Npro: um novo recurso para o processamento computacional do Português

J. Baptista[1,2], F. Batista[1,3], N. Mamede[1,4] e C. Mota[1]

[1] L2F – Laboratório de Sistemas de Língua Falada – INESC ID Lisboa
[2] Universidade do Algarve, Faculdade de Ciências Humanas e Sociais
[3] ISCTE – Instituto de Ciências do Trabalho e da Empresa
[4] Instituto Superior Técnico – Universidade técnica de Lisboa

1. Introdução

A construção de recursos linguístico-informáticos é uma actividade fundamental para o processamento computacional das línguas naturais (PLN). Para várias aplicações, tais como: recuperação de informação, extracção de texto, tradução automática, análise sintáctica automática, reconhecimento de entidades mencionadas e, por exemplo, para a identificação do actor dentro do discurso no contexto de um sistema de diálogo (Fourour, 2002; Traboulsi 2004; McDonald, 1996; Friburger, 2004) pode ser importante que os sistemas de PLN sejam capazes de fazer uma correcta identificação de nomes próprios (*Npr*) e das sequências que constituem a denominação de uma entidade.

A identificação de nomes próprios num texto não etiquetado é convencionalmente realizada ou com recurso a abordagens linguísticas (Fourour, 2002) ou com recurso a abordagens estatísticas (McDonald, 1993; Yarowsky, 1994; Yarowsky, 2000). Em cada um dos casos, um dicionário que contenha informação acerca deste tipo de palavras pode constituir um recurso de PLN importante para o processamento automático de corpora (Piton e Maurel, 2004).

De entre as várias subclasses de nomes próprios destacam-se os antropónimos, os topónimos (e.g. localidades, países, regiões geográficas, hidrónimos, acidentes geográficos, etc.; Piton e Maurel, 2004), os nomes de instituições (Moura, 2000), de marcas comerciais e outros. Cada uma destas subclasses apresenta problemas de representação e reconhecimento particulares, tais como a sua estrutura interna, forma e ambiguidade lexical.

Ao contrário do que acontece com o léxico comum, não é habitual constituírem-se listagens de nomes próprios, talvez porque prevaleça a intuição generalizada de que constituem um conjunto potencialmente infinito, o que efectivamente não é o caso para algumas destas classes, em particular para certas subclasses de antropónimos. Neste artigo, apresentamos o *Npro*, um dicionário electrónico de nomes próprios, que constitui um novo recurso para o processamento computacional do Português. O foco da nossa investigação centra-se apenas no problema do reconhecimento automático de antropónimos, deixando o tratamento das restantes classes para outra ocasião.

XXI Encontro Nacional da Associação Portuguesa de Linguística, Lisboa, APL, 2006, pp. 137-151

Este artigo encontra-se organizado da seguinte forma: a próxima secção sintetiza as principais características linguísticas que podem ser associadas aos antropónimos. A secção 3 apresenta os passos seguidos na construção do dicionário. Na secção 4, a utilidade deste recurso será demonstrada em duas tarefas/aplicações distintas: por um lado, numa tarefa de reconhecimento de entidades mencionadas e, por outro lado, na identificação e capitalização de nomes próprios, de forma a melhorar a qualidade da saída de um reconhecedor de fala. A última secção, apresenta as conclusões e perspectivas para trabalho futuro.

2. Antropónimos

Esta classe de nomes apresenta um comportamento sintáctico e semântico particular (Molino, 1982, Gary-Prieur, 1991; Anderson, 2004), sobretudo no que se refere ao seu valor referencial (Leroy, 2004). É possível estruturá-la em dois grandes subconjuntos, com propriedades morfossintácticas distintas: os *nomes de baptismo* (*nb*) e os *nomes de família* (*nf*). Os *nomes de baptismo* são marcados quanto ao género, podem excepcionalmente apresentar variação em número e, nalguns casos, apresentam inclusive diminutivos formados quer a partir do nome por supressão de sílabas: *Bela* (=*Anabela*), *Tó* (=*António*), quer derivados por afixação de sufixos diminutivos: *Carlinhos* (=*Carlos*), quer por combinação de vários processos: *Zezinho* (= *José*). Os *nomes de família* não são, regra geral, marcados quanto ao género, e não admitem a formação de diminutivos.

A informação quanto ao género (eventualmente quanto ao número) é relevante para o processamento sintáctico, por exemplo para a resolução de anáforas (pronominais e de outros tipos), razão por que deverá ser levada em consideração na construção de recursos que visem o processamento automático deste tipo de unidades lexicais.

Por outro lado, a nomeação de uma pessoa faz-se frequentemente combinando vários nomes próprios (eventualmente com conectores particulares: *e*, *de*, *do*, *da*, *dos*, *das*, *d'* e o hífen) seguindo regras combinatórias culturalmente determinadas. A correcta identificação dessas sequências como uma unidade – designando uma entidade – é necessária, por exemplo, para vários aspectos da análise sintáctica automática.

Finalmente, uma percentagem considerável de nomes próprios (cerca de 43% dos nomes próprios estudados), sobretudo os nomes de família, é constituída por palavras ambíguas com formas do léxico comum. Esta ambiguidade origina uma dificuldade suplementar para o reconhecimento das sequências que constituem a denominação de uma entidade, e poderá ser (pelo menos parcialmente) resolvida por meio de gramáticas locais de desambiguação (Yarowsky, 2000; Silberztein, 1993).

