Análise Morfossintáctica para Português Europeu e Galego: Problemas, Soluções e Avaliação

Marcos Garcia
Universidade de Santiago de Compostela
marcosgg@gmail.com

Pablo Gamallo
Universidade de Santiago de Compostela
pablo.gamallo@usc.es

Resumo

As diferentes tarefas de análise morfossintáctica têm muita importância para posteriores níveis do processamento da linguagem natural. Por isso, estes processos devem ser realizados com ferramentas que garantam bons desempenhos em relação à cobertura, precisão e robustez na análise. FreeLing é uma *suite* com licença GPL desenvolvida pelo Grupo TALP da Universitat Politècnica de Catalunya. Este *software* contém —entre outros— módulos de tokenização, segmentação de orações, reconhecimento de entidades e anotação morfossintáctica. Com o fim de obtermos ferramentas que nos sirvam de base para a análise sintáctica, bem como para disponibilizar *software* livre para o processamento de superfície de Português Europeu e Galego, adaptámos FreeLing para estas variedades. A primeira delas foi desenvolvida com ajuda de recursos linguísticos disponíveis *on-line*, enquanto os ficheiros do Galego tiveram como base a versão anterior de FreeLing (criados pelo Seminario de Lingüística Informática da Universidade de Vigo), que já realizava a análise desta língua. O presente trabalho descreve os principais aspectos da adaptação das ferramentas, com ênfase nos problemas encontrados e nas soluções adoptadas em cada caso. Além disso, são apresentados os resultados de avaliação do módulo PoS-tagger.

1 *Introdução*

Os diferentes processos que compõem a análise morfossintáctica constituem uma etapa com enorme importância no processamento da linguagem natural. Tarefas como a recuperação de informação, a análise sintáctica ou a síntese de voz, por exemplo, precisam de um processamento prévio que seja capaz de segmentar orações, reconhecer tokens e inferir os seus lemas ou atribuir uma categoria morfossintáctica (PoS) a cada um deles.

Variedades linguísticas como o Galego (GA) ou o Português Europeu (PE), que apresentam uma flexão verbal complexa, formas homógrafas, ou contracções de tokens ambíguas, precisam ser tratadas com ferramentas desenvolvidas especificamente para elas (Graña, Barcala e Vilares, 2002; Branco e Silva, 2004). Assim, tanto a criação como a adaptação de recursos para estas variedades devem ter em conta os problemas específicos que apresentam, com o fim de evitar erros de análise em etapas posteriores do processamento.

Tanto o Português Europeu como o Galego dispõem de recursos de análise morfossintáctica de alta precisão (Bick, 2000; Marques e Lopes, 2001; Ribeiro, Oliveira e Trancoso, 2003; Branco e Silva, 2004, por exemplo) e (Graña, Barcala e Vilares, 2002), respectivamente, mas até ao momento desconhecíamos *software* com licença livre (salvo os anteriores ficheiros de treino para Galego de FreeLing, desenvolvidos na Universidade de Vigo (Carreras et al., 2004)) para este fim. Neste sentido, a adaptação de FreeLing para Português, bem como a melhoria da versão galega, implicam a disponibilização de ferramentas livres para a análise morfossintáctica destas variedades.

O presente trabalho tem como objectivos principais (i) mostrar os procedimentos de adaptação de FreeLing para Português Europeu e Galego, indicando os casos problemáticos e as soluções adoptadas em cada um deles e (ii) realizar uma avaliação do módulo PoS-tagger nas duas versões desenvolvidas.

A adaptação realizou-se fundamentalmente com recursos linguísticos de livre distribuição disponíveis *on-line*, uma vez que actualmente, este *software* permite ser adaptado de maneira rápida e relativamente simples. As licenças do próprio FreeLing e de alguns dos dicionários e *corpora* utilizados permitem treinar e adaptar os módulos de análise sem um consumo excessivo de recursos, possibilitando ao mesmo tempo correcções, modificações e ampliações posteriores.

