## Capítulo 15

# Em busca da máxima precisão sem almanaques: O Stencil/NooJ no HAREM

Cristina Mota e Max Silberztein

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 15, p. 191–208, 2007.

A nossa participação no HAREM resulta de uma colaboração que é anterior à avaliação conjunta, enquadrando-se no âmbito do doutoramento da primeira autora. São dois os seus objectivos: (i) estudar as EM, bem como os contextos em que ocorrem, de um ponto de vista diacrónico; (ii) verificar se o desempenho de sistemas de REM é influenciado por variações temporais dos textos. Para tal, a primeira autora está a usar o CETEMPúblico, que abrange 8 anos (de 1991 a 1998), divididos em semestres.

A fim de alcançar o primeiro objectivo, foi necessário ultrapassar o obstáculo do corpus não se encontrar anotado com EM. Sendo inviável proceder à anotação manual do corpus, dada a sua extensão (180 milhões de palavras), a primeira autora optou por utilizar um ambiente de desenvolvimento para PLN que a auxiliasse nessa tarefa, o NooJ, concebido e implementado pelo segundo autor (Silberztein, 2004). Assim, desenhou e construiu uma série de recursos linguísticos (dicionários e gramáticas) para REM, designados Stencil, que são utilizados pelo sistema para produzir um texto anotado com EM. Estes recursos foram construídos manualmente e organizados de modo a serem aplicados numa cadeia de processamento que envolve três fases: (i) extracção de EM com base em regras precisas; (ii) extracção de EM com base em regras combinatórias que usam o almanaque extraído na primeira fase; (iii) anotação do texto por consulta ao almanaque extraído na segunda fase. Tanto a primeira como a segunda fase envolvem revisão manual do almanaque construído nessa fase.

O NooJ, ao ser utilizado com esses recursos, pode ser visto como um reconhecedor de EM, apesar de não ter sido desenvolvido exclusivamente com esse fim em vista. Alguns exemplos de ferramentas criadas com base em sistemas genéricos de desenvolvimento para PLN são: o ELLE (Marcelino, 2005), o AnELL (Mota e Moura, 2003) e o ExtracNP (Friburger, 2002), baseados no INTEX (Silberztein, 1993), o Glossanet (Fairon, 1999), baseado no Unitex (Paumier, 2002), e o MUSE (Maynard et al., 2003b), baseado no GATE (Cunningham et al., 2002). O ELLE (que também participou no HAREM), o ExtracNP e o MUSE são ferramentas de reconhecimento de EM.

A constituição do Stencil e a forma como os recursos que o compõem são usados pelo NooJ na análise de um texto foram condicionadas pelos objectivos do estudo anteriormente referido, sobretudo nos dois aspectos seguintes:

1. Pretende-se optimizar a anotação resultante quanto à precisão, ainda assim garantindo abrangência suficiente. Por outras palavras, é preferível anotar menos entidades, embora com maior certeza quanto à sua correcção em termos da delimitação e classificação, do que anotar mais entidades em detrimento da precisão nas anotações. Esta opção justifica-se pois só desta forma poderão os resultados da análise temporal ser precisos e representativos da totalidade das EM presentes no corpus.

| Categoria | Tipo |
|---|---|
| PESSOA | INDIVIDUAL GRUPOIND CARGO GRUPOCARGO |
| ORGANIZACAO | OUTRO (HAREM) / INSTITUICAO (Mini-HAREM) |
| LOCAL | CORREIO ADMINISTRATIVO GEOGRAFICO VIRTUAL |
| TEMPO | DATA HORA PERIODO |
| VALOR | QUANTIDADE MOEDA |

Tabela 15.1: Categorias e tipos considerados pelo Stencil/NooJ.

2. Não é desejável usar almanaques[1] de nomes próprios, a não ser os criados pelo próprio sistema a partir do texto que estiver a processar, porque isso poderia enviesar o resultado da anotação. Esse enviesamento surgiria, caso os nomes próprios contidos nos almanaques não estivessem igualmente distribuídos pelos vários semestres do corpus de estudo. Esta questão pode ser um problema uma vez que a anotação deve ser feita independentemente por semestre.

A realização do HAREM mostrou-se então a oportunidade de desenvolver e avaliar um etiquetador que produziria a anotação de EM segundo directivas acordadas entre um grupo de investigadores interessados na área em questão.

No entanto, quando começámos a trabalhar no etiquetador Stencil/NooJ tínhamos em vista o reconhecimento de entidades mencionadas no estilo do que foi proposto pelas conferências MUC (Chinchor, 1998b; Grishman e Sundheim, 1995), ou seja, reconhecimento de nomes próprios, em contexto, é certo, mas não o reconhecimento da função das EM no texto, que foi o que acabou por acontecer no HAREM. Por considerarmos a tarefa demasiado complexa, optámos por não readaptar completamente o nosso etiquetador às directivas propostas pela organização da avaliação. Essa complexidade dificultaria não só o trabalho de anotação manual a que teremos de proceder para termos uma colecção dourada por cada semestre do CETEMPúblico, como tornaria mais difícil a um sistema de anotação alcançar uma precisão e uma abrangência acima dos 90% e 40%, respectivamente, que nos permita fazer o estudo diacrónico com algum grau de fiabilidade (ou seja, as entidades que estudaremos cobrirão praticamente metade das entidades existentes no CETEMPúblico e estarão incorrectas em menos de um décimo dos casos).

Tendo em conta os nossos interesses de anotação, optámos por participar na tarefa de classificação em cinco categorias (ver Tabela 15.1): PESSOA, ORGANIZACAO, LOCAL, TEMPO e VALOR. Além disso, participámos na tarefa de classificação morfológica.

[1] Adoptámos aqui o conceito de almanaque (do inglês *gazetteer*) tal como definido por Mikheev et al. (1999): listas de nomes próprios de pessoas, locais, organizações e outra entidadades mencionadas. Note-se, no entanto, que outros autores consideram como almanaques apenas as listas constituídas por nomes próprios de locais (Grishman e Sundheim, 1995) e outros ainda, alargam a sua constituição a indicadores que possam ser úteis na classificação das EM, como por exemplo, os nomes de profissão (Sarmento et al., 2006; Bontcheva et al., 2002), ou distinguem dois tipos de almanaques: almanaques de entidades e almanaques-gatilho (*trigger gazetteers*) (Toral e Muñoz, 2006).