3. Construção do dicionário

O dicionário foi construído semi-automaticamente usando duas abordagens: (i) selecção a partir de listagens de nomes próprios completos constantes de listas oficiais

de instituições; (ii) selecção a partir de listagens extraídas automaticamente a partir de corpora.

A construção do dicionário seguindo a primeira abordagem utilizou sobretudo dois tipos de fontes: (a) uma listagem de formas obtida a partir dos assinantes dos antigos TLP (Trancoso, 1995); (b) uma listagem de nomes completos de pessoas, obtida a partir de listas de alunos inscritos na universidade (IST e UAlg).

As formas da primeira lista incluem várias indicações suplementares, nomeadamente quanto à sua frequência na lista telefónica, ou consoante se trate do primeiro ou do último nome. Procedeu-se a uma selecção manual de cerca de 4.500 antropónimos a partir das 9.660 formas mais frequentes (com frequência igual ou superior a 10), adicionando-se e validando-se por introspecção a informação quanto ao seu emprego como *nb* ou *nf*. Para os primeiros acrescentou-se ainda a informação relativa ao género.

Quanto às listagens de nomes completos (aproximadamente 8.100), após algumas correcções ortográficas e delimitação manual da fronteira entre os *nb* e os *nf* de cada nome, foi possível utilizar a informação posicional para determinar de forma inequívoca o emprego como *nb* ou *nf* de cada forma. À semelhança do que se fez para a lista dos TLP, acrescentou-se aos *nb* a informação relativa ao género bem como alguma informação adicional (e.g. nomes de origem estrangeira: *Yuri*, variantes ortográficas: Mello). O dicionário assim constituído integra aproximadamente 6.200 nomes próprios distintos. Além do processo anteriormente descrito, todas as palavras ambíguas do dicionário foram marcadas, utilizando para isso um dicionário contendo palavras do léxico comum (Ranchod, 1999).

Número de entradas:	6.173	
da Lista 1:	4.533	73,5 %
da Lista 2: (não na Lista 1)	1.640	26,5 %
(Já presentes na Lista 1)	1.756	
nb: nomes de baptismo (apenas):	1.870	30,3 %
nf: nomes de família (apenas):	4.200	68,0 %
Nomes que são *nb* e *nf*:	103	1,7 %
Ambíguos:	2.629	42,6 %
nb ambíguos	468	7,0 %
nf ambíguos	2.196	35,0 %
Nomes *nb* e *nf* ambíguos	35	0,6 %

Tabela 1 – Constituição do dicionário *Npro*

Seguindo a segunda abordagem, aplicou-se uma gramática local de reconhecimento de entidades mencionadas ao CETEMPúblico (Rocha e Santos, 2000), obtendo-se uma listagem com cerca de 20.000 entidades que são (ou incluem) potenciais nomes próprios. A identificação e formalização desses nomes próprios está neste momento em curso.

4. Aplicações

4.1. Reconhecimento de entidades mencionadas

De Setembro de 2004 a Outubro de 2005 decorreu a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas (HAREM) organizada pela Linguateca[1]. Participaram 10 sistemas, num total de 18 resultados (alguns dos quais não oficiais, no sentido em que foram entregues no prazo de uma semana depois de terminado o prazo oficial para envio dos mesmos). Foram caracterizadas três tarefas para avaliação: identificação, classificação semântica e classificação morfológica, que foram executadas numa colecção de textos designados "Colecção Harem". Para efeitos de avaliação, só uma parte das anotações introduzidas pelos sistemas nessa colecção é que foi comparada com as que foram previamente introduzidas manualmente numa sub-colecção designada "Colecção Dourada".

Foi nosso objectivo, com a experiência que a seguir descrevemos, quantificar a utilidade do recurso que desenvolvemos nas actividades de identificação e classificação semântica de entidades com a categoria PESSOA de acordo com o definido no âmbito desta avaliação, comparando os resultados com os obtidos pelos participantes. De notar que, tendo em conta as informações morfológicas constantes no dicionário, este estaria à partida vocacionado para a tarefa de classificação morfológica. No entanto, para já ainda não fizemos essa avaliação.

4.1.1. Caracterização da tarefa

A tarefa vulgarmente designada em inglês por Named Entity Recognition consiste da identificação e classificação de nomes próprios, ou seja, delimitação de uma sequência de maiúsculas num dado texto e atribuição de uma etiqueta semântica (inicialmente PERSON, ORGANIZATION e LOCATION) que classifica o objecto (entidade) referido por essa sequência. Esta tarefa foi definida como uma sub-tarefa de Extracção de Informação na sexta edição da série de conferências de avaliação MUC *Message Understanding Conference.*

Em termos do Harem, uma *entidade mencionada* foi caracterizada como sendo: «Expressões que (1) tenham pelo menos uma palavra em maiúscula. E que só contenham um número muito pequeno (determinado lexicalmente) de palavras em minúscula; (2) tenham pelo menos um algarismo. Nesse caso, as unidades de medida ou monetárias associadas tb podem vir em minúsculas.»[2].

Na tarefa de identificação era avaliada a correcta delimitação dessas entidades. A tarefa de classificação consistia em (1) atribuir uma de 9 etiquetas possíveis e (2) sub-classificar a entidade atribuindo um de vários tipos num total de 40 (ver Tabela 2). A

[1] Para informações mais detalhadas consultar http://www.linguateca.pt, Avaliação Conjunta, Harém.

[2] http://poloxldb.linguateca.pt/ harem.php?l=introducao

classificação morfológica valorizava a atribuição de atributos flexionais (género e número) a todas as entidades excepto as classificadas com TEMPO e VALOR.

