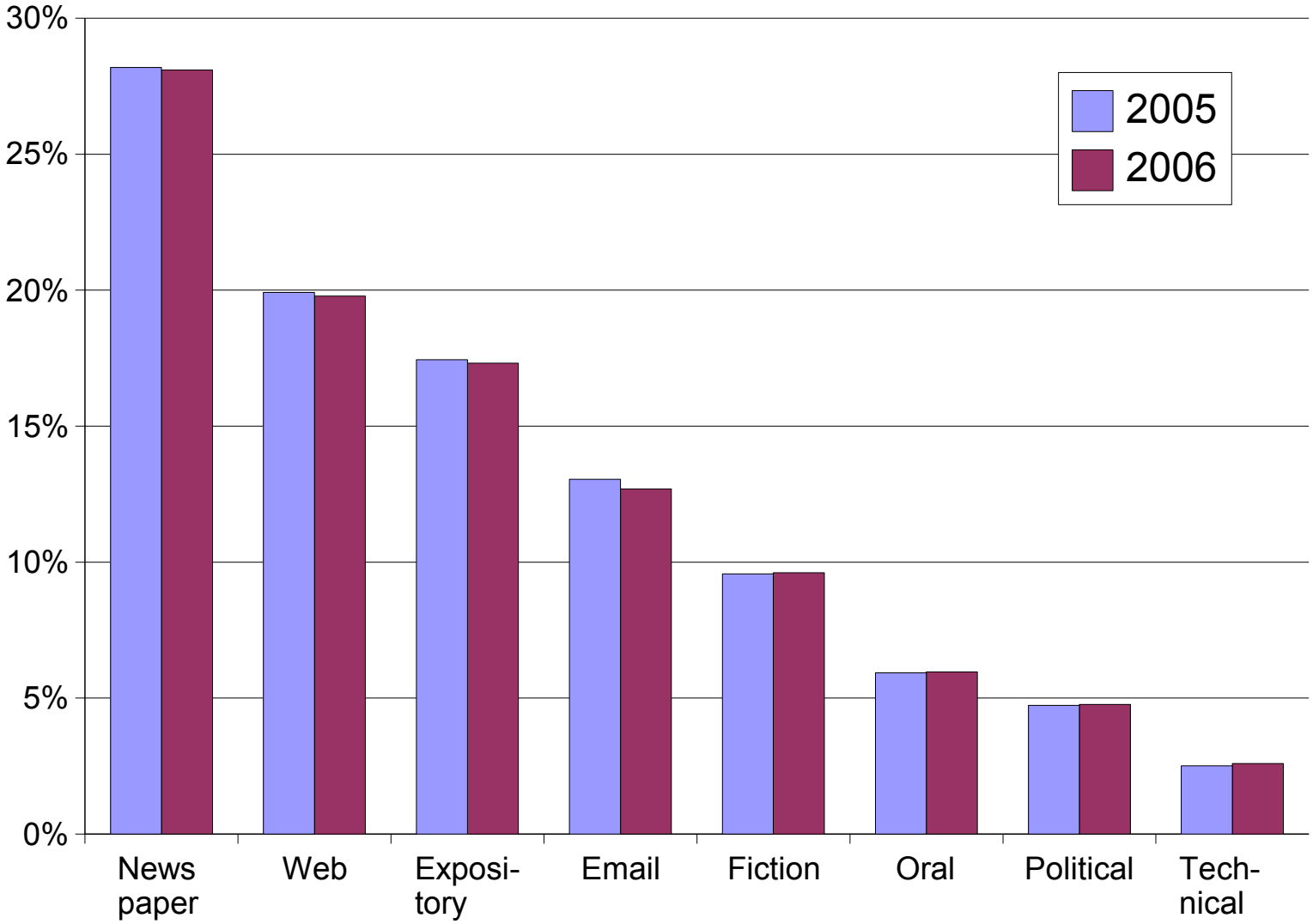
Genre Distribution per Category





NE Density by Genre

$$\text{NE Density}_G = \frac{\text{Number of words belonging to a NE, for genre G}}{\text{Total number of words for genre G}}$$



