

# **Avaliação no HAREM: Métodos e medidas**

Diana Santos, Nuno Cardoso e Nuno Seco

DI-FCUL

TR-06-17

Departamento de Informática  
Faculdade de Ciências da Universidade de Lisboa  
Campo Grande, 1749-016 Lisboa  
Portugal

Technical reports are available at <http://www.di.fc.ul.pt/tech-reports>. The files are stored in PDF, with the report number as filename. Alternatively, reports are available by post from the above address.



# Avaliação no HAREM: Métodos e medidas

Diana Santos<sup>‡</sup>, Nuno Cardoso<sup>†</sup> e Nuno Seco<sup>+</sup>

<sup>‡</sup>SINTEF ICT, Oslo

<sup>†</sup>Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa

<sup>+</sup>Departamento de Engenharia Informática da Universidade de Coimbra

<sup>‡</sup>diana.santos@sintef.no, <sup>†</sup>ncardoso@xldb.di.fc.ul.pt, <sup>+</sup>nseco@dei.uc.pt

## Resumo

Neste relatório técnico apresentam-se os critérios usados na avaliação dos sistemas participantes do HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas (REM) em português, organizada pela Linguateca. Por outras palavras, descreve as pontuações, medidas e métricas usadas para aferir as saídas geradas pelos sistemas de REM dos participantes, além de descrever os diversos relatórios de apresentação de resultados.

A avaliação é feita comparando uma dada colecção de textos etiquetada pelos sistemas, com essa mesma colecção de textos etiquetada manualmente, a denominada colecção dourada. Este relatório, além de apresentar detalhadamente as fórmulas usadas, ilustra o funcionamento da avaliação e o cálculo das medidas com vários exemplos.

Este texto pretende assim ser a referência definitiva em relação à teoria de avaliação empregue no HAREM, permitindo esclarecer os variados resultados tornados acessíveis durante o primeiro evento de avaliação do HAREM, assim como fixando a terminologia usada nesta avaliação conjunta. Um texto complementar [16] indica como é que a implementação das diversas questões foi levada a cabo.

## 1 Enquadramento

Este relatório é um melhoramento e reformulação das directivas de avaliação tornadas públicas no sítio do HAREM, e que estiveram disponíveis durante os dois eventos de avaliação do HAREM.

O HAREM constituiu a primeira avaliação (conjunta) para sistemas de reconhecimento de entidades mencionadas (REM) em textos em português [1, 8, 4, 6, 2, 12, 17, 9, 15, 3], no âmbito da responsabilidade da Linguateca de organizar avaliações conjuntas para a comunidade científica interessada no processamento computacional do português [10]. O HAREM foca a tarefa de identificação e classificação de entidades mencionadas (EM, ou seja, nomes próprios) no texto, uma tarefa relevante para várias áreas de processamento de linguagem natural (PLN) como a resposta automática a perguntas, a extracção de informação e a tradução automática, entre outras.

No HAREM procurámos desenvolver uma nova metodologia de avaliação em REM que abrangesse as especificidades da tarefa, tendo contemplado questões que não tinham sido ainda abordadas com profundidade suficiente pelos anteriores eventos de avaliação internacionais. Exemplos são a vagueza das EM, a caracterização morfológica das mesmas, e a avaliação de expressões apenas parcialmente identificadas.

O HAREM culminou na organização de dois eventos de avaliação, o principal em Fevereiro de 2005 e a sua sequela, o MiniHAREM, em Abril de 2006 [14], e contou com 10 sistemas participantes oriundos de 6 países (Brasil, Dinamarca, Espanha, França, México e Portugal). Os participantes enviaram um total de 38 saídas, ou seja, anotações automáticas da colecção de textos utilizando os seus sistemas REM,

que foram avaliadas. Os resultados foram publicados, e os relatórios detalhados do desempenho de cada sistema foram entregues aos respectivos participantes.

A 15 de Julho de 2006, a Linguateca organizou o Encontro do HAREM na Universidade do Porto, logo a seguir à Primeira Escola de Verão da Linguateca, que reuniu os participantes, organizadores e outros interessados no HAREM. No encontro, os participantes apresentaram os seus sistemas, a organização apresentou detalhadamente o seu trabalho, e todos debateram várias questões sobre o futuro do HAREM, numa sessão final, que demonstrou bem o interesse da comunidade em prosseguir com mais eventos de avaliação em REM. Está assim em curso a organização de um livro (electrónico) sobre o HAREM [13] englobando as comunicações nesse encontro e outra documentação relevante.

Os participantes no HAREM tiveram um papel activo no desenvolvimento da metodologia e na anotação das colecções de texto segundo a metodologia aprovada. Adicionalmente, a organização do HAREM desenvolveu uma plataforma de avaliação de sistemas de REM, para aferir o desempenho dos sistemas participantes, que se encontra publicamente disponível no sítio do HAREM, <http://www.linguateca.pt/HAREM>. Outra das contribuições do HAREM que reputamos valiosa foi a criação manual de uma colecção dourada para avaliação (CD – ver [12]), juntamente com documentação extensa das directivas usadas [11], constituindo a primeira obra dedicada ao tratamento exaustivo em corpora da semântica dos nomes próprios em português.

É essa documentação que pretendemos tornar mais acessível agora, na forma de relatórios técnicos. As directivas de etiquetagem podem ser encontradas em [5, 7], enquanto o presente relatório versa sobre as directivas de avaliação. Em [17, 16] descrevemos detalhadamente a arquitectura informática que concretiza as regras apresentadas neste relatório, e que realizou a medição das saídas dos sistemas participantes. O trabalho de organização do HAREM enquadra-se no projecto da Linguateca, financiada pela Fundação para a Ciência e Tecnologia (FCT) através do projecto POSI/PLP/43931/2001, e co-financiada pelo POSI.

## 2 Introdução

As directivas de avaliação descritas no presente texto apresentam o que definimos como pontuações, medidas e métricas<sup>1</sup> usadas para medir e comparar as saídas dos sistemas de REM em relação a uma colecção dourada. Recordamos que definimos três tarefas no HAREM: identificação, classificação morfológica e classificação semântica.

### 2.1 Pontuação

*Pontuação* é a classificação ou conjunto de classificações atribuída(s) a cada alinhamento entre EM (pondo em correspondência as EM marcada pelo sistema, e as EM constantes da marcação na CD).

Haverá um sistema de pontuação diferente para cada tarefa, que será descrito na secção correspondente.

A pontuação é expressa por um conjunto de nomes, ou seja, é um processo qualitativo, que tem depois uma tradução quantitativa através do conceito de medida.

### 2.2 Medidas

Uma *medida* do HAREM é uma função de combinação da pontuação, de forma a dar um valor que se pretende indicativo de uma dada qualidade.

---

<sup>1</sup>A diferença entre pontuações, medidas e métricas, aplicada neste documento, pode ser ilustrada com um exemplo de uma liga de futebol; a *pontuação* de cada jogo pode ter três valores: vitória, empate e derrota. De acordo com esta pontuação, a *medida* corresponde aos pontos atribuídos a cada resultado (3,1 e 0 nos respectivos casos). A *métrica* corresponde à soma dos pontos ao longo dos jogos todos do campeonato, e o seu valor final é usado para decidir qual a equipa vencedora.

Embora para a tarefa de identificação só exista uma medida (que pode ser considerada simplesmente como a tradução quantitativa da pontuação), para a avaliação das tarefas de classificação definimos várias medidas, que continuam a ser aplicáveis a alinhamentos. (Em termos matemáticos, as medidas têm por domínio o conjunto de alinhamentos de EM.)

## 2.3 Métricas

Uma *métrica* é uma forma de representar o desempenho global de um sistema, de acordo com uma medida, num único valor numérico, entrando portanto em conta com o valor dessa medida para o conjunto de todos os casos.

**Precisão:** a precisão afere a “qualidade” da resposta do sistema, ao calcular a proporção de respostas correctas em relação a todas as respostas fornecidas por este.

**Abrangência:** a abrangência afere a “quantidade” da resposta do sistema, ao calcular a proporção de respostas correctas em relação ao universo de possíveis respostas correctas (no caso presente, as EM da colecção dourada).

**Medida F:** A medida F combina as métricas de precisão e de abrangência para cada tarefa, de acordo com a seguinte fórmula:

$$\text{Medida F} = \frac{2 \times \text{precisão} \times \text{abrangência}}{\text{precisão} + \text{abrangência}}$$

**Sobre-geração:** a sobre-geração representa o excesso de resultados que um sistema produz, ou seja, calcula quantas vezes produz resultados espúrios.

**Sub-geração:** a sub-geração representa a quantidade de resultados que um sistema não analisou, ou seja, calcula quantas vezes faltam resultados.

**Erro combinado:** o erro combinado reúne as métricas de sobre-geração, de sub-geração e o factor de erro nas parcialmente identificadas numa única métrica, de acordo com a seguinte fórmula:

$$\text{Erro combinado} = \frac{\sum \text{em falta} + \sum \text{espúrio} + \sum \text{factor de erro}}{\sum \text{Pontuação máx. sistema} \cup \text{Pontuação máx. CD}}$$

## 2.4 Cenários de avaliação

Os sistemas de REM que participaram no HAREM foram desenvolvidos com diferentes objectivos. Como tal, as directivas de avaliação prevêm a realização de avaliações segundo *cenários selectivos*, de forma a ajustar a avaliação às características ideais de cada sistema de REM. O módulo responsável pela criação de cenários selectivos (e outros) é o sistema dos Véus, que se encontra detalhado em [16].

Além disso, definimos outro parâmetro, relativo à relação entre a tarefa de identificação e as outras duas, conceptualmente separadas de (e posteriores a) esta. Essa distinção criou, para as tarefas de classificação somente, dois tipos de cenários, absoluto e relativo.

**Eixo absoluto–relativo:** O cenário absoluto avalia o desempenho do sistema em relação à tarefa de REM completa, ou seja, a identificação e a classificação de EM. O cenário relativo, por seu lado, restringe a avaliação às EM pontuadas como *correcto* ou *parcialmente correcto* na tarefa de identificação, avaliando o desempenho do sistema apenas na tarefa de classificação (semântica ou morfológica), independentemente do desempenho na tarefa de identificação.

**Eixo total–selectivo** O cenário total abrange todas as categorias de EM da CD, avaliando a tarefa de classificação (morfológica ou semântica) segundo as categorias propostas pelo HAREM. No cenário selectivo, o participante escolhe previamente o sub-conjunto de categorias e de tipos da categorização HAREM em que o seu sistema quer competir, e as três tarefas são avaliadas apenas para esse sub-conjunto de categorias e de tipos.

Assim, a avaliação do HAREM realizou-se segundo dois eixos, produzindo, portanto, para cada saída dos sistemas participantes, resultados de acordo com, no máximo, quatro cenários diferentes.

### **Tarefa de identificação**

A tarefa de identificação é avaliada apenas segundo o eixo total-selectivo:

**Total:** considera todas as etiquetas na CD.

**Selectivo:** considera apenas o leque de categorias e tipos semânticos que o sistema participante se propõe explicitamente identificar.

### **Tarefas de classificação**

As tarefas de classificação (morfológica e semântica) são avaliadas segundo os dois eixos:

**Total:** considera todas as categorias de EM existentes na CD.

**Absoluto:** considera todas as EM identificadas pelo sistema e todas as EM presentes na CD.

**Relativo:** considera apenas as EM correcta ou parcialmente identificadas pelo sistema.

**Selectivo:** considera apenas as EM (na CD e na resposta do sistema) de categorias/tipos que o participante se propôs classificar.

**Absoluto:** considera todas as EM que estejam incluídas na restrição do cenário selectivo.

**Relativo:** considera dentre as EM incluídas na restrição do cenário selectivo, apenas aquelas que foram parcial ou correctamente identificadas pelo sistema.

## **3 Tarefa de identificação**

A avaliação da tarefa de identificação tem por objectivo medir a capacidade dos sistemas em delimitar correctamente os *átomos* que compõem as EM na colecção, em comparação com a CD.

Um átomo é definido no HAREM como sendo qualquer sequência de letras ou dígitos individuais. Para mais pormenores sobre a forma como é processada a atomização no HAREM, veja-se a documentação do AlinhEM [16], o módulo responsável pelo alinhamento que serve de base a toda a pontuação subsequente.

### **3.1 Pontuação**

A avaliação do HAREM atribui a seguinte pontuação para a tarefa de identificação:

**Correcto:** quando as EM alinhadas são iguais, exceptuando diferenças menores numa lista negra de palavras desprezáveis (ver [16]).

**Parcialmente correcto (por defeito):** quando pelo menos um átomo da saída do sistema corresponde a um átomo de uma EM na CD, e o número total de átomos da EM do sistema é menor do que o número de átomos da respectiva EM da CD.

**Parcialmente correcto (por excesso):** quando pelo menos um átomo da saída do sistema corresponde a um átomo de uma EM na CD, e o número total de átomos da EM do sistema é igual ou maior do que o número de átomos da respectiva EM da CD.

**Em falta:** quando a saída do sistema não delimita como EM nenhum dos átomos de uma EM na CD.

**Espúrio:** quando a saída do sistema delimita uma alegada EM que não consta na CD.

De notar que, para alguns alinhamentos, pode haver mais de uma pontuação por alinhamento: é o caso de vários `parcialmente correctos`.

## 3.2 Medidas

Definimos a medida de correcção ( $c$ ) e o factor de erro ( $e$ ).

Para a primeira, às EM pontuadas como `correcto` é atribuído um valor igual a 1. As EM pontuadas como `parcialmente correcto` é atribuído o valor calculado pela equação 1:

$$c = 0,5 \frac{n_c}{n_d} \quad (1)$$

Onde:

$n_c$  representa o número de átomos comuns entre a EM do sistema e a EM da CD, ou seja, a cardinalidade da intersecção dos dois conjuntos de átomos.

$n_d$  representa o número de átomos distintos entre a EM do sistema e a EM da CD, ou seja, a cardinalidade da reunião dos dois conjuntos de átomos.

A medida de correcção determina que a pontuação máxima é limitada a 0,5, de modo para garantir que várias EM pontuadas como `parcialmente correctas` em relação à mesma EM da CD não totalizem o mesmo valor de uma EM pontuada como `correcto`.

