

Capítulo 15

Reconhecimento de entidades mencionadas com o XIP: Uma colaboração entre a Xerox e o L2F do INESC-ID Lisboa

Caroline Hagège, Jorge Baptista e Nuno Mamede

Apresentamos neste capítulo um sistema de reconhecimento de entidades mencionadas (REM) desenvolvido numa colaboração entre o L²F (INESC-ID Lisboa) e o XRCE (Xerox Research Centre Europe, Grenoble, França). Trata-se de um sistema baseado em regras (no que diz respeito à parte do reconhecimento das entidades mencionadas) e integrado numa ferramenta mais geral de análise sintáctica do português, XIP (Xerox Incremental Parser). Uma das características da nossa abordagem é que a parte de REM está completamente integrada numa cadeia mais geral do processamento do português que vai da segmentação à análise sintáctica.

O capítulo desenvolve-se do seguinte modo: Começaremos por apresentar brevemente a ferramenta que usamos (secção 15.1), descrevendo a estratégia adoptada para enriquecer o analisador sintáctico com um módulo de REM. Numa segunda parte, descreveremos em pormenor as várias etapas de processamento e os recursos empregues, léxicos (secção 15.2) e regras (secção 15.3), no reconhecimento das diversas categorias de EM. Procuraremos, sobretudo, dar uma panorâmica geral do funcionamento sistema e explicar a metodologia seguida no seu desenvolvimento. Apresentaremos também o mecanismo de propagação (secção 15.4.4), disponível no XIP, que permite inferir novas EM a partir de EM previamente reconhecidas. Finalmente (secção 15.5), comentaremos os resultados obtidos e faremos um balanço da nossa participação na avaliação conjunta do Segundo HAREM.

15.1 XIP: Uma ferramenta para o processamento lexical, sintáctico e semântico

Começamos por fazer uma breve apresentação do sistema XIP (Xerox Incremental Parser) que utilizamos para a tarefa de reconhecimento de entidades mencionadas.

O XIP (Ait-Mokhtar et al., 2002) é um analisador cujo primeiro objectivo é a extracção de dependências sintácticas. O analisador processa documentos em formato de texto ou XML e produz como saída uma representação sintáctica do conteúdo do documento.

O XIP oferece um formalismo rico, que permite expressar um leque importante de regras, que vão da desambiguação das categorias das palavras, até à construção de dependências, passando pela delimitação de sintagmas nucleares¹.

Importa, no entanto, frisar desde já que o conceito de dependência que adoptámos no XIP não corresponde à noção de dependência considerada, por exemplo, em Tesnières (1959) mas constitui uma perspectiva muito mais abrangente, nomeadamente por não respeitar o princípio de projectividade. Também difere da noção de dependência apresentada em Tapanainen e Järvinen (1997) por se aplicar a relações não necessariamente binárias e por poder permitir relações não só entre unidades lexicais mas também entre sintagmas nucleares.

O XIP é actualmente utilizado no processamento de várias línguas (inglês, francês, japonês, italiano, espanhol e português), estando as gramáticas dessas línguas em diferentes estádios de desenvolvimento.

15.1.1 Ilustração

Para ilustração, veja-se nas figuras 15.1 e 15.2 a análise feita pelo XIP da frase (15.1).

¹ Traduzimos a palavra do inglês “chunk” por “sintagma nuclear”. Mais precisamente, por sintagma nuclear consideramos um grupo sintáctico, não recursivo, cujo limite direito corresponde à cabeça sintáctica dum sintagma tradicional.

```

TOP{
  PP{Na visão}
  PP{do ministro}
  NP{o seguro}
  AP{agrícola}
  VF{desempenhará}
  NP{importante papel}
  PP{no projeto}
  PP{do Governo}
  VINF{de estimular}
  NP{a agricultura}
  PP{através do NOUN{programa Brasil Empreendedor Rural}} .
}

```

Figura 15.1: Construção de sintagmas nucleares

```

MAIN(desempenhará)
DETD(visão ,a)
DETD(ministro ,o)
DETD(seguro ,o)
DETD(projeto ,o)
DETD(Governo ,o)
DETD(agricultura ,a)
DETD(programa Brasil Empreendedor Rural ,o)
PREPD(visão ,Na)
PREPD(ministro ,do)
PREPD(projeto ,no)
PREPD(Governo ,do)
PREPD(programa Brasil Empreendedor Rural ,através do)
MOD-PRE(papel ,importante)
MOD-POST(seguro ,agrícola)
MOD-POST(visão ,ministro)
MOD-POST(projeto ,Governo)
MOD-POST(estimular ,programa Brasil Empreendedor Rural)
SUBJ-PRE(desempenhará ,seguro)
CDIR-POST(desempenhará ,papel)
CDIR-POST(estimular ,agricultura)

```

Figura 15.2: Dependências (principais) extraídas

(15.1) Na visão do ministro, o seguro agrícola desempenhará importante papel no projeto do Governo de estimular a agricultura, através do programa Brasil Empreendedor Rural.