No que resta deste capítulo, começaremos por apresentar sucintamente o NooJ. Em seguida, descreveremos os Stencil e a cadeia de operações que é executada até obter o texto anotado. Na secção seguinte centrar-nos-emos em aspectos relacionados com a participação na avaliação: (i) mostraremos em que tarefas e categorias se focou a nossa participação, ilustrando ainda algumas das opções tomadas; (ii) contrastaremos a participação no HAREM e no Mini-HAREM, e (iii) faremos uma análise dos resultados alcançados, chamando a atenção para alguns problemas e dificuldades na anotação. Finalmente, apresentaremos algumas ideias para trabalho futuro.

## 15.1 O que é o NooJ?

O NooJ é um ambiente de desenvolvimento para PLN. À semelhança do INTEX (Silberztein, 1993), este ambiente permite, por um lado, construir descrições formais (dicionários e gramáticas) de ampla cobertura de linguagens naturais e, por outro, aplicar essas mesmas descrições a textos de grandes dimensões com grande eficiência. Essa eficiência advém do facto de ambos os sistemas manipularem descrições formais representadas por modelos computacionais de estados finitos: autómatos e transdutores, redes de transição recursivas (ou seja, transdutores que integram outros transdutores) e redes de transição recursivas com variáveis (as quais permitem replicar, condicionar e deslocar o seu conteúdo nas saídas dos transdutores).

Ambos os sistemas têm em comum diversas funcionalidades, não só porque ambos têm por objectivo fazer processamento de textos escritos, mas também por se enquadrarem no âmbito da metodologia e princípios estabelecidos por Gross (1975). Contudo, a arquitectura dos sistemas e as opções tomadas aquando do seu desenvolvimento são bastante diferentes, e o NooJ apresenta muitas funcionalidades novas.

O NooJ, cujo desenvolvimento se iniciou em 2002, foi inicialmente concebido para ser um INTEX aperfeiçoado. A primeira versão do sistema INTEX surgiu em 1992, tendo evoluído substancialmente nos 10 anos que se seguiram, sobretudo para dar resposta às necessidades dos utilizadores. Porém, a tecnologia do INTEX tornou-se obsoleta. Desenvolvido em C/C++, trata-se de um sistema monolingue, capaz de lidar com apenas um ficheiro de cada vez, sem suporte para diferentes formatos de texto, e sem suporte para XML.

Assim, em 2002, o NooJ foi desenhado de raiz, usando novas e entusiasmantes tecnologias: programação por componentes em C# para a plataforma .NET e manipulação de XML. Além disso, o seu novo motor linguístico tem a capacidade de processamento multilingue, em cerca de 100 formatos diferentes de ficheiros, incluindo documentos XML.

As funcionalidades do NooJ (das quais se destaca: análise de morfologia flexional e derivacional, elaboração de gramáticas locais, análise transformacional, indexação, localização e extracção de padrões morfo-sintácticos) estão disponíveis através de:

- um programa autónomo (*noojapply.exe*), que pode ser invocado directamente a partir de outros programas mais sofisticados;
- uma biblioteca dinâmica de .NET (*noojengine.dll*), que é constituída por classes e métodos de objectos públicos, os quais podem ser usados por qualquer aplicação .NET, implementada em qualquer linguagem de programação;
- uma aplicação integrada de janelas (*nooj.exe*), que permite executar uma série de funcionalidades num ambiente de janelas, incluindo a edição de gramáticas.

No HAREM utilizámos o ambiente de janelas.

### 15.1.1 Características dos recursos

Uma das principais vantagens do NooJ em relação ao INTEX foi ter unificado a formalização de palavras simples, palavras compostas e tabelas de léxico-gramática. Deste modo, os dicionários do NooJ permitem formalizar indistintamente palavras simples e compostas, e podem ser vistos como tabelas de léxico-gramática em que cada entrada corresponde à descrição de uma unidade lexical seguida das suas propriedades morfológicas, sintácticas e semânticas.

Estes dicionários assemelham-se aos dicionários DELAS-DELAC do INTEX, e, como tal, cada entrada é constituída por um lema seguido das suas propriedades, que no NooJ incluem, entre outras: categoria gramatical (*cat*), no máximo um código de flexão (*codflx*) introduzido por `+FLX`, zero ou mais códigos de derivação (*codderiv*) introduzidos por `+DRV` que poderão ser seguidos por um código de flexão para a forma derivada resultante (*codflex_deriv*), o qual é introduzido por ":", seguido de zero ou mais propriedades de natureza diversa; podem ainda ser especificadas, entre o lema e a categoria, variantes ortográficas ou terminológicas, tal como ilustra a seguinte entrada genérica:

```
lema{,variante}*,cat[+FLX=codflex]{+DRV=codderiv[:<codflex_deriv>]}*{+Prop}*
```

Embora estes dicionários possam ser flexionados automaticamente para efeitos de verificação e correcção (à semelhança do que acontecia no INTEX), para análise de texto não é necessário fazê-lo. Ou seja, a análise morfológica das palavras de um texto é feita directamente a partir da entrada de base (não flexionada) e do seu código de flexão no momento da aplicação do dicionário ao texto. Esta característica permite, por exemplo, a substituição de uma forma verbal que esteja no presente pela correspondente forma participial (o que poderá ser útil para transformar uma frase na forma activa na sua forma passiva).

Relativamente às gramáticas, cada gramática do NooJ corresponde a uma hierarquia de grafos constituída pelo grafo principal e todos os seus sub-grafos. Ou seja, ao contrário do que acontecia no INTEX, os sub-grafos chamados pelo grafo principal não são autónomos. Dado que, como veremos em seguida, as informações produzidas pelas gramáticas

são adicionadas incrementalmente a uma estrutura de anotação, isso torna possível a sua aplicação aos textos em cascata. Estas características permitem uma maior flexibilidade na criação, manutenção e aplicação de gramáticas.

Acrescente-se ainda, no que respeita às tabelas de léxico-gramática, que a sua unificação com os dicionários, bem como a possibilidade de processamento de análise morfológica durante a execução, permitem a sua utilização sem recorrer a meta-grafos. Este factor representa uma vantagem, em termos de descrição, já que os meta-grafos do INTEX tinham tendência a ficar demasiado grandes, e consequentemente difíceis de ler e alterar.