O *factor de erro* é dado pela equação 2:

$$e = 1 - 0,5 \frac{n_c}{n_d} \quad (2)$$

## 3.3 Métricas

Para a tarefa de identificação, as métricas são calculadas da seguinte forma:

### 3.3.1 Precisão

Na tarefa de identificação, a precisão calcula o teor de EM correctas e parcialmente correctas em todas as EM identificadas pelo sistema. Os valores para as EM pontuadas como `parcialmente correctas` são calculados pela equação 1.

$$\text{Precisão}_{\text{identificação}} = \left( \sum \text{EM correctas} + \sum \text{EM parcialmente correctas} \right) / \left( \sum \text{EM identificadas pelo sistema} \right)$$

### 3.3.2 Abrangência

Na tarefa de identificação, a abrangência calcula o teor de EM contidas na CD que o sistema conseguiu identificar. Os valores para as EM pontuadas como parcialmente correctas são novamente calculados pela equação 1.

$$\text{Abrangência}_{\text{identificação}} = (\sum \text{EM correctas} + \sum \text{EM parcialmente correctas}) / (\sum \text{EM na CD})$$

### 3.3.3 Sobre-geração

Na tarefa de identificação, a sobre-geração calcula o teor de EM que foram identificadas pelo sistema, mas que não existem na CD.

$$\text{Sobre-geração}_{\text{identificação}} = (\sum \text{EM espúrias} / \sum \text{EM identificadas pelo sistema})$$

### 3.3.4 Sub-geração

Na tarefa de identificação, a sub-geração calcula o teor de EM que existem na colecção dourada, mas que não foram identificadas pelo sistema.

$$\text{Sub-geração}_{\text{identificação}} = (\sum \text{EM em falta} / \sum \text{EM na CD})$$

### 3.3.5 Erro combinado

Na tarefa de identificação, o erro combinado calcula o teor de erros que o sistema fez. Por “união” denominamos o conjunto de todas as EM, quer as da CD quer as incorrectamente identificadas pelo sistema (a mesma EM só é contada uma vez, mas as EM parcialmente identificadas são contadas de novo).

$$\text{Erro Combinado}_{\text{identificação}} = (\sum \text{EM espúrias} + \sum \text{EM em falta} + \sum \text{Factor de erro}) / (\sum \text{EM identificadas pelo sistema} \cup \text{EM na CD})$$

## 3.4 Exemplo detalhado de atribuição de pontuação

Tomemos uma frase hipotética da colecção dourada na figura 1:

Terminou ontem no <LOCAL TIPO="ALARGADO"> **Laboratório Nacional de Engenharia Civil** </LOCAL>, em <LOCAL TIPO="ADMINISTRATIVO"> **Lisboa** </LOCAL>, o <ACONTECIMENTO TIPO="EVENTO"> **Encontro de Reflexão** </ACONTECIMENTO> sobre a concretização do <ABSTRACCAO TIPO="PLANO"> **Plano Hidrológico** </ABSTRACCAO> espanhol.

Figura 1: Um extracto da CD.

E imaginemos a seguinte saída do sistema que pretendemos avaliar, na figura 2.

A tabela 1 apresenta a pontuação pormenorizada, caso a caso, para cada alinhamento, assim como os valores da medida de correcção.



<PESSOA TIPO="INDIVIDUAL">Terminou</PESSOA> ontem no <LOCAL TIPO="ALARGADO">Laboratório Nacional</LOCAL> de <ABSTRACCAO TIPO="DISCIPLINA">Engenharia Civil</ABSTRACCAO>, em <LOCAL TIPO="ADMINISTRATIVO">Lisboa</LOCAL>, o Encontro de Reflexão sobre a concretização do <ABSTRACCAO TIPO="PLANO">Plano Hidrológico espanhol</ABSTRACCAO>.

Figura 2: O mesmo extracto hipoteticamente analisado por um sistema.

Caso	Colecção dourada	Saída do sistema	Pontuação	Medida
1	-	Terminou	Espúrio	0
2	Laboratório Nacional de Engenharia Civil	Laboratório Nacional	Parcialmente Correcto por Defeito	$0,5 \times \frac{2}{5} = 0,2$
3	Laboratório Nacional de Engenharia Civil	Engenharia Civil	Parcialmente Correcto por Defeito	$0,5 \times \frac{2}{5} = 0,2$
4	Lisboa	Lisboa	Correcto	1
5	Encontro de Reflexão	-	Em Falta	0
6	Plano Hidrológico	Plano Hidrológico espanhol	Parcialmente Correcto Por Excesso	$0,5 \times \frac{2}{3} = 0,333$

Tabela 1: Pontuação da tarefa de identificação e valor da medida de correção, para o exemplo dado.

A tabela 2 calcula os valores das métricas para a tarefa de identificação.

Métrica	Valor
Precisão	$\frac{1+0,2+0,2+0,333}{5} = 34,7\%$
Abrangência	$\frac{1+0,2+0,2+0,333}{4} = 43,3\%$
Medida F	$\frac{2 \times 0,347 \times 0,433}{0,347 + 0,433} = 0,385$
Sobre-geração	$\frac{1}{5} = 20\%$
Sub-geração	$\frac{1}{4} = 25\%$
Erro Combinado	$\frac{(1-0,2)+(1-0,2)+(1-0,333)+1+1}{6} = 0,711$

Tabela 2: Métricas da tarefa de identificação, para o exemplo dado.

Finalmente, a tabela 3 apresenta sete casos particulares de identificação parcial ainda mais complicados, enquanto a tabela 4 ilustra as regras de cálculo da medida de correção para esses casos.

Caso	Sistema participante	Colecção dourada
1	o novo presidente do CNPq, Evando Mirra	o novo presidente do CNPq, Evando Mirra
2	a partir de 1991	a partir de 1991
3	Graduou-se em Engenharia Mecânica e Elétrica	Graduou-se em Engenharia Mecânica e Elétrica
4	Rua 13 de Maio, 733 - Bela Vista - (11) 3262 3256	Rua 13 de Maio, 733 - Bela Vista - (11) 3262 3256
5	Senhores Comandantes das F-FDTL e da PNTL	Senhores Comandantes das F-FDTL e da PNTL
6	secretário-geral do Partido Revolucionário Institucional	secretário-geral do Partido Revolucionário Institucional
7	Estúdio da Oficina Cultural Oswald de Andrade São Paulo, 21 de novembro de 1994	Estúdio da Oficina Cultural Oswald de Andrade São Paulo, 21 de novembro de 1994

Tabela 3: Lista de exemplos para ilustração da avaliação da tarefa de identificação.

Caso	EM na saída e na CD	Medida	Átomos
1	<b>Saída:</b> presidente do CNPq , Evando <b>CD:</b> CNPq	$0,5 \times \frac{1}{4}$	$n_c$ : CPNq $n_d$ : presidente, do, CPNq, Evando
	<b>Saída:</b> presidente do CNPq, Evando <b>CD:</b> Evando Mirra	$0,5 \times \frac{1}{5}$	$n_c$ : Evando $n_d$ : presidente, do, CPNq, Evando, Mirra
2	<b>Saída:</b> 991 <b>CD:</b> 1991	$0,5 \times \frac{3}{4}$	$n_c$ : 9, 9, 1 $n_d$ : 1, 9, 9, 1
3	<b>Saída:</b> Engenharia Mecânica <b>CD:</b> Engenharia Mecânica e Eléctrica	$0,5 \times \frac{2}{4}$	$n_c$ : Engenharia, Mecânica $n_d$ : Engenharia, Mecânica, e, Eléctrica
	<b>Saída:</b> Eléctrica <b>CD:</b> Engenharia Mecânica e Eléctrica	$0,5 \times \frac{1}{4}$	$n_c$ : Eléctrica $n_d$ : Engenharia, Mecânica, e, Eléctrica
4	<b>Saída:</b> Rua <b>CD:</b> Rua 13 de Maio, 733 - Bela Vista	$0,5 \times \frac{1}{10}$	$n_c$ : Rua $n_d$ : Rua, 1, 3, de, Maio, 7, 3, 3, Bela, Vista
	<b>Saída:</b> 13 de Maio <b>CD:</b> Rua 13 de Maio, 733 - Bela Vista	$0,5 \times \frac{4}{10}$	$n_c$ : 1, 3, de, Maio $n_d$ : Rua, 1, 3, de, Maio, 7, 3, 3, Bela, Vista
	<b>Saída:</b> Bela Vista <b>CD:</b> Rua 13 de Maio, 733 - Bela Vista	$0,5 \times \frac{2}{10}$	$n_c$ : Bela, Vista $n_d$ : Rua, 1, 3, de, Maio, 7, 3, 3, Bela, Vista
	<b>Saída:</b> (11) 3262 3256 <b>CD:</b> (11) 3262 3256	1	
5	<b>Saída:</b> Senhores Comandantes das F- <b>CD:</b> Senhores Comandantes das F-FDTL e da PNTL	$0,5 \times \frac{4}{6}$	$n_c$ : Senhores, Comandantes, das, F $n_d$ : Senhores, Comandantes, das, F-, FDTL, PNTL
	<b>Saída:</b> FDTL <b>CD:</b> Senhores Comandantes das F-FDTL e da PNTL	$0,5 \times \frac{1}{6}$	$n_c$ : FDTL $n_d$ : Senhores, Comandantes, das, F-, FDTL, PNTL
	<b>Saída:</b> PNTL <b>CD:</b> Senhores Comandantes das F-FDTL e da PNTL	$0,5 \times \frac{1}{6}$	$n_c$ : PNTL $n_d$ : Senhores, Comandantes, das, F-, FDTL, PNTL
6	<b>Saída:</b> Partido Revolucionário Institucional <b>CD:</b> secretário-geral do Partido Revolucionário Institucional	$0,5 \times \frac{3}{6}$	$n_c$ : Partido, Revolucionário, Institucional $n_d$ : secretário, geral, do, Partido, Revolucionário, Institucional
7	<b>Saída:</b> Oficina Cultural Oswald de Andrade <b>CD:</b> Estúdio da Oficina Cultural Oswald de Andrade	$0,5 \times \frac{5}{6}$	$n_c$ : Oficina, Cultural, Oswald, de, Andrade $n_d$ : Estúdio, Oficina, Cultural, Oswald, de, Andrade
	<b>Saída:</b> São Paulo , 21 <b>CD:</b> São Paulo	$0,5 \times \frac{2}{4}$	$n_c$ : São, Paulo $n_d$ : São, Paulo, 2, 1
	<b>Saída:</b> São Paulo, 21 <b>CD:</b> 21 de novembro de 1994	$0,5 \times \frac{2}{9}$	$n_c$ : 2, 1 $n_d$ : 2, 1, de, Novembro, de, 1, 9, 9, 4
	<b>Saída:</b> novembro de 1994 <b>CD:</b> 21 de novembro de 1994	$0,5 \times \frac{6}{9}$	$n_c$ : Novembro, de, 1, 9, 9, 4 $n_d$ : 2, 1, de, Novembro, de, 1, 9, 9, 4

Tabela 4: Valores da medida de correcção e átomos considerados para os exemplos da tabela 3, na tarefa de identificação.

### 3.5 Identificações alternativas

No caso de considerarmos que há mais do que uma delimitação possivelmente correcta, ou seja, os anotadores humanos da colecção dourada não conseguem decidir naquele contexto, ou concordar entre si, as várias alternativas foram incluídas na colecção dourada, separadas através da etiqueta <ALT> (veja-se [16] para mais exemplos).

Como tal, o avaliador do HAREM irá comparar a CD com a saída do sistema e optar pela melhor alternativa, ou seja, pela alternativa que dá um resultado global melhor para cada sistema.

A escolha entre diferentes alternativas (implementada pelos programas *ALTinaID*, *ALTinaSem* e *ALTinaMor*, ver [16] e secções 4.5 e 5.5 para mais detalhes) é feita independentemente, para cada tarefa. Para a tarefa de identificação, é seguido o seguinte algoritmo:

1. ° – Escolhe-se a alternativa que apresenta melhor medida F.
2. ° – Se os valores da medida F forem iguais, escolhe-se aquela que apresenta menor valor de erro combinado.
3. ° – Se também o valor do erro combinado é igual, escolhe-se a alternativa que apresenta maior número de alinhamentos.

Para permitir que os valores comparados sejam sempre definidos, mesmo em casos de alternativas que não contenham EM, introduz-se sempre no cálculo de cada alternativa um alinhamento correcto, como se exemplificará em seguida. Tal introdução não prejudica a selecção, e evita que alternativas sem EM tenham uma medida F não definida (tal como zero no numerador e no denominador). O texto [16] explica em detalhe este processo.

Tomemos o seguinte exemplo da figura 3, com três alternativas:

```
<ALT> <EM> Governo PSD de Cavaco Silva </EM> |  
<EM> Governo PSD </EM> de <EM> Cavaco Silva </EM> |  
Governo PSD de Cavaco Silva </ALT>
```

ALT1: Governo PSD de Cavaco Silva

ALT2: Governo PSD de Cavaco Silva

ALT3: Governo PSD de Cavaco Silva

Figura 3: Exemplos de alternativas para a tarefa de identificação.

Como já foi referido, o programa irá escolher a alternativa que produz melhores resultados. Seguimos calculando todos os valores envolvidos: as tabelas 5 e 6 apresentam os valores da precisão e da abrangência para o extracto em questão.

O cálculo dos valores da medida F e do erro combinado são detalhados nas tabelas 7 e 8, por sua vez calculadas com base nas tabelas 5 e 6, que se referem respectivamente à precisão e à abrangência.

Finalmente, a tabela 9 apresenta vários exemplos de saídas de sistema (as células a negrito indicam a alternativa escolhida e qual a causa/factor determinante para a escolha), apresentando os valores das métricas respectivas, aplicadas apenas àquele caso de ALT. Note-se que nos casos 4 e 5, a medida F é igual, pelo que a decisão é tomada com base nos valores de Erro Combinado.

De notar que uma vez escolhida essa alternativa, é esse caso que é considerado como certo na CD para essa tarefa e para esse sistema, e é o valor correspondente que aparecerá no cálculo das métricas. Isto faz com que diferentes sistemas sejam avaliados sobre versões ligeiramente diferentes da anotação da CD.

Caso	Precisão		
	ALT1	ALT2	ALT3
1	$(1+1)/(1+1)=100\%$	$(0,4+1)/(1+1)=70\%$	$(0+1)/(1+1)=50\%$
2	$(0+1)/(0+1)=100\%$	$(0+1)/(0+1)=100\%$	$(0+1)/(0+1)=100\%$
3	$(0,4+1)/(1+1)=70\%$	$(0,35+1)/(1+1)=67,5\%$	$(0+1)/(1+1)=50\%$
4	$(0,2+1)/(2+1)=40\%$	$(0,5+1)/(2+1)=50\%$	$(0+1)/(2+1)=33,3\%$
5	$(0,2+1)/(2+1)=40\%$	$(0,5+1)/(2+1)=50\%$	$(0+1)/(2+1)=33,3\%$
6	$(0,2+1)/(1+1)=60\%$	$(1+1)/(1+1)=100\%$	$(0+1)/(1+1)=50\%$
7	$(0,1+1)/(1+1)=55\%$	$(0,25+1)/(1+1)=62,5\%$	$(0+1)/(1+1)=50\%$
8	$(0,3+1)/(1+1)=65\%$	$(0,25+1)/(1+1)=62,5\%$	$(0+1)/(1+1)=50\%$

Tabela 5: Selecção de alternativa - cálculo da precisão.