NE size in words, by category

Categories	2005 Average	2005 Median	2005 Std.Dev.	2006 Average	2006 Median	2006 Std.Dev.
ACONTECIMENTO	3,34	3	2,94	3,76	3	3,16
OBRA	3,26	2	2,89	3,5	3	3,19
VARIADO	2,25	1	2,51	2,23	1	2,48
ABSTRACCAO	2,19	1	2,44	2,21	1	2,01
ORGANIZACAO	2,19	1	1,96	2,21	1	2,45
PESSOA	1,9	2	1,12	1,9	2	1,10
TEMPO	1,81	1	1,34	1,82	1	1,34
VALOR	1,75	2	0,90	1,75	2	0,91
LOCAL	1,65	1	1,43	1,66	1	1,46
COISA	1,45	1	0,83	1,54	1	0,88
TOTAL	**1,97**	**1**	**1,73**	**1,98**	**1**	**1,76**

Vagueness in HAREM

- Vagueness is intrinsic to natural language
- Any NER systems must deal with it:
  - we don't exclude “too difficult cases” from the GC, as NER systems deal with real text, HAREM must represent the systems' real environment
  - If humans can't decide on one annotation, makes no sense to require that systems should decide it. A ceiling must be defined.

How do we deal with vagueness?

- Identification:

  <ALT><EM>Lopes e Silva</EM>|

  <EM>Lopes</EM> e **<EM>Silva</EM></ALT>**

- Morphology classification:

  – Chamo-me <EM **MORF="?,S"**>João</EM>

  (João M ou João F?)

- Semantic classification:

  Os **<PESSOA|ORGANIZACAO TIPO="GRUPOMEMBRO|INSTITUICAO">** **Bombeiros** **</PESSOA|ORGANIZACAO>...**

Vagueness Matrix for categories

2005

Category	ABSTRACCAO	ACONTECIMENTO	COISA	LOCAL	OBRA	ORGANIZACAO	PESSOA	TEMPO	VALOR
ABSTRACCAO	2								
ACONTECIMENTO	3	0							
COISA	2	1	0						
LOCAL	**14**	2	2	**22**					
OBRA	7	0	2	**17**	2				
ORGANIZACAO	**14**	5	0	**38**	4	**36**			
PESSOA	**33**	0	0	3	0	**50**	**82**		
TEMPO	1	3	1	0	0	0	1	4	
VALOR	1	0	0	0	0	0	0	0	0