Etiqueta	Tipo
PESSOA	INDIVIDUAL GRUPOIND CARGO GRUPOCARGO MEMBRO GRUPOMEMBRO
ORGANIZACAO	INSTITUICAO ADMINISTRATIVO EMPRESA SUB
TEMPO	HORA DATA PERIODO CICLICO
LOCAL	GEOGRAFICO VIRTUAL CORREIO ADMINISTRATIVO ALARGAGO
OBRA	PRODUTO REPRODUZIDA ARTE PUBLICACAO
ACONTECIMENTO	EVENTO EFEMERIDE ORGANIZADO
ABSTRACCAO	DISCIPLINA ESCOLA MARCA NOME OBRA PLANO IDEIA ESTADO
COISA	CLASSE SUBSTANCIA OBJECTO
VALOR	MOEDA QUANTIDADE CLASSIFICACAO
VARIADO	Sem subclassificação.

Tabela 2 – Etiquetas e tipos usados na classificação semântica do Harem

As informações pretendidas deveriam ser introduzidas no texto usando anotações expressas em formato SGML, tal como ilustrado no seguinte fragmento da "Colecção Harem":

> Em representação entregue ontem ao **<ORGANIZACAO TIPO="ADMINISTRACAO" MORF="M,S">** Ministério da Justiça**</ORGANIZACAO>**, elas foram acusadas de terem formado cartel para aumentar em até **<VALOR TIPO="QUANTIDADE">**200%**</VALOR>** a taxa cobrada pelos seus serviços.
> A representação foi encaminhada pelo comerciante paulista **<PESSOA TIPO="INDIVIDUAL" MORF="M,S">**Ronaldo Cheguri de Almeida**</PESSOA>**, em nome de cerca de **<VALOR TIPO="QUANTIDADE">**300**</VALOR>** donos de bares e restaurantes de **<LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">**São Paulo**</ /LOCAL>**.
>
> (*Colecção Dourada Versao 3.1, IdDoc: HAREM-284-04226*)

4.1.2. Experiência preliminar

Na nossa experiência, limitámo-nos a reconhecer entidades PESSOA sem fazer subclassificação, embora a utilização do *Npro* seja sobretudo adequada para identificar entidades do tipo INDIVIDUAL e GRUPOIND, que são aquelas que normalmente envolvem antropónimos. Para tal, usámos regras simples que combinam apenas as informações do Npro. Essas regras são de dois tipos:

- combinação dos nomes de baptismo e/ou família com palavras em maiúsculas, em que pelo menos um do elementos da sequência é um nome de baptismo ou um nome de família; caso ambos ocorram, os nomes de baptismos precedem sempre os de família. Por exemplo: "<N+nb> <CAP>*" ou "<CAP>* <N+nf>";

– coordenação dos nomes de baptismo e/ou família com palavras em maiúsculas. Por exemplo, "<N+nb>, <CAP> (e+ou) <CAP>" ou "<CAP> <CAP> e <N+nb>".

As regras descritas em grafos do sistema Nooj (Silberztein, 2004) são usadas em combinação com duas versões alternativas do dicionário: uma completa (*Npro*) e outra onde apenas as entradas não ambíguas com léxico comum permaneceram (*Npro-NA*).

Convém chamar a atenção para algumas directivas do Harem que afectam a utilização só por si do *Npro* nas tarefas avaliadas. No que respeita à delimitação das entidades, é necessário incluir palavras que não são nomes próprios. Por exemplo, no caso da frase:

A tia Maria ofereceu-me um livro

o grau de parentesco deve fazer parte da entidade (directiva 2.2.1). Por outro lado, sequências de nomes próprios que não são antropónimos também deverão ser identificadas e classificadas com a categoria PESSOA, tipo CARGO, como é o caso dos cargos que se encontrem escritos em maiúsculas no texto (directiva 2.3.).

Além disso, uma terceira fonte de problemas advém do facto da classificação ter por objectivo atribuir uma categoria de acordo com o papel que a entidade desempenha no texto em questão. Ou seja, tomando como exemplo o nome próprio Antonieta que *a priori* é um antropónimo, caso ele se encontre numa frase como:

A minha tia chama-se Antonieta

deverá ser etiquetado com a categoria ABSTRACCAO, tipo NOME, em vez de PESSOA (directiva 2.2.5).

4.1.3. Avaliação

As regras simples que representámos em grafos foram aplicadas pelos sistema NooJ à Colecção Dourada V3.1; a avaliação dos resultados produzidos foi feita utilizando os programas de avaliação desenvolvidos pela Linguateca no âmbito do HAREM[3], que calculam várias medidas de desempenho, entre as quais precisão (P), abrangência (A) e medida-F (m-F). Essas medidas permitem comparar os resultados que obtivemos com os resultados oficiais obtidos pelos participantes no Harem[4].

O corpus de avaliação é constituído por 129 documentos num total de 89.241 átomos. Cada documento está classificado quanto à sua origem em termos de variante do português (Portugal, Brasil, Angola, Moçambique, Cabo Verde, Macau, Timor-Leste e Índia) e quanto a género (jornalístico, literário, expositivo, político, web, transcrito de entrevistas, correio electrónico e técnico). Das 3.851 entidades mencionadas existentes

[3] Trata-se da versão que estava disponível *online* em Setembro de 2005. Posteriormente foram disponibilizadas versões mais recentes dos programas, mas, por questões de tempo, não nos foi possível reproduzir os resultados com esses novos programas. Chamamos também a atenção para o facto de os resultados dos participantes apresentados nas tabelas Tabela 3, Tabela 4 e Tabela 5 são relativos a 14 de Setembro de 2005.

