HAREM e MiniHAREM: Uma análise comparativa

Nuno Cardoso

Encontro HAREM

15 de Julho de 2006

FLUP, Porto

Linguateca

Participação no HAREM

HAREM (14-2-**2005**)

- 10 participantes
- 6 países
  - PT, BR, MX, DK, ES e FR
- 18 saídas (3 não-oficiais)

MiniHAREM (3-4-**2006**)

- 5 participantes
- 2 países
  - PT e BR
- 20 saídas

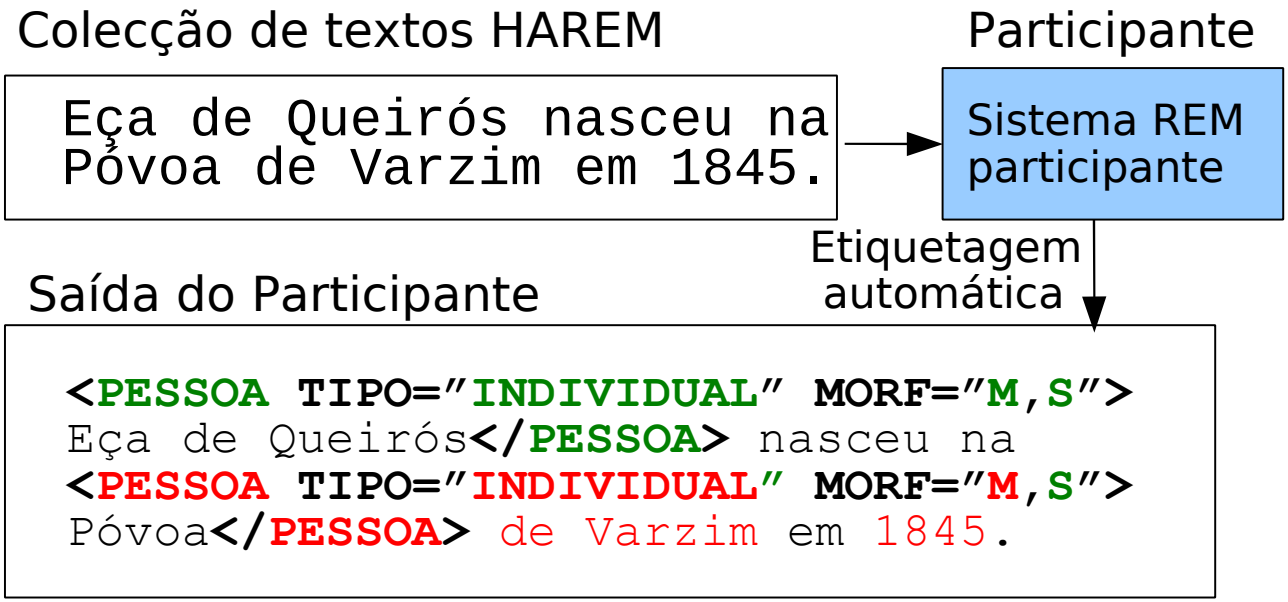
- Só para os participantes do primeiro HAREM

HAREM numa casca de noz

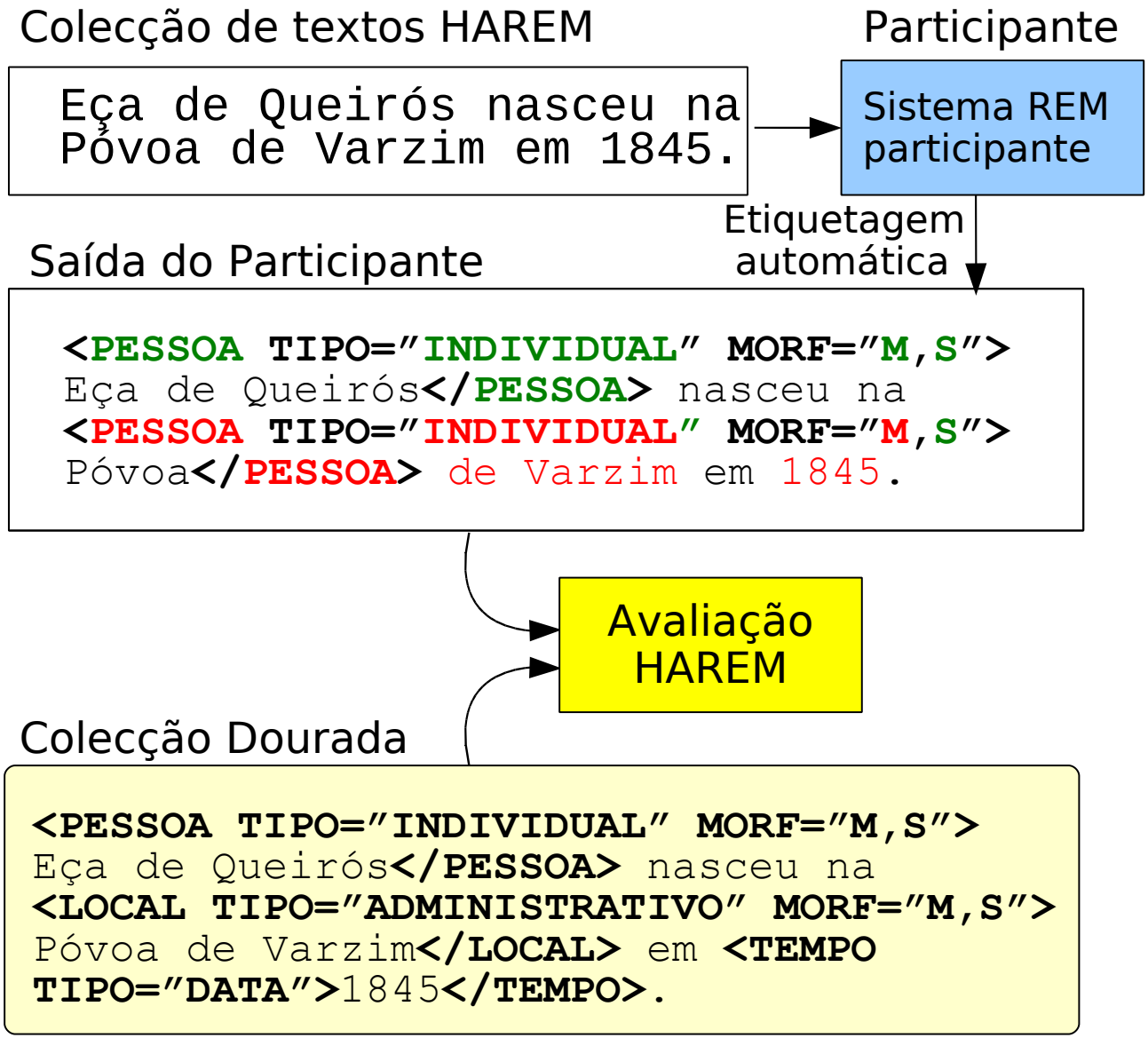
Colecção de textos HAREM

```
Eça de Queirós nasceu na
Póvoa de Varzim em 1845.
```

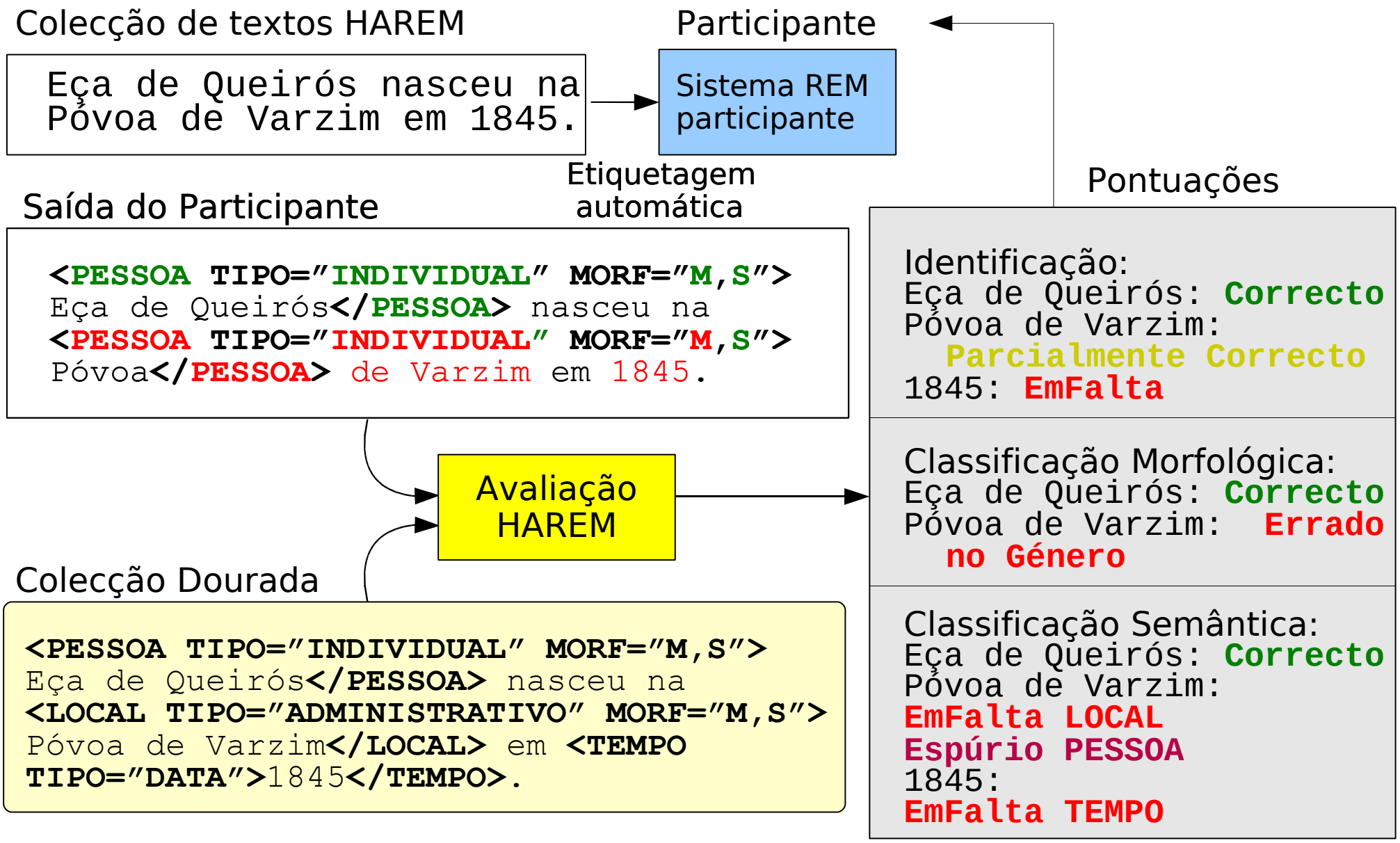
HAREM numa casca de noz



HAREM numa casca de noz



HAREM numa casca de noz



Cenários selectivos em 2005

- HAREM

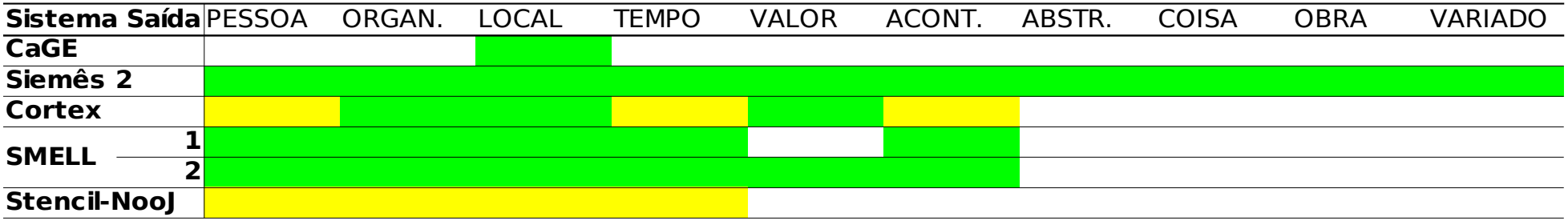
Sistema	Saída	PESSOA	ORGAN.	LOCAL	TEMPO	VALOR	ACONT.	ABSTR.	COISA	OBRA	VARIADO
CaGE	1										
	2										
	3										
Cortex	1										
	2										
	3										
ELLE											
Malinche											
Nerua	1										
	2										
	3										
RSN-NILC											
PALAVRAS-NER											
RENA											
SIEMÊS	1										
	2										

= alguns tipos

= todos os tipos

Cenários selectivos em 2006

- MiniHAREM



= alguns tipos

= todos os tipos

Colecção Dourada (CD)

Propósito da Colecção Dourada

- Incluir todas as EMs relevantes em texto português, obtendo uma marcação “ideal” de EMs no texto
- Categorias das EMs criadas empiricamente a partir de análise do texto. Categorização feita em dois níveis, categorias e tipos.
- Não representa o que os sistemas REM devem obter hoje, mas permitem:
  - avaliar a dificuldade da tarefa REM
  - estabelecer um limite superior
- Há muito mais em REM do que pessoas, organizações, locais e números...

Colecções Douradas usadas

- HAREM: Colecção Dourada de 2005 (CD 2005)
- MiniHAREM: Colecção Dourada de 2006 (CD 2006)
- Ambas as CDs foram retiradas da Colecção HAREM

Tamanhos	Colecção HAREM	CD 2005	CD 2006	Ambas
Palavras	600 086	92 830	62 461	155 291
Documentos	1 202	129	128	257
EMs	~ 40 000	5 270	3 858	9 128
EMs vagas (class.)	~ 1 000	133	142	275
EMs vagas (ident.)	~ 500	71	56	127

Categorias e Tipos em 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

Categorias e Tipos em 2006

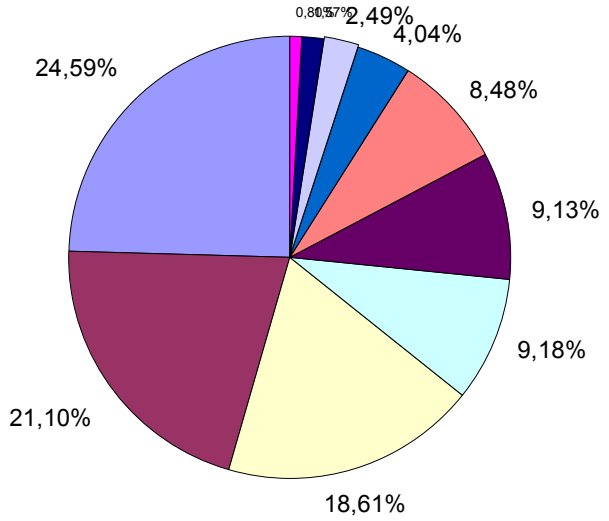
- **ABSTRACCAO**
  - DISCIPLINA
  - ESTADO
  - ESCOLA
  - OBRA
  - PLANO
  - IDEIA
  - NOME
- **OBRA**
  - ARTE
  - REPRODUZIDA
  - ~~**PRODUTO**~~
  - PUBLICACAO
- **ACONTECIMENTO**
  - EFEMERIDE
  - ORGANIZADO
  - EVENTO
- **COISA**
  - OBJECTO
  - SUBSTANCIA
  - CLASSE
  - **MEMBROCLASSE**
- **ORGANIZACAO**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