### 15.1.2 Processamento linguístico de textos

O motor linguístico do NooJ é baseado numa estrutura de anotação. Uma anotação é um par (posição, informação) que determina que uma certa posição no texto tem certas propriedades. Quando o NooJ processa um texto, produz um conjunto de anotações que são guardadas na Estrutura de Anotação do Texto (*Text Annotation Structure*, TAS) e estão sincronizadas com o mesmo. Portanto, a aplicação de dicionários ou de gramáticas ao texto nunca é destrutiva. Além disso, as gramáticas podem ser aplicadas em cascata, uma vez que vão sendo incrementalemente incluídas informações no TAS que podem ser usadas pelos recursos de níveis seguintes[2].

A partir das informações adicionadas ao TAS é possível criar um novo texto anotado em formato XML com essas informações integradas. Inversamente, também é possível abrir um documento XML no NooJ e integrar as anotações que nele existirem na estrutura de anotação do texto.

O sistema permite ainda a criação de colecções de textos. Esta funcionalidade torna possível aplicar a mesma operação (ou série de operações) a todos os textos de forma independente. Ou seja, a operação é aplicada a cada um dos textos individualmente, em vez de à união dos textos.

## 15.2 O que é o Stencil?

Antes do HAREM ser organizado, construímos uma série de grafos simples que faziam a anotação de nomes de pessoas, organizações e lugares no sistema INTEX. Essa classificação não tinha tipos, não tinha atributos morfológicos, mas estabelecia co-referência entre os nomes completos de organizações e as respectivas siglas ou acrónimos. Toda a informação necessária para fazer a anotação encontrava-se integrada nos grafos, não fazendo portanto uso de informações adicionais que estivessem formalizadas em dicionários, e também não tinha almanaques de nomes próprios a auxiliá-los na anotação.

[2] Saliente-se que a aplicação de gramáticas em cascata era possível no INTEX usando, por exemplo, a ferramenta CasSys (Friburger, 2002). No entanto, esta aplicação era destrutiva, pois em cada aplicação era criado um novo texto anotado.

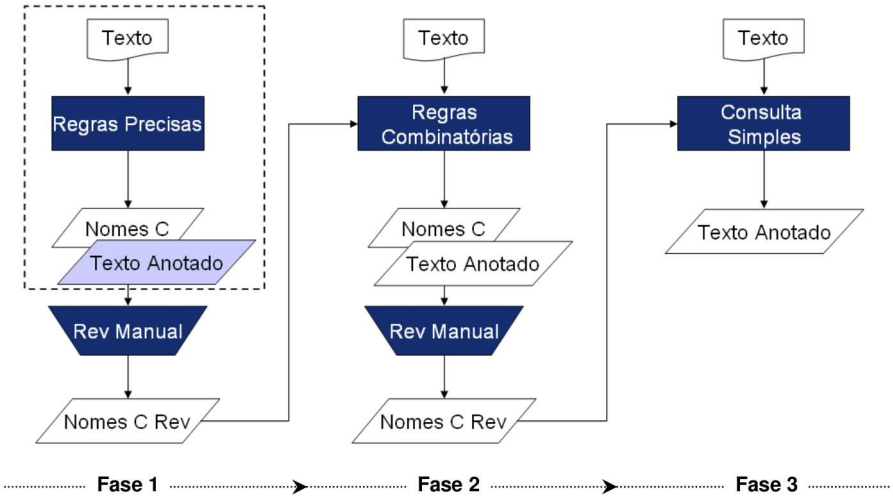


Figura 15.1: Arquitectura do etiquetador.

Uma vez que o NooJ apresentava várias vantagens em relação ao INTEX, tal como já referido na secção anterior, demos início à integração desses grafos no NooJ. Essa integração resultou praticamente numa reformulação dos grafos, pois tivemos de fazer várias modificações de acordo com as directivas do HAREM, nomeadamente: (i) prever novas categorias, (ii) fazer sub-categorização (iii) integrar classificação morfológica, e (iv) omitir a co-referência já que esta não foi contemplada na avaliação. Como o tempo era limitado, não nos aventurámos a fazer uma reestruturação completa dos grafos mais condizente com a filosofia do NooJ de construção de pequenas gramáticas para aplicação em cascata. A reformulação dos grafos também passou por uma simplificação do seu conteúdo, uma vez que muitas das informações que se encontravam explicitadas lexicalmente nos nós das gramáticas foram formalizadas em dicionários, e consequentemente essas informações lexicais passaram a ser categoriais. Por exemplo, em vez do nó conter os nomes de várias profissões (por exemplo, jornalista, linguista, pedreiro ou actor), passou a constar no nó apenas `<K+Profissão>`.

Este conjunto de recursos linguísticos, na forma de dicionários e gramáticas locais, que tem por fim fazer a anotação de EM, foi baptizado com o nome Stencil.

### 15.2.1 Organização dos recursos e forma de aplicação

Os recursos estão organizados de forma a serem aplicados em três fases distintas, como ilustrado na Figura 15.1.

Em cada uma das fases obtém-se não só um texto anotado, mas também uma lista

de nomes próprios classificados correspondentes às entidades que foram identificadas no texto. Uma vez que a última fase consiste apenas na anotação dos nomes que constarem na lista de nomes obtidos com o segundo passo, não é necessário extrair uma nova lista de nomes, pois seria idêntica à anterior. Dado que estamos interessados em fazer uma anotação optimizada quanto à precisão, as listas de nomes resultantes de cada um dos passos são revistas manualmente de modo a excluir potenciais fontes de erro nas fases seguintes. Por exemplo, se uma dada entidade for classificada com duas etiquetas distintas, em geral será eliminada da lista, pois quando a lista for reutilizada será criada uma falsa ambiguidade, que neste momento o Stencil "resolve" arbitrariamente; os nomes de pessoas ambíguos com nomes comuns também serão removidos, uma vez que a sua permanência não beneficia a análise ou poderá mesmo prejudicá-la (Baptista et al., 2006).

Através de experiências que fizemos com o CETEMPúblico, esta reutilização dos nomes encontrados no texto, sobretudo depois de revistos, permite o aumento da abrangência sem diminuir a precisão, mas apenas quando se trata da anotação de nomes próprios ao estilo das MUC. Isto porque, de uma forma geral, o nome de um local, por exemplo, não passa a ser o nome de uma organização dependendo do contexto, tal como acontece no HAREM. Este aumento de abrangência deve-se ao facto de as EM que foram encontradas pelas regras precisas poderem ocorrer noutros contextos que não foram previstos pelo primeiro conjunto de regras. Ao fazer a realimentação das EM irão ser encontradas essas ocorrências.