[4] Para mais detalhes sobre os programas avaliadores e as medidas produzidas, consultar as páginas sobre o Harem em http://www.linguateca.pt.

no texto, 1.073 estão categorizadas como podendo ser PESSOA (dessas, 856 têm tipo INDIVIDUAL e 10 GRUPOIND, estando as restantes distribuídas por CARGO (79), GRUPOCARGO (19), MEMBRO (10) e GRUPOMEMBRO (137). Convém referir que algumas das entidades marcadas como PESSOA tinham mais do que um tipo associado na Colecção Dourada, bem como podiam em alternativa ser de outra classe.

Tarefa de identificação. A avaliação da tarefa de identificação é feita através da análise de dois cenários: o *total*, no qual são consideradas todas as entidades presentes na Colecção Dourada e também todas as categorias possíveis, e o *selectivo*, onde são apenas tidas em conta as entidades com a categoria que os sistemas se propõem identificar. No nosso caso, pretendemos avaliar, tal como já referido, exclusivamente a atribuição da categoria PESSOA (optámos por atribuir sempre o tipo INDIVIDUAL).

Resultado	**P (%)**	**A (%)**	**m-F**
doha	58.75	72.72	**64.99**
marraquexe	66.65	53.78	59.53
riad	64.97	53.69	58.79
bahrein	64.27	51.55	57.21
mascate	**69.35**	35.49	46.95
manama	**69.35**	35.49	46.95
sana	59.23	35.8	44.63
Npro-NA	40.38	37.5	38.89
Npro	26.71	49.83	34.78
argel	24.62	26.52	25.53
abudhabi	18.82	25.45	21.64
qatar	0	0	0

Tabela 3 – Avaliação da tarefa de identificação no cenário selectivo

Dado que só nos interessa avaliar as entidades PESSOA, os resultados constantes na Tabela 3 são referentes apenas ao cenário selectivo. De acordo com a referida tabela, podemos concluir que as entradas do *Npro* completo participam em cerca de 50% das entidades PESSOA presentes na Colecção Dourada. Retirando as entradas ambíguas, a abrangência decresce para 37,5% (de notar que o Npro sem entradas ambíguas tem cerca de metade do tamanho). Apesar desse decréscimo, a precisão aumenta cerca de 14%, o que acaba por compensar em 4 pontos em termos de medida-F.

No entanto, convém salientar que cerca de 28% das entidades PESSOA não são do tipo INDIVIDUAL e que portanto o *Npro* não cobre essas entidades. Por outro lado, uma inspecção às restantes entidades não identificadas permite concluir que se tivermos em conta apenas os diferentes nomes próprios que participam na sua constituição (cerca de 200 nomes diferentes), apenas 7 são evidentemente nomes portugueses: Carlinhos,

Moniquinha, Otelo, Parrela, Salinas, Tonico e Zeca, enquanto que os restantes ou são brasileiros, como Adílson e Romário, ou estrangeiros como Beethoven, Rodríguez, e Steinberger.

Tarefa de classificação semântica. A avaliação da classificação semântica é também feita segundo os cenários *total* e *selectivo*, sendo ambos vistos segundo dois novos eixos, *absoluto* e *relativo*. Assim, no cenário *total/absoluto* são consideradas todas as entidades presentes na Colecção Dourada e todas as entidades identificadas pelo sistema, enquanto no cenário *total/relativo* são consideradas todas as etiquetas da Colecção Dourada e apenas as entidades bem (ou parcialmente bem) identificadas pelo sistema; nos cenários *selectivo/absoluto* e *selectivo/relativo* são apenas avaliadas as entidades cujas categorias o sistema pretende etiquetar, sendo que no primeiro caso (absoluto) são tidas todas as entidades identificadas pelo sistema, enquanto que no segundo é apenas avaliada a correcta atribuição de categorias (e tipos) de entidades bem (ou parcialmente bem) identificadas pelo sistema.

Dentro de cada um dos cenários são produzidas várias medidas, segundo quatro perspectivas diferentes: por categorias (tem apenas em conta a categoria atribuída), por tipos (avalia os tipos de categorias bem atribuídas), combinada (combina as pontuações das duas anteriores) e plana (combina a categoria e o tipo numa só etiqueta). Neste caso, interessa-nos apenas avaliar a classificação por categorias no cenário selectivo tanto na vertente absoluta como relativa.

Saída	**P (%)**	**A (%)**	**m-F**
damasco	61.07	**75.23**	**0.6742**
marraquexe	69.46	55.61	0.6176
bahrein	68.07	55.82	0.6134
casablanca	68.47	54.92	0.6095
rabat	72.88	37.29	0.4934
sana	72.88	37.29	0.4934
dakar	63.69	38.5	0.4799
Npro-NA	45.84	37.09	0.4100
Npro	31.79	49.99	0.3887
teerão	27.09	28.98	0.2801
bagdad	20.78	28.09	0.2389
oman	0	0	0

Saída	**P (%)**	**A (%)**	**m-F**
bagdad	**95.23**	**94.91**	**95.07**
argel	94.33	93.86	0.941
gaza	92.72	94.07	0.9339
damasco	93.18	92.91	0.9305
bengazi	93.18	92.91	0.9305
sana	91.33	90.74	0.9103
Npro-NA	90.61	90.14	0.9038
Npro	90.36	90.19	0.9027
eritreia	89.39	88.79	0.8909
asmara	87.57	87.83	0.877
riad	87.03	83.74	0.8535
oman	0	0	0

Tabela 4 e Tabela 5 – Avaliação da tarefa de classificação semântica no cenário selectivo nos eixos absoluto e relativo, respectivamente.