Principais alterações HAREM => MiniHAREM

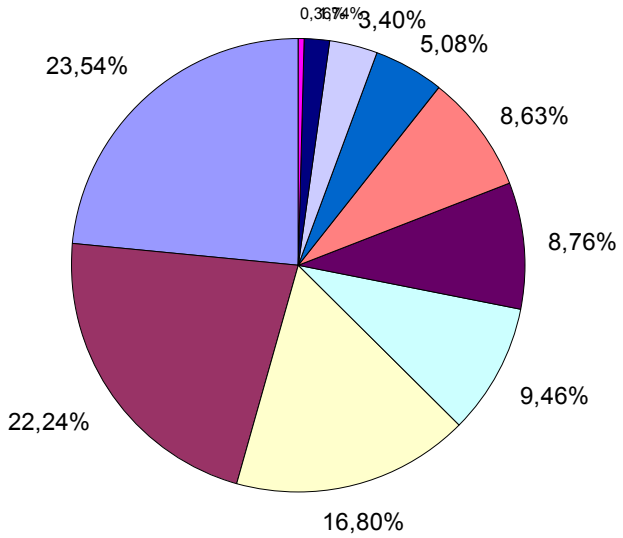
- OBRA TIPO=”PRODUTO” eliminado
- COISA TIPO=”MEMBROCLASSE” criado
- Emails e URLs deixam de ser marcados
- Referências anafóricas mantêm significado semântico
  - ex: Revolução de 1830 (...) e a de 1832...
- PESSOA TIPO=”MEMBRO” para referências a pessoas a partir de uma organização

CD: Distribuição por Categorias

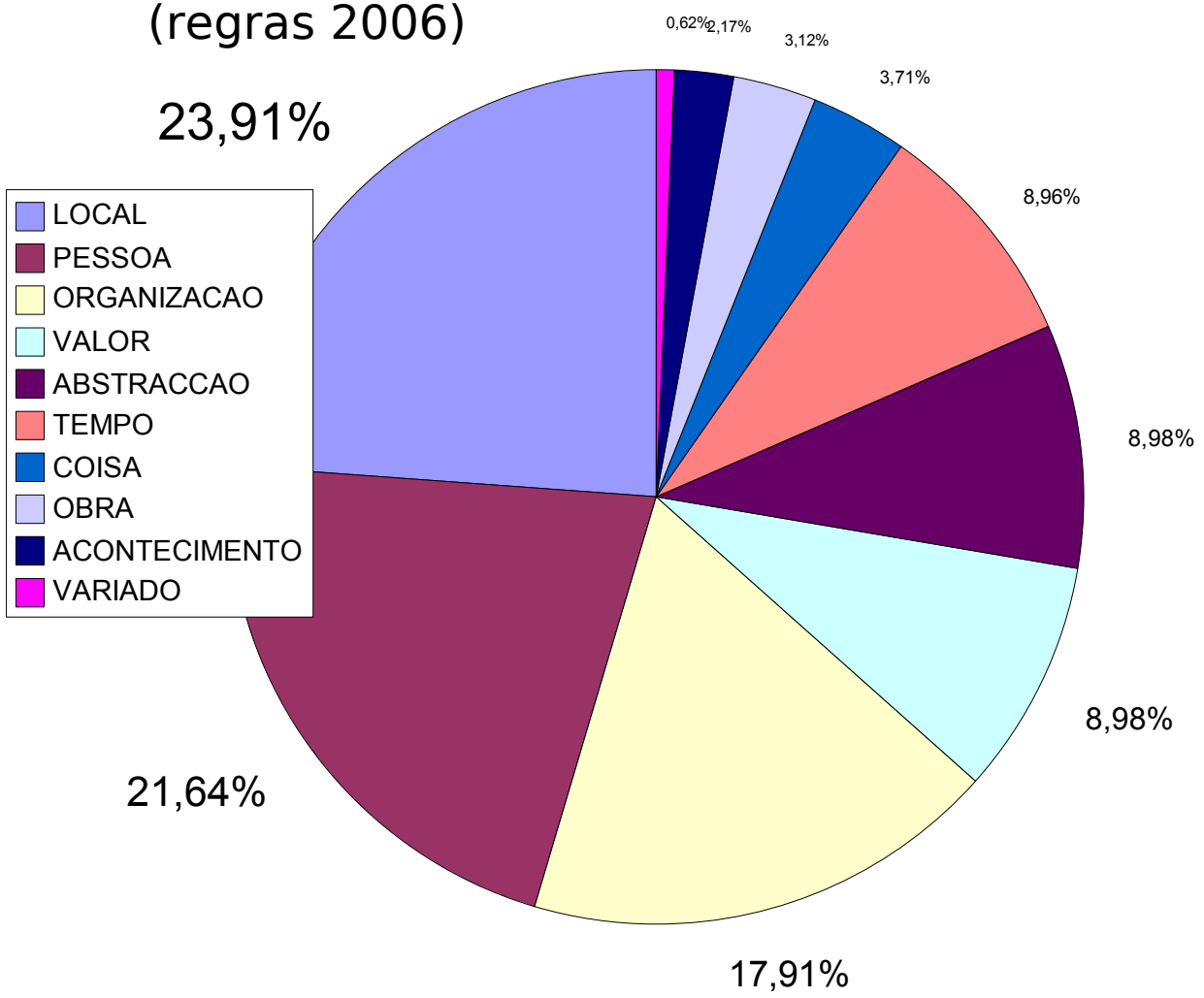
Colecção Dourada de 2005 (regras 2005)



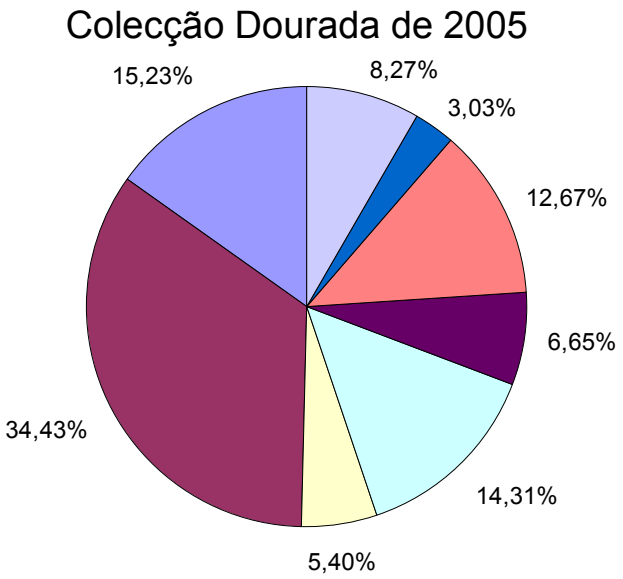
Colecção Dourada de 2006 (regras 2006)

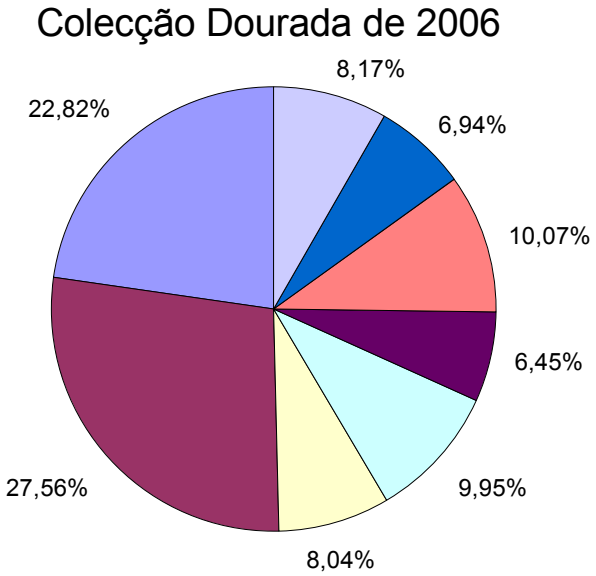


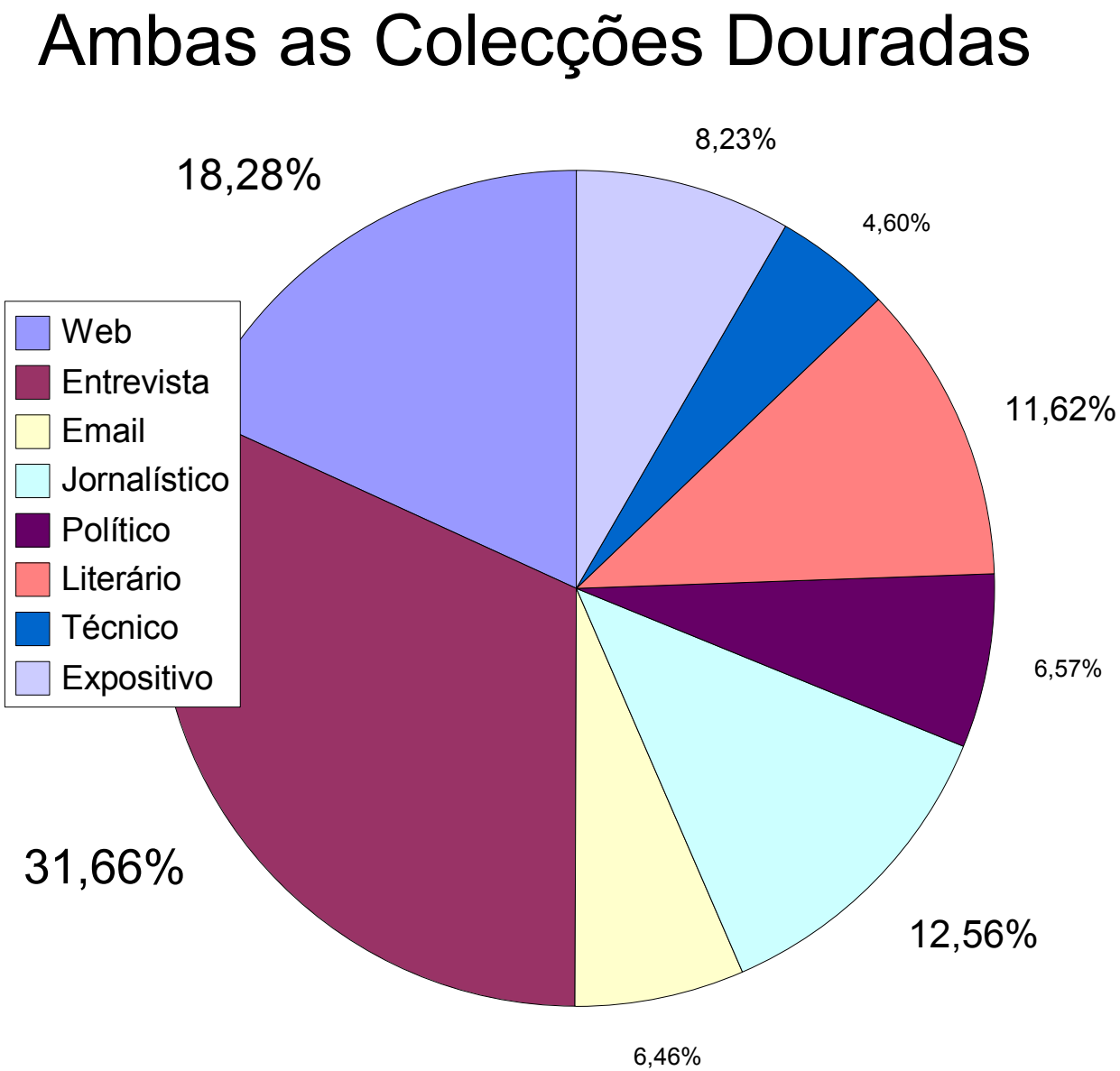
Ambas as Colecções Douradas (regras 2006)



CD: Distribuição por Género Textual (nº de palavras)

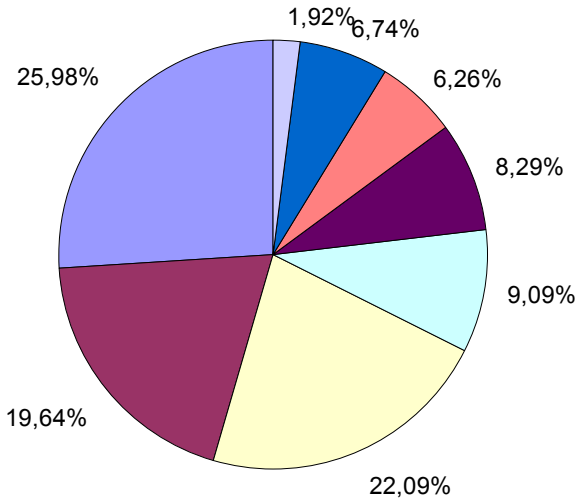




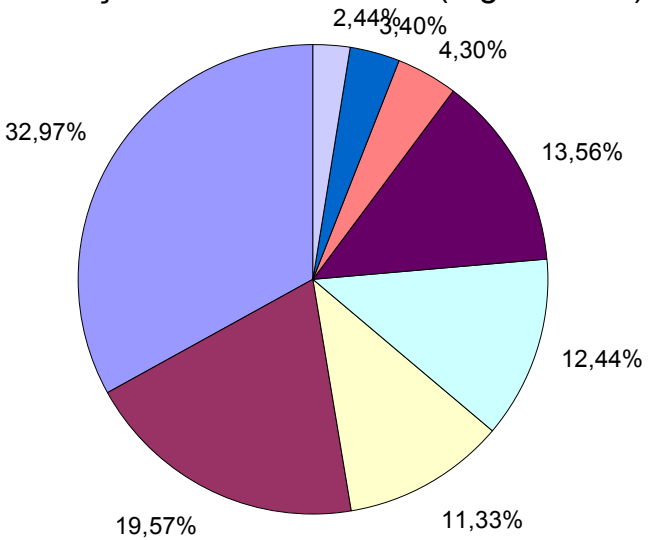


CD: Distribuição por Género Textual (nº de EMs)

Colecção Dourada de 2005 (regras 2005)