Nesta saída do sistema, pode-se observar que, além da delimitação dos sintagmas nucleares (NP, PP, AP, etc.), o XIP permite também extrair relações gramaticais entre constituintes, tais como sujeito (SUBJ) ou complemento directo (CDIR), além de identificar o núcleo da frase (MAIN) e várias outras relações sintácticas, tais como a relação entre o determinante e o núcleo nominal por ele determinado (DETD) ou entre preposição e núcleo nominal (PREPD). Finalmente, são também extraídas algumas relações genéricas de modificação (MOD) entre um núcleo (seja ele verbal, nominal ou de outra categoria) e um argumento ou modificador deste núcleo².

² Nesta fase do desenvolvimento da gramática do português, por falta de informação lexical sistemática, ainda não fazemos a distinção entre complementos (argumentos de um operador) e adjuntos.

15.1.2 Desenvolvimento do módulo de REM

O desenvolvimento do módulo de REM³ seguiu uma metodologia que obedece a duas orientações gerais:

- integração do módulo de REM no âmbito mais abrangente do processamento morfo-sintáctico do português.
- tratamento incremental da informação linguística.

15.1.2.1 Integração do REM no processamento geral do português

A integração do módulo de REM na cadeia de processamento é motivada por vários factores: em particular, o reconhecimento das EM permite melhorar os resultados dos outros módulos de processamento linguístico. Com efeito, as entidades mencionadas constituem superficialmente uma estrutura sintáctica por vezes complexa. No entanto, enquanto EM, elas correspondem muitas vezes a um nome. Por exemplo, a expressão *E tudo o vento levou*, que corresponde ao título de uma obra, tem a estrutura superficial de uma frase. Contudo, para proceder a uma correcta análise sintáctica da frase (15.2), é necessário determinar que *E tudo o vento levou* corresponde a um nome, para, por exemplo, não se interpretar *E* como uma coordenação e se estabelecer adequadamente a relação de dependência entre *ontem* e *fomos rever* e não entre o advérbio e o verbo *levou*.

(15.2) Fomos rever *E tudo o vento levou ontem*

O facto de se ter acesso à estrutura sintáctica permite ir mais longe na tarefa de REM. De facto, é possível, graças à informação sintáctica, resolver certos casos de uso metonímico de EM. Por exemplo, pode-se determinar o emprego metonímico de *Portugal* como PESSOA (GRUPOIND) em vez de LOCAL em exemplos como *Portugal respondeu...* sabendo que, aqui, *Portugal* - por defeito um nome de lugar - é o sujeito de um *verbum dicendi*.

Sobre as vantagens de integrar o REM na análise sintáctica, veja-se por exemplo Brun e Hagège (2004).

15.1.2.2 Tratamento incremental da informação linguística

Apresentamos na figura 15.3 a arquitectura geral, ilustrando a forma como o módulo de REM está integrado no sistema desenvolvido para o português (XIP-PT).

15.2 Léxico e pré-processamento

15.2.1 O que é uma entrada lexical no XIP?

Uma entrada lexical no XIP corresponde a um conjunto de traços (atributos-valores). Todos os traços e todos os valores possíveis têm de ser declarados explicitamente, com excepção de alguns traços geridos pelo sistema que são os traços *lemma*, *surface*, *maj* e *toutmaj*.

Os traços *lemma* e *surface*, cujo valor é uma cadeia de caracteres, correspondem, respectivamente, ao lema da unidade linguística e à forma de superfície da unidade linguística;

³ Este desenvolvimento foi iniciado pelos trabalhos de Loureiro (2007) e Silva Romão (2007).