Caso	Abrangência		
	ALT1	ALT2	ALT3
1	$(1+1)/(1+1)=100\%$	$(0,4+1)/(2+1)=46,7\%$	$(0+1)/(0+1)=100\%$
2	$(0+1)/(1+1)=50\%$	$(0+1)/(2+1)=33,3\%$	$(0+1)/(0+1)=100\%$
3	$(0,4+1)/(1+1)=70\%$	$(0,35+1)/(2+1)=45\%$	$(0+1)/(0+1)=100\%$
4	$(0,2+1)/(1+1)=60\%$	$(0,5+1)/(2+1)=50\%$	$(0+1)/(0+1)=100\%$
5	$(0,2+1)/(1+1)=60\%$	$(0,5+1)/(2+1)=50\%$	$(0+1)/(0+1)=100\%$
6	$(0,2+1)/(1+1)=60\%$	$(1+1)/(2+1)=66,7\%$	$(0+1)/(0+1)=100\%$
7	$(0,1+1)/(1+1)=55\%$	$(0,25+1)/(2+1)=41,7\%$	$(0+1)/(0+1)=100\%$
8	$(0,3+1)/(1+1)=65\%$	$(0,25+1)/(2+1)=41,7\%$	$(0+1)/(0+1)=100\%$

Tabela 6: Selecção de alternativa - cálculo da abrangência.

Caso	Medida F		
	ALT1	ALT2	ALT3
1	$2 \times 1 \times 1 / (1+1) = 1$	$2 \times 0,7 \times 0,467 / (0,7+0,467) = 0,56$	$2 \times 0,5 \times 1 / (0,5+1) = 0,666$
2	$2 \times 1 \times 0,5 / (1+0,5) = 0,66$	$2 \times 1 \times 0,33 / (1+0,33) = 0,5$	$2 \times 1 \times 1 / (1+1) = 1$
3	$2 \times 0,7 \times 0,7 / (0,7+0,7) = 0,7$	$2 \times 0,675 \times 0,45 / (0,675+0,45) = 0,54$	$2 \times 0,5 \times 1 / (0,5+1) = 0,666$
4	$2 \times 0,4 \times 0,6 / (0,4+0,6) = 0,48$	$2 \times 0,33 \times 1 / (1+0,33) = 0,5$	$2 \times 0,5 \times 0,5 / (0,5+0,5) = 0,5$
5	$2 \times 0,4 \times 0,6 / (0,4+0,6) = 0,48$	$2 \times 0,5 \times 0,5 / (0,5+0,5) = 0,5$	$2 \times 0,33 \times 1 / (1+0,33) = 0,5$
6	$2 \times 0,6 \times 0,6 / (0,6+0,6) = 0,6$	$2 \times 1 \times 0,666 / (1+0,666) = 0,8$	$2 \times 0,5 \times 1 / (1+0,5) = 0,667$
7	$2 \times 0,55 \times 0,55 / (0,55+0,55) = 0,55$	$2 \times 0,625 \times 0,417 / (0,625+0,417) = 0,5$	$2 \times 0,5 \times 1 / (1+0,5) = 0,667$
8	$2 \times 0,65 \times 0,65 / (0,65+0,65) = 0,65$	$2 \times 0,625 \times 0,417 / (0,625+0,417) = 0,5$	$2 \times 0,5 \times 1 / (1+0,5) = 0,667$

Tabela 7: Selecção de alternativa - cálculo da medida F.

Caso	Erro Combinado		
	ALT1	ALT2	ALT3
1	$0/(0+1)=0\%$	$(2 \times (1-0,2))/(2+1)=53,3\%$	$1/(1+1)=50\%$
2	$1/(1+1)=50\%$	$(2 \times 1)/(2+1)=66,6\%$	$0/(0+1)=0\%$
3	$0,6/(1+1)=30\%$	$((1-0,1)+(1-0,25))/(2+1)=55,0\%$	$1/(1+1)=50\%$
4	$(2 \times (1-0,1))/(2+1)=60\%$	$(2 \times (1-0,25)+1)/(3+1)=62,5\%$	$2/(2+1)=66,7\%$
5	$(2 \times (1-0,1))/(2+1)=60\%$	$(2 \times (1-0,25))/(2+1)=50\%$	$2/(2+1)=66,7\%$
6	$(1-0,2)/(1+1)=40\%$	$1/(2+1)=33,3\%$	$1/(1+1)=50\%$
7	$(1-0,1)/(1+1)=45\%$	$(1+(1-0,25))/(2+1)=58,3\%$	$1/(1+1)=50\%$
8	$(1-0,3)/(1+1)=35\%$	$(2 \times (1-0,125))/(2+1)=58,3\%$	$1/(1+1)=50\%$

Tabela 8: Selecção de alternativa - cálculo do erro combinado.

Caso	Saída do sistema	ALT1	ALT2	ALT3
1	<EM>Governo PSD de Cavaco Silva</EM>	<b>1 Correcto</b> <b>Medida-F: 1</b> <b>Erro Combinado: 0%</b>	2 Parc. Correcto Medida-F: 0,56 Erro Combinado: 53,3%	1 Espúrio Medida-F: 0,67 Erro Combinado: 50,0%
2	Governo PSD de Cavaco Silva	1 Em Falta Medida-F: 0,67 Erro Combinado: 50,0%	2 Em Falta Medida-F: 0,5 Erro Combinado: 66,7%	<b>Sem pontuação</b> <b>Medida-F: 1</b> <b>Erro Combinado: 0%</b>
3	Governo <EM>PSD de Cavaco Silva</EM>	<b>1 Parc.Cor. por Def.</b> <b>Medida-F: 0,7</b> <b>Erro Combinado: 30%</b>	2 Parc.Cor. por Exc. Medida-F: 0,54 Erro Combinado: 55%	1 Espúrio Medida-F: 0,67 Erro Combinado: 50%
4	<EM>Governo</EM> <EM>PSD</EM> de Cavaco Silva	2 Parc. Correcto Medida-F: 0,48 Erro Combinado: 60%	<b>2 Parc.Cor.+1 Em Falta</b> <b>Medida-F: 0,5</b> <b>Erro Combinado: 62,5%</b>	2 Espúrio Medida-F: 0,5 Erro Combinado: 66,7%
5	Governo <EM>PSD</EM> de Cavaco <EM>Silva</EM>	2 Parc. Correcto Medida-F: 0,48 Erro Combinado: 60%	<b>2 Parc. Correcto</b> <b>Medida-F: 0,5</b> <b>Erro Combinado: 50%</b>	2 Espúrio Medida-F: 0,5 Erro Combinado: 66,7%
6	<EM>Governo PSD</EM> de Cavaco Silva	1 Parc. Correcto Medida-F: 0,6 Erro Combinado: 40%	<b>1 Correcto, 1EmFalta</b> <b>Medida-F: 0,8</b> <b>Erro Combinado: 33,3%</b>	1 Espúrio Medida-F: 0,67 Erro Combinado: 50%
7	Governo PSD de Cavaco <EM>Silva</EM>	1 Parc. Correcto Medida-F: 0,55 Erro Combinado: 45%	1 Parc. Cor., 1 Em Falta Medida-F: 0,5 Erro Combinado: 58,3%	<b>1 Espúrio</b> <b>Medida-F: 0,67</b> <b>Erro Combinado: 50%</b>
8	Governo <EM>PSD de Cavaco</EM> Silva	1 Parc. Correcto Medida-F: 0,65 Erro Combinado: 35%	2 Parc. Correcto Medida-F: 0,5 Erro Combinado: 58,3%	<b>1 Espúrio</b> <b>Medida-F: 0,67</b> <b>Erro Combinado: 50%</b>

Tabela 9: Exemplos de selecção de alternativa na tarefa de identificação.

## 4 Tarefa de classificação semântica

A tarefa de classificação semântica avalia até que ponto os sistemas participantes conseguem classificar a EM na hierarquia de categorias e de tipos definidos no HAREM, que foi especialmente criada para o português e foi revista conjuntamente pelos participantes e pela organização [12, 5].

O método de avaliação aqui descrito, contudo, é genérico e independente das categorias particulares usadas. A única coisa que assume é que existe um conjunto disjunto de categorias, que por sua vez possuem um conjunto distinto de tipos.

### 4.1 Pontuação

A pontuação na classificação semântica é feita para a categoria e para o tipo, em separado, e tem três valores possíveis:

**Correcto:** quando a categoria (ou tipo) da EM da saída é igual à categoria (ou tipo) da EM da CD.

**Em Falta:** quando a categoria (ou tipo) da EM da CD está ausente da categoria (ou tipo) da EM da saída.

**Espúria:** quando a categoria (ou tipo) da EM da saída está ausente da categoria (ou tipo) da EM da CD.

Contudo, como as EM podem ter mais do que uma categoria e tipo ( $\langle A|B|C|\dots$  TIPO= $"X|Y|Z|\dots"$ ), temos de detalhar qual o procedimento seguido nesses casos, tal como fizemos para o caso do ALT anterior.

Em relação às categorias, veja-se como lidamos no HAREM com mais do que uma categoria por EM, detalhado na tabela 10:

Caso	Saída Sistema	Solução	Correcta	Em Falta	Espúria
1	$\langle A \rangle$	$\langle A \rangle$	A	-	-
2	$\langle B \rangle$	$\langle A \rangle$	-	A	B
3	$\langle A \rangle$	$\langle A B C \rangle$	A	-	-
4	$\langle D \rangle$	$\langle A B C \rangle$	-	A, B e C	D
5	$\langle A \rangle$		-	-	A

Tabela 10: Pontuação pormenorizada de categorias na classificação semântica.

**Correcta:** Quando o sistema atribui à EM uma categoria, e se essa categoria for igual à da EM na CD, é pontuada como *correcto* (caso 1). Contudo, se a respectiva EM da CD possui um conjunto de categorias, basta a categoria da EM da saída corresponder a uma desse conjunto para que, além de ser pontuado igualmente como *correcto*, o sistema não será prejudicado por faltarem as outras. Ou seja, o caso 3 resulta na mesma pontuação que o caso 1.

**Em falta:** Se a categoria da EM de saída não corresponde à categoria da EM da CD, no caso de esta ter uma classificação única (caso 2), ou não corresponder a nenhuma das classificações múltiplas (caso 4), cada uma das categorias da EM da CD é pontuada como *em falta*. Contudo, se a categoria que o sistema classificou estiver incluída no conjunto presente na EM da CD, nada é considerado *em falta* (caso 3), como já explicado acima.

**Espúria:** no caso de o sistema atribuir uma categoria a uma EM que não existe com essa categoria na CD, ou a uma EM que não existe simplesmente na CD, essa categoria é pontuada como *espúria* (casos 2, 4 e 5). Esta marcação é atribuída, quer em conjunção com *em falta* (casos 2 e 4), quer se o sistema identificou algo como EM que não o seja (caso 5). Há portanto duas causas distintas possíveis para uma categoria ter associada a pontuação *espúria*.

Uma situação semelhante se passa na pontuação dos tipos:

Caso	Saída do sistema	Solução	Correcta	Em falta	Espúria
1	<A>	<A TIPO="X">	-	X	-
2	<A TIPO="OUTRO">	<A TIPO="X">	-	X	-
3	<A TIPO="OUTRO">	<A   A   A TIPO="X   Y   Z">	-	X, Y e Z	-
4	<A TIPO="X">	<A TIPO="X">	X	-	-
5	<A TIPO="X">	<A TIPO="Y">	-	Y	X
6	<A TIPO="X">	<A   B   C TIPO="X   Y   Z">	X	-	-
7	<A TIPO="X">	<A   A   A TIPO="X   Y   Z">	X	-	-
8	<A TIPO="X">	<A   A   A TIPO="W   Y   Z">	-	W, Y e Z	X

Tabela 11: Pontuação pormenorizada dos tipos na classificação semântica.

A tabela 11 resume a pontuação atribuída nos diversos casos, sendo o raciocínio análogo ao referente às categorias, que acabámos de expor.

Caso	Saída Sistema	CD	Correcta	Em Falta	Espúria
1	<A TIPO="X">	<A TIPO="X">	(A,X)	-	-
2	<A TIPO="Y">	<A TIPO="X">	-	(A,X)	(A,Y)
3	<A TIPO="Y">	<A   A   A TIPO="X   Y   Z">	(A,Y)	-	-
4	<A TIPO="W">	<A   A   A TIPO="X   Y   Z">	-	(A,X Y Z)	(A,W)
5	<B TIPO="Z">	<A TIPO="X">	-	(A,X)	(B,Z)

Tabela 12: Pontuação pormenorizada do par categoria-tipo, na classificação semântica.



## 4.2 Medidas

A classificação semântica é avaliada através de quatro medidas, que fornecem diferentes tipos de informação aos participantes sobre o desempenho dos seus sistemas:

**Por categorias:** mede apenas a atribuição da categoria, ignorando o tipo.

**Por tipos:** mede apenas a capacidade de discriminação do sistema em escolher o tipo correcto, se a categoria tiver sido correctamente atribuída.

**Combinada:** mede a atribuição da categoria e do tipo, combinando as duas através da equação 3, que pretende reflectir a entropia associada à escolha do tipo dentro da categoria.

**Plana:** avalia os pares categoria-tipo, considerando apenas como certos os casos que tenham acertado tanto na categoria como no tipo.

Em todos os casos, se a EM em causa tiver sido identificada como apenas parcialmente correcta, a pontuação respectiva será multiplicada pelo factor  $\frac{n_c}{n_d}$ .

Além disso, ainda se considera mais duas medidas: Falta e Excesso, correspondendo sempre respectivamente ao número de casos em falta e em excesso.

### 4.2.1 Medida por categorias

A medida por categorias avalia a classificação semântica das EM cingindo-se apenas aos valores das categorias, sendo irrelevante qual a atribuição de tipo proposta pelo sistema.

### 4.2.2 Medida por tipos

A medida por tipos avalia apenas uma parte da classificação semântica, sendo uma medida relativa por excelência, visto que pressupõe a categoria certa, ou seja, o seu domínio são as EM correctamente classificadas pelo sistema em relação à sua categoria (e daí ainda previamente correctamente (ou parcialmente) identificadas pelo sistema).