A *suite* disponibiliza, desde a versão 2.1, os ficheiros de treino apresentados no presente trabalho. Uma vez que FreeLing contém módulos para diferentes níveis de análise, é preciso referir que até ao momento o desenvolvimento centrou-

se nos seguintes tópicos: (i) Tokenização, (ii) segmentação de orações, (iii) lematização (com base em léxicos) e (iv) PoS-tagging. Outros módulos foram também treinados, mas o seu desenvolvimento ainda está em processo, pelo que não serão apresentados pormenorizadamente: Estes são os (v) reconhecedores de numerais, datas e quantidades e (vii) identificadores de expressões multipalavra de classe fechada.

Actualmente, os módulos de análise do Português Europeu e do Galego incluídos em FreeLing têm desempenhos próximos do estado-da-arte, quer em comparação com outro *software* para as mesmas variedades, quer em relação às avaliações de sistemas para outras línguas.

Para além desta secção introdutória, este artigo apresenta aqueles módulos de FreeLing que foram adaptados (Secção 2), os recursos utilizados (Secção 3), os resultados de diferentes avaliações dos PoS-tagger (Secção 4) bem como as conclusões finais (Secção 5).

2 Módulos de FreeLing

Nesta secção será realizada uma apresentação dos módulos de FreeLing que foram adaptados para o Português Europeu e Galego, destacando os principais problemas encontrados durante o seu desenvolvimento; como foi dito, o *software* fornece mais serviços dos aqui descritos, sendo que não todos eles foram adaptados para as duas variedades referidas.

2.1 Tokenizador

O primeiro módulo adaptado foi o tokenizador, que converte, através de regras, um texto plano num vector de palavras. Uma vez que é uma tarefa relativamente simples, o principal aspecto a ter em conta tem a ver com a ordem de aplicação entre o próprio tokenizador e o PoS-tagger, a qual influencia o modo como as contracções ambíguas são tratadas (Graña, Barcala e Vilares, 2002; Branco e Silva, 2003). Assim, formas como `desse` (em PE, ou `dese` em GA), que pode ser um verbo ou uma contracção de preposição e demonstrativo, provocam uma circularidade entre o etiquetador morfossintáctico e o tokenizador. Este último não poderá decidir se separar `desse` em `de+esse` sem conhecer a sua categoria morfossintáctica, mas o PoS-tagger não pode ser aplicado sob um texto não tokenizado. As soluções que FreeLing permite adoptar nestes casos têm a ver com a análise morfossintáctica (dicionário, afixos e PoS-tagger), pelo que neste primeiro processo o tokenizador não separará as contracções. O *output* do tokenizador, portanto, manterá ainda a ambiguidade nas formas contraídas.

Outro aspecto a considerar é a interacção com o segmentador de orações. Na ordem de aplicação proposta (tokenizador > segmentador), o primeiro dos módulos deve reconhecer as abreviaturas (identificando o ponto como parte da abreviatura: `Sr.` e não `Sr .`) para evitar as ambiguidades mais comuns no *input* do segmentador de orações. A diferença entre as configurações do tokenizador de PE e de Galego está, portanto, na lista de abreviaturas.

2.2 Segmentador de Orações

O segmentador de orações recebe o *output* do tokenizador e devolve uma nova oração cada vez que detecta uma fronteira. As línguas românicas não apresentam muitas diferenças nos marcadores ortográficos, pelo que a adaptação para Português Europeu e Galego não apresentou grandes dificuldades. Uma vez que o tokenizador eliminou já as ambiguidades mais frequentes entre os pontos finais e os pontos de abreviação, o segmentador não precisa tratar especificamente estes casos.

Entre as duas variedades tratadas, as diferenças de segmentação não são significativas, e dizem respeito a especificidades ortográficas, como a utilização dos pontos de interrogação e exclamação para abrir orações (não utilizados em PE, mas facultativos em GA).