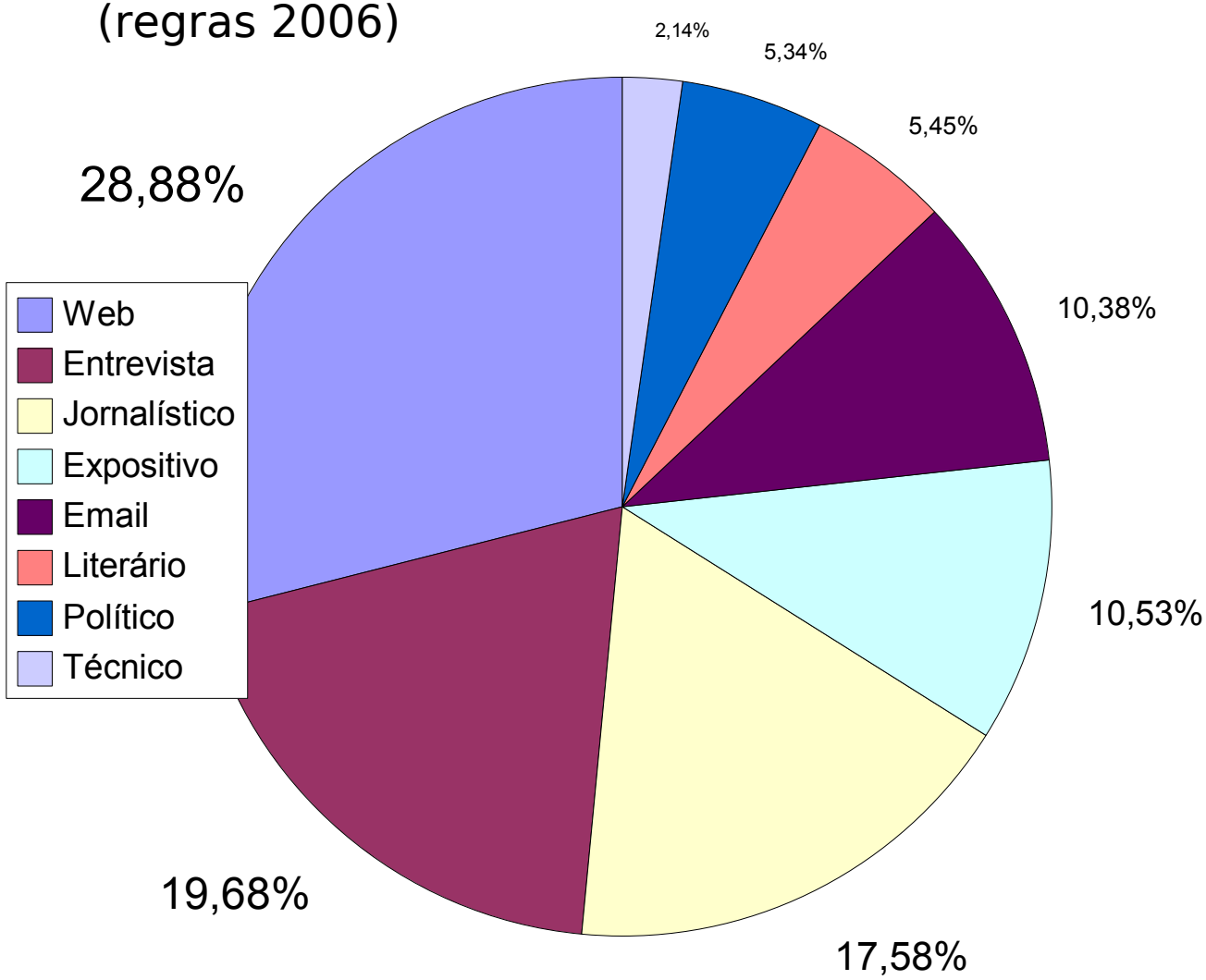


Colecção Dourada de 2006 (regras 2006)



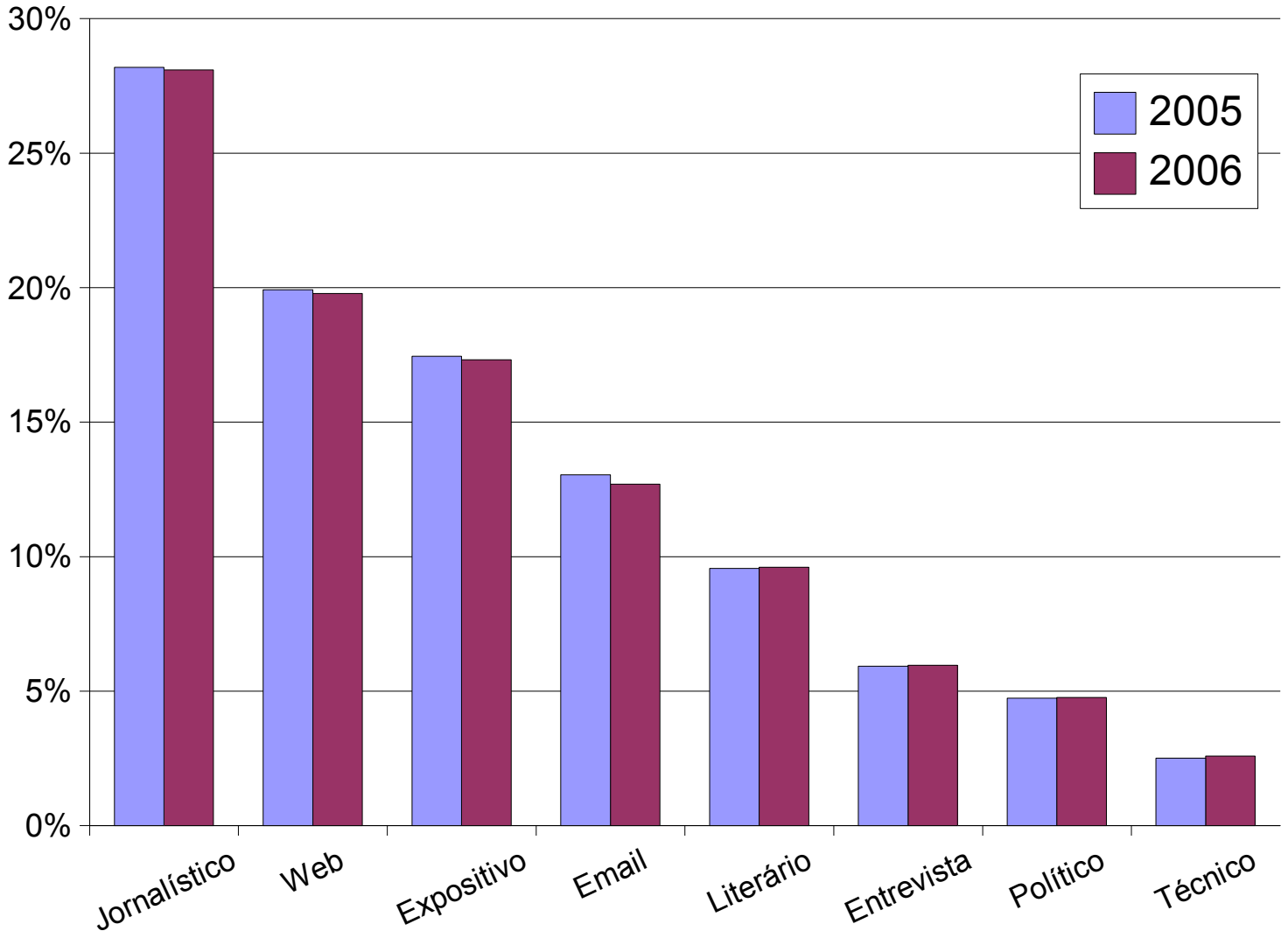
Ambas as Colecções Douradas

(regras 2006)



Densidade de EMs, por Género Textual

$$\text{Densidade EM}_G = \frac{\text{Nº palavras que pertencem e EMs, para o género G}}{\text{Nº total de palavras, para o género G}}$$



Nº de palavras das EMs, por Categoria

	2005			2006		
Categorias	Média	Mediana	Desv.Pad.	Média	Mediana	Desv.Pad.
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**

Teor de Sobreposição de EMs

Quantas EMs aparecem em ambas as Colecções Douradas?

	CD 2005	CD 2006
Número Total de EMs	5132	3712
Número de EMs Distintas	3060	2434
Rácio	59,63%	65,57%

Nº Total de EMs comuns	**623**	
Nº EMs Distintas comums	**380**	
Rácio de total EMs	12,14%	16,78%
Rácio de EMs distintas	12,42%	15,61%

Análise Estatística ao HAREM e MiniHAREM

vs

Objectivos da análise estatística

- Determinar o nível de confiança dos resultados do HAREM e MiniHAREM
- Distinguir os sistemas / estratégias realmente diferentes
- Verificar se o tamanho da CD é suficiente
- **Validar os eventos de avaliação**

Requisitos dos testes estatísticos

- Não paramétrico -- distribuição de EMs (quase) impossível de determinar --
- Comparação directa entre pares de saídas
- Simples e robusto
- Nível de confiança calculado facilmente

Teste estatístico escolhido: permutação

- Escolha: Testes de permutação
  - *Approximate Randomization*
- Sinopse do teste de permutação*:*
  - Permutar aleatoriamente observações entre duas saídas
  - Se a diferença inicial diminuir, é provável que as duas saídas sejam diferentes
  - Se a diferença inicial se mantiver, é provável que seja obra do acaso

Teste de Permutação - exemplo

Saída A 

$P_A = 16/20 = 80\%$

Saída B 

$P_B = 6/20 = 30\%$

Teste de Permutação - exemplo

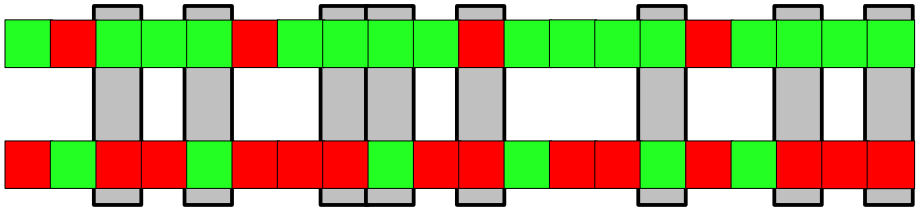
Saída A  $P_A = 16/20 = 80\%$

Saída B  $P_B = 6/20 = 30\%$

Fase 1: Permutação

- Troca de observações, com ½ de probabilidade

Saída A

Saída B 

Teste de Permutação - exemplo

Saída A  $P_A = 16/20 = 80\%$

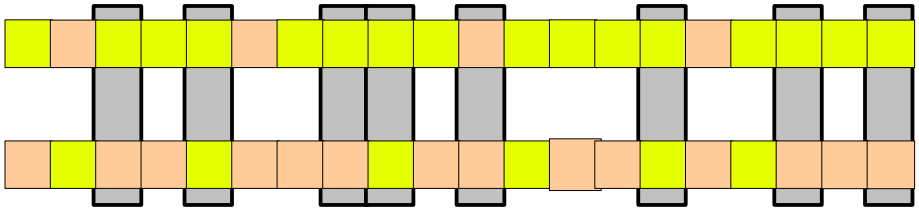
Saída B  $P_B = 6/20 = 30\%$

Fase 1: Permutação

- Troca de observações, com ½ de probabilidade

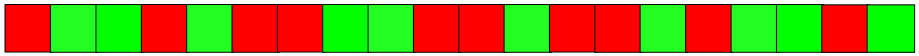
Saída A

Saída B



Fase 2: Cálculo da nova diferença

Pseudo-Saída A  $P^*_A = 12/20 = 60\%$

Pseudo-Saída B  $P^*_B = 10/20 = 50\%$

Teste de Permutação - exemplo

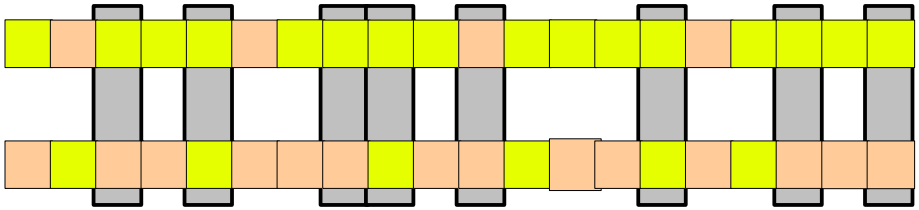
Saída A  $P_A = 16/20 = 80\%$

Saída B  $P_B = 6/20 = 30\%$

50%

Fase 1: Permutação

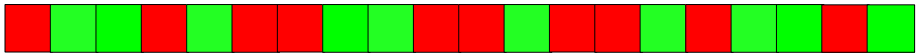
- Troca de observações, com ½ de probabilidade

Saída A 

Saída B

Fase 2: Cálculo da nova diferença

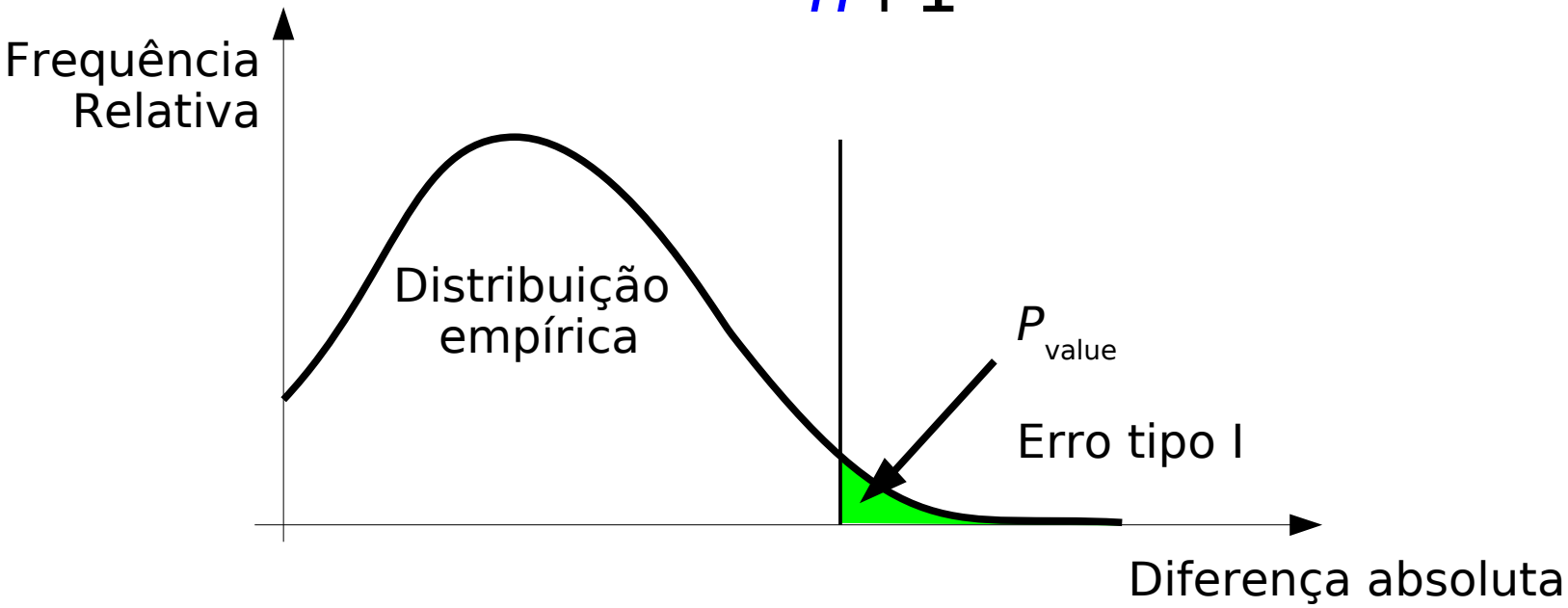
Pseudo-Saída A  $P^*_A = 12/20 = 60\%$

Pseudo-Saída B  $P^*_B = 10/20 = 50\%$

10%

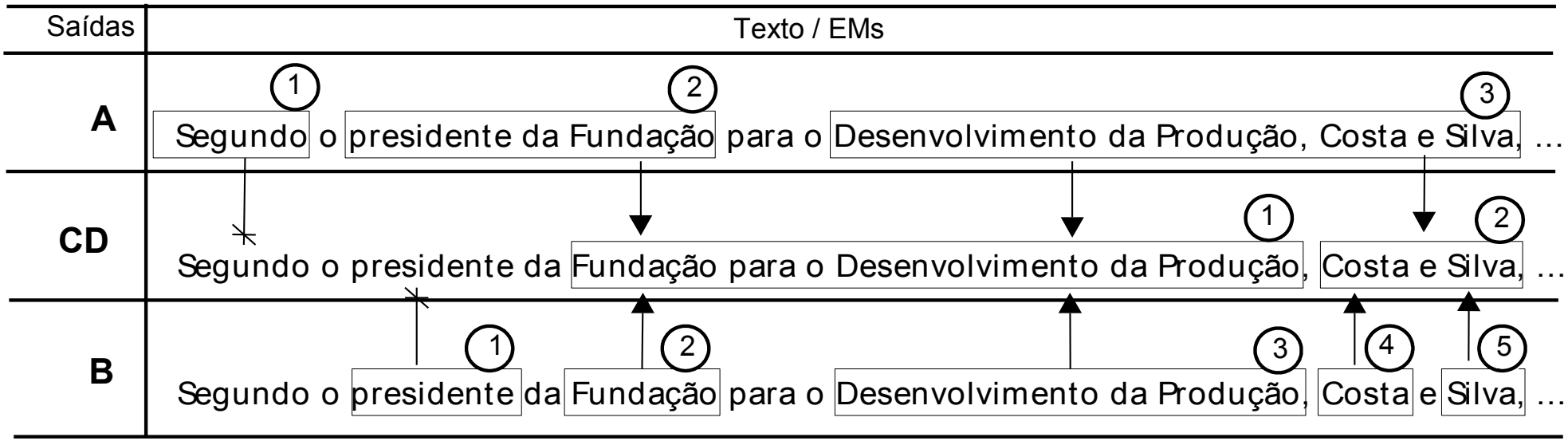
Teste de Permutação - exemplo

- Repetir para *n* iterações (ou seja, gerar *pseudo-saídas*)
  - Contar o nº de vezes *m* que a diferença aumentou em vez de diminuir
- Valor de P: $P_{value} = \frac{m+1}{n+1}$



Mas... como permutar no HAREM?