### 4.2.3 Medida combinada

A medida semântica combinada combina a pontuação da categoria e do tipo através da seguinte fórmula:

$$P_{CSC} = \begin{cases} 0 & \text{se a categoria não estiver correcta} \\ 1 & \text{se a categoria estiver correcta mas o tipo não estiver correcto} \\ 1 + \left(1 - \frac{n_c}{n_t}\right) - \frac{n_e}{n_t} & \text{se a categoria estiver correcta e pelo menos um tipo correcto} \end{cases} \quad (3)$$

Onde  $n_c$  representa o número de tipos correctos,  $n_e$  o número de tipos espúrios, e  $n_t$  o número de tipos possível nessa categoria. Note-se que para calcular estes últimos valores, é preciso naturalmente conhecer quantos TIPOS diferentes cada categoria pode ter, o que está descrito na tabela 13. Como o número de tipos de certas categorias foi alterado do HAREM para o MiniHAREM, apresentamos os valores para cada evento, destacando a negrito aqueles que mudaram.

Veja-se a tabela 14 com alguns exemplos, assumindo que a categoria *A* tem quatro tipos distintos.

Categoria	HAREM		MiniHAREM	
	Número de tipos distintos	Valor máximo	Número de tipos distintos	Valor máximo
ABSTRACCAO	8	1,875	8	1,875
ACONTECIMENTO	3	1,666	3	1,667
COISA	3	1,666	<b>4</b>	<b>1,75</b>
LOCAL	5	1,8	5	1,8
OBRA	4	1,75	<b>3</b>	<b>1,667</b>
ORGANIZACAO	4	1,75	4	1,75
PESSOA	6	1,833	6	1,833
TEMPO	4	1,75	4	1,75
VALOR	3	1,667	3	1,667

Tabela 13: Quantidade de tipos distintos que uma categoria semântica pode ter, e valor máximo correspondente para o cálculo da medida combinada, para o HAREM e o MiniHAREM

Caso	CD	Saída do Sistema	Medida combinada
1	<A TIPO="B">	<A TIPO="C">	1
2	<A TIPO="B">	<A TIPO="B">	$1+(1-\frac{1}{4}) = 1,75$
3	<A TIPO="B">	<A   A TIPO="B   Y">	$1+(1-\frac{1}{4})-\frac{1}{4} = 1,5$
4	<A TIPO="B">	<A   A TIPO="C   D">	1

Tabela 14: Exemplo para a classificação semântica na medida combinada, para uma categoria A com quatro tipos ( $n_t = 4$ ).

#### 4.2.4 Medida plana

A classificação semântica na medida plana tem como objecto de estudo o par (CATEGORIA, TIPO). Pode ser considerada como a "intersecção" (ou multiplicação) das medidas por categorias e por tipos, ou calculada tomando o par categoria-tipo como unidade.

Por exemplo, se as EM em análise fossem <LOCAL TIPO="GEOGRAFICO">Coimbra</LOCAL> e <PESSOA TIPO="INDIVIDUAL">Magalhães</PESSOA>, então os pares a serem avaliados seriam (LOCAL, GEOGRAFICO) e (PESSOA, INDIVIDUAL), respectivamente. Um par é pontuado como correcto quando a categoria e o tipo são o mesmo na entidade correspondente da CD.

A tabela 12 ilustra as regras de pontuação, e a tabela 15 os respectivos valores das medidas.

Caso	Medida plana	Falta	Excesso
1	1	0	0
2	0	1	1
3	1	0	0
4	0	3	1
5	0	1	1

Tabela 15: Medidas plana, falta e excesso para os casos da tabela 12.

## 4.3 Métricas

### 4.3.1 Precisão

A precisão apresenta-se sobre dois cenários: absoluto (para todas as EM) e relativo (às EM correctamente identificadas).

Para a medida por categorias, a precisão é dada pela fórmula:

**Absoluto:**  $\text{Precisão}_{\text{medida categorias}} = (\sum \text{EM correctamente identificadas e com categoria correcta} + Y) / (\sum \text{EM classificadas pelo sistema})$

**Relativo:**  $\text{Precisão}_{\text{medida categorias}} = (\sum \text{EM correctamente identificadas e com categoria correcta} + Y) / (\sum \text{EM que o sistema reconheceu, total ou parcialmente})$

Em que Y corresponde ao somatório dos valores obtidos para as EM parcialmente identificadas e com categoria correcta, valores esses pesados pela fórmula  $\frac{n_c}{n_d}$ .

A medida por tipos é, por definição, sempre relativa:

**Relativo:**  $\text{Precisão}_{\text{medida tipos}} = (\sum \text{EM correctamente identificadas e com categoria e tipo correctos} + Z) / (\sum \text{EM que o sistema reconheceu, total ou parcialmente})$

Em que Z corresponde ao somatório dos valores obtidos para as EM parcialmente identificadas e com categoria e tipo correctos, valores esses pesados pela fórmula  $\frac{n_c}{n_d}$ .

Para a medida semântica combinada, a precisão mede o grau de sucesso relativo à classificação máxima (calculada assumindo todas as categorias e tipos propostos pelo sistema como correctos):

**Absoluto:**  $\text{Precisão}_{\text{medida CSC}} = (\text{Valor de CSC obtida pelo sistema} / \text{Valor máximo da CSC para a saída do sistema})$

**Relativo:**  $\text{Precisão}_{\text{medida CSC}} = (\text{Valor da CSC obtida pelo sistema} / \text{Valor máximo da CSC para a saída do sistema só considerando as EM que o sistema reconheceu, total ou parcialmente})$

Para a medida plana, a precisão é calculada da seguinte forma:

**Absoluto:**  $\text{Precisão}_{\text{medida plana}} = (\sum \text{EM correctamente identificadas e com categoria e tipo correctos} + Z) / (\sum \text{EM classificadas pelo sistema})$

**Relativo:**  $\text{Precisão}_{\text{medida plana}} = (\sum \text{EM correctamente identificadas e com categoria e tipo correctos} + Z) / (\sum \text{EM que o sistema reconheceu, total ou parcialmente})$

Em que Z corresponde ao somatório dos valores obtidos para as EM parcialmente identificadas e com categoria e tipo correctos, valores esses pesados pela fórmula  $\frac{n_c}{n_d}$ .

### 4.3.2 Abrangência

A abrangência define-se de forma diferente para cada uma das quatro medidas, e de forma diferente para os cenários absoluto e relativo.

Para a medida por categorias, a abrangência é calculada da seguinte forma:

**Absoluto:**  $Abrangência_{medida\ categorias} = (\sum EM\ correctamente\ identificadas\ e\ com\ categoria\ correcta + Y) / (\sum EM\ classificadas\ na\ CD)$

**Relativo:**  $Abrangência_{medida\ categorias} = (\sum EM\ correctamente\ identificadas\ e\ com\ categoria\ correcta + Y) / (\sum EM\ da\ CD\ que\ o\ sistema\ reconheceu,\ total\ ou\ parcialmente)$

Relembramos que Y corresponde ao somatório dos valores obtidos para as EM parcialmente identificadas e com categoria correcta, valores esses pesados pela fórmula  $\frac{n_c}{n_d}$ .

A medida por tipos é, por definição, sempre relativa:

**Relativo:**  $Abrangência_{medida\ tipos} = (\sum EM\ correctamente\ identificadas\ e\ com\ categoria\ e\ tipo\ correctos + Z) / (\sum EM\ da\ CD\ que\ o\ sistema\ reconheceu,\ total\ ou\ parcialmente)$

Relembramos que Z corresponde ao somatório dos valores obtidos para as EM parcialmente identificadas e com categoria e tipo correctos, valores esses pesados pela fórmula  $\frac{n_c}{n_d}$ .

Usando a medida semântica combinada, a abrangência mede o nível de cobertura de acordo com a classificação máxima (se tanto as categorias como os tipos enviados estiverem correctos). Mais uma vez, no cenário absoluto usam-se todas as EM na CD, e no relativo apenas o subconjunto das que foram parcial ou correctamente identificadas.

**Absoluto:**  $Abrangência_{medida\ CSC} = (Valor\ da\ medida\ semântica\ combinada\ obtida\ pelo\ sistema / Valor\ máximo\ da\ medida\ semântica\ combinada\ na\ CD)$

**Relativo:**  $Abrangência_{medida\ CSC} = (Valor\ da\ medida\ semântica\ combinada\ obtida\ pelo\ sistema / Valor\ máximo\ da\ medida\ semântica\ combinada\ na\ CD\ usando\ apenas\ as\ EM\ da\ CD\ que\ o\ sistema\ reconheceu,\ total\ ou\ parcialmente)$

Para a medida plana, a abrangência é calculada da seguinte forma:

**Absoluto:**  $Abrangência_{medida\ plana} = (\sum EM\ correctamente\ identificadas\ e\ com\ categoria\ e\ tipo\ correctos + Z) / (\sum EM\ na\ CD)$

**Relativo:**  $Abrangência_{medida\ plana} = (\sum EM\ correctamente\ identificadas\ e\ com\ categoria\ e\ tipo\ correctos + Z) / (\sum EM\ da\ CD\ que\ o\ sistema\ reconheceu,\ total\ ou\ parcialmente)$

Em que Z corresponde ao somatório dos valores obtidos para as EM parcialmente identificadas e com categoria e tipo correctos, valores esses pesados pela fórmula  $\frac{n_c}{n_d}$ .

### 4.3.3 Sobre-geração

A sobre-geração na classificação semântica mede o número de EM com uma classificação semântica espúria, em comparação com a CD. A sobre-geração é calculada de forma diferente, de acordo com o cenário usado (absoluto ou relativo).

Para a medida por categorias, a sobre-geração é calculada da seguinte forma:

**Absoluto:**  $\text{Sobre-geração}_{\text{medida categorias}} = (\sum \text{EM com classificação semântica espúria na categoria}) / (\sum \text{EM classificadas com categoria pelo sistema})$

**Relativo:**  $\text{Sobre-geração}_{\text{medida categorias}} = (\sum \text{EM parcial ou correctamente identificadas com classificação semântica espúria na categoria}) / (\sum \text{EM parcial ou correctamente identificadas classificadas com categoria pelo sistema})$

A medida plana implica que a sobre-geração seja calculada da seguinte forma:

**Absoluto:**  $\text{Sobre-geração}_{\text{medida plana}} = (\sum \text{EM com classificação semântica espúria na categoria ou no tipo}) / (\sum \text{EM classificadas com categoria e tipo pelo sistema})$

**Relativo:**  $\text{Sobre-geração}_{\text{medida plana}} = (\sum \text{EM parcial ou correctamente identificadas com classificação semântica espúria na categoria ou no tipo}) / (\sum \text{EM parcial ou correctamente identificadas e classificadas com categoria e tipo pelo sistema})$

### 4.3.4 Sub-geração

A sub-geração na classificação semântica mede o número de EM com uma classificação semântica em falta, em comparação com a saída. A sub-geração é calculada de forma diferente, de acordo com o cenário usado (absoluto ou relativo).

Para a medida por categorias, a sub-geração é calculada da seguinte forma:

**Absoluto:**  $\text{Sub-geração}_{\text{medida categorias}} = (\sum \text{EM com classificação semântica em falta na categoria}) / (\sum \text{EM na CD})$

**Relativo:**  $\text{Sub-geração}_{\text{medida categorias}} = (\sum \text{EM parcial ou correctamente identificadas com classificação semântica em falta na categoria}) / (\sum \text{EM parcial ou correctamente identificadas com categoria na CD})$

A medida por tipos é, por definição, sempre relativa:

**Relativo:**  $\text{Sub-gera\c{c}\~{a}o}_{\text{medida tipos}} = (\sum \text{EM parcial ou correctamente identificadas com classifica\c{c}\~{a}o sem\~{a}ntica em falta no tipo}) / (\sum \text{EM parcial ou correctamente identificadas com tipo na CD})$

A medida plana implica que a subgera\c{c}\~{a}o \c{e} calculada da seguinte forma:

**Absoluto:**  $\text{Sub-gera\c{c}\~{a}o}_{\text{medida plana}} = (\sum \text{EM com classifica\c{c}\~{a}o sem\~{a}ntica em falta na categoria ou no tipo}) / (\sum \text{EM com categoria na CD})$

**Relativo:**  $\text{Sub-gera\c{c}\~{a}o}_{\text{medida plana}} = (\sum \text{EM parcial ou correctamente identificadas com classifica\c{c}\~{a}o sem\~{a}ntica em falta na categoria ou no tipo}) / (\sum \text{EM parcial ou correctamente identificadas com categoria e tipo na CD})$

#### 4.4 Exemplo detalhado de avalia\c{c}\~{a}o da classifica\c{c}\~{a}o sem\~{a}ntica

Apresentamos um exemplo de texto na figura 4, etiquetado por um sistema hipot\c{e}tico, e a respectiva CD na figura 5. Para n\~{a}o sobrecarregar o presente documento, todas as EM da CD s\~{a}o identificadas correctamente ou parcialmente (portanto, os cen\~{a}rios relativo e absoluto produzem os mesmos resultados).

```
<LOCAL TIPO="GEOGRAFICO"> Plano hidrol\c3{o}gico de Espanha
</LOCAL> analisado em <LOCAL TIPO="ADMINISTRATIVO"> Lisboa
</LOCAL>. Terminou ontem no <LOCAL TIPO="ALARGADO">
Laborat\c3{o}rio Nacional </LOCAL> de <ORGANIZACAO
TIPO="SUB"> Engenharia Civil </ORGANIZACAO>, em <LOCAL
TIPO="ADMINISTRATIVO"> Lisboa </LOCAL>, o <ABSTRACCAO
TIPO="PLANO"> Encontro de Reflex\c3{o} </ABSTRACCAO> sobre a
concretiza\c3{\c{a}}o do <ABSTRACCAO TIPO="PLANO"> Plano Hidrol\c3{o}gico
</ABSTRACCAO> espanhol. <ABSTRACCAO TIPO="DISCIPLINA">
Em an\~{a}lise </ABSTRACCAO> esteve um documento que prev\c3{e}
a transfer\c3{e}ncia de significativos volumes de \c3{a}gua dos
r\c3{ios} <LOCAL TIPO="GEOGRAFICO"> Douro </LOCAL> e <LOCAL
TIPO="GEOGRAFICO"> Tejo </LOCAL> para a bacia hidrogr\c3{a}fica
do rio <ABSTRACCAO TIPO="PLANO"> Jucar </ABSTRACCAO>.
```

Figura 4: Exemplo de uma sa\c3{ida}.

##### 4.4.1 Medida por categorias

Na tabela 16 apresentamos a pontua\c3{\c{c}\~{a}o por categorias para a classifica\c{c}\~{a}o sem\~{a}ntica. Na tabela 17 apresentamos as medidas, e na tabela 18 os valores das m\c3{e}tricas. No caso das identifica\c3{\c{c}\~{a}o}es parciais, coloc\~{a}mos entre par\c3{e}nteses o factor  $\frac{n_c}{n_d}$ .