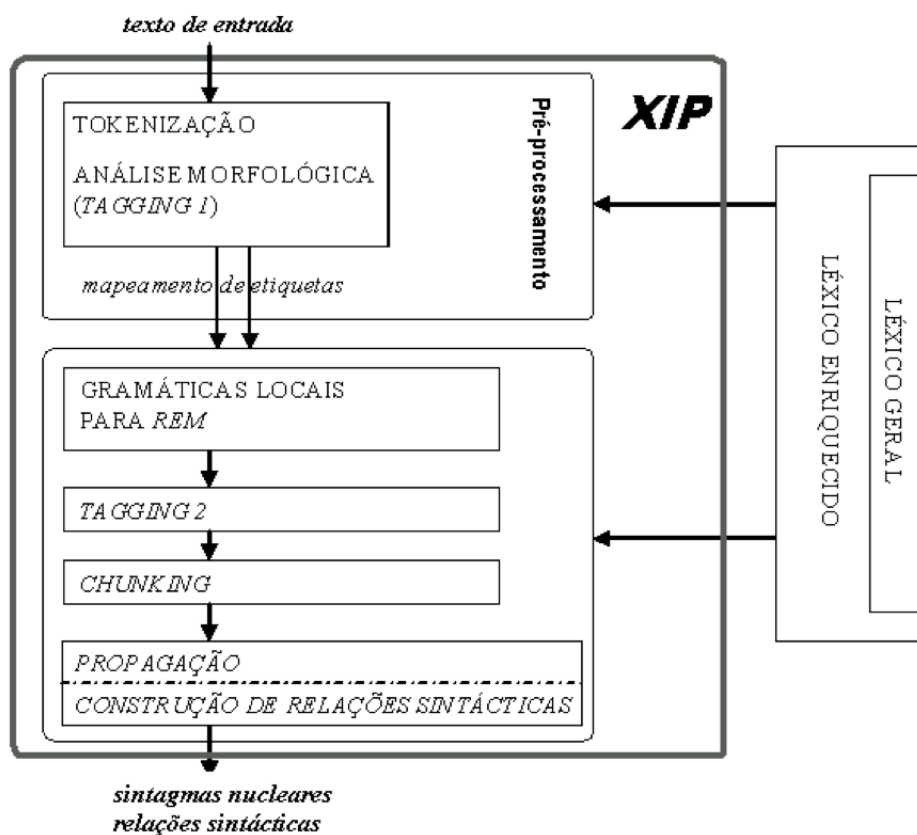


Figura 15.3: Arquitectura geral do sistema XIP-PT

maj e *toutmaj* são traços booleanos que indicam, respectivamente, se a forma de superfície começa por uma maiúscula, ou se a forma de superfície é totalmente em maiúscula. Todos os outros traços associados ao léxico são traços sintáticos ou semânticos e são definidos e declarados na gramática.

15.2.2 Dois tipos de léxicos

Consideramos dois tipos de léxico no XIP:

- léxico pré-existente
- léxico definido no XIP

15.2.2.1 Léxico pré-existente

Chamamos léxico pré-existente ao léxico oriundo da ferramenta de análise morfológica (e, possivelmente, do processo de anotação morfossintática (em inglês, *POS tagging*), a que chamamos pré-processamento sintático. Para integrar este léxico no XIP, é necessário

definir o mapeamento entre as categorias e traços do pré-processamento sintáctico (ver figura 15.3) e as categorias e traços que vão ser manipulados dentro do XIP⁴.

15.2.2.2 Léxico definido no XIP

No XIP, também é possível definir directamente entradas lexicais ou, então, modificar entradas lexicais pré-existentes. Para o Segundo HAREM, privilegiámos esta última abordagem. Sendo assim, os dados lexicais necessários para a tarefa de REM foram acrescentados, sob a forma de léxico XIP, aos léxicos gerais utilizados pelo analisador.

A construção de novos recursos lexicais, definidos no XIP especificamente para a tarefa de REM, consistiu, pois, essencialmente nos seguintes passos:

- introdução de novas entradas que constituem EM:

```
Herbert += [people:+, individual:+, firstname:+].
```

Aqui, definiu-se uma nova entrada *Herbert*, à qual se associaram novos traços booleanos, a saber: *people:+*, *individual:+*, *firstname:+*. São estes traços que permitem marcar esta entrada como o primeiro nome de uma pessoa.

- enriquecimento de entradas do léxico pré-existente com novos traços:

Este processo consistiu basicamente na marcação de elementos linguísticos que funcionam como pistas contextuais, isto é, que servirão em seguida para a identificação e classificação de EM (ver mais adiante a apresentação das regras locais). Exemplo:

```
arcebispo: noun += [cargo:+].
```

Aqui, acrescentámos à entrada (lema) *arcebispo*, já existente nos léxicos, o traço *cargo:+*. Este traço será depois utilizado por diversas regras (ver adiante).

É de notar que tanto a parte de enriquecimento como a parte de novas entradas foi feita com base em listas de palavras já existentes ou criadas manualmente para o efeito. Por outras palavras, não utilizámos processos automáticos para enriquecimento lexical.

15.2.3 Adaptação do pré-processamento

Além do enriquecimento do léxico, as regras de desambiguação categorial da gramática geral foram adaptadas especificamente para a tarefa de REM. Para desambiguação categorial, adoptámos também uma abordagem incremental que pode ser resumida da maneira seguinte:

- uma primeira fase de desambiguação por regras (a que chamamos *Tagging1* na figura 15.3);
- uma segunda fase de desambiguação por regras (integrada no módulo a que chamamos *Tagging2*, na figura 15.3);

⁴ Nesta participação no HAREM, trabalhamos com duas ferramentas de pré-processamento distintas (pertencentes a cada uma das instituições – INESC-ID e XEROX – que colaboraram neste trabalho). O resultado de cada mapeamento permitiu que, mesmo sendo dois léxicos distintos, fosse possível utilizar uma gramática e módulo de REM comuns no XIP.