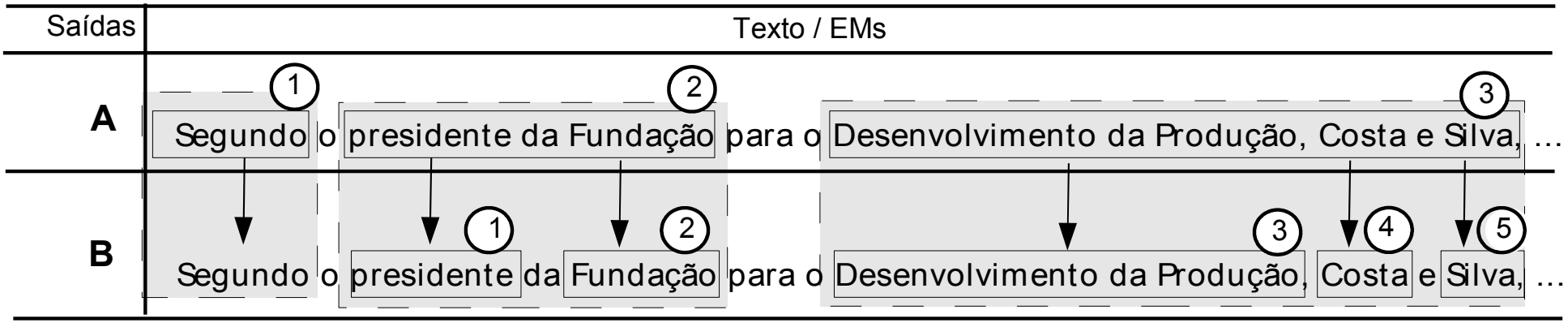
- Observações das saídas são dependentes!



- Como permutar “Costa”? E se “Silva” não for permutado?

Permutação por blocos

- Solução: agrupar EMs em blocos independentes



  - Mantém pontuações
  - Resolve diferenças dadas pelo <ALT>
- Ver a permutação como a troca de desempenho entre saídas, para uma determinada frase ou pedaço de texto

Tabelas de valores P

0,7977		**1,00000**	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00050	0,00020	0,00090	0,00010
0,8700		**1,00000**	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
0,8323		**1,00000**	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
0,7977	**1,00000**		0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00060	0,00010	0,00040	0,00010
0,8700	**1,00000**		0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
0,8323	**1,00000**		0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
0,7284	0,00010	0,00010		**1,00000**	0,00640	0,00520	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00040	0,00010
0,6958	0,00010	0,00010		**1,00000**	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
0,7117	0,00010	0,00010		**1,00000**	**0,00870**	**0,00930**	**0,00890**	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
0,7284	0,00010	0,00010	**1,00000**		0,00610	0,00560	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00040	0,00010
0,6958	0,00010	0,00010	**1,00000**		0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
0,7117	0,00010	0,00010	**1,00000**		**0,00900**	**0,00840**	**0,00900**	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
0,7320	0,00010	0,00010	0,0047	0,00610		**1,00000**	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00080	0,00010
0,6852	0,00010	0,00010	0,00010	0,00010		**1,00000**	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
0,7078	0,00010	0,00010	0,00650	0,00810		**1,00000**	**0,03210**	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
0,7320	0,00010	0,00010	0,00520	0,00540	**1,00000**		0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00100	0,00010
0,6852	0,00010	0,00010	0,00010	0,00010	**1,00000**		0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
0,7078	0,00010	0,00010	**0,01**	0,00750	**1,00000**		**0,02980**	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
0,8733	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010		0,00010	0,00010	0,00010	0,00010	0,00010	0,00100	**0,01160**
0,5707	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010		0,00010	0,00720	0,00010	0,00010	0,00010	0,00010	0,00010
0,6903	0,00010	0,00010	**0,00960**	**0,00930**	**0,03430**	**0,03200**		**0,84950**	0,00530	0,00010	0,00010	0,00010	0,00010	0,00010
0,7584	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010		**0,42330**	0,00010	0,00010	0,00010	**0,48480**	0,00010
0,6310	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010		0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
0,6888	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	**0,85220**		0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
0,7570	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	**0,42050**		0,00010	0,00010	0,00010	**0,41450**	0,00010
0,5969	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00710	0,00010		0,00010	0,00010	0,00010	0,00010	0,00010
0,6675	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00700	0,00010		0,00010	0,00010	0,00010	0,00010	0,00010
0,6835	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010		0,00010	0,00010	0,00010	0,00010
0,5317	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00020	0,00010	0,00010		0,00010	0,00010	0,00010	0,00010
0,5982	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010		0,00010	0,00010	0,00010	0,00010
0,8311	0,00060	0,00040	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010		**0,64320**	0,00010	**0,07550**
0,4165	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010		**0,31040**	**0,10520**	0,00010
0,5549	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010		**0,43020**	0,00290	0,00010
0,8366	0,00010	0,00020	0,00010	0,00010	0,00010	0,00010	0,00020	0,00010	0,00010	0,00010	**0,61220**		0,00010	0,00030
0,4072	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	**0,31530**		**0,49480**	0,00010
0,5477	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	**0,43520**		**0,01730**	0,00010
0,7655	0,00080	0,00110	0,00010	0,00030	0,00080	0,00120	0,00010	**0,46020**	**0,39540**	0,00010	0,00010	0,00010		0,00010
0,4011	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	**0,10620**	**0,49600**		0,00010
0,5264	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00260	**0,01770**		**0,05280**
0,8511	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	**0,01250**	0,00010	0,00010	0,00010	**0,06780**	0,00030	0,00010	
0,3617	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	
0,5076	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	**0,05060**	
0,8288	0,00140	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
0,3606	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
0,5026	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010
	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010

Testes estatísticos com tamanho da CD

Saída A			***Saída B***			***Diferença***		
Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF
0.728	**0.696**	**0.712**	**0.798**	**0.870**	**0.832**	**0.069**	**0.174**	**0.121**

Testes estatísticos com tamanho da CD

Saída A			Saída B			Diferença		
Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF
0.728	**0.696**	**0.712**	**0.798**	**0.870**	**0.832**	**0.069**	**0.174**	**0.121**

- Utilizando cada vez menos blocos no teste...

n iterações = 9999

	Valor de P			PSEUDO-SAÍDAS de A: Média			PSEUDO-SAÍDAS de A: Desvio Padrão			PSEUDO-SAÍDAS de B: Média			PSEUDO-SAÍDAS de B: Desvio Padrão		
# Blocos	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF
Todos	0.0001	0.0001	0.0001	0.765	0.783	0.774	0.003	0.004	0.003	0.765	0.783	0.774	0.003	0.004	0.003
2000	0.0001	0.0001	0.0001	0.765	0.772	0.769	0.008	0.009	0.007	0.765	0.772	0.769	0.008	0.009	0.007
1000	0.0001	0.0001	0.0001	0.765	0.765	0.765	0.013	0.015	0.012	0.765	0.765	0.765	0.013	0.014	0.012
500	0.0004	0.0001	0.0001	0.766	0.761	0.763	0.019	0.022	0.017	0.765	0.761	0.763	0.018	0.021	0.017
250	**0.0181**	0.0001	0.0001	0.765	0.759	0.762	0.027	0.031	0.025	0.766	0.760	0.763	0.027	0.031	0.025
200	**0.0351**	0.0001	0.0001	0.765	0.759	0.762	0.030	0.034	0.028	0.765	0.759	0.762	0.030	0.035	0.028
100	**0.1391**	0.0009	0.0047	0.766	0.759	0.761	0.043	0.049	0.040	0.765	0.758	0.761	0.044	0.049	0.041
75	**0.1912**	0.0034	**0.0123**	0.767	0.759	0.762	0.050	0.057	0.047	0.767	0.759	0.762	0.050	0.057	0.047
50	**0.2900**	**0.0181**	**0.0453**	0.766	0.759	0.761	0.062	0.069	0.058	0.766	0.760	0.761	0.062	0.068	0.057
25	**0.4488**	**0.0843**	**0.1505**	0.767	0.762	0.762	0.087	0.093	0.079	0.766	0.760	0.760	0.086	0.093	0.079

- ... aumenta o desvio padrão das pseudo-saídas...

Testes estatísticos com tamanho da CD

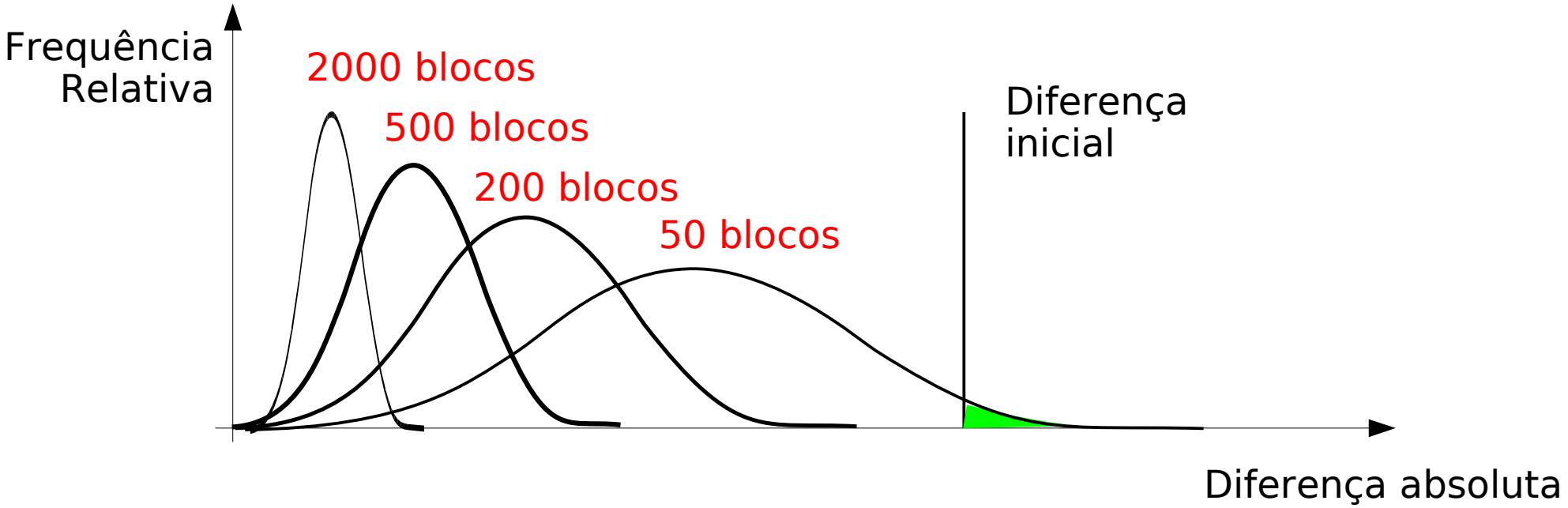
Saída A			*Saída B*			*Diferença*		
Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF
0.728	**0.696**	**0.712**	**0.798**	**0.870**	**0.832**	**0.069**	**0.174**	**0.121**

- ... e aumenta a média e desvio padrão das diferenças entre pseudo-saídas!

PSEUDO-DIFERENÇAS

	Valor de P			Média			Desvio Padrão		
# Blocos	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF
Todos	0.0001	0.0001	0.0001	0.006	0.006	0.005	0.004	0.005	0.004
2000	0.0001	0.0001	0.0001	0.008	0.009	0.008	0.006	0.007	0.006
1000	0.0001	0.0001	0.0001	0.012	0.013	0.011	0.009	0.010	0.008
500	0.0004	0.0001	0.0001	0.017	0.018	0.015	0.013	0.014	0.011
250	**0.0181**	0.0001	0.0001	0.024	0.026	0.021	0.018	0.020	0.016
200	**0.0351**	0.0001	0.0001	0.026	0.029	0.024	0.020	0.022	0.018
100	**0.1391**	0.0009	0.0047	0.037	0.041	0.034	0.028	0.031	0.026
75	**0.1912**	0.0034	**0.0123**	0.043	0.048	0.039	0.032	0.036	0.029
50	**0.2900**	**0.0181**	**0.0453**	0.053	0.058	0.048	0.040	0.045	0.036
25	**0.4488**	**0.0843**	**0.1505**	0.073	0.081	0.066	0.056	0.061	0.051

Variação do nº de blocos *vs* valor de P



PSEUDO-DIFERENÇAS

	Valor de P			Média			Desvio Padrão		
# Blocos	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF
Todos	0.0001	0.0001	0.0001	0.006	0.006	0.005	0.004	0.005	0.004
2000	0.0001	0.0001	0.0001	0.008	0.009	0.008	0.006	0.007	0.006
1000	0.0001	0.0001	0.0001	0.012	0.013	0.011	0.009	0.010	0.008
500	0.0004	0.0001	0.0001	0.017	0.018	0.015	0.013	0.014	0.011
250	**0.0181**	0.0001	0.0001	0.024	0.026	0.021	0.018	0.020	0.016
200	**0.0351**	0.0001	0.0001	0.026	0.029	0.024	0.020	0.022	0.018
100	**0.1391**	0.0009	0.0047	0.037	0.041	0.034	0.028	0.031	0.026
75	**0.1912**	0.0034	**0.0123**	0.043	0.048	0.039	0.032	0.036	0.029
50	**0.2900**	**0.0181**	**0.0453**	0.053	0.058	0.048	0.040	0.045	0.036
25	**0.4488**	**0.0843**	**0.1505**	0.073	0.081	0.066	0.056	0.061	0.051

Prova dos nove...