Plano hidrológico de <ORGANIZACAO|LOCAL TIPO="ADMINISTRACAO|ADMINISTRATIVO"> **Espanha** </ORGANIZACAO|LOCAL> analisado em <LOCAL TIPO="ADMINISTRATIVO"> **Lisboa** </LOCAL>. Terminou ontem no <LOCAL TIPO="ALARGADO"> **Laboratório Nacional de Engenharia Civil** </LOCAL>, em <LOCAL TIPO="ADMINISTRATIVO"> **Lisboa** </LOCAL>, o <ACONTECIMENTO TIPO="EVENTO"> **Encontro de Reflexão** </ACONTECIMENTO> sobre a concretização do <ABSTRACCAO TIPO="PLANO"> **Plano Hidrológico** </ABSTRACCAO> espanhol. Em análise esteve um documento que prevê a transferência de significativos volumes de água dos rios <LOCAL TIPO="GEOGRAFICO"> **Douro** </LOCAL> e <LOCAL TIPO="GEOGRAFICO"> **Tejo** </LOCAL> para a bacia hidrográfica do rio <LOCAL TIPO="GEOGRAFICO"> **Jucar** </LOCAL>.

Figura 5: Exemplo da marcação na CD do texto da figura anterior.

Caso	Saída do Sistema	Correcta	Em Falta	Espúria
1	<LOCAL TIPO="GEOGRAFICO">Plano hidrológico de Espanha</LOCAL>	LOCAL (0,25)	-	-
2	<LOCAL TIPO="ADMINISTRATIVO">Lisboa</LOCAL>	LOCAL	-	-
3	<LOCAL TIPO="ALARGADO">Laboratório Nacional</LOCAL> de <ORGANIZACAO TIPO="SUB">Engenharia Civil</ORGANIZACAO>	LOCAL (0,4)		ORGANIZACAO
4	<LOCAL TIPO="ADMINISTRATIVO">Lisboa</LOCAL>	LOCAL	-	-
5	<ABSTRACCAO TIPO="PLANO">Encontro de Reflexão</ABSTRACCAO>	-	ACONTECIMENTO	ABSTRACCAO
6	<ABSTRACCAO TIPO="PLANO">Plano Hidrológico</ABSTRACCAO>	ABSTRACCAO	-	-
7	<ABSTRACCAO TIPO="DISCIPLINA">Em análise</ABSTRACCAO>	-	-	ABSTRACCAO
8	<LOCAL TIPO="GEOGRAFICO">Douro</LOCAL>	LOCAL	-	-
9	<LOCAL TIPO="GEOGRAFICO">Tejo</LOCAL>	LOCAL	-	-
10	<ABSTRACCAO TIPO="PLANO">Jucar</ABSTRACCAO>	-	LOCAL	ABSTRACCAO

Tabela 16: Pontuação da classificação semântica medida por categorias, para o exemplo dado.

Caso	Medida por categorias	Falta	Excesso
1	0,25	0	0
2	1	0	0
3	0,4	0	1
4	1	0	0
5	0	1	1
6	1	0	0
7	0	0	1
8	1	0	0
9	1	0	0
10	0	1	1
Total	5,65	2	4

Tabela 17: Medida por categorias, falta e excesso para o exemplo dado.

Métrica	Valor
Precisão	$\frac{5,65}{11} = 51,36\%$
Abrangência	$\frac{5,65}{9} = 62,77\%$
Medida F	$\frac{2 \times 0,5136 \times 0,6277}{0,5136 + 0,6277} = 0,565$
Sobre-geração	$\frac{4}{11} = 36,36\%$
Sub-geração	$\frac{2}{9} = 22,2\%$

Tabela 18: Valores das métricas para a tarefa de classificação semântica, medida por categorias, para o exemplo dado.



#### 4.4.2 Medida por tipos

Na tabela 19 apresentamos a pontuação para a classificação semântica segundo a medida por tipos, na tabela 20 apresentamos as medidas e na tabela 21 os valores das métricas. De notar que os casos 3 (na segunda EM), 5, 7 e 10 não são classificados, porque não foram pontuados como correctos na tabela 16.

Caso	Saída do Sistema	Correcta	Em Falta	Espúria
1	<LOCAL TIPO="GEOGRAFICO"> Plano hidrológico de Espanha</LOCAL>	-	ADMINISTRATIVO	GEOGRAFICO
2	<LOCAL TIPO="ADMINISTRATIVO"> Lisboa</LOCAL>	ADMINISTRATIVO	-	-
3	<LOCAL TIPO="ALARGADO"> Laboratório Nacional</LOCAL>	ALARGADO (0,4)	-	-
4	<LOCAL TIPO="ADMINISTRATIVO"> Lisboa</LOCAL>	ADMINISTRATIVO	-	-
6	<ABSTRACCAO TIPO="PLANO"> Plano Hidrológico</ABSTRACCAO>	PLANO	-	-
8	<LOCAL TIPO="GEOGRAFICO"> Douro</LOCAL>	GEOGRAFICO	-	-
9	<LOCAL TIPO="GEOGRAFICO"> Tejo</LOCAL>	GEOGRAFICO	-	-

Tabela 19: Pontuação da classificação semântica por tipos, para o exemplo dado

Caso	Medida por tipos	Falta	Excesso
1	0	1	1
2	1	0	0
3	0,4	0	0
4	1	0	0
6	1	0	0
8	1	0	0
9	1	0	0
Total	5,4	1	1

Tabela 20: Medida portipos, falta e excesso para o exemplo dado da tabela 19

Métrica	Valor
Precisão	$\frac{5,4}{7} = 77,14\%$
Abrangência	$\frac{5,4}{7} = 77,14\%$
Medida F	$\frac{2 \times 0,7714 \times 0,7714}{0,7714 + 0,7714} = 0,7714$
Sobre-geração	$\frac{1}{7} = 14,28\%$
Sub-geração	$\frac{1}{7} = 14,28\%$

Tabela 21: Valores das métricas para a classificação semântica, medida por tipos, para o exemplo dado.

### 4.4.3 Medida combinada

Na tabela 22 apresentamos a pontuação para a classificação semântica segundo a medida combinada, e na tabela 23 os valores das métricas. Salientamos que os casos 1 e 3 são multiplicados pelo factor de correcção  $\frac{n_c}{n_d}$ , respectivamente, 0,25 e 0,4.

Caso	Classificação
1	$(1) \times 0,25 = 0,25$
2	$1 + (1 - \frac{1}{5}) = 1,80$
3	$(1 + (1 - \frac{1}{5})) \times 0,4 = 0,72$
4	$1 + (1 - \frac{1}{5}) = 1,80$
5	0,0
6	$1 + (1 - \frac{1}{8}) = 1,875$
7	0,0
8	$1 + (1 - \frac{1}{5}) = 1,80$
9	$1 + (1 - \frac{1}{5}) = 1,80$
10	0
Total	10,045

Tabela 22: Avaliação da classificação semântica segundo a medida combinada, para o exemplo dado.

Métrica	Valor
Precisão	$\frac{10,045}{20,05} = 50,1\%$
Abrangência	$\frac{10,045}{16,14} = 62,2\%$
Medida F	$\frac{2 \times 0,501 \times 0,6223}{0,501 + 0,6223} = 0,555$

**Nota:** os denominadores usados no cálculo da precisão e da abrangência recorrem aos conceitos de precisão máxima do sistema (correspondendo ao somatório do cálculo da classificação semântica combinada assumindo que as classificações atribuídas pelo sistema estão totalmente correctas) e de abrangência máxima na CD (correspondendo ao somatório do cálculo da classificação semântica combinada para as EM da CD). Para melhor perceber o primeiro conceito, imagine que as categorias da tabela 16 e os tipos (agora com os restantes casos 4, 6, 8 e 11) da tabela 19 estivessem a ser sempre considerados correctos. Analogamente, o cálculo da abrangência máxima da CD utiliza a mesma fórmula para calcular o somatório das classificações combinadas para cada uma das entidades na CD.

Tabela 23: Valores das métricas para a tarefa de classificação semântica, segundo a medida combinada, para o exemplo dado.

#### 4.4.4 Medida plana

Na tabela 24 apresentamos a pontuação para a classificação semântica segundo a medida plana, na tabela 25 as medidas e na tabela 26 os valores das métricas. Mais uma vez o caso 3 é multiplicado pelo factor de correcção 0,4.

Caso	Saída do Sistema	Correcta	Em Falta	Espúria
1	<LOCAL TIPO="GEOGRAFICO">Plano hidrológico de Espanha</LOCAL>	-	(LOCAL, ADMINISTRATIVO)	(LOCAL, GEOGRAFICO)
2	<LOCAL TIPO="ADMINISTRATIVO">Lisboa</LOCAL>	(LOCAL, ADMINISTRATIVO)	-	-
3	<LOCAL TIPO="ALARGADO">Laboratório Nacional</LOCAL> de <ORGANIZACAO TIPO="SUB">Engenharia Civil</ORGANIZACAO>	(LOCAL, ALARGADO) (0,4)	-	(ORGANIZACAO, SUB)
4	<LOCAL TIPO="ADMINISTRATIVO">Lisboa</LOCAL>	(LOCAL, ADMINISTRATIVO)	-	-
5	<ABSTRACCAO TIPO="PLANO">Encontro de Reflexão</ABSTRACCAO>	-	(ACONTECIMENTO, EVENTO)	(ABSTRACCAO, PLANO)
6	<ABSTRACCAO TIPO="PLANO">Plano Hidrológico</ABSTRACCAO>	(ABSTRACCAO, PLANO)	-	-
7	<ABSTRACCAO TIPO="DISCIPLINA">Em análise</ABSTRACCAO>	-	-	(ABSTRACCAO, DISCIPLINA)
8	<LOCAL TIPO="GEOGRAFICO">Douro</LOCAL>	(LOCAL, GEOGRAFICO)	-	-
9	<LOCAL TIPO="GEOGRAFICO">Tejo</LOCAL>	(LOCAL, GEOGRAFICO)	-	-
10	<ABSTRACCAO TIPO="PLANO">Jucar</ABSTRACCAO>	-	(LOCAL, GEOGRAFICO)	(ABSTRACCAO, PLANO)

Tabela 24: Pontuações da classificação semântica, segundo a medida plana, para o exemplo dado.

Caso	Medida por categorias	Falta	Excesso
1	0	1	1
2	1	0	0
3	0,4	0	1
4	1	0	0
5	0	1	1
6	1	0	0
7	0	0	1
8	1	0	0
9	1	0	0
10	0	1	1
Total	5,4	3	5

Tabela 25: Medida por categorias, falta e excesso para o exemplo dado da tabela 24.

Métrica	Valor
Precisão	$\frac{5,4}{11} = 49,09\%$
Abrangência	$\frac{5,4}{9} = 60,00\%$
Medida F	$\frac{2 \times 0,4909 \times 0,6000}{0,4909 + 0,6000} = 0,5400$
Sobre-geração	$\frac{5}{11} = 45,45\%$
Sub-geração	$\frac{3}{9} = 33,33\%$

Tabela 26: Avaliação global da tarefa de classificação semântica segundo a medida plana.

## 4.5 Identificações alternativas

Conforme já mencionado na secção 3.5, quando existem diferentes alternativas sobre quais as correctas EM na CD, fazemos de forma a que o sistema seja avaliado de acordo com a alternativa que o favorece mais.

Para a classificação semântica, o algoritmo (melhor documentado em [16]) é o seguinte:

1. ° – Seleccionar a alternativa que maximiza a medida F segundo a medida CSC.
2. ° – No caso de empate segundo o critério anterior, escolhe-se a alternativa que maximiza o somatório dos valores da medida CSC para cada alinhamento.
3. ° – Finalmente, no caso de empate dos dois critérios anteriores, escolhe-se a alternativa com o maior número de alinhamentos.

De notar que segundo esta estratégia, é possível que tenham sido, para o mesmo sistema, escolhidos três conjuntos de alternativas diferentes, ao avaliar o seu desempenho para as três tarefas diferentes.

Na tabela 27 apresentamos alguns casos difíceis, com as respectivas saídas do sistema, e as alternativas marcadas na CD. Na tabela 28 estão os valores calculados, que servem de base para a selecção da melhor alternativa. Nos casos 1 a 5, a última alternativa é a escolhida, uma vez que apresenta a medida F mais elevada, de acordo com a 1ª condição. Nos casos 6 a 8, no entanto, a medida F é igual nos dois casos, o desempate é dado pela condição 2, ou seja, o valor da medida CSC multiplicado pelo peso (ou seja,  $\frac{n_c}{n_d}$ ).

Caso	Sistema participante	Colecção dourada
1	Agrupamento de Escuteiros	<b>ALT1:</b> <ORGANIZACAO PESSOA TIPO="SUB GRUPOMEMBRO"> Agrupamento de Escuteiros </ORGANIZACAO PESSOA> <b>ALT2:</b> <ORGANIZACAO TIPO="INSTITUICAO"> Agrupamento de Escuteiros </ORGANIZACAO> <b>ALT3:</b> Agrupamento de Escuteiros
2	<ORGANIZACAO TIPO="INSTITUICAO"> Agrupamentos de Escuteiros </ORGANIZACAO>	<b>ALT1:</b> Agrupamento de Escuteiros <b>ALT2:</b> <ORGANIZACAO TIPO="INSTITUICAO"> Escuteiros </ORGANIZACAO> <b>ALT3:</b> <ORGANIZACAO PESSOA TIPO="SUB GRUPOMEMBRO"> Agrupamento de Escuteiros </ORGANIZACAO PESSOA>
3	<VALOR TIPO="MOEDA">98 anos</MOEDA>	<b>ALT1:</b> <VALOR TIPO="MOEDA">98 anos e meio</MOEDA> <b>ALT2:</b> <VALOR TIPO="MOEDA">98 anos</MOEDA>
4	<ORGANIZACAO TIPO="INSTITUICAO"> Oficina Cultural Oswald </ORGANIZACAO> de <PESSOA TIPO="INDIVIDUAL"> Andrade </PESSOA>	<b>ALT1:</b> <LOCAL TIPO="ALARGADO"> Estúdio da Oficina Cultural Oswald de Andrade </LOCAL> <b>ALT2:</b> <ORGANIZACAO TIPO="INSTITUICAO"> Oficina Cultural Oswald de Andrade </ORGANIZACAO>
5	<ORGANIZACAO TIPO="ADMINISTRACAO"> Governo de Cavaco Silva </ORGANIZACAO>	<b>ALT1:</b> <ORGANIZACAO TIPO="ADMINISTRACAO"> Governo </ORGANIZACAO> de <PESSOA TIPO="INDIVIDUAL"> Cavaco Silva </PESSOA> <b>ALT2:</b> <PESSOA ORGANIZACAO TIPO="GRUPOCARGO ADMINISTRACAO"> Governo de Cavaco Silva </PESSOA ORGANIZACAO>
6	<ORGANIZACAO TIPO="ADMINISTRACAO"> Faculdade de Ciências e Tecnologia </ORGANIZACAO>	<b>ALT1:</b> <ORGANIZACAO TIPO="INSTITUICAO"> Faculdade de Ciências </ORGANIZACAO> <b>ALT2:</b> <ORGANIZACAO TIPO="ADMINISTRACAO"> Ciências e Tecnologia </ORGANIZACAO>
7	<ORGANIZACAO TIPO="INSTITUICAO"> Faculdade </ORGANIZACAO> de Ciências e <COISA TIPO="CLASSE"> Tecnologia </COISA>	<b>ALT1:</b> <ABSTRACCAO TIPO="ESCOLA"> Faculdade </ABSTRACCAO> de <COISA TIPO="CLASSE"> Ciências e Tecnologia </COISA> <b>ALT2:</b> <ORGANIZACAO TIPO="INSTITUICAO"> Faculdade de Ciências </ORGANIZACAO> e <ABSTRACCAO TIPO="DISCIPLINA"> Tecnologia </ABSTRACCAO>
8	<ORGANIZACAO TIPO="INSTITUICAO"> Ordem Nacional do Mérito Científico do Governo Federal </ORGANIZACAO>	<b>ALT1:</b> <VARIADO TIPO="OUTRO"> Ordem Nacional do Mérito Científico do Governo Federal </VARIADO> <b>ALT2:</b> <VARIADO TIPO="OUTRO"> Ordem Nacional do Mérito Científico </VARIADO> do <ORGANIZACAO TIPO="ADMINISTRACAO"> Governo Federal </ORGANIZACAO>