- uma escolha por defeito realizada por um HMM (*Hidden Markov Model*) igualmente integrado em `Tagging2`.

A primeira fase de desambiguação é muito específica e corresponde à desambiguação particular de certas formas linguísticas. Por exemplo, a regra:

```
5> verb<lemma:podar>, verb<lemma:poder> = verb<lemma:poder> |
verb[inf:+] | .
```

pode ser descrita por: antes de uma forma verbal no infinitivo não flexionado, a unidade lexical *pode* é uma forma do verbo modal *poder*, e não a forma do presente do conjuntivo do verbo *podar*.

Para a tarefa de REM, além das regras existentes, foram acrescentadas novas regras que dizem respeito explicitamente aos acréscimos lexicais que foram feitos para o REM. Por exemplo, a regra seguinte permite desambiguar a entrada `Natal` (quadra festiva ou estado do Brasil):

```
20> noun[maj:+, surface:Natal] %= | noun[denot_time:],
prep[lemma:de], art | noun[one_day=+,maj=+,proper=+] .
```

Esta regra determina que, depois de uma palavra como *altura* ou *tempo* seguida pela preposição *de* e um artigo, a palavra *Natal* corresponde à quadra festiva (a interpretação como estado do Brasil é então excluída).

15.3 Gramáticas locais para o REM

15.3.1 Expressão de gramáticas locais em XIP

O XIP oferece um formalismo que permite, entre outras coisas, exprimir regras de reescrita tomando em consideração, facultativamente, os contextos à esquerda e à direita da expressão regular a reescrever:

```
LHS = | <reg_expr_esq> | <reg_expr> | <reg_expr_dir> |
```

LHS vai corresponder a um novo nó resultante do emparelhamento de `reg_expr` (expressão regular de categorias, aumentadas pela possibilidade de fazer restrições sobre os traços associados a estas categorias), com, eventualmente, a verificação dos contextos à esquerda e à direita (`reg_expr_esq` e `reg_expr_dir` respectivamente) que correspondem também a expressões regulares de categorias.

A este novo nó, representado por LHS, podem ser acrescentados novos traços (na medida em que eles forem previamente declarados).

Este formalismo é usado para definir regras de gramáticas locais para as EM em dois tipos de situações:

- para a delimitação de EM constituídas por mais de uma unidade lexical;
- na utilização do contexto imediato para delimitar e classificar EM.

15.3.2 Delimitação de EM complexas

Algumas das EM a reconhecer são constituídas por mais de uma palavra gráfica (unidades), possibilidade que, naturalmente, está contemplada nas directivas do Segundo HAREM. É o caso de *Oceano Atlântico, senhor Pedro da Conceição* e muitos outros.

Contudo, nas fases preliminares de pré-processamento, as várias unidades que constituem estas expressões produtivas apenas foram considerados individualmente, sendo função das gramáticas locais juntar agora esses elementos numa única EM.

Por exemplo, a regra a seguir constrói um nome complexo ao qual se acrescentam os traços *cargo:+* e *people:+*, para uma sequência que começa por um elemento lexical que tem o traço *cargo:+*, seguido ou pelo adjectivo *honorário* ou pelo adjectivo *mor* em maiúscula.

```
1> noun[cargo=+,mwe=+,people=+] @=
   ?[cargo,maj], (punct[hifen]),
   adj[lemma:"honorário", maj]; adj[lemma:mor].
```

Assim, sequências como *Cônsul Honorário* ou *Sargento-mor* vão, graças a esta regra, ser consideradas como nomes de cargo.

15.3.3 Utilização de contexto imediato

Para algumas unidades lexicais, é o contexto imediato que permite reconhecer ou classificar uma EM. As regras locais utilizando o contexto adaptam-se perfeitamente a esta tarefa. No exemplo seguinte, apresentamos uma dessas regras:

```
1> NOUN[org=+, institution=+] @= |[lemma:governo, maj: ],
   prep[lemma:de], (art)| ?[location].
```

Esta regra faz com que, num contexto à direita constituído por *governo* seguido da preposição *de* e eventualmente seguido por um artigo, um elemento lexical marcado com o traço *location* passe a ser classificado como uma organização institucional.