Saída A			Saída B			Diferença		
Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF
0.728	**0.696**	**0.712**	**0.798**	**0.870**	**0.832**	**0.069**	**0.174**	**0.121**

PSEUDO-DIFERENÇAS

	Valor de P			Média			Desvio Padrão		
# Blocos	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF
Todos	0.0001	0.0001	0.0001	0.006	0.006	0.005	0.004	0.005	0.004
2000	0.0001	0.0001	0.0001	0.008	0.009	0.008	0.006	0.007	0.006
1000	0.0001	0.0001	0.0001	0.012	0.013	0.011	0.009	0.010	0.008
500	0.0004	0.0001	0.0001	0.017	0.018	0.015	0.013	0.014	0.011
250	**0.0181**	0.0001	0.0001	0.024	0.026	0.021	0.018	0.020	0.016
200	**0.0351**	0.0001	0.0001	0.026	0.029	0.024	0.020	0.022	0.018
100	**0.1391**	0.0009	0.0047	0.037	0.041	0.034	0.028	0.031	0.026
75	**0.1912**	0.0034	**0.0123**	0.043	0.048	0.039	0.032	0.036	0.029
50	**0.2900**	**0.0181**	**0.0453**	0.053	0.058	0.048	0.040	0.045	0.036
25	**0.4488**	**0.0843**	**0.1505**	0.073	0.081	0.066	0.056	0.061	0.051

Teste T-Student (99%) = 2,58					
Média - t * desvPad			Média + t * desvPad		
-0.005	-0.006	-0.005	0.016	0.019	0.015
-0.008	-0.009	-0.007	0.024	0.028	0.022
-0.011	-0.013	-0.01	0.035	0.039	0.032
-0.016	-0.017	-0.014	0.050	0.054	0.044
-0.022	-0.025	-0.02	**0.069**	0.078	0.063
-0.025	-0.028	-0.023	**0.078**	0.086	0.070
-0.036	-0.039	-0.033	**0.110**	0.122	0.100
-0.041	-0.045	-0.037	**0.126**	0.140	0.114
-0.05	-0.057	-0.046	**0.155**	**0.173**	**0.142**
-0.072	-0.078	-0.065	**0.218**	**0.239**	**0.198**

Influencia do nº de iterações no valor de P

n blocos = 2000

	PSEUDO-SAÍDAS de A									PSEUDO-SAÍDAS de B									PSEUDO-DIFERENÇAS					
	Valor de P			Média			Desvio Padrão			Média			Desvio Padrão			Média			Desvio Padrão		
# Iter	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF
9999	0.0001	0.0001	0.0001	0.765	0.772	0.769	0.008	0.009	0.007	0.765	0.772	0.769	0.008	0.009	0.007	0.00831	0.00937	0.00750	0.00626	0.00707	0.00570
999	0.0010	0.0010	0.0010	0.765	0.772	0.768	0.008	0.009	0.007	0.765	0.771	0.768	0.008	0.009	0.007	0.00836	0.00930	0.00744	0.00609	0.00692	0.00562
99	0.0100	0.0100	0.0100	0.764	0.771	0.768	0.009	0.009	0.007	0.766	0.772	0.769	0.009	0.009	0.008	0.00984	0.00952	0.00855	0.00750	0.00642	0.00609

n blocos = 200

	Valor de P			Média			Desvio Padrão			Média			Desvio Padrão			Média			Desvio Padrão		
# Iter	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF
9999	**0.0351**	0.0001	0.0001	0.765	0.759	0.762	0.030	0.034	0.028	0.765	0.759	0.762	0.030	0.035	0.028	0.02650	0.02904	0.02379	0.01991	0.02212	0.01804
999	**0.0290**	0.0010	0.0010	0.767	0.758	0.762	0.030	0.036	0.029	0.767	0.759	0.762	0.030	0.035	0.028	0.02516	0.02896	0.02360	0.01866	0.02212	0.01753
99	**0.0500**	0.0100	0.0100	0.772	0.760	0.766	0.032	0.035	0.029	0.767	0.760	0.763	0.032	0.039	0.032	0.02613	0.03271	0.02526	0.02005	0.02390	0.01957

n blocos = 25

	Valor de P			Média			Desvio Padrão			Média			Desvio Padrão			Média			Desvio Padrão		
# Iter	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF
9999	**0.4488**	**0.0843**	**0.1505**	0.767	0.762	0.762	0.087	0.093	0.079	0.766	0.760	0.760	0.086	0.093	0.079	0.07302	0.08053	0.06637	0.05618	0.06136	0.05085
999	**0.4330**	**0.0930**	**0.1580**	0.766	0.761	0.761	0.086	0.094	0.079	0.764	0.761	0.760	0.089	0.096	0.082	0.07373	0.08299	0.06646	0.05823	0.06107	0.05393
99	**0.4800**	**0.0800**	**0.1200**	0.762	0.764	0.760	0.090	0.088	0.077	0.765	0.763	0.762	0.088	0.096	0.082	0.07849	0.08180	0.06811	0.05604	0.06272	0.04987

Uma experiência com nº iterações

n blocos = 2000

	PSEUDO-SAÍDAS de A									PSEUDO-SAÍDAS de B									PSEUDO-DIFERENÇAS					
	Valor de P			Média			Desvio Padrão			Média			Desvio Padrão			Média			Desvio Padrão				
# Iter	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF		
9999	0.0001	0.0001	0.0001	0.765	0.772	0.769	0.008	0.009	0.007	0.765	0.772	0.769	0.008	0.009	0.007	0.00831	0.00937	0.00750	0.00626	0.00707	0.00570		
999	0.0010	0.0010	0.0010	0.765	0.772	0.768	0.008	0.009	0.007	0.765	0.771	0.768	0.008	0.009	0.007	0.00836	0.00930	0.00744	0.00609	0.00692	0.00562		
99	0.0100	0.0100	0.0100	0.764	0.771	0.768	0.009	0.009	0.007	0.766	0.772	0.769	0.009	0.009	0.008	0.00984	0.00952	0.00855	0.00750	0.00642	0.00609		

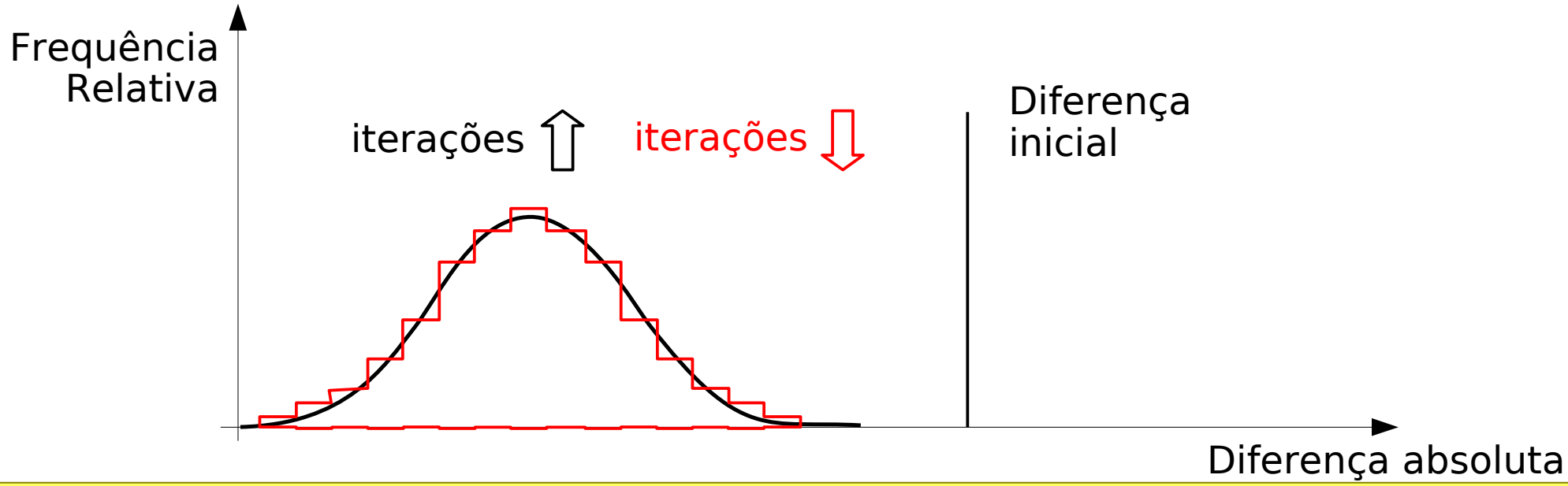
n blocos = 200

	Valor de P			Média			Desvio Padrão			Média			Desvio Padrão			Média			Desvio Padrão				
# Iter	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF		
9999	**0.0351**	0.0001	0.0001	0.765	0.759	0.762	0.030	0.034	0.028	0.765	0.759	0.762	0.030	0.035	0.028	0.02650	0.02904	0.02379	0.01991	0.02212	0.01804		
999	**0.0290**	0.0010	0.0010	0.767	0.758	0.762	0.030	0.036	0.029	0.767	0.759	0.762	0.030	0.035	0.028	0.02516	0.02896	0.02360	0.01866	0.02212	0.01753		
99	**0.0500**	0.0100	0.0100	0.772	0.760	0.766	0.032	0.035	0.029	0.767	0.760	0.763	0.032	0.039	0.032	0.02613	0.03271	0.02526	0.02005	0.02390	0.01957		

n blocos = 25

	Valor de P			Média			Desvio Padrão			Média			Desvio Padrão			Média			Desvio Padrão				
# Iter	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF	Prec	Abr	MedF		
9999	**0.4488**	**0.0843**	**0.1505**	0.767	0.762	0.762	0.087	0.093	0.079	0.766	0.760	0.760	0.086	0.093	0.079	0.07302	0.08053	0.06637	0.05618	0.06136	0.05085		
999	**0.4330**	**0.0930**	**0.1580**	0.766	0.761	0.761	0.086	0.094	0.079	0.764	0.761	0.760	0.089	0.096	0.082	0.07373	0.08299	0.06646	0.05823	0.06107	0.05393		
99	**0.4800**	**0.0800**	**0.1200**	0.762	0.764	0.760	0.090	0.088	0.077	0.765	0.763	0.762	0.088	0.096	0.082	0.07849	0.08180	0.06811	0.05604	0.06272	0.04987		

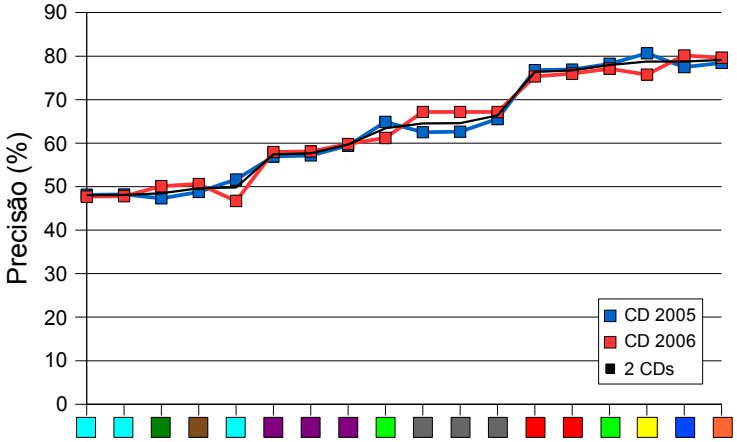
- Sem grandes diferenças. Média e desvio padrão na mesma.