2006

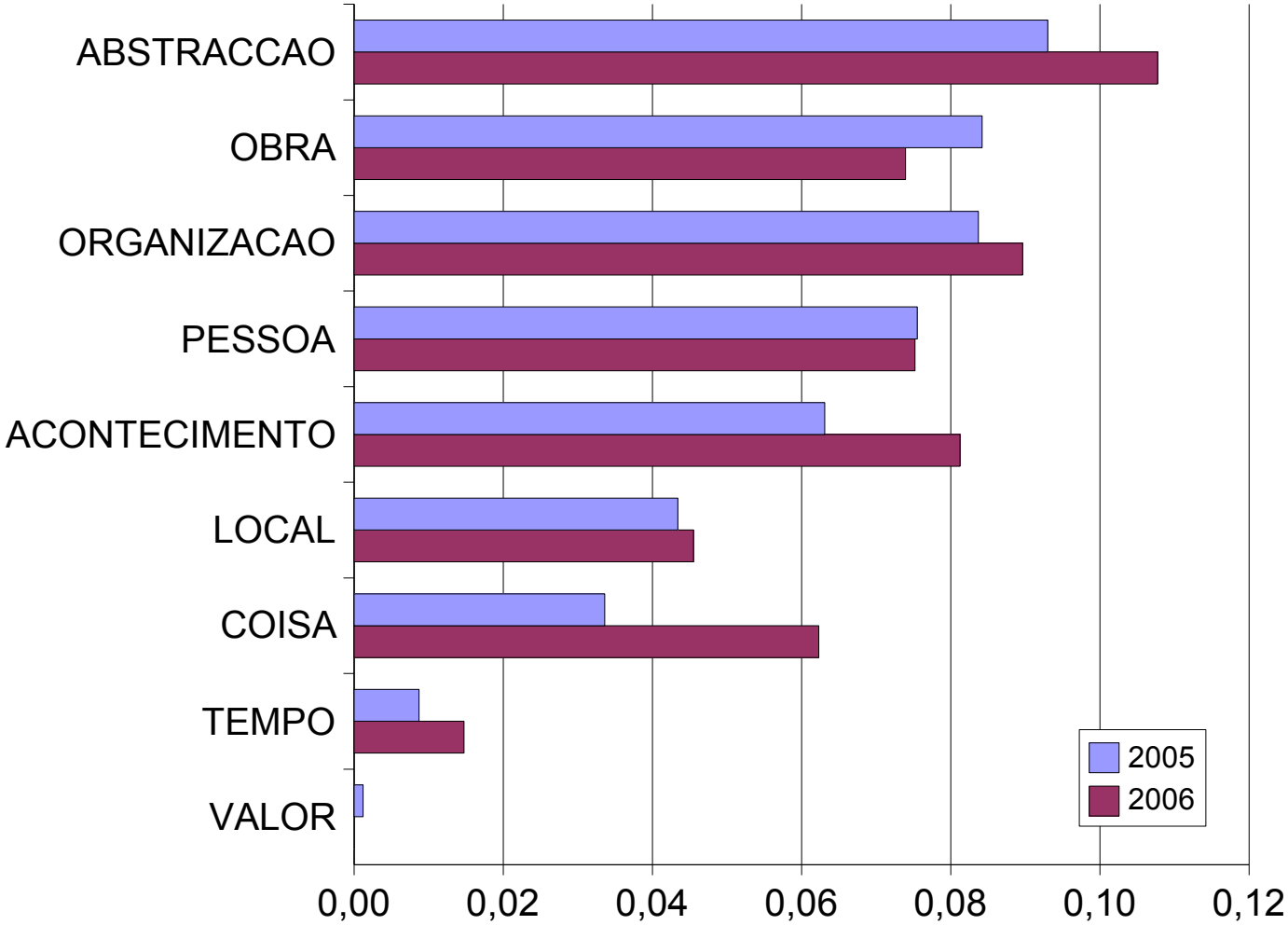
Category	ABSTRACCAO	ACONTECIMENTO	COISA	LOCAL	OBRA	ORGANIZACAO	PESSOA	TEMPO	VALOR
ABSTRACCAO	2								
ACONTECIMENTO	4	0							
COISA	3	0	8						
LOCAL	**19**	3	4	**24**					
OBRA	**19**	0	2	**11**	0				
ORGANIZACAO	**33**	4	4	**39**	0	**52**			
PESSOA	0	0	0	3	0	**46**	**84**		
TEMPO	3	5	0	0	0	0	2	2	
VALOR	0	0	0	0	0	0	0	0	0