Note-se aqui a restrição (*maj:~*, isto é, a palavra não deverá começar por maiúscula) associada ao lema *governo* no contexto à direita: esta restrição é devida às actuais directivas do Segundo HAREM que estipulam que os nomes de organizações devem sempre começar por maiúscula. Assim, numa expressão como *o governo de Lisboa*, só a palavra *Lisboa* será marcada como uma organização.

É de salientar que estas regras se podem aplicar a sequências de categorias ambíguas. Relembramos que, na arquitectura que definimos (v. figura 15.3), a aplicação das regras locais para EM se faz depois da aplicação de um primeiro módulo de desambiguação, mais específico, e que a maior parte das ambiguidades categoriais ainda não foram inteiramente resolvidas. Estas gramáticas locais procedem, pois, a uma desambiguação suplementar, na medida em que, se houver emparelhamento com uma regra, serão seleccionadas as categorias com que essas regras emparelharem.

A seguir a estas regras locais será, então, aplicado o módulo de desambiguação (misto, isto é, combinando HMM e regras), que permitirá resolver as ambiguidades restantes (Tagging2, na figura 15.3).

15.4 Últimas fases de processamento das EM

A possibilidade de utilizar expressões regulares contextuais (ver secção acima) permite atingir um certo grau de generalização na formulação e representação desse contexto. No entanto, para um grau de generalização ainda maior, utilizam-se outras técnicas que necessitam de uma análise linguística mais complexa: no caso da tarefa de REM, são necessárias a análise do texto em sintagmas nucleares (particionamento, em inglês *chunking*) e o cálculo de relações sintácticas (dependências) entre potenciais entidades mencionadas e outros constituintes da frase (Brun e Hagège, 2004). As últimas fases do processamento das EM consistem, assim, no aproveitamento dos módulos de particionamento, de construção de dependências e de propagação de traços. Todos estes módulos se encontram integrados no XIP (v. figura 15.3). É deles que falaremos a seguir.

15.4.1 Particionamento

O módulo de particionamento do XIP permite fazer uma análise sintáctica preliminar do texto, construindo para cada frase uma sequência de sintagmas nucleares. As entidades mencionadas reconhecidas nas fases anteriores (por codificação lexical ou através das gramáticas locais) recebem, de um modo geral, a etiqueta *NOUN*, o que permite que elas se integrem naturalmente nas regras gerais de construção dos sintagmas nucleares (em inglês, *chunks*) nominais da gramática. Por outras palavras, EM delimitadas (simples ou complexas) vão ser núcleos nominais de sintagmas nominais nucleares.

15.4.2 Dependências

A construção das dependências permite exprimir relações sintácticas, como as de sujeito, objecto directo, etc., entre os diversos constituintes das frases. Além das relações gramaticais clássicas, construídas pela gramática, cria-se para as EM uma nova relação unária (*NE*), cujo argumento consiste na EM reconhecida. A esta relação são associados os traços que permitem classificar o tipo de EM.

Ilustramos a análise em sintagmas nucleares e em dependências com o exemplo `ex:xip:joaninha`.

(15.3) Joaninha Sampaio vivia na Lourinhã

A figura 15.4 apresenta a interface gráfica do XIP, na qual se mostra a análise em sintagmas nucleares (NP, VF, PP). A primeira parte da saída representa a árvore de sintagmas nucleares (i.e., os sintagmas nucleares estão todos ligados a um nó *TOP*).

A sequência *Joaninha Sampaio* foi correctamente delimitada e etiquetada como um único nome (*NOUN*), tendo sido também classificada como um nome de pessoa (*NE_INDIVIDUAL_PEOPLE*(Joaninha Sampaio)). Este nome complexo constitui o núcleo do sintagma nuclear nominal (NP) *A Joaninha Sampaio*. Verificamos ainda que foram construídas várias dependências gramaticais, como, por exemplo, a relação de sujeito (*SUBJ*) entre *viver* e *Joaninha Sampaio*.

As dependências unárias *NE** correspondem às EM que foram encontradas. O nome genérico da relação unária é *NE*, a que se juntam, ligados por caracteres de sublinhado, os traços associados à dependência e que, neste caso, consistem na classificação destas EM.

Estas dependências são criadas graças aos traços que foram previamente associados aos nomes *Joaninha Sampaio* e *Lourinhã*.

As outras relações correspondem à relações sintácticas calculadas entre diversos constituintes. É o aproveitamento destas relações que permite generalizar contextos para o cálculo de novas EM (ver ponto seguinte).