Dado que do ponto de vista da identificação, as regras simples em combinação com o *Npro* não são suficientes para obter um bom desempenho na delimitação das entidades, isso vai-se reflectir em termos de precisão e abrangência ao nível da classificação como mostram as tabelas 4 e 5, embora os resultados sejam ligeiramente

melhores. De qualquer maneira em termos da classificação das entidades total e parcialmente bem identificadas tanto a precisão como a abrangência estão na ordem dos 90% o que significa que as entidades que foram bem delimitadas são efectivamente entidades PESSOA do tipo INDIVIDUAL

4.1.4. Discussão

No cenário em que todas as entidades e etiquetas da Colecção Dourada são tidas em conta, o estabelecimento de regras simples que façam uso das informações contidas no dicionário *Npro* é insuficiente na tarefa de identificação, não só do ponto de vista da Precisão (entre cerca de 26 e 40%) como da Abrangência (entre cerca de 37 e 50%), distanciando-se da maioria dos participantes no Harém.

No entanto, num cenário em que são apenas tidas em conta as etiquetas consideradas e as EMs que foram bem identificadas, a classificação com base nesse processo simples obtém resultados ao nível dos participantes.

Embora na maioria dos casos não haja uma diferença notória entre a utilização do *Npro* e do *Npro* só com entradas não ambíguas, o facto de este último ter consistentemente melhores resultados sugere que não há necessidade de manter as entradas ambíguas no dicionário.

Naturalmente que a experiência levada a cabo foi apenas preliminar e é nosso objectivo vir a combinar as informações do *Npro* (tanto da versão completa como da versão com entradas não ambíguas) com as regras de um dos sistemas participantes, o STENCIL que é baseado apenas em regras que descrevem evidências internas e externas no sentido de (McDonald, 1996), e verificar de que modo é que os resultados são alterados. Por outro lado, iremos utilizar o *Npro* para identificar novos contextos a serem integrados nas gramáticas.

4.2. Capitalização

Foram realizadas várias experiências de capitalização de nomes próprios, de forma a avaliar a utilidade do recurso *Npro*. Foram consideradas duas subtarefas: na *subtarefa 1* apenas se avalia a capitalização de antropónimos; na *subtarefa 2* avalia-se a capitalização de todos os nomes próprios, independentemente do seu valor antroponímico.

4.2.1. Corpus

A experiência consistiu em aplicar o dicionário a um *corpus* de notícias televisivas, produzido automaticamente por um sistema de reconhecimento de fala (Caseiro e Trancoso, 2002) e corrigido manualmente. Nesta versão corrigida, cada nome próprio (antropónimos e outros) existente no texto foi capitalizado e precedido do sinal '^'. Pelo contrário, o resultado produzido automaticamente pelo sistema não oferece pistas, tais como pontuação e palavras capitalizadas, que podem ser de outra forma utilizadas para identificar nomes próprios. Este aspecto resulta numa menor qualidade de leitura dos

textos assim produzidos. A tarefa de capitalização aqui descrita e avaliada visa justamente melhorar a saída do reconhecedor.

O *corpus* contém aproximadamente 500.000 palavras e foi dividido em dois *subcorpora*: (i) um *corpus* de treino com aproximadamente 498.000 palavras (27.513 diferentes); e (ii) um *corpus* de avaliação com 49.306 palavras (7.513 diferentes). Por sua vez, no *corpus* de avaliação, os antropónimos foram posteriormente distinguidos dos outros nomes próprios, substituindo manualmente as marcas '^' por '#'. A Figura 1 apresenta um excerto do *corpus* de avaliação.

```
Jornal Dois, a informação com #Manuel #Menezes.
Boa noite.
A Comissão Europeia decidiu pedir a ^Portugal que explique
alguns aspectos do traçado da auto-estrada do ^Algarve. Em
causa está o projectado troco da -A dois, que atravessa a zona
de protecção especial de ^Castro ^Verde, e que poderá
constituir uma violação da directiva comunitária sobre
protecção das aves selvagens.
```

Figura 1 – Excerto do *corpus* de avaliação. Os antropónimos e os outros nomes próprios encontram-se destacados a negrito.

O *corpus* de avaliação contém 3.001 nomes próprios, dos quais 1.101 são antropónimos.

4.2.2. Informação probabilística

A fim de determinar em que medida é que o dicionário poderia melhorar a tarefa de capitalização dos nomes próprios, quando comparado com o simples uso da informação associada a uma dada palavra quanto à probabilidade de esta surgir escrita com maiúscula, produzimos a partir do *corpus* de treino um dicionário probabilístico. A cada palavra deste *corpus* e dependendo do número de vezes que ocorreu no *corpus* como maiúscula e como minúscula, foi associada uma probabilidade de ocorrer como maiúscula. Atribuiu-se a etiqueta CAP a uma palavra, se ela ocorreu capitalizada mais do que 50%,das vezes, caso contrário atribuiu-se a etiqueta MIN. A lista de formas do *corpus* de treino, com esta informação probabilística, constitui o nosso dicionário probabilístico. Posteriormente verificámos que 15% das formas deste dicionário foram classificadas com a etiqueta CAP. A lista de palavras deste dicionário cobrem cerca de 80% das formas do *corpus* de avaliação. Posteriormente, a informação obtida para o dicionário probabilístico também foi integrada nas entradas do *Npro*, de forma a enriquecer este recurso. A Tabela 6 resume a constituição do dicionário probabilístico.