Tabela 27: Lista de exemplos para ilustrar a selecção de alternativas para a tarefa de classificação semântica

Caso	ALT	CSC	Peso	Precisão	Abrang	Med.F	CSC×Peso
1	ALT1	0	0	100%	50%	0,667	
	ALT2	0,0	0,0	100,0%	50,0%	0,667	
	ALT3	-	-	100,0%	100,0%	1,0	
2	ALT1	0,0	0,0	50,0%	100,0%	0,667	
	ALT2	1,75	0,33	66,7%	66,7%	0,667	
	ALT3	1,0	0,0	100,0%	100,0%	1,0	
3	ALT1	1,0	0,5	75,0%	75,0%	0,75	
	ALT2	1,0	1,0	100,0%	100,0%	1,0	
4	ALT1	0,0+0,0	0,429+0,143	33%	50,0%	0,40	
	ALT2	1,75+0,0	0,6+0,2	53,3%	80,0%	0,64	
5	ALT1	1,75+0,0	0,25+0,50	62,5%	41,7%	0,50	
	ALT2	1,75	1,0	100,0%	100,0%	1,0	
6	ALT1	1,0	0,6	80,0%	80,0%	0,80	0,60
	ALT2	1,75	0,60	80,0%	80,0%	0,80	1,05
7	ALT1	0,0+1,67	1,0+0,33	44,4%	44,4%	0,44	0,556
	ALT2	1,75+0,0	0,33+1,0	44,4%	44,4%	0,44	0,583
8	ALT1	0,0	1,0	50,0%	50,0%	0,50	0,0
	ALT2	0,0+1,0	0,625+0,25	62,5%	41,67%	0,50	0,25

Tabela 28: Valores calculados para os exemplos da tabela 27, para a selecção das alternativas.

## 5 Tarefa de classificação morfológica

A tarefa de classificação morfológica tem por objectivo avaliar a aptidão do sistema em identificar o género e o número das EM identificadas, em comparação com as respectivas classificações morfológicas feitas manualmente na CD.

### 5.1 Pontuação

**correcto:** quando exactamente semelhante à classificação da CD.

**incorrecto:** quando está especificado um valor específico diferente do que está na CD

**em falta:** quando falta a classificação morfológica ou está marcada “?”, ou seja, não especificada, e na CD tem um valor específico

**espúrio:** quando uma EM é espúria e o sistema lhe atribui classificação morfológica específica

**sobre-especificado:** quando na CD está “?” e o sistema marcou um valor específico

**ignorado:** como o nome indica, os valores são para ser ignorados. Usámos esta “pontuação” para evitar penalizar casos em que o HAREM não considerou classificação morfológica na CD mas que o sistema possa classificar.

Assim, casos como as EM de categoria TEMPO são simplesmente ignoradas no processamento subsequente.

### 5.2 Medidas

A tarefa de classificação morfológica é avaliada segundo três medidas:

**número:** só é considerada a pontuação relativamente ao número.

**género:** só é considerada a pontuação relativamente ao género.

**combinada:** combina-se as pontuações para género e para o número.

Quando uma EM é imperfeitamente reconhecida (ou seja, foi classificada na tarefa de identificação como parcialmente correcta), apenas contamos os casos em que essa identificação parcial concordava na primeira palavra da EM, multiplicando por um peso de 0,5.

A tabela 29 ilustra alguns exemplos de pontuação, quer em relação ao género, quer ao número, quer ao par género-número (a que chamámos combinada).



Caso	Classificação		Pontuação		
	CD	Sistema	Género	Número	Combinada
1	M,S	M,S	Correcto	Correcto	Correcto
2	M,S	F,S	Incorrecto	Correcto	Incorrecto
3	M,S	M,P	Correcto	Incorrecto	Incorrecto
4	M,S	F,P	Incorrecto	Incorrecto	Incorrecto
5	M,S	?,S	Em Falta	Correcto	Em Falta
6	?,S	M,S	Sobre-especificado	Correcto	Incorrecto
7	?,S	?,S	Correcto	Correcto	Correcto
8	M,S	Nada	Em Falta	Em Falta	Em Falta
9	sem identificação (TEMPO, etc)		Ignorado	Ignorado	Ignorado
10	sem identificação		Espúrio	Espúrio	Espúrio

Tabela 29: Pontuação para a classificação morfológica, segundo as três medidas.

## 5.3 Métricas

### 5.3.1 Precisão

Na tarefa de classificação morfológica, a precisão mede o teor de classificações em género/número correctas de todas as produzidas pelo sistema (que tenham classificação morfológica na CD). Ou seja, excluindo sempre os casos em que a EM da CD não se encontra marcada morfológicamente.

Apresentamos a precisão para as três medidas (género, número e combinada), e para os dois cenários de avaliação: independente da identificação (absoluto), ou apenas para os casos em que a identificação obteve pontuação correcta ou parcialmente correcta (relativo).

**Absoluto:**  $Precisão_{género} = (\sum EM \text{ identificadas correctamente e com género correcto} + 0,5 \sum EM \text{ identificadas parcialmente correctamente e com género correcto}) / (\sum EM \text{ com classificações de género produzidas pelo sistema})$

**Relativo:**  $Precisão_{género} = (\sum EM \text{ identificadas correctamente e com género correcto} + 0,5 \sum EM \text{ identificadas parcialmente correctamente e com género correcto}) / (\sum EM \text{ com classificações de género produzidas pelo sistema em EM identificadas correctamente ou parcialmente})$

**Absoluto:**  $Precisão_{número} = (\sum EM \text{ identificadas correctamente e com número correcto} + 0,5 \sum EM \text{ identificadas parcialmente correctamente e com número correcto}) / (\sum EM \text{ com classificações de número produzidas pelo sistema})$

**Relativo:**  $Precisão_{número} = (\sum EM \text{ identificadas correctamente e com número correcto} + 0,5 \sum EM \text{ identificadas parcialmente correctamente e com número correcto}) / (\sum EM \text{ com classificações de número produzidas pelo sistema em EM identificadas correctamente ou parcialmente})$

**Absoluto:**  $Precisão_{combinada} = (\sum EM \text{ identificadas correctamente e com género e número correcto} + 0,5 \sum EM \text{ identificadas parcialmente correctamente e com género e número correcto}) / (\sum EM \text{ com classificações de número e género produzidas pelo sistema})$

**Relativo:**  $Precisão_{combinada} = (\sum EM \text{ identificadas correctamente e com género e número correcto} + 0,5 \sum EM \text{ identificadas parcialmente correctamente e com género e número correcto}) / (\sum EM \text{ com classificações de número e género produzidas pelo sistema em EM identificadas correctamente ou parcialmente})$

### 5.3.2 Abrangência

Na tarefa de classificação morfológica, a abrangência mede o teor de classificações em género/número que se encontram na CD e que o sistema conseguiu acertar. Tal como para a precisão, mede-se a abrangência no género morfológico, no número morfológico, e na combinação de ambos.

**Absoluto:**  $Abrang\hat{e}ncia_{g\acute{e}n\acute{e}r\acute{o}} = (\sum EM\ correctamente\ identificadas\ com\ classifica\c{c}\tilde{o}es\ de\ g\acute{e}n\acute{e}r\acute{o}\ correctas + 0,5\sum EM\ identificadas\ parcialmente\ correctamente\ com\ classifica\c{c}\tilde{o}es\ de\ g\acute{e}n\acute{e}r\acute{o}\ correctas) / (\sum EM\ com\ classifica\c{c}\tilde{o}es\ de\ g\acute{e}n\acute{e}r\acute{o}\ na\ CD)$

**Relativo:**  $Abrang\hat{e}ncia_{g\acute{e}n\acute{e}r\acute{o}} = (\sum EM\ correctamente\ identificadas\ com\ classifica\c{c}\tilde{o}es\ de\ g\acute{e}n\acute{e}r\acute{o}\ correctas + 0,5\sum EM\ identificadas\ parcialmente\ correctamente\ com\ classifica\c{c}\tilde{o}es\ de\ g\acute{e}n\acute{e}r\acute{o}\ correctas) / (\sum EM\ parcial\ ou\ correctamente\ identificadas\ com\ classifica\c{c}\tilde{o}es\ de\ g\acute{e}n\acute{e}r\acute{o}\ na\ CD)$

**Absoluto:**  $Abrang\hat{e}ncia_{n\acute{u}m\acute{e}r\acute{o}} = (\sum EM\ correctamente\ identificadas\ com\ classifica\c{c}\tilde{o}es\ de\ n\acute{u}m\acute{e}r\acute{o}\ correctas + 0,5\sum EM\ identificadas\ parcialmente\ correctamente\ com\ classifica\c{c}\tilde{o}es\ de\ n\acute{u}m\acute{e}r\acute{o}\ correctas) / (\sum EM\ com\ classifica\c{c}\tilde{o}es\ de\ n\acute{u}m\acute{e}r\acute{o}\ na\ CD)$

**Relativo:**  $Abrang\hat{e}ncia_{n\acute{u}m\acute{e}r\acute{o}} = (\sum EM\ correctamente\ identificadas\ com\ classifica\c{c}\tilde{o}es\ de\ n\acute{u}m\acute{e}r\acute{o}\ correctas + 0,5\sum EM\ identificadas\ parcialmente\ correctamente\ com\ classifica\c{c}\tilde{o}es\ de\ n\acute{u}m\acute{e}r\acute{o}\ correctas) / (\sum EM\ parcial\ ou\ correctamente\ identificadas\ com\ classifica\c{c}\tilde{o}es\ de\ n\acute{u}m\acute{e}r\acute{o}\ na\ CD)$

**Absoluto:**  $Abrang\hat{e}ncia_{combinada} = (\sum EM\ correctamente\ identificadas\ com\ classifica\c{c}\tilde{o}es\ de\ n\acute{u}m\acute{e}r\acute{o}\ e\ g\acute{e}n\acute{e}r\acute{o}\ correctas + 0,5\sum EM\ identificadas\ parcialmente\ correctamente\ com\ classifica\c{c}\tilde{o}es\ de\ n\acute{u}m\acute{e}r\acute{o}\ e\ g\acute{e}n\acute{e}r\acute{o}\ correctas) / (\sum EM\ com\ classifica\c{c}\tilde{o}es\ morfol\acute{o}gicas\ na\ CD)$

**Relativo:**  $Abrang\hat{e}ncia_{combinada} = (\sum EM\ correctamente\ identificadas\ com\ classifica\c{c}\tilde{o}es\ de\ n\acute{u}m\acute{e}r\acute{o}\ e\ g\acute{e}n\acute{e}r\acute{o}\ correctas + 0,5\sum EM\ identificadas\ parcialmente\ correctamente\ com\ classifica\c{c}\tilde{o}es\ de\ n\acute{u}m\acute{e}r\acute{o}\ e\ g\acute{e}n\acute{e}r\acute{o}\ correctas) / (\sum EM\ parcial\ ou\ correctamente\ identificadas\ com\ classifica\c{c}\tilde{o}es\ morfol\acute{o}gicas\ na\ CD)$

Note-se que os denominadores para as tr\^es medidas (g\^enero, n\^umero e combinada), embora formulados de maneira diferente, s\~ao exactamente iguais, visto que n\~ao consideramos como legal a marca\c{c}\tilde{o}es\ s\~o de g\^enero ou s\~o de n\^umero.

### 5.3.3 Sobre-gera\c{c}\tilde{o}es

Relembramos que n\~ao se considera, para efeitos de avalia\c{c}\tilde{o}es, esp\^urios morfol\acute{o}gicos, ou seja, s\~o contam para avalia\c{c}\tilde{o}es os casos que tamb\^em cont\^em classifica\c{c}\tilde{o}es morfol\acute{o}gicas na CD. Assim, s\~o no cen\~ario absoluto \^e que h\~a medida de sobre-gera\c{c}\tilde{o}es, uma vez que num cen\~ario relativo, n\~ao existem EM com morfologia identificadas como esp\^urias, sendo portanto o valor desta medida sempre 0.

**Absoluto:**  $Sobre-gera\c{c}\tilde{o}es_{g\acute{e}n\acute{e}r\acute{o}} = (\sum EM\ com\ classifica\c{c}\tilde{o}es\ em\ g\acute{e}n\acute{e}r\acute{o}\ esp\^urias) / (\sum EM\ com\ classifica\c{c}\tilde{o}es\ em\ g\acute{e}n\acute{e}r\acute{o}\ produzidas\ pelo\ sistema\ e\ que\ tenham\ tamb\^em\ classifica\c{c}\tilde{o}es\ morfol\acute{o}gicas\ na\ CD)$

**Absoluto:** Sobre-geração<sub>número</sub> =  $(\sum \text{EM com classificações em número ou género espúrias}) / (\sum \text{EM com classificações de número produzidas pelo sistema e que tenham também classificação morfológica na CD})$

**Absoluto:** Sobre-geração<sub>combinada</sub> =  $(\sum \text{EM com classificações em número ou género espúrias}) / (\sum \text{EM com classificações de número ou género produzidas pelo sistema e que tenham também classificação morfológica na CD})$

### 5.3.4 Sobre-especificação

Para a tarefa de classificação morfológica, consideramos também a medida de sobre-especificação, que mede a percentagem dos casos sobre-especificados em todos os casos analisados pelo sistema. Por sobre-especificado entendemos os casos em que na CD está “?” e o sistema escolheu um determinado valor concreto.