2.3 Analisador Morfológico

O módulo de análise morfológica de FreeLing é na verdade um conjunto de módulos que realizam tarefas como a identificação de numerais e de datas, o reconhecimento de entidades nomeadas e de expressões multipalavra, bem como a pesquisa no dicionário e o tratamento dos afixos.

Até ao momento, o maior esforço foi dedicado à adaptação do módulo de pesquisa em dicionário, tranformando o formato dos léxicos disponíveis e criando regras de lematização de afixos verbais e nominais.

Este módulo compõe-se de dois sub-módulos, que actuam em paralelo: Um deles procura no dicionário todas as possibilidadess de análise de cada um dos tokens encontrados no *input*, enquanto o outro aplica as regras de lematização de afixos, que permitem que tokens que não estejam no dicionário sejam analisados pelo sistema.

O dicionário de Português Europeu contém mais de 908.000 entradas (mais de 1.257.000 formas, tendo em conta as entradas com várias análises), enquanto o de Galego supera as 428.000 (577.000 formas). Note-se que o sistema não possui um lematizador próprio: O lema de cada token é procurado no léxico. Isto implica a necessidade de léxicos amplos, com o fim de atingir

níveis altos de precisão nesta tarefa. A avaliação deste processo realizou-se dividindo o número de lemas correctamente atribuídos pelo número total de lemas de um *corpus* de teste. Em PE, o sistema foi avaliado sobre um *corpus* de 50.000 tokens, obtendo uma precisão de 98,583%. Em Galego, o *corpus* de teste foi de 6.200 tokens e o resultado de 99,41%.

O sub-módulo de tratamento de afixos permite criar regras de lematização de formas com prefixos e sufixos. Deste modo não é preciso incluir no dicionário todas as possibilidades de combinação de formas verbais com clíticos, nem diminutivos, aumentativos, advérbios acabados em `mente`, ou formas prefixadas.

A pesquisa em dicionário e o tratamento de afixos permitem que na execução do etiquetador morfossintáctico sejam tratadas as contracções, não divididas pelo tokenizador. O funcionamento é o seguinte:

As contracções não ambíguas (por exemplo `do`: preposição `de` + artigo `o`), estão presentes no dicionário com o formato `do de+o SPS00+DA`, pelo que estas formas serão divididas em dois tokens na saída final.[1] Os casos de ambiguidade (`desse/dese, destes, pelo/polo`, etc.), porém, podem ser tratados —fundamentalmente— de duas maneiras: Incluindo-as no dicionário, ou acrescentando regras de lematização destas formas ao sub-módulo de tratamento de afixos.

As duas soluções referidas permitem evitar a circularidade referida no Ponto 2.1, uma vez que todas as alternativas existentes no módulo de análise morfológica são avaliadas pelo desambiguador morfossintáctico e pelo PoS-tagger que, se for preciso, realizará uma retokenização. A única diferença que encontrámos entre as duas propostas de tratamento diz respeito ao formato do *output*: Enquanto as entradas ambíguas do dicionário são apresentadas sem serem divididas (`desse de+esse SPS00+DD0MS0`), as regras de lematização permitem separar a saída: `de de SPS00 / esse esse DD0MS0` (o formato de saída é `token lema TAG`).[2] Para manter a coerência com o formato das contracções não ambíguas (que são divididas), decidimos utilizar a segunda das opções. Esta solução é similar às adoptadas em (Branco e Silva, 2003) ou em (Graña, Barcala e Vilares, 2002), deixando a decisão de realizar *split* aos módulos de análise morfossintáctica, e não ao tokenizador. Note-se, contudo, que em (Graña, Barcala e Vilares, 2002) a desambiguação de locuções é realizada no mesmo processo.

Dentro do conjunto de módulos de análise morfológica, a seguinte ferramenta de FreeLing (o desambiguador morfossintáctico) assigna uma probabilidade para cada um dos possíveis tags de cada token e, com base na análise das terminações, tenta saber que tags são possíveis nas formas desconhecidas. Esta análise é realizada com base no treino em *corpus*.