Dado que o nosso maior interesse era fazer a anotação do CETEMPúblico com vista à análise temporal das EM que nele ocorrem, não seria adequado o uso de almanaques de nomes próprios externos ao texto que está a ser analisado. Tal como justificado anteriormente, isso restringiria as EM encontradas, mesmo que em combinação com regras de reconhecimento com base em contexto. Embora possa parecer obscura essa opção, ela justifica-se porque, por um lado, não dispomos de recursos que estejam anotados em relação à época em que foram recolhidos e, por outro, queremos também estudar o aparecimento de novos nomes que não tenham sido previstos nos recursos.

### 15.2.2 Utilização de regras precisas

Na primeira fase, são aplicadas ao texto gramáticas locais que descrevem contextos muito restritivos que identificam e classificam EM com base em indícios internos e externos de acordo com a definição de McDonald (1996). Dado que não usámos almanaques, os indícios internos restringem apenas superficialmente a constituição interna do nome próprio dependendo da sua classificação. Por exemplo, o nome de pessoa é uma sequência de palavras em maiúsculas, eventualmente intercaladas por *de*, *do*, *das* e *dos*, não permitindo a ocorrência de *para*, como no caso das organizações; além disso, os indícios internos condicionam a primeira palavra do nome das organizações. Os indícios externos estabelecem

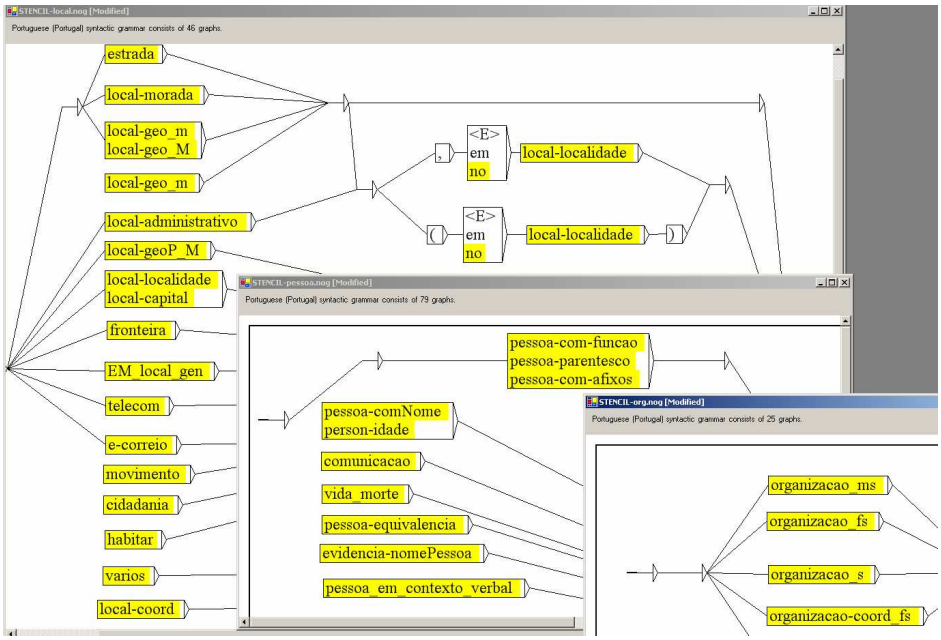


Figura 15.2: Primeiro nível das gramáticas aplicadas na primeira fase (apenas se mostra para ORGANIZACAO, PESSOA e LOCAL). O nome das sub-gramáticas encontra-se sombreado; alguns nós encontram-se desligados dos restantes por diminuirem a precisão.

contextos que com algum grau de certeza garantem a classificação da sequência em causa. Por exemplo, se uma sequência de palavras em maiúsculas que tem a constituição interna de nome de pessoa, for imediatamente precedida pelo nome de um cargo, então essa sequência será etiquetada como nome de pessoa.

As gramáticas utilizadas nesta fase estão organizadas de acordo com o tipo de entidade que reconhecem (ver Figura 15.2).

Nos casos em que era necessário fazer a classificação morfológica, os caminhos foram desdobrados de acordo com a flexão em género e número (i) do determinante que precede a sequência candidata a EM, ou (ii) do nome (no caso de ser um cargo, função, parentesco, etc.) que precede ou sucede a sequência, ou (iii) da primeira palavra que constitui a sequência, no caso dessa palavra ser um nome comum. Esse desdobramento permite atribuir a informação morfológica adequada à sequência que estiver a ser analisada como candidata a entidade mencionada. Este desdobramento deixou de ser necessário em versões do NooJ posteriores à realização do HAREM, pois passou a ser possível atribuir implicitamente atributos de elementos constituintes de uma sequência, a toda a sequência.

Adicionalmente, as gramáticas que classificam as entidades com a categoria PESSOA, tipos INDIVIDUAL e GRUPOIND, segmentam a sequência identificada como sendo nome de

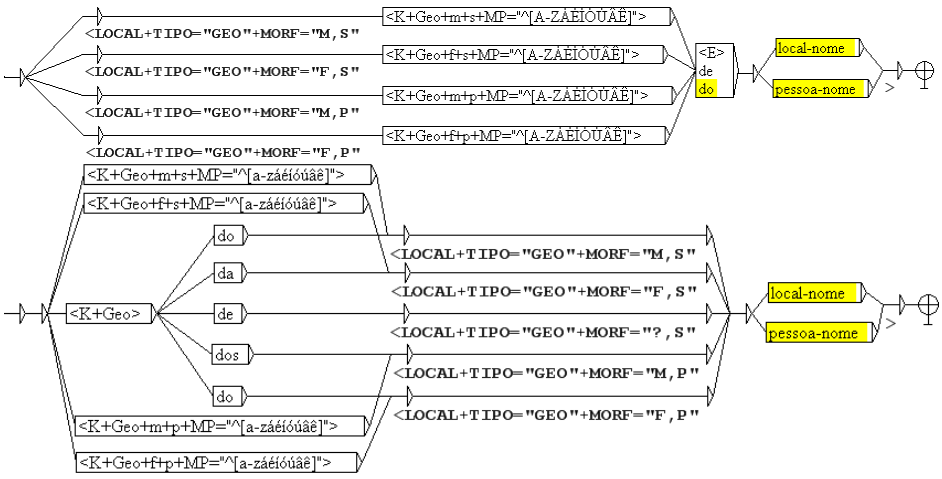


Figura 15.3: Detalhes da gramática de classificação de entidades de categoria LOCAL.

pessoa, associando a cada nome constituinte a etiqueta NOMEP. As regras combinatórias do passo seguinte usam os nomes classificados com essa etiqueta para descobrir novos nomes.