15.4.3 Generalizando o contexto para classificar EM

Um dos problemas com que se defronta a tarefa de REM consiste na resolução dos casos de metonímia, aspecto que, naturalmente, também está contemplado nas directivas do Segundo HAREM. Um exemplo típico desta situação consiste no uso de nomes de países para se referir ou ao conjunto dos habitantes/o povo ou às instituições da respectiva organização política. Muitos destes fenómenos de transferência metonímica seguem padrões regulares (cf. avaliação conjunta de detecção de metonímia de SemEval 2007 (Markert e Nissim, 2007)). Contudo, a fim de capturar estes fenómenos de metonímia, é necessário levar em consideração um contexto relativamente alargado. As regras contextuais das gramáticas locais não seriam, então, o formalismo mais adequado ou mais eficiente para representar esse tipo de contexto.

Para ilustrar o que dizemos, tomemos o exemplo (15.4).

(15.4) Portugal ratificou o tratado.

O nome *Portugal* está marcado no léxico como nome de país. A priori, no fim da cadeia de processamento, esta unidade lexical seria classificada como uma EM de tipo LOCAL. No entanto, neste contexto sintáctico, isto é, como sujeito de um verbo como *ratificou*, *Portugal* não designa aqui o espaço geográfico de um país mas sim a entidade que representa a sua organização político-administrativa (ou eventualmente, mas noutros contextos, um grupo de pessoas). Ora, é graças às dependências previamente calculadas, nomeadamente à relação de sujeito (ou de agente da passiva) entre *Portugal* e o verbo *ratificar* que é possível que o sistema corrija a classificação cega do nome de país como LOCAL e, tal como se indica nas directivas gerais de classificação do Segundo HAREM, passe a tratá-lo, pois, como ORGANIZACAO.

A grande vantagem de levar a cabo esta correcção ao nível das dependências resulta da possibilidade de, com apenas uma regra (ver adiante) dar conta de casos como: *ex:xip:naoratificou*, *ex:xip:foiratificado*, *ex:xip:queratificou*, etc. pois, para todos estes casos, a palavra *Portugal* é analisada como estando numa relação sujeito ou agente com o verbo *ratificar*.

(15.5) Portugal ainda não ratificou o tratado

(15.6) O tratado foi ratificado por Portugal

(15.7) Portugal, que ratificou este tratado

Eis um exemplo de uma regra que permite transformar uma EM de tipo geográfico em EM de tipo organização quando ela é sujeito (ou agente) de *ratificar*.