Para além destas ferramentas, a análise morfológica contém, como foi dito, outros módulos que realizam tarefas diversas. Até ao momento, com base nas configurações para outras línguas românicas, estão a ser adaptados para Português Europeu e Galego os módulos que têm a ver com o reconhecimento de numerais, de pontuação, de datas, de quantidades, de nomes próprios e de expressões multipalavra. Este último compõe-se de uma lista de expressões multipalavra de classe fechada, extraída dos *corpora* utilizados para o treino do PoS-tagger, e ampliada de modo manual.

2.4 PoS-Tagger

FreeLing permite utilizar dois métodos de anotação morfossintáctica: O modelo probabilístico com base nos Hidden Markov Models (Brants, 2000), e um método híbrido que permite combinar informação estatística com informação gerada manualmente (Padró, 1998).

Com este último modelo, um utilizador pode criar um conjunto de restrições para tratar certos tipos de erros produzidos pela abordagem estatística. Estas restrições estabelecem probabilidades de atribuição de uma etiqueta a um token em função do contexto, das propriedades do token, etc.

O modelo híbrido é —apesar de ligeiramente mais lento— de maior precisão do que os HMM, mas requer a criação manual das restrições, pelo que no treino realizado utilizámos o método puramente estatístico.

Para além do *corpus*, um dos aspectos de maior incidência no desempenho de um PoS-tagger deriva do seu tagset; se este for muito complexo, a informação fornecida pelo etiquetador é maior, mas a sua precisão será mais baixa. Pelo contrário, se o tagset for reduzido, a sua precisão

[1] Pode entender-se que a análise de `do` contém ambiguidade relativa à categoria de `o`, que além de artigo, poderia ser pronome nos casos em que o núcleo da frase nominal não está preenchido: `O homem do qual ele falou` (pelo que a entrada do dicionário incluiria SPS00+DA/PD). No caso que nos ocupa, unicamente nos referimos à ambiguidade em que uma única forma pode ser analisada como contraída ou não: `deste` como contracção de preposição+demonstrativo ou como verbo.

[2] A análise das contracções ambíguas não deve ser inserida como a primeira das opções do dicionário, já que desta maneira FreeLing selecciona-a sem aplicar o PoS-tagger, ignorando as restantes hipóteses e, não resolvendo, portanto, a circularidade.

será mais elevada, mas pode correr-se o risco de que a informação obtida não seja suficiente para os objectivos do PoS-tagger.

Um dos propósitos da adaptação de FreeLing para Português Europeu e Galego foi o de utilizá-lo como etiquetador morfossintáctico base para um analisador de dependências multilíngue (Gamallo e González, 2009). Tendo em conta que este sistema requer informação morfológica (género, número, pessoa, tempo e modo verbal, etc.), e com base nos tagsets utilizados nas outras línguas de FreeLing, decidimos usar as recomendações propostas pelo Grupo EAGLES (Leach e Wilson, 1996). O tagset definido para Português Europeu contém 255 tags, enquanto em Galego empregaram-se 277 etiquetas.[3]

O tagset utilizado contém informação morfossintáctica detalhada, mas não todos estes dados são utilizados propriamente pelo PoS-tagger; este usa unicamente os dois primeiros elementos da etiqueta, sendo os restantes extraídos do léxico. O primeiro elemento do tag (`D`, Determinante, `P`, Pronome, `N`, Nome, etc.) indica a categoria morfossintáctica; o segundo (`D` Demonstrativo, `P`, Possessivo, etc., variando em função do primeiro elemento) refere a subclasse da categoria à que pertence. O resto de entradas das etiquetas variam em função da categoria principal, e englobam aspectos como o possuidor (singular ou plural) dos possessivos o grau (aumentativo ou diminutivo) dos nomes, o caso dos pronomes ou informação sobre modo, pessoa e número dos verbos.