A funcionalidade do NooJ que permite combinar expressões regulares com sintaxe semelhante à do Perl com as palavras-chave do sistema permitiu simplificar e melhorar o processo de análise. Por exemplo, como se vê na Figura 15.3, dependendo de um marcador geográfico (K+Geo) começar por letra maiúscula (+MP= "[A-ZÁÉÍÓÚÂÊ]") ou minúscula (+MP="[a-záéíóúâê]") levará a que o mesmo seja ou não incluído dentro da anotação. A Figura 15.3 também ilustra o desdobramento das regras.

Estas gramáticas são aplicadas após a aplicação de um dicionário auxiliar que fornece as informações necessárias às gramáticas. Esse dicionário contém entradas nominais e adjectivais que se encontram sub-categorizadas de forma a poderem ser usadas na descrição tanto de indícios internos como externos. A constituição desse dicionário encontra-se descrita e exemplificada na Tabela 15.2.

De modo a flexionar estas formas, foram criados 51 paradigmas, dos quais 16 servem para flexionar compostos.

### 15.2.3 Utilização de regras combinatórias

A partir da anotação feita na primeira fase são geradas listas de nomes próprios classificados. Os que forem associados à etiqueta NOMEP são utilizados em regras combinatórias que identificam sequências de palavras em maiúsculas em que pelo menos um dos elementos tem essa classificação. Por exemplo, se a sequência *Jorge Sampaio* for identificada no primeiro passo como sendo PESSOA será integrada no almanaque do texto; além disso, tanto *Jorge* como *Sampaio* serão igualmente adicionados a essa lista com a classificação NOMEP.

| Tipo | Formas canónicas | Formas flexionadas | Exemplo |
|---|---|---|---|
| Adjectivos patronímicos e gentílicos | 530 | 2110 | alentejano,A+FLX=Pato+Pátrio |
| Substantivos que designam profissões e funções | 1581 | 6180 | actor,K+FLX=Actor+Profissão |
| Substantivos que designam cargos | 26 | 104 | ministro,K+FLX=Cantor+Cargo |
| Parentescos | 29 | 86 | cunhado,K+FLX=Pato+Parentesco |
| Substantivos que introduzem instituições (mais 6 que introduzem departamentos) | 81 | 162 | escola,K+FLX=Mesa+Org+Cabeça |
| Substantivos que introduzem empresas | 25 | 50 | café,K+FLX=Carro+Emp+Cabeça |
| Substantivos geográficos, dos quais 8 são geopolíticos | 39 | 78 | comarca,K+FLX=Mesa+GeoP<br>lago,K+FLX=Carro+Geo |
| TOTAL | 2311 | 8770 | |

Tabela 15.2: Constituição do dicionário auxiliar.

Se neste passo, surge a sequência *Daniel Sampaio*, mesmo que esta não tenha sido identificada pelo passo anterior, então por conter *Sampaio* passará toda ela a ser identificada como PESSOA também. Por outro lado, mesmo que esses nomes ocorram isolados também serão classificados com essa categoria.

As restantes entidades que foram igualmente colocadas no almanaque do texto serão utilizadas directamente para identificar ocorrências dessas entidades em contextos que não foram previstos pelo primeiro passo.

Com excepção da abrangência dos nomes completos de organizações cuja classificação depende exclusivamente de indícios internos (e como tal, todas as ocorrências são encontradas no primeiro passo), a abrangência dos restantes tipos de nomes vai aumentar com a execução deste passo; a abrangência das organizações só aumenta ao nível das siglas e acrónimos que no primeiro passo apenas são identificadas quando estão no contexto do nome completo da organização.

### 15.2.4 Consulta simples dos dicionários de nomes próprios extraídos

Finalmente, na terceira fase, as listas de nomes classificados extraídos a partir da anotação feita no segundo passo, são aplicadas directamente ao texto sem recurso a novas regras de combinação nem de contexto. Ou seja, este passo consiste apenas numa consulta aos almanaques revistos (manualmente) de nomes próprios gerados a partir do próprio texto com as fases anteriores.

Esta fase tem sobretudo por objectivo aumentar a abrangência dos nomes de pessoa, uma vez que com o passo anterior mais alguns novos nomes de PESSOA passaram a constar

da lista de nomes próprios.

## 15.3 Participação no HAREM

O Stencil foi desenhado a pensar numa tarefa mais simples do que a que foi proposta pelo HAREM, ou seja, a classificação dos nomes das EM. Por esse motivo, como previamente referido, não fizemos algumas distinções estabelecidas nas directivas. Eis alguns exemplos em que não respeitámos as directivas:

- Independentemente de uma organização, como seja *Hotel Alfa*, estar a ser usada como locativo (*O congresso decorrerá no Hotel Alfa*) considerámo-la como `ORGANIZACAO`.
- Mesmo que um nome geográfico, como *Moçambique*, esteja na posição de um sujeito humano (*Moçambique fornecia muito café*) considerámo-lo como `LOCAL`.
- A uma data como *6 de Novembro* que em *No dia 6 de Novembro comemora-se...* devia ser considerada do tipo `CICLICO`, foi atribuído o tipo `DATA`.

Mesmo assim, adaptámos alguns aspectos de modo a que a participação não fosse completamente desadequada:

a) Alargámos a classificação às categorias `TEMPO` e `VALOR`;

b) Integrámos a atribuição de tipos;

c) Introduzimos a classificação morfológica;

d) Adaptámos algumas regras. Por exemplo, em alguns casos, os cargos, formas de tratamento e parentescos passaram a fazer parte das entidades classificadas como `PESSOA`, tipo `INDIVIDUAL`.