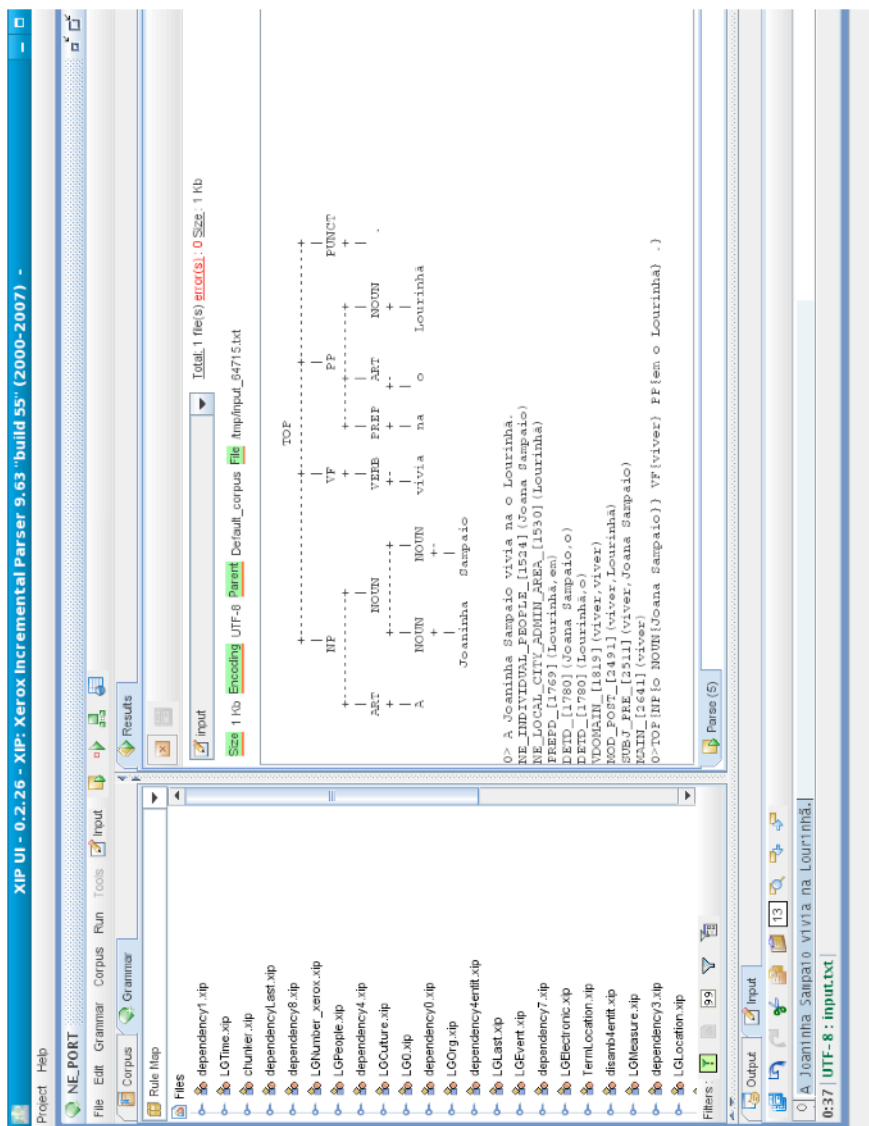


Figura 15.4: Exemplo de análise das EM, dos sintagmas nucleares das dependências pelo XIP

```

if ( ^NE[local:+,admin_area:+] (#1) &
    ( SUBJ(?[lemma:ratarificar],#1)
      | AGENT(?[lemma:ratarificar],#1)
    )
)
NE[features=\~,org=+,administration=+] (#1)

```

Para esta avaliação conjunta do HAREM, limitámo-nos a uma primeira abordagem, ainda muito preliminar, do tratamento da metonímia. A maior parte dos casos de metonímia previstos nas directivas não estão ainda contemplados no nosso sistema, seja por falta de tempo seja por discordância com algumas das escolhas da organização. Com efeito, consideramos que o emprego metonímico de uma EM deveria passar a ser explicitamente anotado, de modo a ter em paralelo a classificação semântica básica (literal ou por defeito) e a classificação (ou interpretação semântica) que resulta do fenómeno de transferência metonímica de acordo com Brun et al. (2007).

Note-se que as nossas regras de correcção para dar conta dos empregos metonímicos foram feitas manualmente. No entanto, tendo a possibilidade de dispor de um corpus anotado, a aquisição automática deste tipo de regra pode ser explorada.

15.4.4 Propagação

A propagação é um mecanismo que permite conservar a informação sobre EM previamente calculadas e propagar essa informação ao resto da análise de um texto. Este processo parte do pressuposto de que, num mesmo documento, novas EM são introduzidas num contexto suficientemente rico para que possam ser classificadas de forma não ambígua; no entanto, muitas vezes, essas EM são retomadas nesse mesmo texto mas já sem se apresentarem nessa distribuição característica. Trata-se, tipicamente, do caso de nomes de pessoas mas pode também acontecer com outro tipo de entidades.

O XIP oferece, além das operações habituais para o processamento linguístico, uma linguagem dedicada (em inglês, *scripting language*), que pode ser usada para algumas tarefas simples (contagens de ocorrências ou de relações por exemplo). A propagação é realizada graças a esta linguagem dedicada e processa-se em dois passos:

1. Marcação da EM
Se houver uma EM atestada (marcada no léxico ou calculada por regras locais), então a totalidade ou parte da sequência correspondente a esta EM será marcada graças às variáveis particulares geridas pelo XIP.
2. Restituição da EM
Se, no decorrer da análise, for encontrada uma sequência de caracteres previamente marcada, pode-se então associar-lhe uma operação qualquer (neste caso a construção de uma dependência NE unária associada a esta sequência).

O exemplo (15.8) (extraído da colecção do Segundo HAREM) ilustra o processo de propagação:

(15.8) Um capitão norueguês chamado *Trygve Petersen* conduziu o Mira de novo a Portugal... <frase intermédia>... *Petersen* não trazia carga nenhuma.

Nem *Trygve* nem *Petersen* estavam previamente codificados no léxico. No entanto, graças ao contexto, no primeiro caso, e ao mecanismo de propagação, no segundo caso, é possível obter os seguintes nomes de entidades:

```
NE_INDIVIDUAL_PEOPLE(Trygve Petersen)
NE_INDIVIDUAL_PEOPLE_PROPAG(Petersen)
```

Num primeiro passo, *Trygve Petersen* é classificado como uma entidade de tipo `PESSOA`, por se tratar de um complemento de *chamado*. A cadeia *Trygve Petersen* recebe a categoria `NOUN`, à qual vão estar associados os traços `individual:+` e `people:+`.

Num segundo passo, graças à linguagem dedicada, integrada no XIP, variáveis numéricas, que são indexadas aos lemas *Trygve* e *Petersen* (representados respectivamente por `PERSON##2` e `PERSON##3`) e que foram anteriormente inicializadas a 0, vão passar a ter o valor 1.

```
Script:
noun#1[people,individual]{?* , noun#2[title:~,location:~,org:~,initial:~,maj:+] ,
  ?* , noun#3[last,title:~,location:~,initial:~,maj:+] } |
if (NE[people](#1) )
{ PERSON##2=1; PERSON##3=1; }
```

Esta instrução faz com que cada lema correspondente a sub-sequências de um nome complexo de tipo `PESSOA` (traços `people:+` e `individual:+`) venha a ser marcado de igual modo graças a estas variáveis. Neste exemplo concreto, os lemmas *Trygve* e *Petersen* são, pois, associados às variáveis `PERSON`, cujo valor será igual a 1 (as variáveis de lemas são inicializadas a 0 por defeito).