Vagueness Ratio, by Category

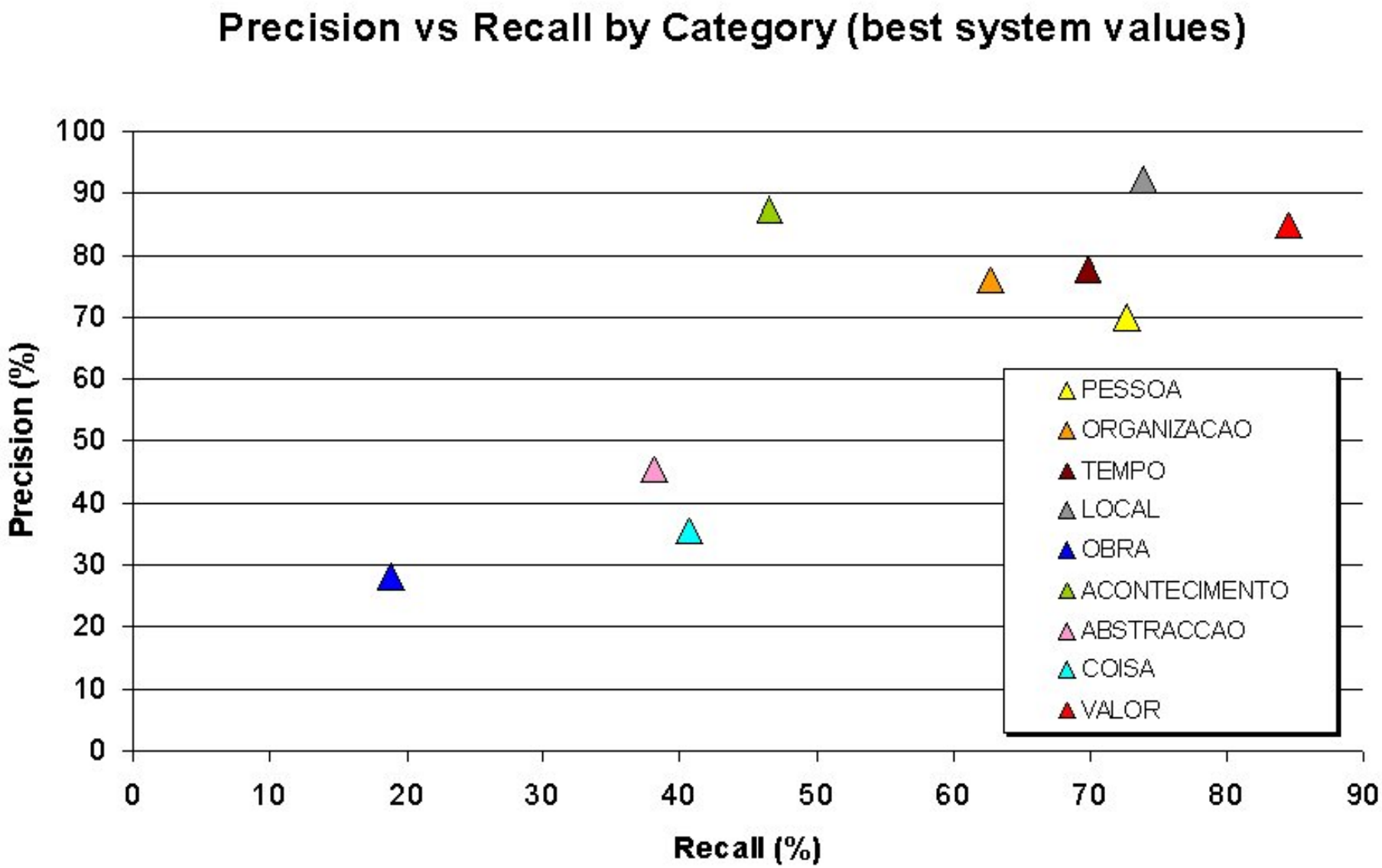
Ne_c - Number of NEs with exact category C