Dado que não se espera numa avaliação conjunta que exista intervenção humana durante o processo de anotação, a Colecção HAREM foi anotada apenas com base no primeiro passo descrito anteriormente. Poderíamos ter considerado automatizar o processo de revisão ou eliminá-lo, antes de fazer a realimentação. Porém, tendo em conta que no HAREM a classificação de uma entidade varia com a função que desempenha na frase, o processo de realimentação tal como está desenhado seria desastroso (já que esse processo assume exactamente que a função da entidade não varia). Naturalmente, que um processo de realimentação mais sofisticado poderia ajudar a resolver esta questão, como por exemplo o descrito por Mikheev et al. (1999), mas não tivemos tempo para o fazer. Além disso, as experiências que fizemos com a colecção dourada do HAREM, enquanto preparávamos o sistema para o Mini-HAREM, mostraram que o primeiro passo de extracção de EM não era suficientemente preciso para fazer a reutilização, como se poderá confirmar pelos valores

de precisão por categoria do resultado da experiência `stencil_1`, que foram ligeiramente superiores a 70% no caso da categoria `LOCAL` e entre 60% e 70% no caso das categorias `PESSOA` e `ORGANIZACAO` (ver secção 15.3.2).

### 15.3.1 HAREM vs. Mini-HAREM

Aquando do HAREM apenas a primeira fase do Stencil estava concluída. Existia apenas uma gramática principal organizada em sub-gramáticas de acordo com as entidades que classificava e foi construído o dicionário auxiliar. O NooJ não tinha sido sequer divulgado oficialmente, e muitas funcionalidades que existem agora, na altura ainda não estavam implementadas ou aperfeiçoadas[3]. Ao HAREM foram submetidos dois resultados, um oficial e outro não-oficial (ou seja, fora de prazo). Estes dois resultados distinguem-se pelo facto de ter sido corrigido um problema que nada tinha a ver com a análise das EM: na versão oficial, as anotações adicionadas ao TAS com base em contexto (por exemplo, indícios externos) não foram consideradas aquando da criação do texto anotado. Por exemplo, se no texto existisse a sequência *a irmã de Maria*, seria adicionada ao TAS a informação de que *Maria* tinha a categoria `PESSOA:INDIVIDUAL`; no entanto, essa informação não seria adicionada ao ficheiro anotado final.

No Mini-HAREM usámos a versão 1.21/b0322 do NooJ e as três fases do Stencil já estavam concluídas. Todavia, tal como anteriormente referido, a Colecção HAREM foi anotada apenas usando o primeiro passo (o qual corresponde à zona destacada com o rectângulo tracejado na Figura 15.1). Tendo em vista a aplicação em cascata, começámos a reestruturar a gramática que usámos no HAREM, dividindo-a em quatro gramáticas de acordo com as categorias: `PESSOA`, `ORGANIZACAO`, `LOCAL` e outra que reunia `TEMPO` e `VALOR`. Além disso, corrigimos alguns erros que as gramáticas tinham, restringimos os contextos descritos e introduzimos algumas regras novas. Com o objectivo de observar a diferença de desempenho com e sem almanaques submetemos, além do resultado anterior (que designaremos por `stencil_1`), mais três resultados:

- `stencil_pol`: obtido utilizando as gramáticas do passo 1 combinadas com a consulta simples de almanaques de nomes próprios extraídos do CETEMPúblico (extractos da secção de Política dos semestres 91a, 91b e 98b) usando o primeiro passo do Stencil com revisão. Este almanaque contém 14314 nomes de locais, 31764 nomes de pessoas, e 28510 nomes de organizações, num total de 75588 nomes próprios. Por lapso, os nomes de pessoa incluídos no dicionário não estavam a ser reconhecidos (por esse motivo, nos resultados seguintes mostra-se e comenta-se apenas o resultado corrigido, `stencil_polcor`);

[3] A primeira versão pública do NooJ (1.10) foi lançada em Março de 2005.

| | Precisão | Abrangência | Medida F | Lugar |
|---|---|---|---|---|
| Identificação (cenário total) | 78,25 | 58,83 | 0,6716 | 8º |
| Identificação (cenário selectivo) | 64,09 | 63,17 | 0,6363 | 9º |
| Classificação combinada (cenário selectivo absoluto) | 40,85 | 39,63 | 0,4023 | 9º |

Tabela 15.3: Resumo das pontuações obtidas com o resultado não oficial no HAREM.

- `stencil_polcor`: obtido utilizando as gramáticas do passo 1 combinadas com o almanaque do passo anterior, com o reconhecimento de nomes de pessoas presentes no almanaque corrigido;
- `stencil_dic`: obtido utilizando as gramáticas do passo 1 e 2 em que o almanaque usado é o Npro (versão 5 sem nomes próprios ambíguos com nomes comuns) que contém 3544 nomes simples de pessoas classificados quanto a género e número, e quanto a serem nome de baptismo ou apelido (**?**).

### 15.3.2 Resultados

Relativamente à participação no HAREM, as pontuações obtidas ficaram muito aquém das expectativas, correspondendo a medida F do resultado oficial a cerca de metade do valor alcançado pelo resultado não oficial (por exemplo, no cenário total e absoluto a medida F foi 0,2073 e 0,4073, respectivamente). Essa diferença deveu-se ao facto de algumas anotações terem sido adicionadas ao TAS, sem terem sido integradas posteriormente no texto anotado oficial. Por esse motivo, não vamos sequer analisar esse resultado em mais detalhe, focando a análise de resultados no HAREM apenas nas classificações obtidas com o resultado não oficial, que acabou por não ser satisfatório devido a uma falha na gramática de reconhecimento. Por lapso, um dos caminhos da gramática que identifica as entidades do tipo `LOCAL` permaneceu demasiado genérico, o que levou a que boa parte das entidades do tipo `PESSOA` e `ORGANIZACAO`, bem como outras entidades que não pretendíamos identificar, fossem identificadas incorrectamente como `LOCAL` no resultado não oficial. Essa falha é sobretudo visível comparando a pontuação da identificação no cenário total com as pontuações da identificação no cenário selectivo e da classificação combinada no cenário selectivo (cf. Tabela 15.3). De notar que, corrigindo este erro, a medida F na classificação combinada seria inferior (23%). No entanto, observar-se-ia uma melhoria significativa em termos de precisão (66%). Como optámos por optimizar a precisão, essa correcção foi tida em conta no Mini-HAREM.