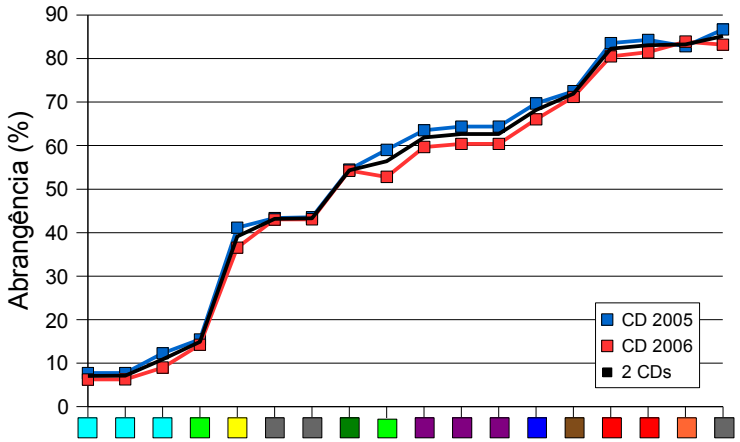


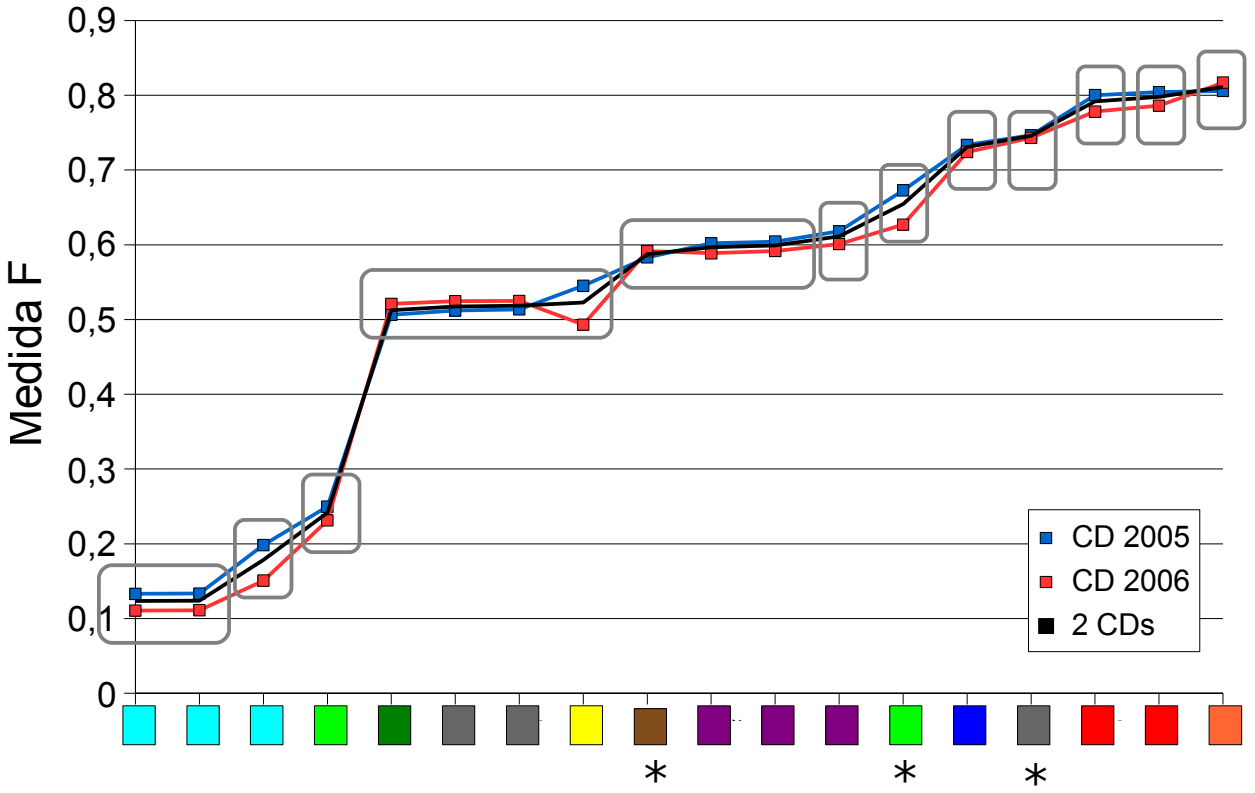
Resultados finais

HAREM **2005**

Tarefa de **Identificação**



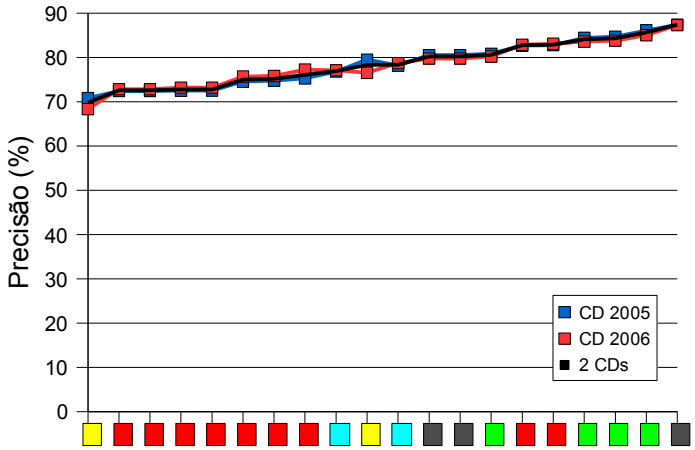


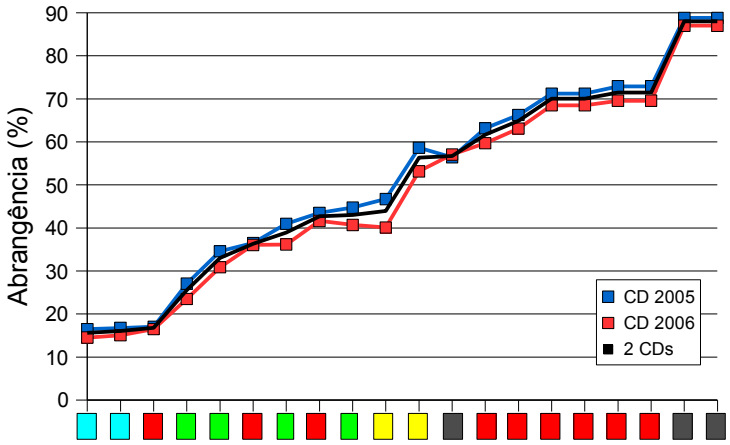


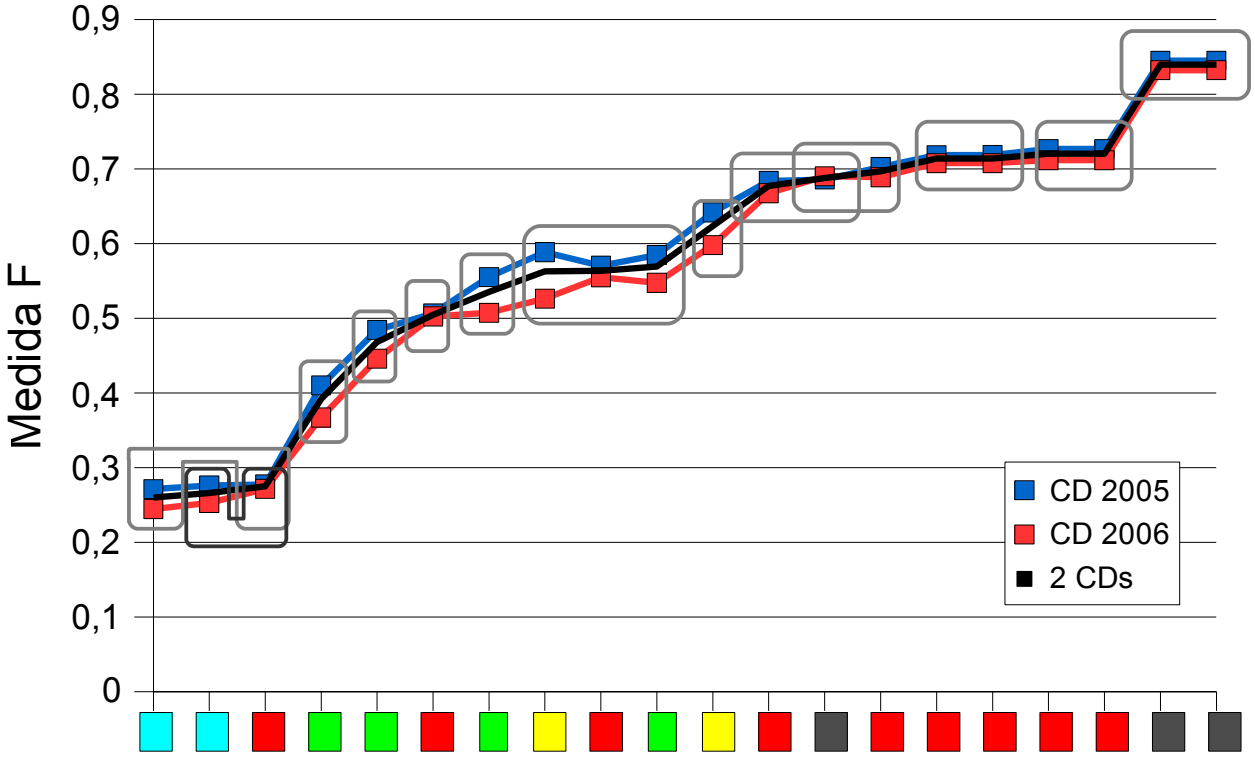
* - Saídas não oficiais

MiniHAREM **2006**

Tarefa de **Identificação**

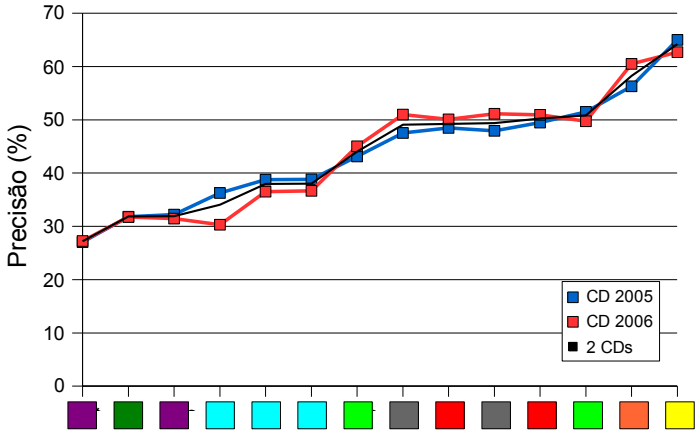


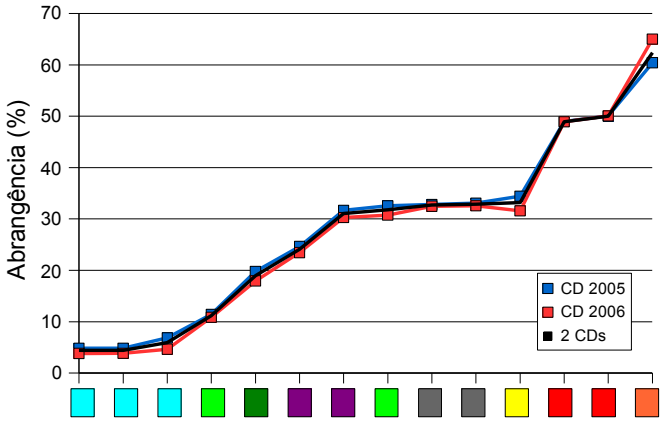


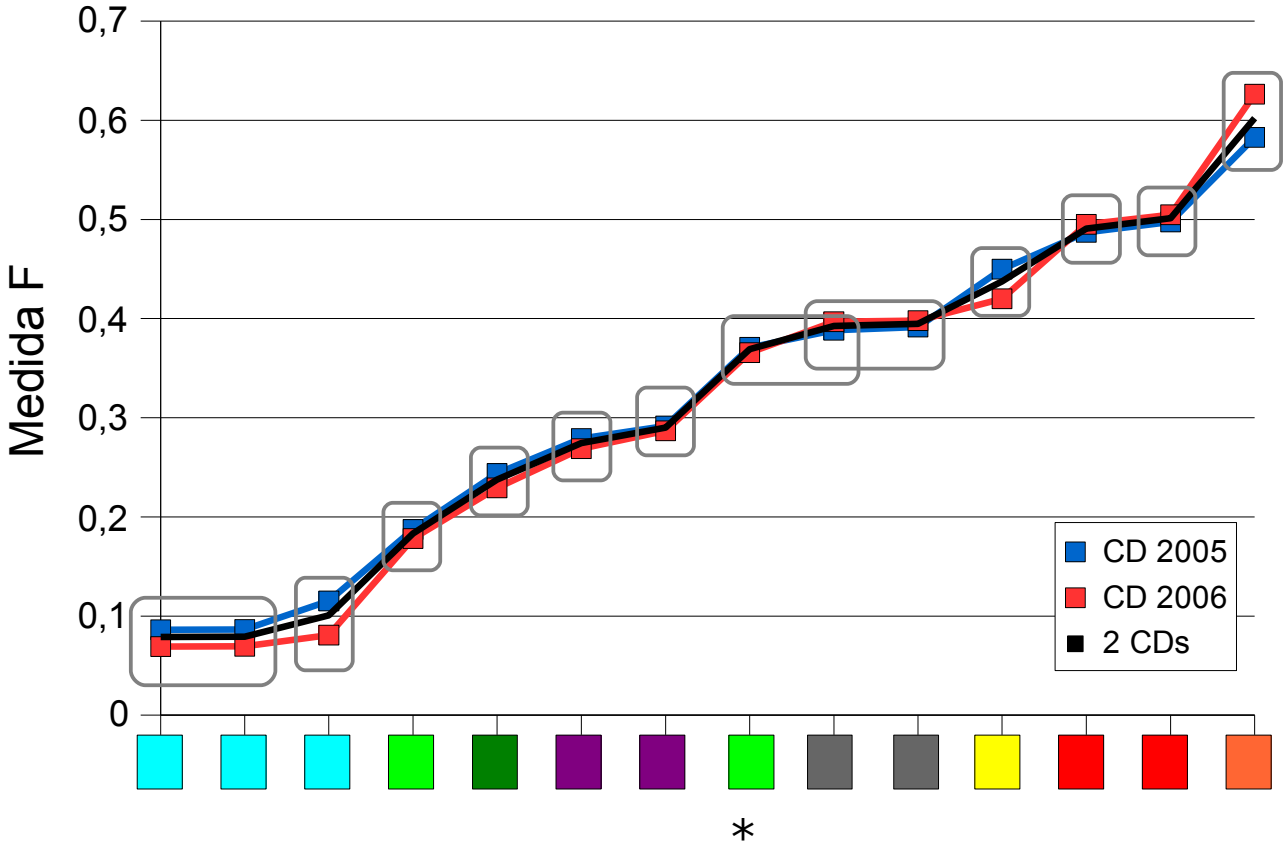


HAREM **2005**

Tarefa de **Classificação Semântica** (Combinada)



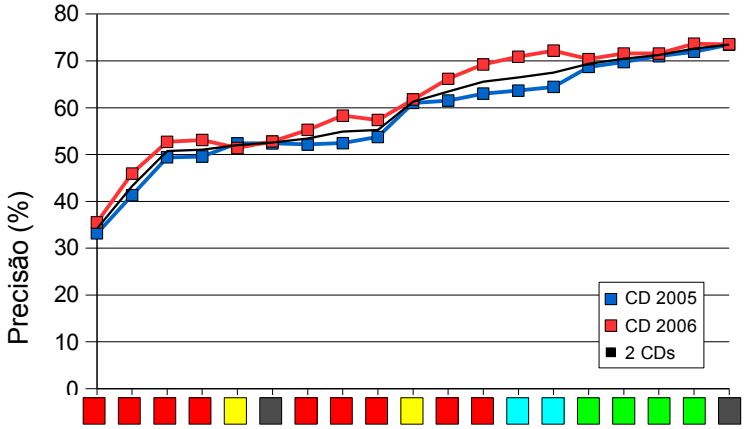


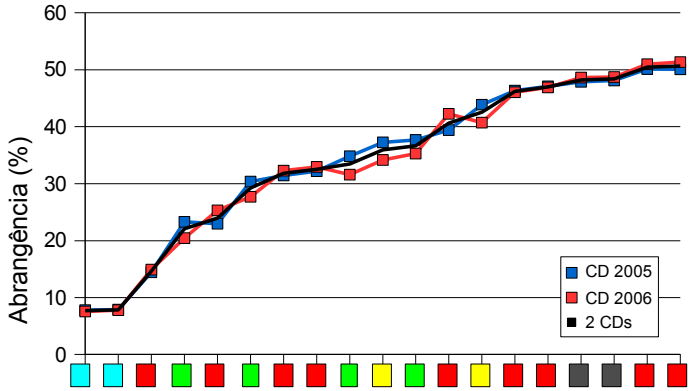


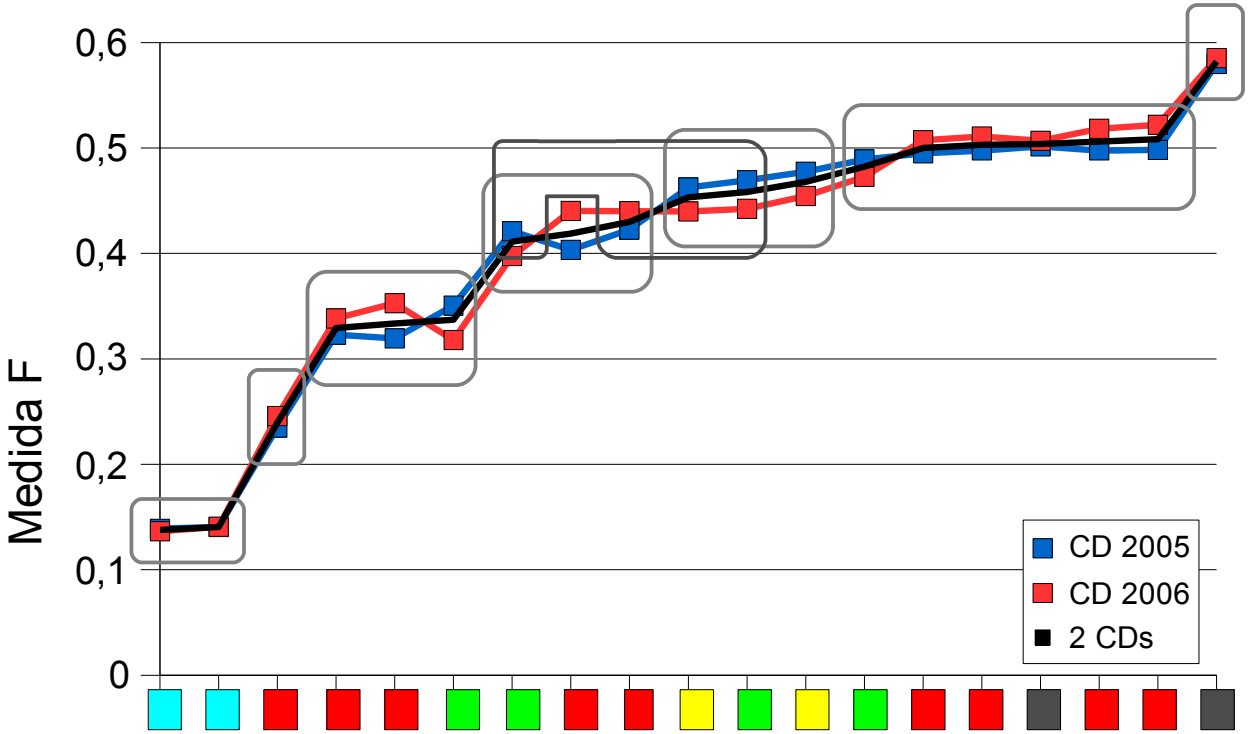
* - Saída não-oficial

MiniHAREM **2006**

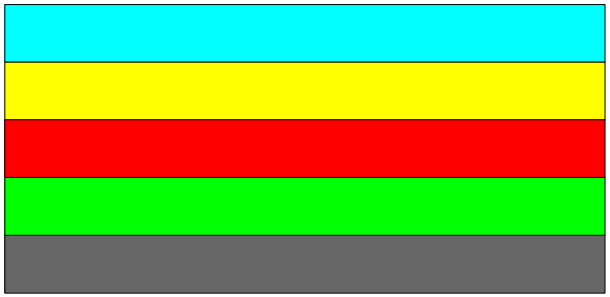
Tarefa de **Classificação Semântica** (Combinada)

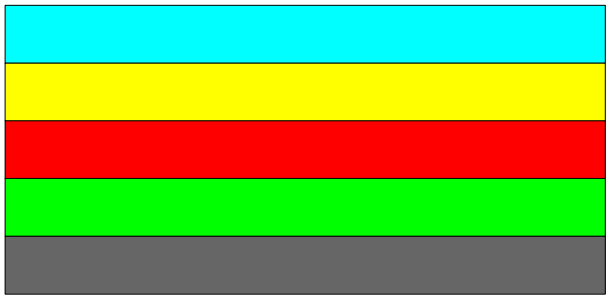






Evolução dos Sistemas, em 1 ano?