Número total de entradas	25,528[5]	
CAP:	3,822	15.0 %
MIN	21,706	85.0 %
Em comum com o *Npro*	1,373	
CAP	899	65.5 %
MIN	474	34.5 %

Tabela 6 – Constituição do dicionário probabilístico.

4.2.3. Gramáticas locais

Foram construídas uma série de gramáticas locais que combinam informação posicional e lexical. Estas gramáticas, que consistem num conjunto de autómatos, podem ser também vistas como regras que permitem reconhecer sequências de palavras (e eventuais conectores) candidatas ao estatuto de nomes próprios e que, consequentemente, deveriam ser escritos com maiúscula inicial.

Para este artigo iremos considerar apenas os resultados produzidos por 10 destas regras. Cada regra pode ser utilizada separadamente, contudo, o propósito final é o de construir eventualmente uma única gramática que integre todas estas regras e que, possivelmente, fornecerá melhores resultados. A Figura 2 ilustra dois exemplos dessas regras. Podemos verificar que a *regra 8* identifica sequências de pelo menos um *nome de baptismo*: <N+Npr+nb>, seguido de pelo menos um *nome de família*: <N+Npr+nf>. A *regra 9* é menos específica do que a regra 8 e identifica qualquer sequência de nomes próprios (e qualquer eventual conector), sem ter em conta as etiquetas de nome de baptismo ou de família. Em ambos os grafos, o grafo auxiliar *sw* (=*stopword*) representa os eventuais conectores – sem considerar a conjunção *e*.

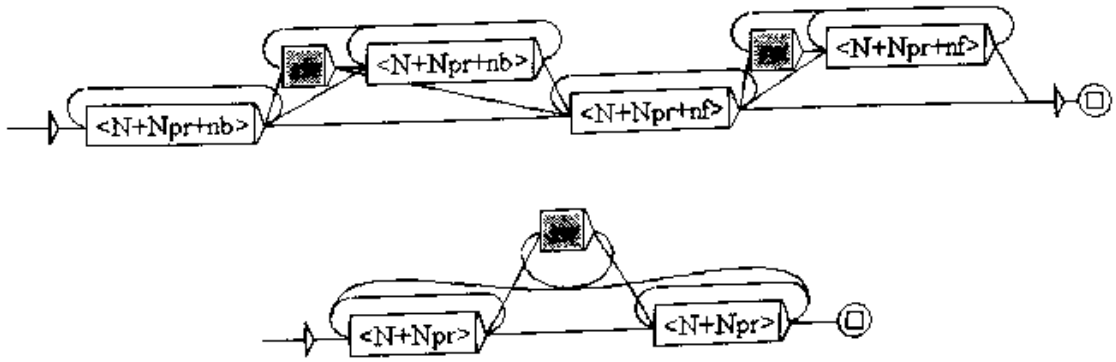

Figura 2 – Regras 8 e 9 – dois autómatos finitos que reconhecem sequências de palavras, candidatas a nomes próprios

[5] A diferença entre o número de entradas do dicionário probabilístico e o número de formas diferentes do *corpus* de treino prende-se com os diferentes critérios de segmentação utilizados. Por exemplo, as formas com hífen foram consideradas uma só no dicionário probabilístico.

4.2.4. Resultados e discussão

De forma a encontrar a melhor forma de identificação de antropónimos e nomes próprios em geral, várias experiências foram realizadas. Nestas experiências, foram comparados métodos diferentes de capitalização, nomeadamente: a) usando apenas a informação do dicionário *Npro* (experiência 1); b) usando apenas informação probabilística com base no uso de maiúsculas no *corpus* de treino (experiência 2); c) usando o dicionário *Npro* juntamente com informação contextual (posicional) (experiências 3, 4 e 5); d) combinando os diferentes métodos (experiências 6 e 7). Os resultados encontram-se na Tabela 7:

As primeiras duas experiências ajudam a definir a *baseline* no que diz respeito a precisão e cobertura, para os dois métodos de capitalização principais, nomeadamente o uso separado do *Npro* por oposição ao uso separado do dicionário probabilístico. Para a subtarefa 1, os resultados são aproximadamente equivalentes, mesmo que o dicionário probabilístico mostre uma medida F ligeiramente melhor. Em termos gerais, para a tarefa de identificação de antropónimos, deve-se esperar pelo menos uma precisão de cerca de 30% e uma cobertura de 80% para ambos os métodos. Contudo, tal como seria de esperar para a subtarefa 2, o dicionário probabilístico consegue resultados muito melhores, uma vez que a cobertura do *Npro* está limitada apenas aos antropónimos.