Saliente-se, no entanto, que o Stencil/NooJ obteve as melhores pontuações na identificação e classificação da categoria `TEMPO`, tendo alcançado a segunda melhor medida F e a melhor abrangência tanto na identificação como na classificação da categoria `VALOR` (ver Tabela 15.4, nos cenários total no caso da identificação, e total absoluto no caso da classificação combinada).

| | Precisão | Abrangência | Medida F | Lugar |
|---|---|---|---|---|
| Identificação (cenário total) de TEMPO | 85,74 | 76,65 | 0,8094 | 1º |
| Class. combinada (cenário total absoluto) de TEMPO | 83,24 | 74,61 | 0,7869 | 1º |
| Identificação (cenário total) de VALOR | 52,88 | 86,44 | 0,6562 | 2º |
| Class. combinada (cenário total absoluto) de VALOR | 53.63 | 87,78 | 0,6659 | 2º |

Tabela 15.4: Resumo das pontuações obtidas com o resultado não oficial no HAREM nas categorias TEMPO e VALOR.

Como se pode ver na Figura 15.4, o desempenho do Stencil/NooJ melhorou do HAREM (*stencil_no* – não oficial) para o Mini-HAREM (stencil_1, stencil_polcor e stencil_dic), em consequência de um aumento significativo da precisão.

Fazendo a análise por categoria (Figura 15.5), todas melhoraram excepto VALOR[4] que piorou em termos de medida F por ter havido uma diminuição da abrangência em troca de uma aumento de precisão. Também é possível observar que a categoria TEMPO melhorou ligeiramente a medida F como reflexo de um aumento da precisão, porém o sistema não conseguiu manter a melhor classificação nesta categoria, passando para terceiro lugar. Naturalmente, estas duas categorias não sofrem alterações nas experiências stencil_1, stencil_polcor e stencil_dic, uma vez que não dependem de almanaques. Na categoria LOCAL, em relação ao HAREM, houve uma descida da medida F, com as experiências stencil_1 e stencil_dic, como consequência da diminuição na abrangência compensada por um aumento significativo da precisão; com a experiência stencil_polcor, a medida F aumenta porque com base nos almanaques, que incluem nomes de locais, foi possível aumentar a abrangência sem prejudicar a precisão da experiência stencil_1. Com a categoria PESSOA, pelo contrário, a utilização de almanaques de nomes próprios de pessoas (quer simples, como na experiência stencil_dic, quer simples e compostos, como na experiência stencil_polcor) embora faça aumentar a abrangência, penaliza a precisão. No que respeita à categoria ORGANIZACAO, verifica-se um aumento mais significativo da medida F na experiência stencil_polcor, devido a um ligeiro aumento da abrangência, não tendo a precisão praticamente variado; esse aumento resulta sobretudo do reconhecimento de siglas que fazem parte do almanaque. O facto de não serem encontradas novas organizações para além das siglas deve-se ao facto das EM que estão no almanaque terem sido extraídas do CETEMPúblico com base em regras que dependem essencialmente dos mesmos indícios internos que estão a ser usados no reconhecimento de EM desse tipo na colecção HAREM. De acordo com as experiências de Wakao et al. (1996) esta categoria tem a beneficiar com o uso de indícios externos, nomeadamente porque muitas organizações são nomes de empresas, os quais não contêm em geral indícios internos bem definidos.

[4] No resultado stencil_1 a categoria VALOR, embora tenha sido adicionada ao TAS não foi exportada acidentalmente para o texto anotado final. Caso essas anotações tivessem sido exportadas, obter-se-ia para a categoria VALOR na classificação combinada uma precisão de 93,82%, uma abrangência de 37,18% e uma medida F de 53,26%. Estes valores são naturalmente semelhantes aos obtidos nas restantes experiências do Mini-HAREM.

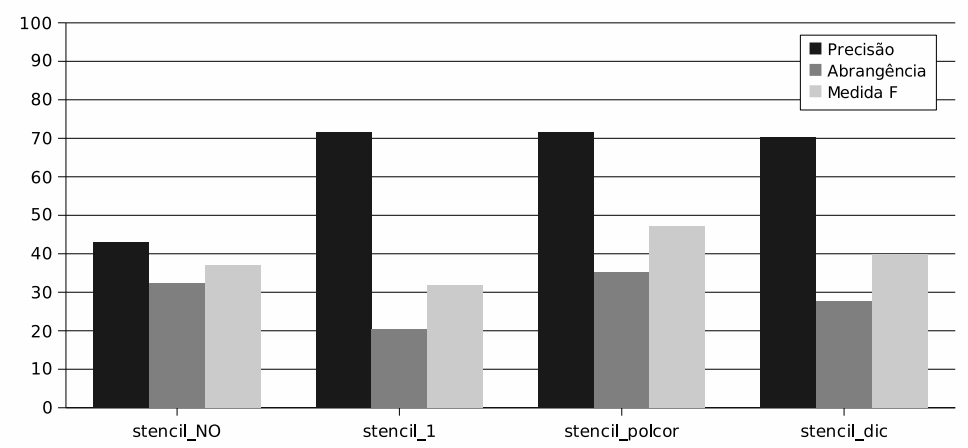


Figura 15.4: Classificação combinada no cenário total absoluto.