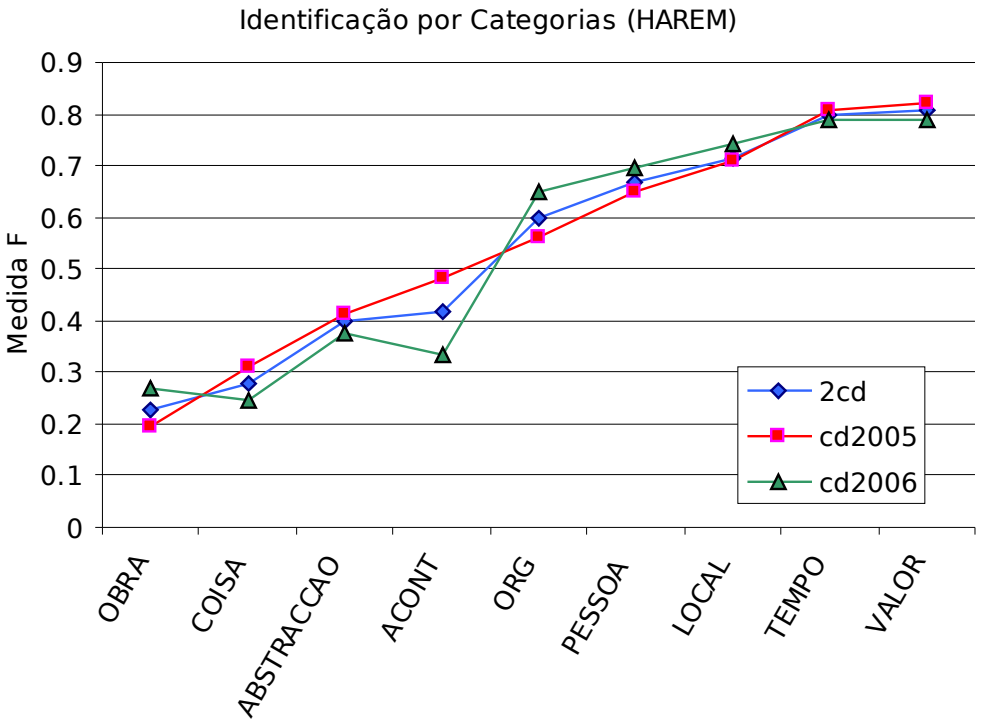
Identificação	2005	2006	%
	0,178	0,266	49,4%
	0,523	0,624	19,2%
	0,798	0,720	-9,7%
	0,655	0,569	-13,0%
	0,746	**0,839**	12,6%

Semântica (CSC)	2005	2006	%
	0,101	0,141	39,8%
	0,438	0,468	6,9%
	0,501	0,508	1,4%
	0,369	0,482	30,7%
	0,395	**0,582**	47,5%