**Absoluto:** Sobre-especificação<sub>género</sub> =  $(\sum \text{EM com classificações de género sobre-especificadas em EM identificadas correctamente} + 0,5 \sum \text{EM com classificações de género sobre-especificadas em EM identificadas parcialmente correctamente}) / (\sum \text{EM com classificações de género produzidas pelo sistema})$

**Relativo:** Sobre-especificação<sub>género</sub> =  $(\sum \text{EM com classificações de género sobre-especificadas em EM identificadas correctamente} + 0,5 \sum \text{EM com classificações de género sobre-especificadas em EM identificadas parcialmente correctamente}) / (\sum \text{EM com classificações de género produzidas pelo sistema em EM identificadas parcial ou correctamente})$

**Absoluto:** Sobre-especificação<sub>número</sub> =  $(\sum \text{EM com classificações de número sobre-especificadas em EM identificadas correctamente} + 0,5 \sum \text{EM com classificações de número sobre-especificadas em EM identificadas parcialmente correctamente}) / (\sum \text{EM com classificações de número produzidas pelo sistema})$

**Relativo:** Sobre-especificação<sub>número</sub> =  $(\sum \text{EM com classificações de número sobre-especificadas em EM identificadas correctamente} + 0,5 \sum \text{EM com classificações de número sobre-especificadas em EM identificadas parcialmente correctamente}) / (\sum \text{EM com classificações de número produzidas pelo sistema em EM identificadas parcial ou correctamente})$

**Absoluto:** Sobre-especificação<sub>combinada</sub> =  $(\sum \text{EM com classificações de número ou género sobre-especificadas em EM identificadas correctamente} + 0,5 \sum \text{EM com classificações de número ou género sobre-especificadas em EM identificadas parcialmente correctamente}) / (\sum \text{EM com classificações morfológicas produzidas pelo sistema})$

**Relativo:** Sobre-especificação<sub>combinada</sub> =  $(\sum \text{EM com classificações de número ou género sobre-especificadas em EM identificadas correctamente} + 0,5\sum \text{EM com classificações em número ou género sobre-especificadas em EM identificadas parcialmente correctamente}) / (\sum \text{EM com classificações morfológicas produzidas pelo sistema em EM identificadas parcial ou correctamente})$

### 5.3.5 Sub-geração

Na tarefa de classificação morfológica, a subgeração mede o número de classificações em falta comparadas com a informação morfológica na CD. Classificações em falta incluem tanto casos em que nenhuma classificação foi dada, como casos em que o sistema atribuiu “?” à classificação do género ou número enquanto na CD existe um valor mais específico. Como anteriormente, apresentamos separadamente as fórmulas para o cenário absoluto e relativo.

**Absoluto:** Sub-geração<sub>género</sub> =  $(\sum \text{EM com classificações em género em falta} / \sum \text{classificações em género na CD})$

**Relativo:** Sub-geração<sub>género</sub> =  $(\sum \text{EM parcial ou correctamente identificadas com classificações em género em falta}) / (\sum \text{EM parcial ou correctamente identificadas com classificações em género na CD})$

**Absoluto:** Sub-geração<sub>número</sub> =  $(\sum \text{EM com classificações em número em falta}) / (\sum \text{classificações em número na CD})$

**Relativo:** Sub-geração<sub>número</sub> =  $(\sum \text{EM parcial ou correctamente identificadas com classificações em número em falta}) / (\sum \text{EM parcial ou correctamente identificadas com classificações em número na CD})$

**Absoluto:** Sub-geração<sub>combinada</sub> =  $(\sum \text{EM com classificações em género ou número em falta} / \sum \text{classificações morfológicas na CD})$

**Relativo:** Sub-geração<sub>combinada</sub> =  $(\sum \text{EM parcial ou correctamente identificadas com classificações em género em falta} / \sum \text{EM parcial ou correctamente identificadas com classificações morfológicas na CD})$

## 5.4 Exemplo detalhado do cálculo das métricas na classificação morfológica

Nas tabelas seguintes, vamos considerar, para simplicidade de exposição, que os exemplos são relativos a EM que o participante queria classificar (cenário selectivo), ou então a todas as etiquetas da CD (cenário total), e que todas as identificações estavam correctas.

Se estivermos num cenário relativo (ou seja, só considerando as EM com valor de pontuação maior que 0 na tarefa de identificação) e os 10 exemplos da tabela 29 como um exemplo de saída do sistema participante (note-se que os casos 9 e 10 serão ignorados e não contabilizados), a avaliação global produziria os resultados apresentados na tabela 30:

Cenário Absoluto			
	Gênero	Número	Combinada
Precisão	$\frac{3}{8} = 37,5\%$	$\frac{5}{8} = 62,5\%$	$\frac{2}{8} = 25,0\%$
Abrangência	$\frac{3}{8} = 37,5\%$	$\frac{5}{8} = 62,5\%$	$\frac{2}{8} = 25,0\%$
Medida F	$\frac{2 \times 0,375 \times 0,375}{(0,375 + 0,375)} = 0,375$	$\frac{2 \times 0,625 \times 0,625}{(0,625 + 0,625)} = 0,625$	$\frac{2 \times 0,25 \times 0,25}{(0,25 + 0,25)} = 0,25$
Sobre-geração	$\frac{1}{8} = 12,5\%$	$\frac{1}{8} = 12,5\%$	$\frac{1}{8} = 12,5\%$
Sobre-especificação	$\frac{1}{8} = 12,5\%$	$\frac{0}{8} = 0\%$	$\frac{1}{8} = 12,5\%$
Sub-geração	$\frac{2}{8} = 25,0\%$	$\frac{1}{8} = 12,5\%$	$\frac{2}{8} = 25,0\%$

Cenário Relativo			
	Gênero	Número	Combinada
Precisão	$\frac{3}{7} = 42,8\%$	$\frac{5}{7} = 71,4\%$	$\frac{2}{7} = 28,3\%$
Abrangência	$\frac{3}{8} = 37,5\%$	$\frac{5}{8} = 62,5\%$	$\frac{2}{8} = 25,0\%$
Medida F	$\frac{2 \times 0,428 \times 0,375}{(0,428 + 0,375)} = 0,40$	$\frac{2 \times 0,714 \times 0,625}{(0,714 + 0,625)} = 0,666$	$\frac{2 \times 0,283 \times 0,25}{(0,283 + 0,25)} = 0,266$
Sobre-geração	-	-	-
Sobre-especificação	$\frac{1}{7} = 14,3\%$	$\frac{0}{7} = 0\%$	$\frac{1}{7} = 14,3\%$
Sub-geração	$\frac{2}{8} = 25,0\%$	$\frac{1}{8} = 12,5\%$	$\frac{2}{8} = 25,0\%$

Tabela 30: Valor das métricas para as três medidas da classificação morfológica, considerando os dez casos da tabela 29.

## 5.5 Identificações alternativas

Conforme já mencionado nas secções 3.5 e 4.5, quando existem diferentes alternativas na CD sobre quais as EM correctas, fazemos de forma a que o sistema seja avaliado de acordo com a alternativa que o favorece mais.

Para a classificação morfológica, o algoritmo (melhor documentado em [16]) é:

1. ° – Seleccionar a alternativa que maximiza a soma das três medidas F (género, número, e combinada).
2. ° – No caso de empate segundo o critério anterior, escolhe-se a alternativa com o maior número de pares em caso de empate da medida F para a classificação semântica, segundo a medida CSC.

De notar que segundo esta estratégia, é possível que tenham sido, para o mesmo sistema, escolhidos três conjuntos de alternativas diferentes, ao avaliar o seu desempenho para as três tarefas diferentes.

## 6 Apresentação dos resultados

Os resultados da avaliação foram depois apresentados sob duas formas no HAREM:

**Globais:** centrados sobre os diversos aspectos da avaliação (por uma determinada categoria, um cenário ou um género textual, por exemplo). Aqui, os desempenhos das várias saídas (devidamente anonimizadas) são reunidas em torno de tabelas e/ou gráficos, para permitir uma análise global do comportamento dos sistemas para cada aspecto da avaliação.

**Individuais:** centrados sobre o desempenho de uma saída. As tabelas e gráficos mostram a posição que a saída ocupou em relação às restantes saídas (devidamente anonimizadas). Estes relatórios possuem dados adicionais sobre o desempenho da saída que não são usados nos relatórios globais.

Após termos recebido autorização de desanonimização das saídas, criámos ainda novas tabelas com o nome dos sistemas.

### 6.1 Resultados globais

Para os resultados globais, apresentam-se várias tabelas comparativas do desempenho dos sistemas. Cada tabela diz respeito a um conjunto dos seguintes parâmetros:

**Tarefa:** pode ser identificação, classificação morfológica ou classificação semântica.

**Por critérios:** pode ser global, ou discriminado por categorias, por género textual ou por variante.

**Cenário:** pode ser total (absoluto ou relativo, nas tarefas de classificação) ou selectivo (absoluto ou relativo, nas tarefas de classificação).

**Medida:** género, número ou combinada (classificação morfológica), ou por categorias, por tipos, combinada ou plana (na classificação semântica).

De reparar que, nos relatórios globais, os sistemas são devidamente anonimizados, tendo os nomes das saídas sido substituídos por pseudónimos, que não se mantêm constantes de tabela para tabela...

As tabelas apresentam os valores para as métricas para cada medida / cenário usado. Um exemplo de tabela de resultados globais para a tarefa de identificação num cenário total pode ser visto na tabela 31.

	Precisão (%)	Abrangência (%)	Medida F	Erro combinado	Sobre-geração	Sub-geração
riad	78,50	82,84	0,8061	0,2752	0,07913	0,07329
casablanca	77,15	84,35	0,8059	0,2721	0,09134	0,03575
ancara	76,85	83,56	0,8006	0,2781	0,08966	0,04035
sana	77,43	69,57	0,7329	0,3796	0,09524	0,2079
bahreïn	59,45	64,39	0,6182	0,5056	0,2018	0,1607
asmara	56,95%	64,39%	0,6044	0,5230	0,2353	0,1607

Tabela 31: Exemplo de apresentação dos resultados globais, comparando o desempenho de várias saídas (sistemas) para uma determinada tarefa.

Os resultados globais acompanham a tabela de gráficos. Os valores são apresentados em forma de gráfico de barras (ver Figura 6) e em forma de gráfico de pontos (ver Figura 7). Nos gráficos de barras, as saídas ficam no eixo das abcissas, e nos gráficos de pontos, cada ponto representa uma saída.



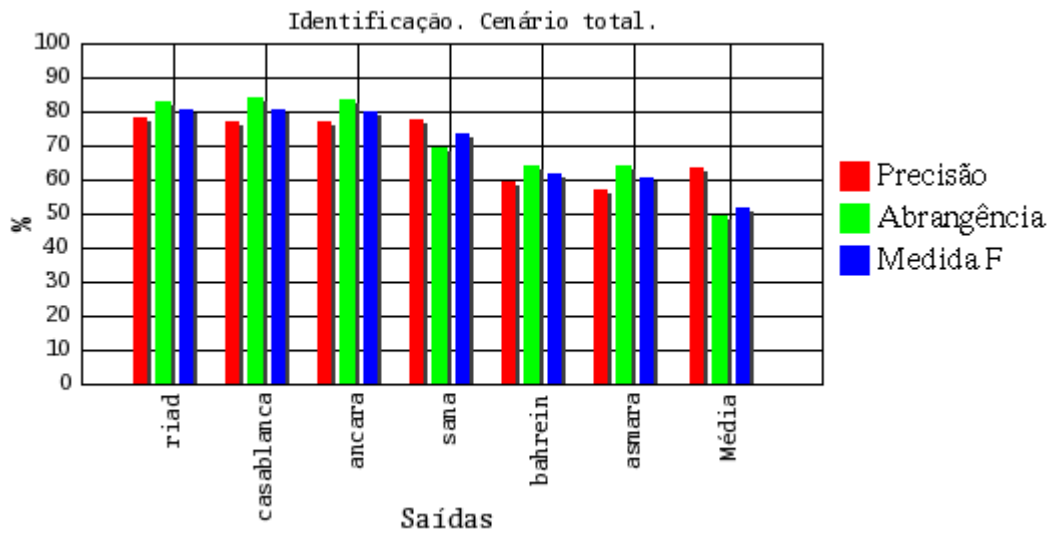


Figura 6: Exemplo de um gráfico de barras para o relatório global da tarefa de identificação (cenário total), apresentando os valores da precisão, abrangência e medida-F.

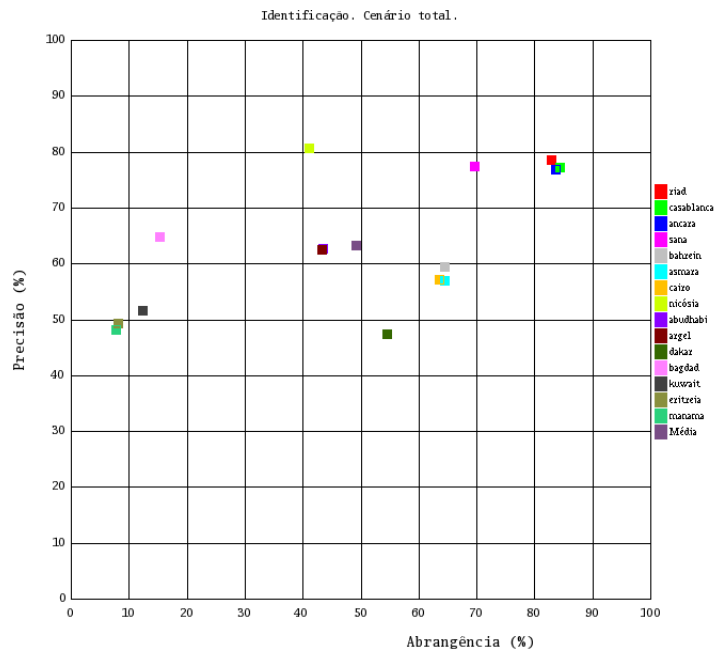


Figura 7: Exemplo de um gráfico de pontos para o relatório dos resultados globais da tarefa de identificação (cenário total).

Total na CD: 5002. Identificadas: 4494. Correctas: 3305 (66,07%).  
 Parcialmente Correctas: 836 (16,71%). Espúrias: 428 (8,56%). Em Falta: 1040 (20,79%).

Posição	Precisão (%)	Abrangência (%)	Medida F	Erro combinado	Sobre-geração	Sub-geração
4º	77,43	69,57	0,7329	0,3796	0,09524	0,2079

Tabela 32: Tabela do relatório individual para a saída RENA, para a tarefa de identificação.