	Subtarefa 1 (antropónimos)			Subtarefa 2 (nomes próprios)			
Experiência	Precisão	Cobertura	Medida F	Precisão	Cobertura	Medida F	Max F
1. <N+Npr>	32,6%	79,3%	0,462	60,1%	53,6%	0,566	0,509
2. <WORD+CAP>	30,3%	79,9%	0,439	72,7%	70,4%	0,715	0,544
3. regra 8	86,6%	51,0%	0,641	97,3%	20,5%	0,338	0,443
4. regra 9	70,5%	65,5%	0,679	92,2%	31,4%	0,468	0,554
5. regras 1-9	63,6%	74,8%	0,687	93,6%	40,4%	0,564	0,619
6. regra 10	58,4%	69,5%	0,634	89,2%	39,0%	0,542	0,585
7. regras 1-10	30,5%	87,0%	0,451	71,8%	75,2%	0,734	0,559

Tabela 7 – Resultados obtidos de 7 experiências de identificação de antropónimos e nomes próprios. Max F = Medida F usando as Medidas F das duas subtarefas

As experiências 3, 4 e 5 ilustram o uso combinado de *Npro* com várias regras contextuais. Nestas experiências destacam-se os resultados obtidos isoladamente pelas regras 8 e 9, que obtiveram o melhor desempenho. A experiência 5 mostra o resultado da combinação de todas as regras.

Fica claro que, para a subtarefa 1, o uso combinado de regras contextuais e o *Npro*, conduzem a melhores resultados, quando comparados com as duas primeiras experiências, que servem de *baseline*. As regras 1 a 7, que aqui não apresentamos, são altamente específicas, conduzindo a uma elevada precisão mas a uma cobertura muito baixa. De uma forma geral, cada regra parece capturar diferentes fenómenos combinatórios, mas a experiência 5, que é a conjunção de todas as regras, consegue a melhor medida-F.

Para a subtarefa 2, estas experiências conseguiram os melhores resultados em termos de precisão, com o custo de uma redução da cobertura. Contudo, a medida-F da

experiência 5 é apenas ligeiramente menor do que a da experiência 1. Ainda assim, dever-se-á ter em conta que a informação presente no *Npro* apenas contempla antropónimos, sendo por isso esperada uma baixa cobertura na subtarefa 2.

A experiência 6 combina os três métodos, usando a regra 10, que está representada na Figura 3.

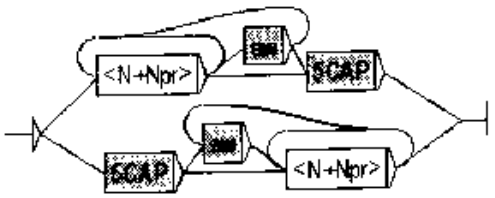

Figura 3 – Regra 10 – Um autómato finito que reconhece sequências de (até 5) palavras que não estão incluídas no *Npro*, mas foram marcadas como CAP no dicionário probabilístico, seguidas ou precedidas de antropónimos

Nas duas subtarefas 1 e 2, a adição de informação probabilística não parece melhorar significativamente os resultados, quando comparada com as experiências 3 a 5. Contudo, a adição dessa informação melhora muito a precisão obtida nas experiências 1 e 2 para a tarefa 1, ao custo de apresentar uma cobertura mais baixa.

A experiência 7, que combina as experiências 2, 5 e 6, consegue o melhor desempenho na cobertura das duas subtarefas 1 e 2 (87% e 75%, respectivamente), ainda que a precisão mantenha valores semelhantes aos atingidos nas experiências 1 e 2. Esta experiência também consegue a melhor medida-F para a subtarefa 2.

A última coluna apresenta o cálculo da medida-F relativa às duas subtarefas, de acordo com a fórmula seguinte:

$$MaxF = \frac{2 * F_1 * F_2}{F_1 + F_2}$$, sendo F_x o valor da medida F obtido para a tarefa x.

Desses valores podemos concluir que a utilização do *Npro* em combinação com regras contextuais é uma boa escolha para identificar antropónimos, enquanto também se consegue uma boa precisão na tarefa de detecção de nomes próprios. O valor da cobertura obtido na subtarefa 2 constitui um valor de base, o qual pode ser subsequentemente melhorado através de outros métodos. A introdução de informação estatística tornou possível obter melhores resultados para a subtarefa 3, mas sob pena de empobrecer os resultados da subtarefa 1. Tal parece constituir uma confirmação das principais motivações que nos levaram à construção do dicionário *Npro* e que justificaram a realização destas experiências.

5. Conclusões e trabalho futuro

Em Processamento da Linguagem Natural, a construção de bases de dados lexicais para tratamento de nomes próprios não tem sido a abordagem mais usada para lidar com estes tipos de elementos linguísticos. Uma das razões que possivelmente justificam tal atitude tem a ver com a noção generalizada de que o conjunto dos nomes próprios é

potencialmente infinito, o que talvez não seja exacto para todas as classes de nomes próprios, e muito provavelmente não o é no que diz respeito aos antropónimos (especialmente no caso dos nomes de baptismo).

Este artigo descreveu a metodologia empregue na construção de um dicionário electrónico de nomes próprios. No contexto do artigo, foi avaliada a utilidade deste recurso em duas tarefas particulares: o reconhecimento de entidades mencionadas; e a capitalização de antropónimos e nomes próprios em geral.

Relativamente à tarefa de reconhecimento de entidades mencionadas, considerando apenas a natureza do recurso aqui apresentado, ou seja, num cenário mais simples em que são apenas tidas em conta as etiquetas consideradas e as EMs que foram bem identificadas, a classificação obtém resultados ao nível dos atingidos pelos restantes participantes do Harem.