A regra seguinte exemplifica a propagação destes traços:

```
DependencyRules:
| noun#1[toutmaj:~,maj:+] |
  if ( PERSON##1:1 & ~NE[people](#1) )
NE[people=+,individual=+,propag=+](#1)
```

Esta regra determina que: (i) se a um nome começando por maiúscula (traço `maj:+`) mas não totalmente em maiúsculas (traço `toutmaj` negado) for associada uma variável `PERSON` cujo valor é 1; e (ii) se este nome ainda não estiver marcado como sendo `NE` de tipo `PERSON` deve, então, ser criada uma relação unária `NE` com traços `people:+` e `individual:+` à qual se acrescenta a informação de que se trata do resultado de uma propagação (`propag=+`) desses traços.

Assim, na medida em que a variável `PERSON` foi inicializada a 1 durante o processamento da primeira frase, a ocorrência isolada de *Petersen* na segunda frase do exemplo será também considerada como um nome de pessoa e será marcada com uma dependência unária que classifica os nomes de pessoas individuais.

As variáveis são associadas de forma consistente aos lemas durante toda a fase de análise (i.e. todo o documento que estiver a ser processado). No entanto, é possível durante o processamento reinicializar o valor destas variáveis. Tipicamente, no caso do Segundo `HAREM`, consideramos que não é desejável propagar variáveis além do âmbito de um documento.

Uma vez que os documentos estão delimitados por balizas `</DOC>`, produzimos a regra seguinte:

```
Script:
| #1[lemma:"</DOC>"]; #1[lemma:"</doc>"] |
{ CleanAllLemmas; }
```

Esta regra determina que, cada vez que o segmento `</DOC>` ou `</doc>` for encontrado num texto, todas as variáveis de lemas sejam reinicializadas a 0.

A propagação é um mecanismo extremamente poderoso, que permite aumentar a abrangência (em inglês, *recall*) de um sistema de REM. No entanto, pode também ter efeitos perversos, sobretudo se a entidade inicial não tiver sido correctamente classificada.

Para o Segundo HAREM, utilizámos o mecanismo de propagação de uma maneira limitada e exclusivamente para os nomes de pessoa⁵. Deixamos para uma próxima avaliação conjunta do HAREM a complexa tarefa de desenvolver e estender a outros casos as regras de propagação, contando levar, então, em linha de conta o grau de confiança associado à classificação de uma EM reconhecida na propagação dos seus traços a outras EM.

15.5 Resultados e perspectivas

Os resultados que obtivemos no Segundo HAREM foram bastante encorajadores. Nesta primeira participação numa avaliação conjunta de REM, fomos o terceiro sistema em termos de medida F (obtendo até os melhores resultados em medida F para o cenário selectivo 2), considerando que não levámos em conta algumas das categorias previstas para o Segundo HAREM como `ABSTRACCAO` e `COISA`. Investimos bastante energia na tarefa de reconhecimento das expressões temporais e, tanto para o TEMPO clássico, como para a tarefa específica do TEMPO de acordo com a proposta por nós apresentada (Hagège et al., 2008), obtivemos os melhores resultados.

No nosso trabalho, favorecemos claramente a precisão em relação à abrangência. Achamos, no entanto, que a abrangência poderá ainda ser bastante aumentada e com relativa facilidade uma vez que, por falta de tempo, deixámos por codificar muito léxico (já identificado) e apenas utilizámos de forma incipiente e exploratória o mecanismo de propagação de traços.

A experiência desta nossa participação no HAREM mostrou-nos que negligenciámos alguns aspectos, como o da formatação dos resultados finais, tarefa que nos ocupou muito mais tempo do que esperávamos e que terá prejudicado duas das nossas corridas. Certamente levaremos isto em consideração numa próxima avaliação conjunta.

Temos consciência de que muito ficou ainda por fazer, tanto do ponto de vista da codificação do léxico, como do ponto de vista do desenvolvimento das regras das gramáticas locais e de propagação de traços. Estamos confiantes de que o sistema ainda tem bastante margem para melhoramento.

⁵ Ainda não fizemos a avaliação quantitativa do benefício da propagação para o sistema mas, a título indicativo, foram detetados 1700 nomes de entidades da categoria `PESSOA` na colecção do Segundo HAREM graças a este mecanismo sobre um total de 10017 entidades com esta categoria.