Nv_c - Number of NEs with category C, among others

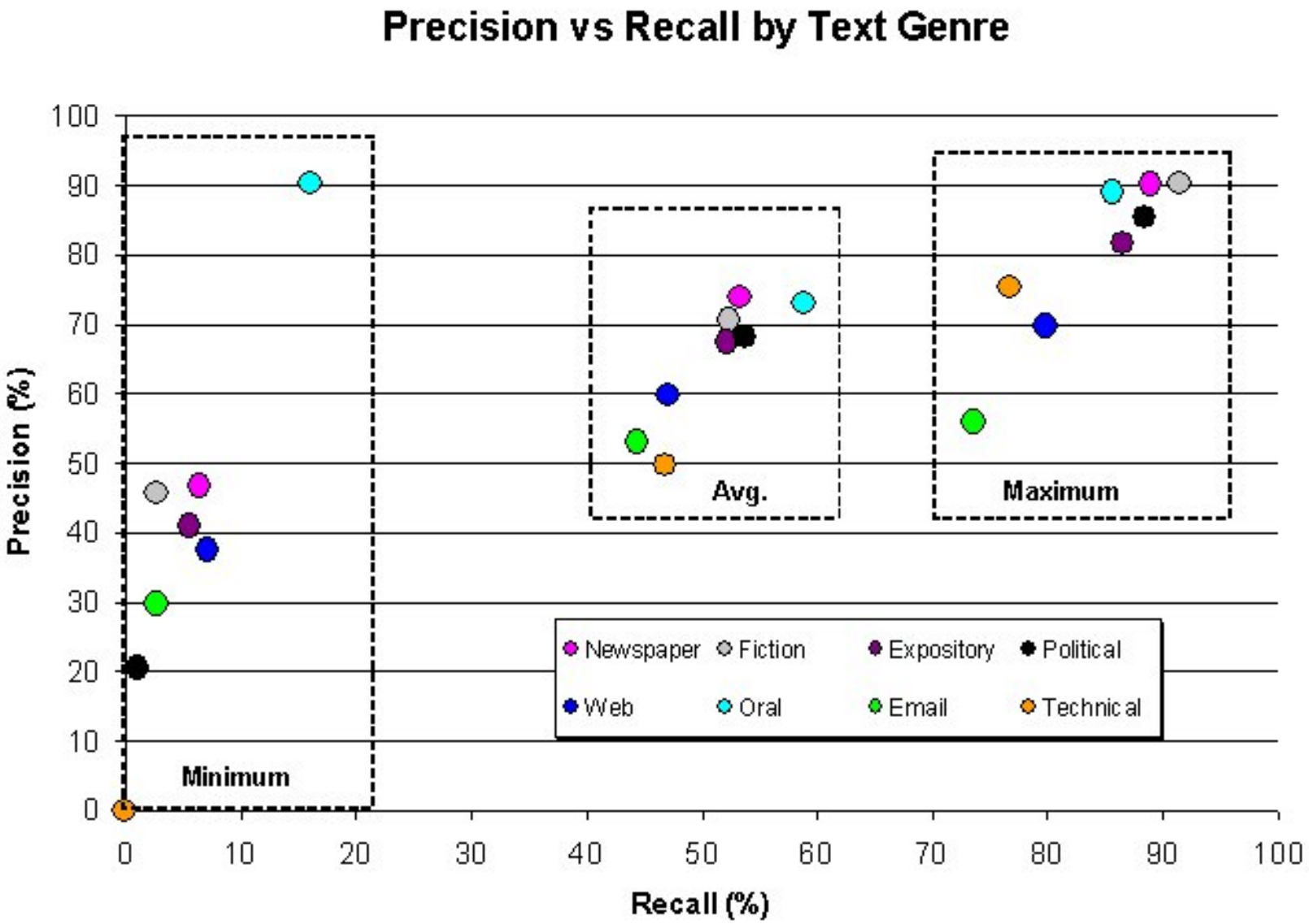
$$Ratio_C = \frac{Nv_C}{(Ne_C + Nv_C)}$$



Results of HAREM 2005



Official results of HAREM 2005



Outlook of HAREM

- We ran mini-HAREM with a subset of the HAREM participants for statistical testing (GC 2006)
- The final workshop of the 1st HAREM will take place on 15th July in Porto, after Linguateca's summer school. People are welcome to participate!
- Next editions: we intend to organize further contests, hoping to join more research groups working in different areas, such as GIR, ontology learning, semantic interpretation, ...
- See also our poster about the HAREM evaluation architecture.

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