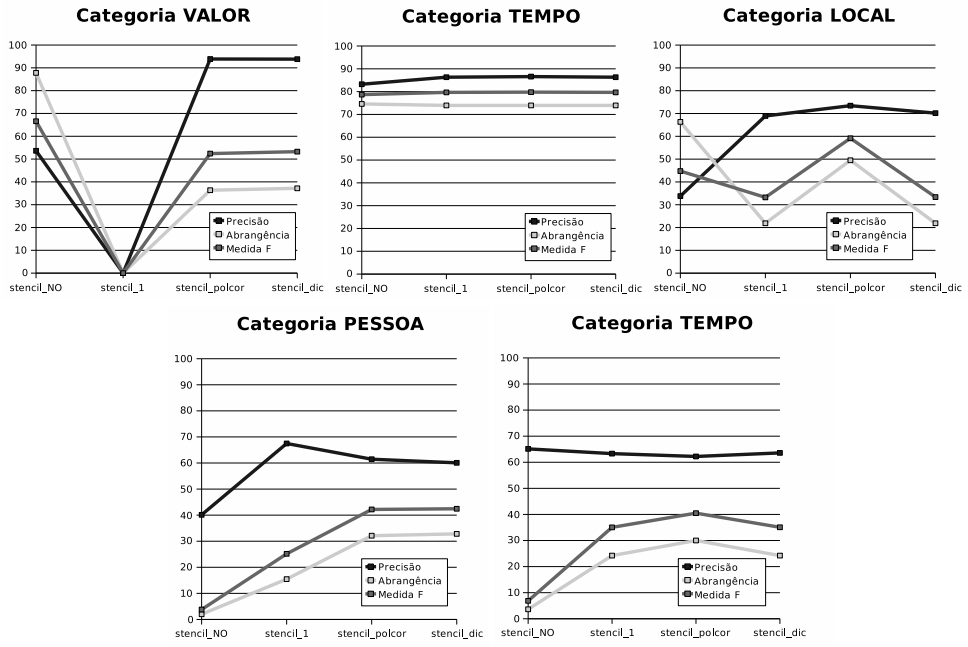


Figura 15.5: Classificação combinada por categoria no cenário total absoluto

Saliente-se ainda que os resultados obtidos para PESSOA, ORGANIZACAO e LOCAL, apesar de significativamente piores do que os de Mikheev et al. (1999), não são de espantar e sugerem a mesma conclusão: o reconhecimento da categoria LOCAL não consegue tirar partido tão facilmente do contexto e por isso o uso de almanaques ajuda, sobretudo, a melhorar o reconhecimento de entidades deste tipo.

Comparativamente com os outros sistemas participantes, apesar de não termos uma medida F tão boa devido à falta de abrangência, conseguimos mesmo assim estar entre os sistemas com melhor precisão.

No que diz respeito à classificação morfológica, acabámos por ser o único sistema a submeter resultados ao Mini-HAREM que a integrassem. Todavia, esses resultados não foram positivos. Para além da falta de abrangência (que não foi superior a 15% no melhor cenário total absoluto e mesmo no cenário total selectivo não ultrapassou os 20%), sobretudo nos resultados que foram obtidos com auxílio do almanaque do CETEMPúblico (stencil_polcor), a precisão foi baixa (61% no melhor caso no cenário total absoluto), sendo, no entanto, ligeiramente melhor em termos de número (no melhor caso, 35% de medida F no cenário total absoluto) do que em género (25% de medida F no cenário total absoluto, no melhor caso). Mesmo assim, tendo apenas em conta as entidades que são bem identificadas, os resultados são bem melhores (a medida F, passa de 25% no cenário total absoluto para 58% no cenário total relativo).

### 15.3.3 Problemas e dificuldades

Apesar de estarmos à espera de uma abrangência baixa, esta poderia ter sido mais alta se alguns pequenos lapsos na descrição das regras não tivessem ocorrido. Por exemplo, a regra que atribuía a categoria PESSOA a uma sequência de maiúsculas que ocorre após um cargo iniciado por letra minúscula tinha uma pequena falha que impediu a anotação das entidades neste contexto. Na experiência stencil_1, por exemplo, a correcção deste pequeno erro faria aumentar a abrangência de 15,46% para 16,89% e a precisão de 67,48% para 69,03% na classificação combinada da categoria PESSOA. Por outro lado, regras que em termos de precisão pudessem ser arriscadas por envolverem algum grau de ambiguidade não foram previstas. Por exemplo, se entre o nome de um cargo e uma sequência de maiúsculas existir a preposição *de* eventualmente contraída com um artigo definido, então é possível que essa sequência seja uma ORGANIZACAO (*o presidente da Sonae*); no entanto, também pode ser um LOCAL (*o presidente da China*); note-se, porém, que segundo as directivas do HAREM o segundo caso deve também ser anotado como ORGANIZACAO, mas terão tipos diferentes: EMPRESA no primeiro caso e ADMINISTRACAO no segundo.

O facto de termos dividido a gramática que tínhamos inicialmente em quatro gramáticas também trouxe algumas dificuldades. Por exemplo, com uma única gramática dada a sequência *o professor Ribeiro da Silva* que permite fazer a análise de *Ribeiro da Silva* como

PESSOA (por ocorrer a seguir a *professor*) bem como de LOCAL (por conter *ribeiro*), apenas a primeira anotação como PESSOA vai ser adicionada ao TAS por fazer parte de um caminho mais longo que tem precedência sobre análises mais curtas. Pelo contrário, usando as gramáticas separadas ambas as anotações são adicionadas ao TAS, o que leva a que no momento da geração do texto anotado o NooJ opte arbitrariamente por uma delas. Chamamos a atenção para o facto de neste momento já poderem ser geradas as duas anotações, o que, seja como for, não é a solução que pretendemos pois trata-se de uma falsa ambiguidade.

## 15.4 Comentários finais

Apesar de não termos seguido à risca as directivas da avaliação conjunta e termos acabado por concorrer com um sistema preparado para uma tarefa mais simples e com menos categorias, consideramos a participação positiva. Em particular, conseguimos uma precisão equiparável à do melhor sistema no Mini-HAREM (acima de 70%, enquanto o melhor sistema teve 73,55%), e por vezes ligeiramente melhor, apesar de ter tido uma medida F que variou entre 20% e 47%, quando o melhor sistema obteve quase 59%, no cenário total absoluto.

Contamos, numa futura edição do HAREM, caso se mantenham os objectivos de anotação da função das entidades, ser mais fiéis às directivas, mesmo que isso nos obrigue a manter dois sistemas diferentes: um para fins de anotação do CETEMPúblico com nomes próprios no âmbito da tese da primeira autora, e outro com o objectivo de competir conjuntamente na avaliação.

Mais do que a questão de quão bons foram os resultados na avaliação, interessa-nos saber quão melhores é que eles se tornarão no futuro. Para isso os programas avaliadores criados pela organização do HAREM (capítulo 19) são um instrumento fundamental para poder ir desenvolvendo e testando o sistema.

### Agradecimentos

Os autores estão gratos ao grupo *Text Analysis and Language Engineering* do centro de investigação da IBM, T. J. Watson Research Center, por lhes terem dado a oportunidade de em 2001 trabalharem em conjunto em REM, o que serviu, em parte, de fonte inspiradora para o trabalho aqui apresentado. Os autores estão igualmente gratos ao Nuno Seco pelo apoio dado na utilização dos programas avaliadores, bem como ao Nuno Mamede, à Diana Santos, ao Nuno Cardoso, aos autores do CaGE, ao Luís Costa e ao Jorge Baptista pelas sugestões que nos deram para melhorar a versão final deste capítulo.

O trabalho da primeira autora foi financiado pela Fundação para a Ciência e a Tecnologia através da bolsa de doutoramento com a referência SFRH/BD/3237/2000.