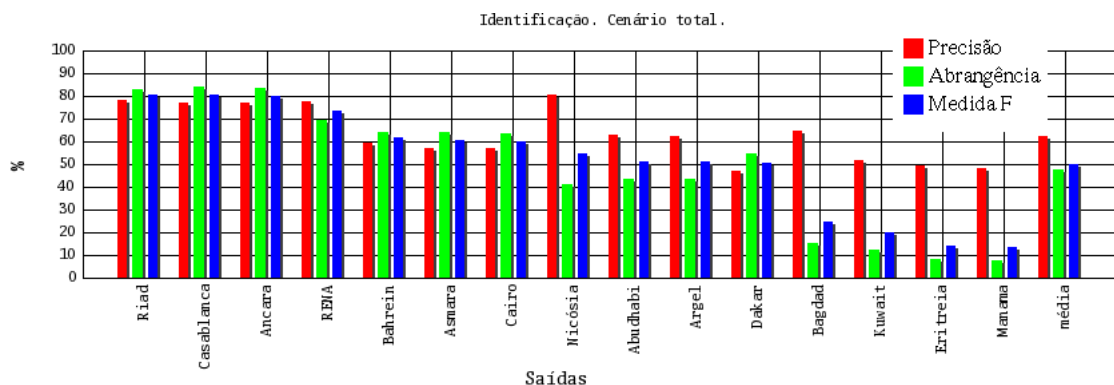


Figura 8: Exemplo de um gráfico de barras para o relatório individual da tarefa de identificação (cenário total) para a saída RENA, apresentando os valores da precisão, abrangência e medida-F.

## 6.2 Resultados individuais

Os resultados individuais de cada saída são gerados pelo módulo Alcaide [16] com base nos resultados globais, mas com os seguintes melhoramentos:

**Resultados filtrados:** Nas tabelas de resultados, só se mostra o desempenho das saídas do sistema. A tabela é complementada com informação adicional dos valores de avaliação detalhados. Nos respectivos gráficos de barras, mostra-se também o desempenho de todas as saídas, mas na legenda mostra-se o nome real das saídas do sistema em causa. A tabela 32 e as figuras 8 e 9 exemplificam o desempenho do sistema RENA.

**Agrupamento de resultados por vários critérios:** Enquanto que, nos relatórios globais, os resultados são discriminados por cada item (ou seja, há uma tabela para os desempenhos para cada categoria, género textual ou variante), nos relatórios individuais os desempenhos de uma mesma saída são reunidos numa única tabela. Adicionalmente, os valores de avaliação detalhados são agrupados em tabelas novas (ver tabelas 33 e 34).

**Gráficos precisão-abrangência individuais:** No caso de resultados por critérios (categoria, género textual ou variante), os gráficos de pontos apresentam o desempenho da saída para cada item, em vez de comparar para as restantes saídas como no relatório global (ver Figura 10).

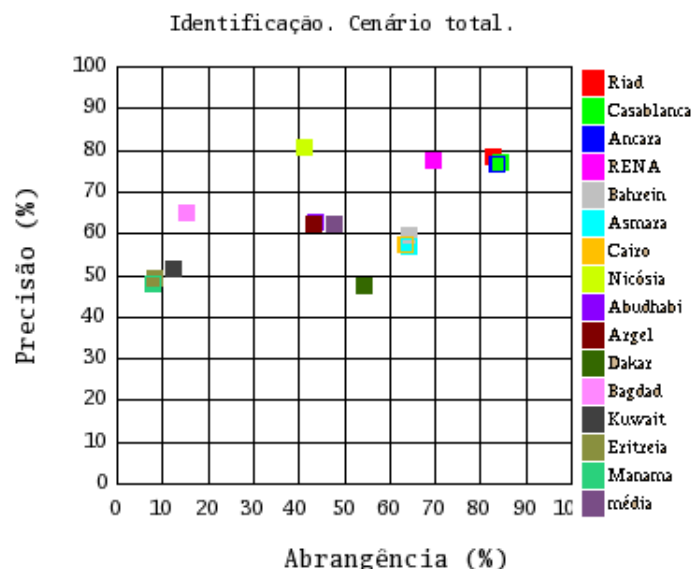


Figura 9: Exemplo de um gráfico de precisão x abrangência para o relatório individual da tarefa de identificação (cenário total) para a saída RENA.

Categoria	Total CD	Identificadas	Correctas		Parcialmente correctas		Espúrias		Em Falta	
			Total	%	Total	%	Total	%	Total	%
PESSOA	1024	619	339	33,11%	108	10,55%	178	17,38%	580	56,64%
ORGANIZACAO	955	242	176	18,43%	33	3,46%	36	3,77%	746	78,12%
TEMPO	434	264	96	22,12%	11	2,53%	161	37,10%	327	75,35%
LOCAL	1244	713	521	41,88%	47	3,78%	145	11,66%	678	54,50%
OBRA	215	4	0	0,00%	1	0,47%	3	1,40%	214	99,53%
ACONTECIMENTO	109	8	7	6,42%	0	0,00%	1	0,92%	102	93,58%
ABSTRACCAO	453	0	0	0,00%	0	0,00%	0	0,00%	453	100,00%
COISA	81	0	0	0,00%	0	0,00%	0	0,00%	81	100,00%
VALOR	479	0	0	0,00%	0	0,00%	0	0,00%	479	100,00%

Tabela 33: Exemplo de uma tabela com valores de avaliação detalhados, de um relatório individual. No caso presente, os valores referem-se ao desempenho da saída do RENA para a tarefa de identificação, discriminado por categorias (cenário total).

Categoria	Posição	Precisão (%)	Abrangência (%)	Medida F	Erro combinado	Sobre-geração	Sub-geração
PESSOA	5°	59,23	35,80	0,4463	0,6958	0,2876	0,5664
ORGANIZACAO	7°	76,03	19,27	0,3074	0,8143	0,1488	0,7812
TEMPO	7°	37,44	22,77	0,2832	0,8339	0,6098	0,7535
LOCAL	7°	74,55	42,73	0,5432	0,6179	0,2034	0,5450
OBRA	5°	9,375	0,1744	0,003425	0,9983	0,7500	0,9953
ACONTECIMENTO	5°	87,50	6,422	0,1197	0,9364	0,1250	0,9358

Tabela 34: Exemplo de uma tabela de desempenho discriminado do relatório individual. No caso presente, os valores referem-se ao desempenho da saída do RENA para a tarefa de identificação, discriminado por categorias (cenário total).

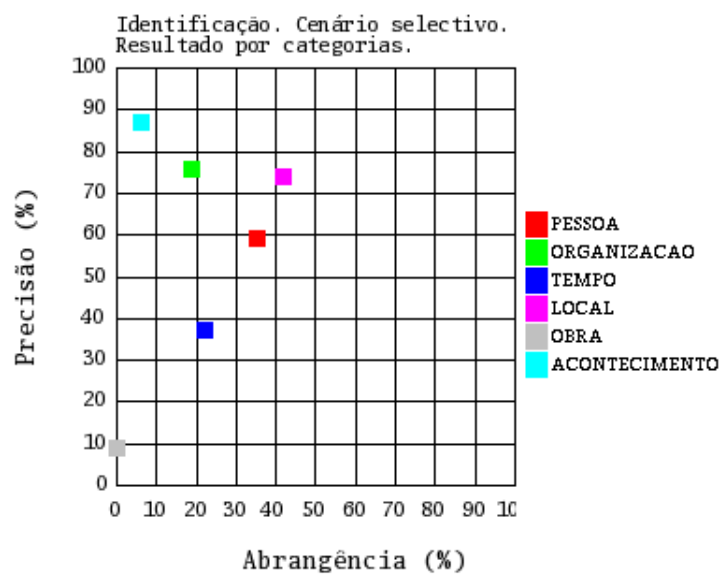


Figura 10: Exemplo do gráfico precisão x abrangência do relatório individual da tarefa de identificação (cenário total) para a saída do RENA, discriminado por categorias.

## Referências

- [1] Nuno Cardoso. Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas. Proposta de tese de mestrado. Faculdade de Engenharia da Universidade do Porto. Janeiro de 2006. <http://www.linguateca.pt/documentos/NCardosoPropostaTese.pdf>.
- [2] Nuno Cardoso. Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas. Apresentação no 2º Simpósio Doutoral da Linguateca, FCUL, Lisboa, Portugal, 10–11 de Abril de 2006. <http://www.linguateca.pt/documentos/CardosoSDL2006.pdf>.
- [3] Nuno Cardoso. Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas. Tese de Mestrado, Faculdade de Engenharia da Universidade do Porto, Outubro 2006. Também disponível como Relatório Técnico DI-FCUL TR–06–26.
- [4] Nuno Cardoso e Diana Santos. Directivas para identificação e classificação semântica na colecção dourada do HAREM. 29 de Março de 2006. Republicado como Relatório técnico DI-FCUL TR–06–18.
- [5] Nuno Cardoso e Diana Santos. “Directivas para identificação e classificação semântica na colecção dourada do HAREM”. Relatório Técnico DI-FCUL TR–06–18, Departamento de Informática da Faculdade, de Ciências da Universidade de Lisboa, Outubro 2006.
- [6] Nuno Cardoso, Diana Santos e Rui Vilela. Directivas para identificação e classificação morfológica na colecção dourada do HAREM. 29 de Março de 2006. Republicado como Relatório técnico DI-FCUL TR–06–19.
- [7] Nuno Cardoso, Diana Santos e Rui Vilela. “Directivas para identificação e classificação morfológica na colecção dourada do HAREM”. Relatório Técnico DI-FCUL TR–06–19, Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa, Outubro 2006.
- [8] Diana Santos. HAREM: the first evaluation contest for Named Entity Recognition in Portuguese. IST, Lisboa, Portugal. 24 de Fevereiro de 2006. <http://www.linguateca.pt/documentos/SantosISTFev2006.pdf>.
- [9] Diana Santos. Reconhecimento de entidades mencionadas. Palestra convidada na PUC, Rio de Janeiro, Brasil, 18 de Maio de 2006. <http://www.linguateca.pt/documentos/SantosPalestraPUCRio2006.pdf>.
- [10] Diana Santos, editora. *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press. 2007.
- [11] Diana Santos e Nuno Cardoso. “Balanço do primeiro HAREM e perspectivas de trabalho futuro”. Em Diana Santos e Nuno Cardoso, editores, *HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para português: documentação e actas do encontro*, Linguateca, 2006.
- [12] Diana Santos e Nuno Cardoso. “A Golden Resource for Named Entity Recognition in Portuguese”. Em Renata Vieira, Paulo Quaresma, Maria das Graças Volpe Nunes, Nuno J. Mamede, Cláudia Oliveira e Maria Carmelita Dias, editores, *Proceedings of the 7th International Workshop on Computational Processing of the Portuguese Language, PROPOR 2006*, volume 3960 de *Lecture Notes in Computer Science*, págs. 69–79, Itatiaia, Brasil, 13-17 Maio 2006. Springer.

- [13] Diana Santos e Nuno Cardoso, editores. *HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para português: documentação e actas do encontro*. Linguatca. 2007. Em preparação.
- [14] Diana Santos, Nuno Cardoso, Nuno Seco e Rui Vilela. “Breve introdução ao HAREM”. Em Diana Santos e Nuno Cardoso, editores, *HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para português: documentação e actas do encontro*, Linguatca, 2007.
- [15] Diana Santos, Nuno Seco, Nuno Cardoso e Rui Vilela. “HAREM: An Advanced NER Evaluation Contest for Portuguese”. Em Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik e Daniel Tapias, editores, *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, págs. 1986–1991, Génova, Itália, 22–28 Maio 2006. ELRA.
- [16] Nuno Seco, Diana Santos, Nuno Cardoso e Rui Vilela. “A arquitectura dos programas de avaliação do HAREM”. Em Diana Santos e Nuno Cardoso, editores, *HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para português: documentação e actas do encontro*, Linguatca, 2006.
- [17] Nuno Seco, Diana Santos, Nuno Cardoso e Rui Vilela. “A Complex Evaluation Architecture for HAREM”. Em Renata Vieira, Paulo Quaresma, Maria das Graças Volpe Nunes, Nuno J. Mamede, Cláudia Oliveira e Maria Carmelita Dias, editores, *Proceedings of the 7th International Workshop on Computational Processing of the Portuguese Language, PROPOR 2006*, volume 3960 de *Lecture Notes in Computer Science*, págs. 260–263, Itatiaia, Brasil, 13–17 Maio 2006. Springer.

# Índice

<b>1</b>	<b>Enquadramento</b>	<b>1</b>
<b>2</b>	<b>Introdução</b>	<b>2</b>
2.1	Pontuação . . . . .	2
2.2	Medidas . . . . .	2
2.3	Métricas . . . . .	3
2.4	Cenários de avaliação . . . . .	3
<b>3</b>	<b>Tarefa de identificação</b>	<b>4</b>
3.1	Pontuação . . . . .	4
3.2	Medidas . . . . .	5
3.3	Métricas . . . . .	5
3.3.1	Precisão . . . . .	5
3.3.2	Abrangência . . . . .	6
3.3.3	Sobre-geração . . . . .	6
3.3.4	Sub-geração . . . . .	6
3.3.5	Erro combinado . . . . .	6
3.4	Exemplo detalhado de atribuição de pontuação . . . . .	6
3.5	Identificações alternativas . . . . .	10
<b>4</b>	<b>Tarefa de classificação semântica</b>	<b>13</b>
4.1	Pontuação . . . . .	13
4.2	Medidas . . . . .	15
4.2.1	Medida por categorias . . . . .	15
4.2.2	Medida por tipos . . . . .	15
4.2.3	Medida combinada . . . . .	15
4.2.4	Medida plana . . . . .	16
4.3	Métricas . . . . .	17
4.3.1	Precisão . . . . .	17
4.3.2	Abrangência . . . . .	18
4.3.3	Sobre-geração . . . . .	19
4.3.4	Sub-geração . . . . .	19
4.4	Exemplo detalhado de avaliação da classificação semântica . . . . .	20
4.4.1	Medida por categorias . . . . .	20
4.4.2	Medida por tipos . . . . .	23
4.4.3	Medida combinada . . . . .	24
4.4.4	Medida plana . . . . .	25
4.5	Identificações alternativas . . . . .	27
<b>5</b>	<b>Tarefa de classificação morfológica</b>	<b>30</b>
5.1	Pontuação . . . . .	30
5.2	Medidas . . . . .	30
5.3	Métricas . . . . .	32
5.3.1	Precisão . . . . .	32
5.3.2	Abrangência . . . . .	32
5.3.3	Sobre-geração . . . . .	33
5.3.4	Sobre-especificação . . . . .	34

5.3.5	Sub-geração . . . . .	35
5.4	Exemplo detalhado do cálculo das métricas na classificação morfológica . . . . .	35
5.5	Identificações alternativas . . . . .	37
<b>6</b>	<b>Apresentação dos resultados</b>	<b>38</b>
6.1	Resultados globais . . . . .	38
6.2	Resultados individuais . . . . .	40
	<b>Referências</b>	<b>43</b>