A tarefa de capitalização teve como motivação principal a melhoria da qualidade da saída de um reconhecedor de fala, na qual as palavras, incluindo nomes próprios, surgem em minúsculas. A utilidade desta nova ferramenta poderia ser estendida para outros cenários, tais como a correcção automática de textos em que a informação sobre nomes próprios não existe ou foi eliminada e deve ser recuperada. Comparámos e combinámos várias abordagens, nomeadamente, a utilização de um dicionário probabilístico, construído com base na forma como as palavras foram encontradas escritas num *corpus* de treino, e o uso de regras contextuais, baseadas na informação constante do *Npro*. Os valores obtidos nos resultados das experiências aqui apresentadas mostram que a utilização do *Npro* permite melhorar os resultados. Apesar disso, espera-se ser possível obter ainda melhores resultados através da aplicação de técnicas de aprendizagem automática, tais como listas de decisão. A metodologia descrita em [5,6] demonstrou que se podem atingir melhores resultados para problemas de natureza semelhante. Esperamos ainda que a aplicação de tais técnicas possam contribuir para a melhoria dos dados linguísticos deste recurso.

De futuro, prevemos expandir o dicionário *Npro*, de forma a contemplar também outros tipos de nomes próprios (topónimos, hidrónimos, etc.), e a integrar tanto formas simples como formas compostas.

6. Agradecimentos

A colaboração sempre pronta do Nuno Seco foi um factor que muito contribuiu para a rápida automatização da execução dos programas de avaliação de reconhecimento de entidades mencionadas. Este trabalho foi parcialmente financiado por uma bolsa de doutoramento atribuída pela Fundação para a Ciência e Tecnologia (referência SFRH/BD/3237/2000).

7. Referências

Aït-Mokhtar, S. (1998) *L'analyse Présyntaxique en une seule étape* (PhD. Thesis), Univ. Blaise Pascal, Fevereiro 1998.

Anderson, J. (2005) *On the Grammar of names*. (to appear in *Language* 2004/05)

Caseiro, D. e Trancoso, I. (2002) Using dynamic wfst composition for recognizing broadcast news, em *Proceedings of ICSLP '2002*. Denver, Colorado, EUA.

Fourour, N., Morin, E, Daille, B. (2002) Incremental recognition and referential categorization of French proper names. In: *Proceedings of the Third International Conference on Language Resources and Evaluation* (LREC 2002), vol. III, pp. 1068-1074.

Friburger, N., Maurel, D. (2004) Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, Vol. 313(1): pp. 93-104

Gary-Prieur, Marie-Noël (ed.) (1991) Syntaxe et sémantique des noms propres. *Langue Française* 92. Larousse: Paris.

Leroy, S. (2004) *Le nom propre en français*. Paris: Ophrys.

McDonald, David D (1996). Internal and External Evidence in the Identification and Semantic Categorization of Proper Names, in Boguraev, B., Pustejowsky, J. (eds.): *Corpus Processing for Lexical Acquisition*, The MIT Press, Cambridge, MA & London, England, pp. 21-39.

Meinedo, H. *et al*, (2003) AUDIMUS.media: A Broadcast News Speech Recognition System for the European Portuguese Language. *Computational Processing of the Portuguese Language* – Proc. of the 6th Intl. Workshop, PROPOR 2003, Lecture Notes in Artificial Inteligence, pp 9-17, Faro, Portugal.

Molino, Jean (ed.) (1982) Le nom propre. *Langue Française* 66. Paris: Larousse

Moura, P. (2000) *Dicionário electrónico de siglas e acrónimos* (Tese de Mestrado), Faculdade de Letras da Universidade de Lisboa.

Piton, O., Maurel, D. (2004) Les noms propres géographiques et le dictionnaire Prolintex. C. Muller, J. Royauté e M. Silberztein (eds.) *INTEX Pour la linguistique et le traitement automatique des langues*. Cahiers MSH Ledoux 1, pp. 53-76. Presses Universitaires de Franche-Comté: Besançon.

Ranchhod, E., Mota, C., Baptista, J. (1999) A Computational Lexicon of Portuguese for Auto-matic Text Parsing. SIGLEX-99: Standardizing Lexical Resources, pp. 74-80. ACL/Maryland Univ., Maryland

Rocha P. e Santos D. (2000) "CETEMPúblico: Um *corpus* de grandes dimensões de linguagem jornalística portuguesa", in M. G. Volpe Nunes (ed.), *Actas do V Encontro para o processamento computacional da lingua portuguesa escrita e falada* (PROPOR'2000), pp. 131-140.

Silberztein, M. (1993) *Dictionnaires électroniques et analyse automatique de texts. Le systeme INTEX*. Masson, Paris.

Silberztein, M. (2004) *NooJ: A Cooperative, Object-Oriented Architecture for NLP*. In INTEX pour la Linguistique et le traitement automatique des langues. Cahiers de la MSH Ledoux, Presses Universitaires de Franche-Comté.

Traboulsi, H. (2004) *A Local Grammar for Proper Names* (MPh. Thesis). Surrey University.

Trancoso, I (1995) "The ONOMASTICA Inter-Language Pronunciation Lexicon" Presented by I. M. Trancoso, on behalf of the ONOMASTICA Consortium. *Proceedings of EUROSPEECH'95 – 4th European Conference on Speech Communication and Technology* – Madrid, Spain, Setembro 1995.

Yarowsky, D. (1994) Decision lists for lexical ambiguity resolution: Application to accent restora-tion in Spanish and French. *Proceedings of ACL '94*, pp. 88-95.

Yarowsky, D. (2000) Hierarchical Decision Lists for Word Sence Disambiguation. *Computers and the Humanities*, 34 (1-2): pp. 179-186.