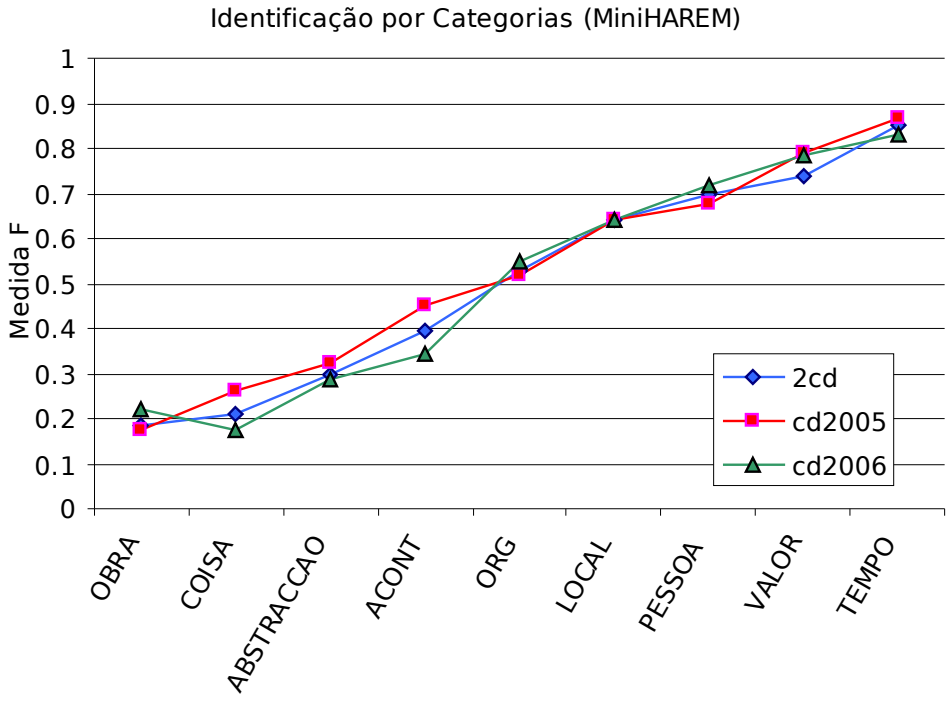
Estado da Arte em REM: Categorias

Panorama de Identificação, por Categorias

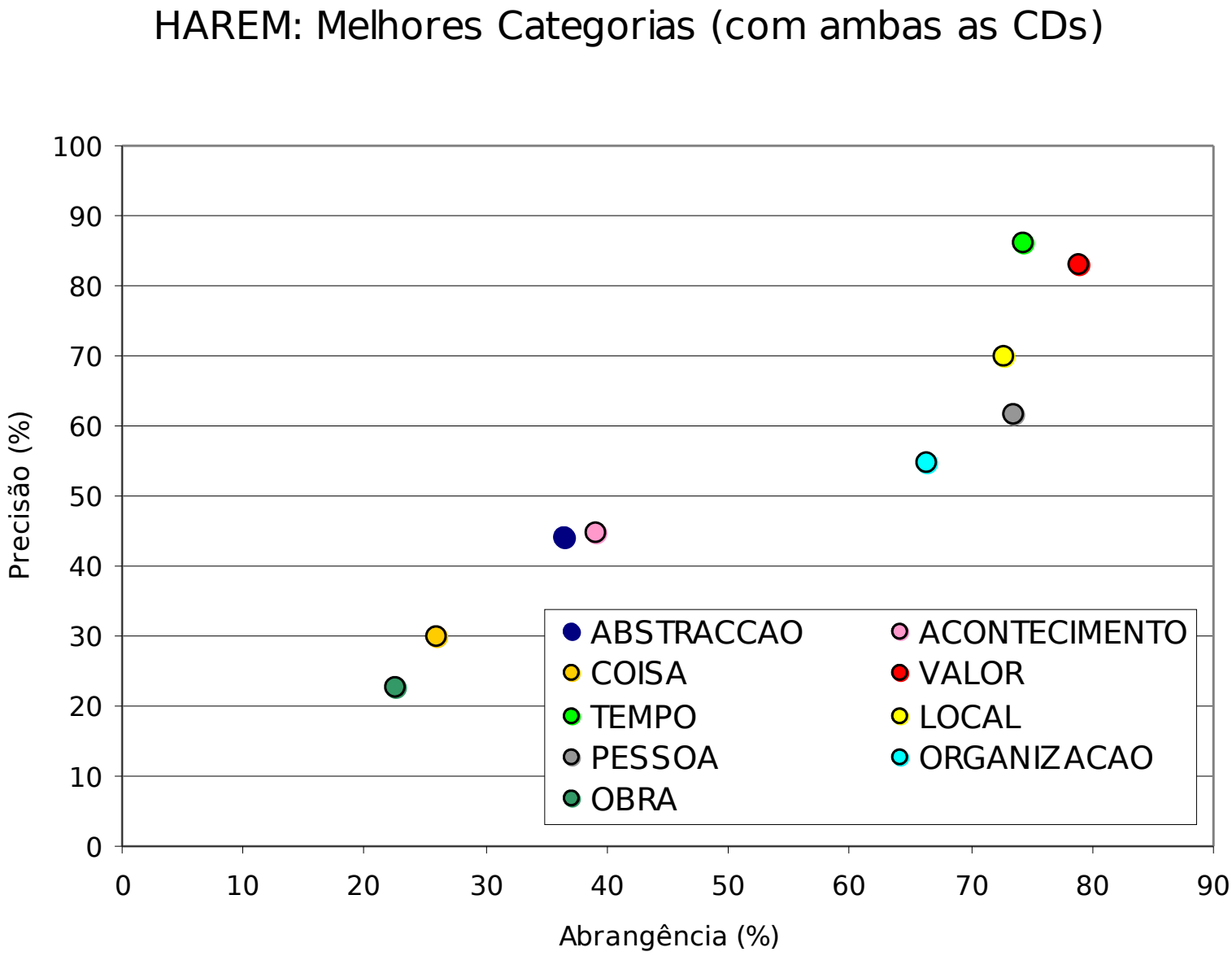
2005



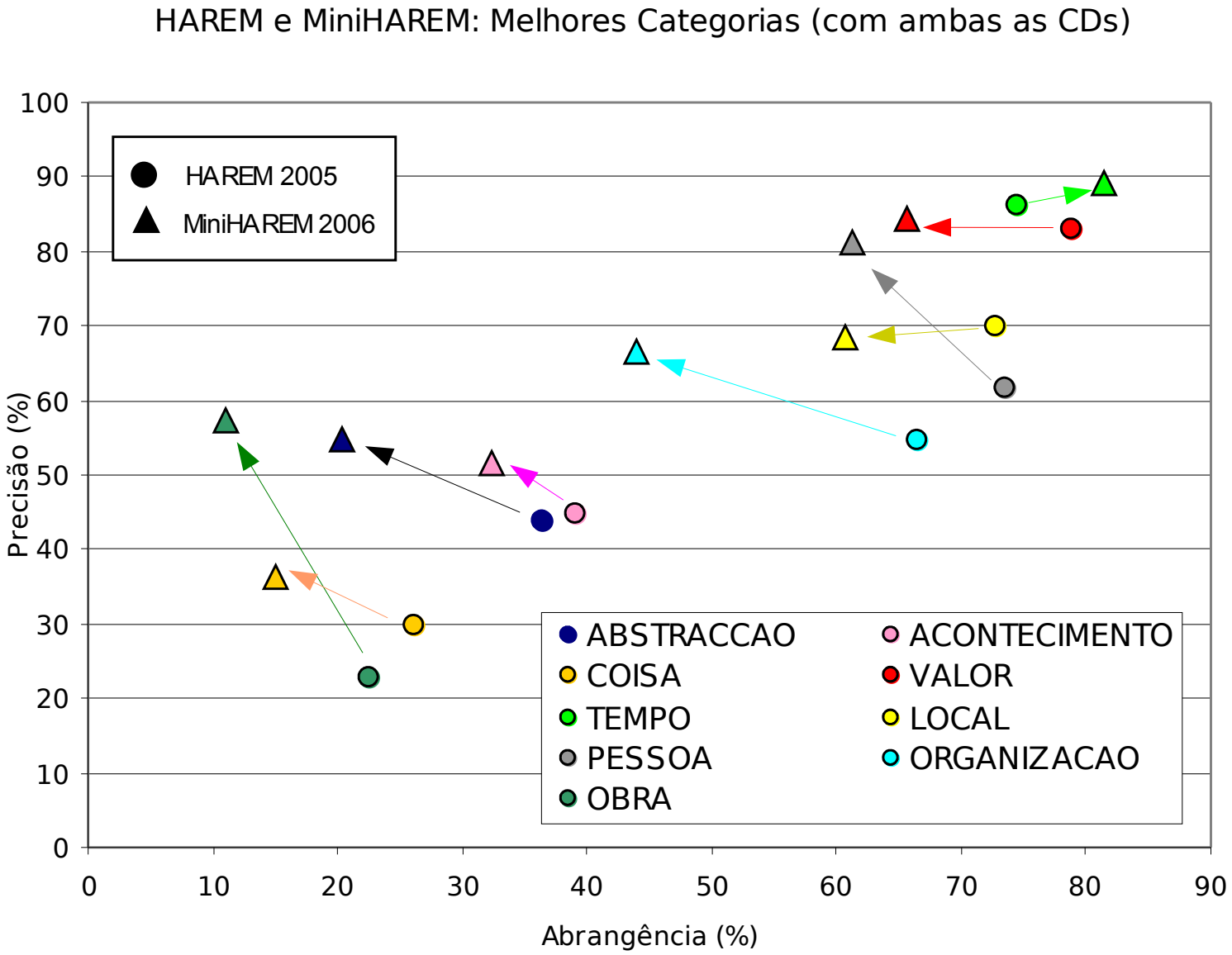
2006



Panorama de Identificação, por Categorias

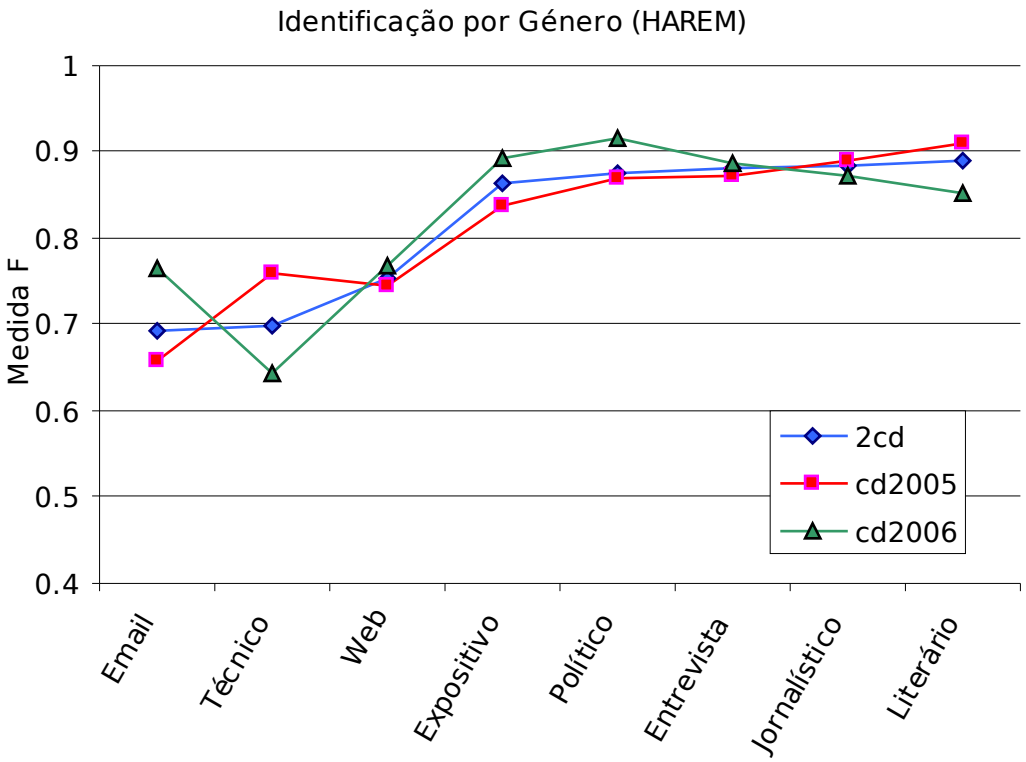


Panorama de Identificação, por Categorias

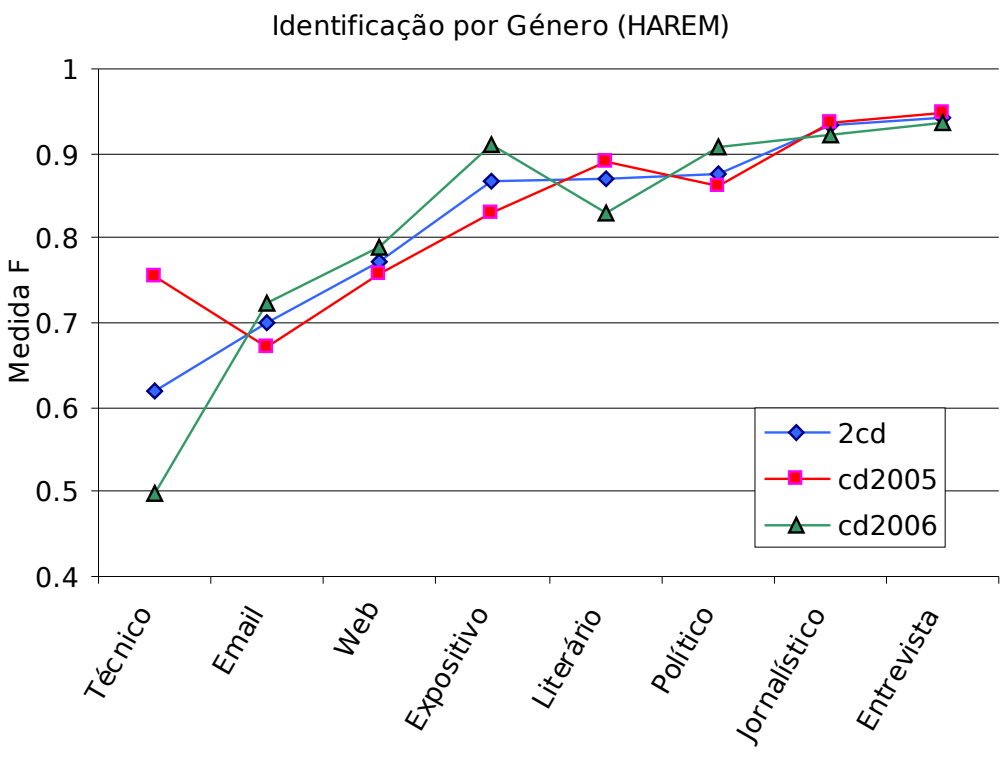


Panorama de Identificação, por Género

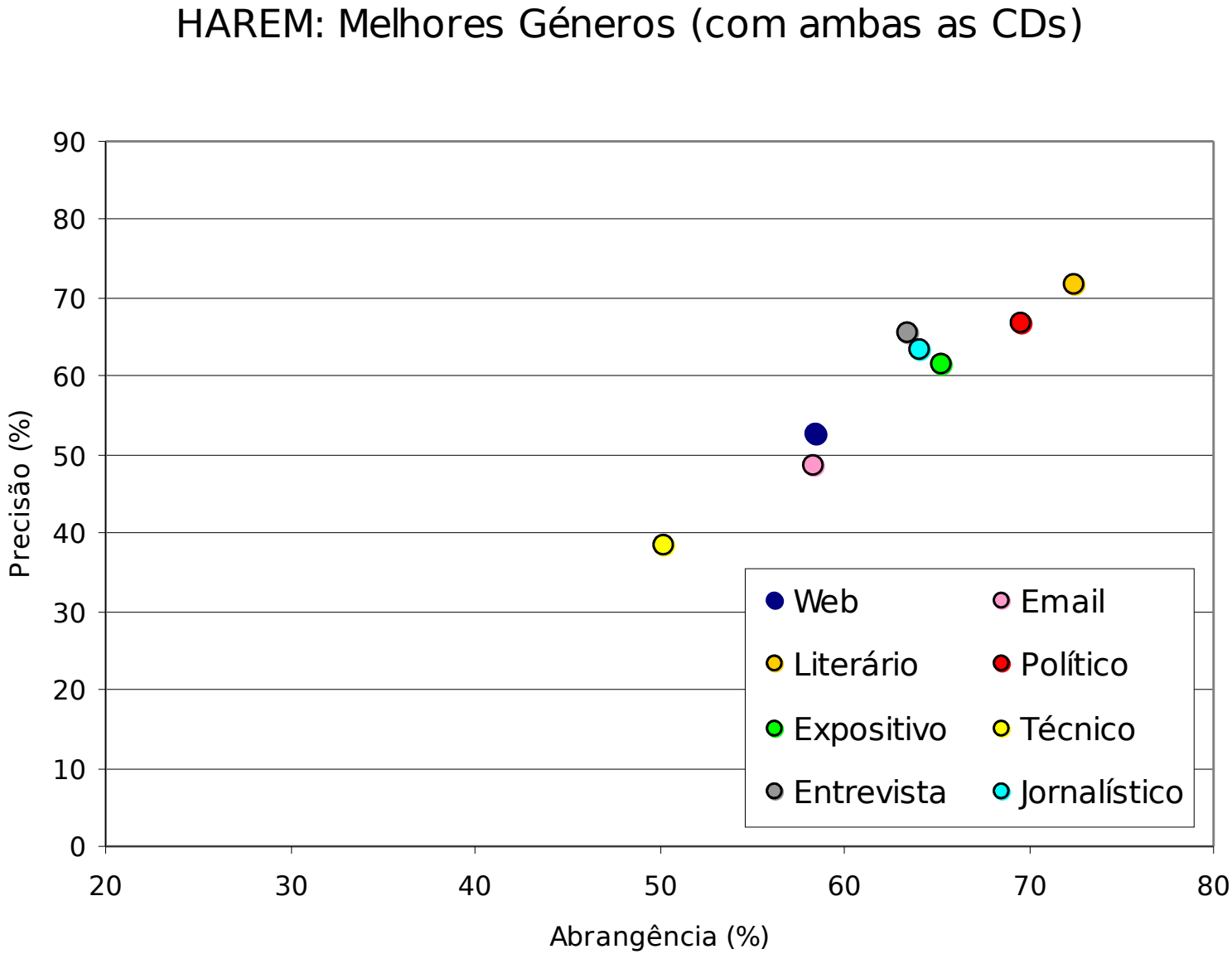
2005



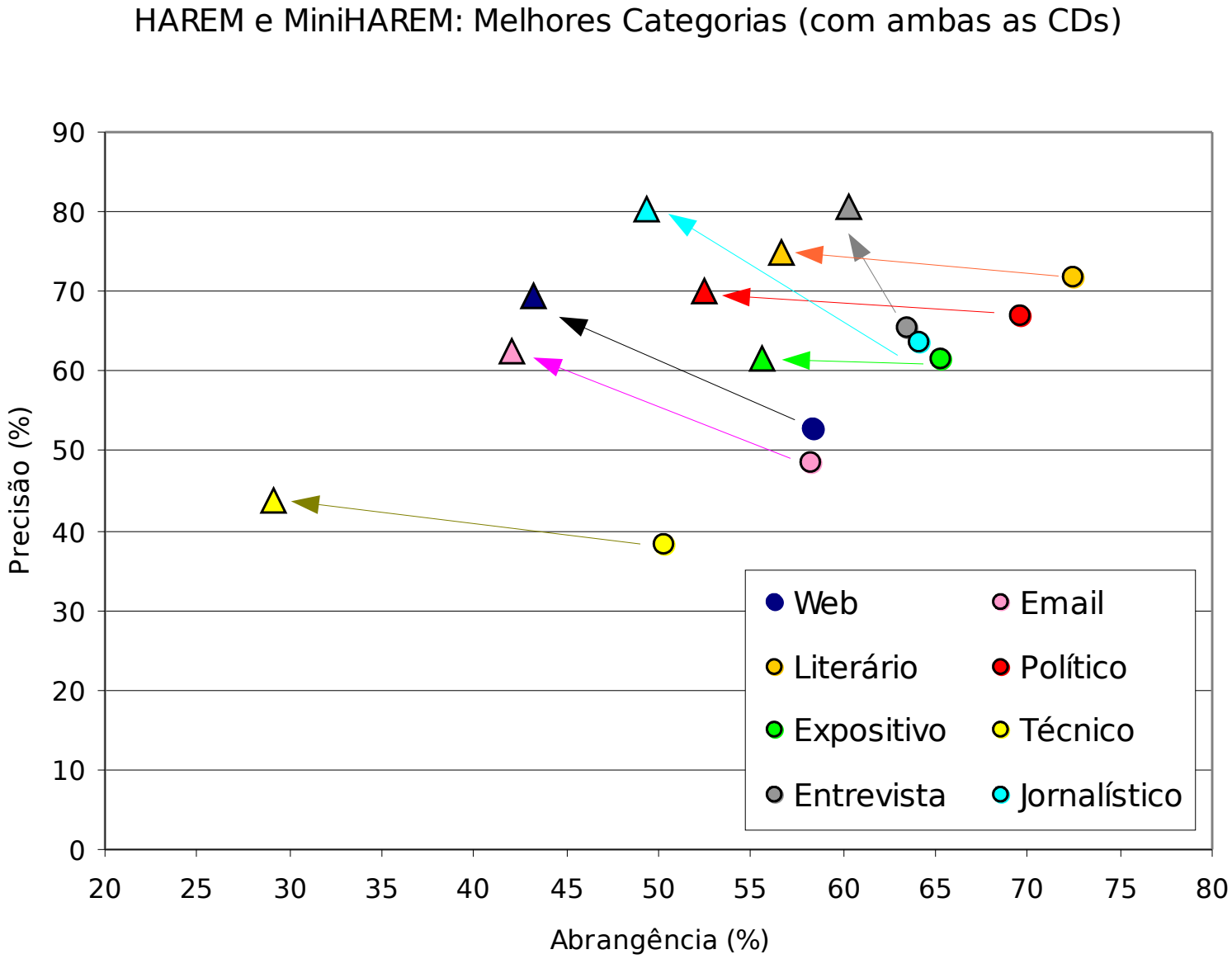
2006



Panorama de Identificação, por Género



Panorama de Identificação, por Género



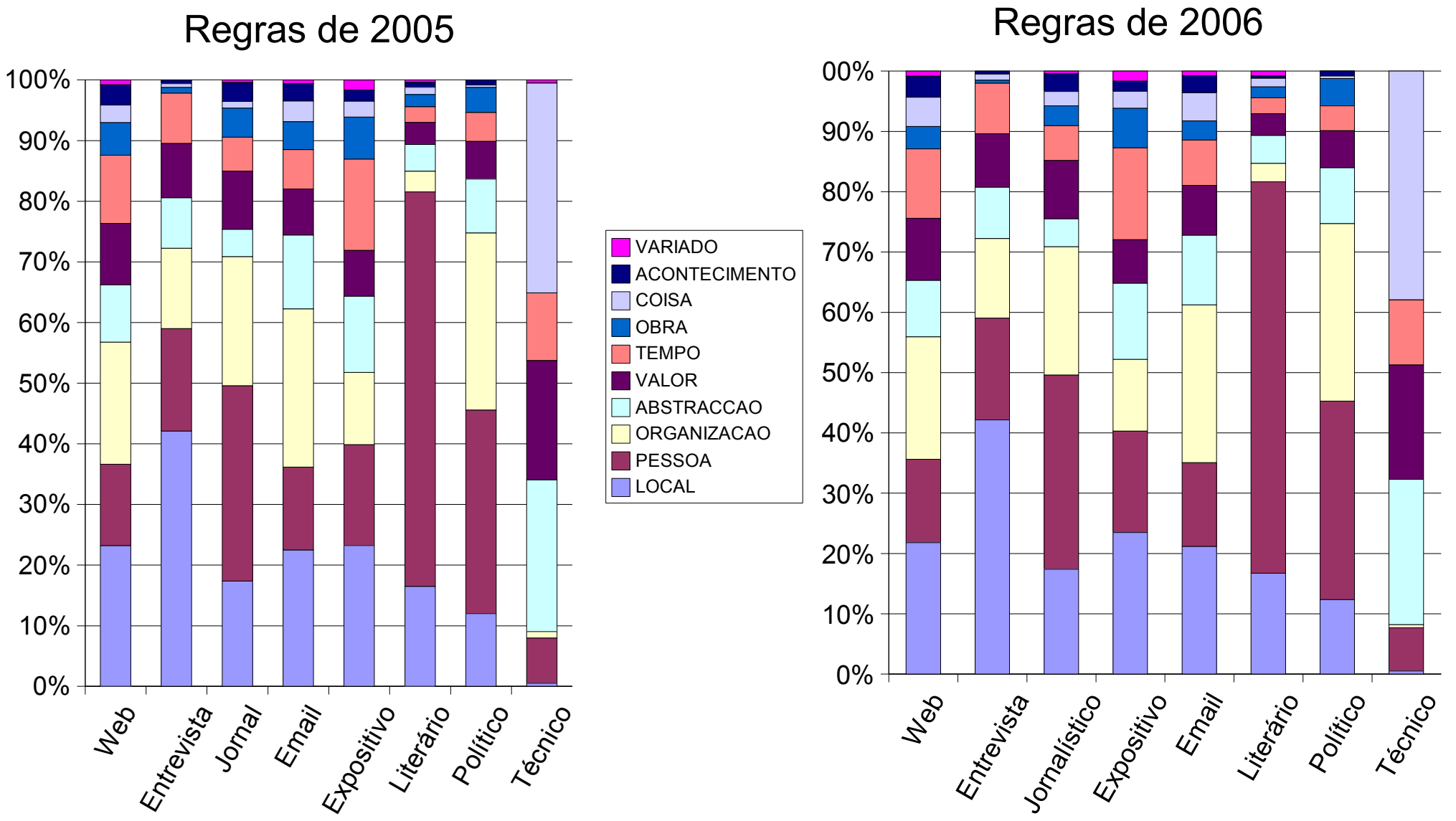
Conclusões

- HAREM e MiniHAREM: São comparáveis
  - CDs são semelhantes
  - Produzem desempenhos semelhantes
  - Tamanho da CD mais que adequado
- A avaliação HAREM está validada.
- As melhores estratégias para REM foram aferidas e comparadas
- Os sistemas REM melhoraram em 1 ano!
- REM em PT com futuro risonho.

Fim

Obrigado pela atenção.

Distribuição de Categorias, por Género Textual



Distribuição de Género Textual, por Categoria