3 Recursos Utilizados

Uma vez que alguns dos recursos linguísticos utilizados na adaptação de FreeLing para as duas variedades em causa são de livre distribuição, nesta secção apresentaremos sucintamente a sua origem, bem como as modificações feitas durante o desenvolvimento.

O processo de aprendizagem do módulo de anotação morfossintáctica precisa de um *corpus* etiquetado de alta qualidade. Com este fim, para Português Europeu o *corpus* utilizado foi criado a partir do Bosque 8.0 (Bosque. Uma floresta integralmente revista por linguistas,), o único que conhecemos disponível livremente com informação morfossintáctica detalhada. Este *corpus* contém aproximadamente os 1.000 primeiros extractos dos *corpora* CETEMPúblico e do CETEMFolha (este último, não empregue, do Português do Brasil), o que faz um total de mais de 138.000 tokens.

O Bosque foi anotado automaticamente e, posteriormente, revisto de forma manual por linguistas. Sendo um *corpus* com informação sintáctica, esta foi eliminada na conversão para o formato requerido por FreeLing.

Para além do *corpus*, o treino de FreeLing requer um dicionário de formas flexionadas (que contenha os lemas e tags possíveis para cada token). Para este fim, utilizou-se o léxico de formas simples LABEL-LEX (SW) (Eleutério et al., 2003), que contém mais de 1.257.000 formas, geradas a partir de perto de 120.000 lemas.

Os dois recursos referidos têm características diferentes em relação à anotação morfossintáctica, pelo que foi preciso fazer uma conversão de cada um deles para o formato utilizado. Neste processo surgiram algumas incoerências que implicaram tomadas de decisão do ponto de vista linguístico. Assim, tags como "pron-indp: pronome independente" (utilizado no Bosque), não tinham correspondente directo nas etiquetas do léxico, pelo que não foi possível uma transferência automática entre os formatos. A conversão destes casos teve de ser incluída no *script* de transformação de maneira individual, e decidir em cada ocorrência dos tokens no *corpus* qual era o tag que lhes correspondia de acordo com o dicionário.

Para além das inconsistências no nível morfossintáctico, a conversão do *corpus* e do dicionário apresenta problemas em termos de lematização nominal. Assim, enquanto o Bosque lematiza os adjectivos superlativos como elementos não derivados (`altíssimo` é o lema de `altíssimo/a/(s)`), o LABEL-LEX (SW) opta por decisões mais coerentes do ponto de vista teórico: `altíssimo/a/(s)` > `alto`. De modo similar, outras diferenças notórias entre as lematizações do *corpus* e do dicionário foram as relacionadas com derivação semântica: O LABEL-LEX (SW) considera que formas como `mulher` (nome) e `melhor` (adjectivo) derivam de `homem` e `bom`, respectivamente, enquanto o Bosque atribui `mulher` e `melhor` como lemas dos mesmos tokens.

Nestes casos, a solução adoptada foi de modo geral aquela que tivesse como base processos morfológicos e não semânticos. Assim, no primeiro dos casos, optou-se por considerar os adjectivos superlativos como derivados do adjectivo simples; no segundo exemplo, a decisão tomada foi consistente com a lematização utilizada no Bosque, que diferencia as formas que não apresentam uma relação morfológica directa.

[3] A estas quantidades são acrescentados 24 tags de símbolos de pontuação (atribuídos não pelo PoS-tagger, mas pelo identificador de pontuação). Na Tabela 4 pode ver-se o formato do tagset. Note-se que, para manter a compatibilidade com outros tagsets de FreeLing, os elementos que não sejam precisos em PE e GA serão marcados com um <0> (veja-se como exemplo os valores semânticos dos nomes, que ocupam os elementos 5 e 6).

No tratamento das locuções e dos nomes próprios compostos por mais de um elemento, o Bosque apresenta algumas inconsistências que, em função dos nossos objectivos, não permitiram avaliar o desempenho do reconhecedor de expressões multipalavra com precisão. Assim, enquanto `Conselho de Administração da PEC-Alimentação` é dividido em quatro tokens, expressões como `director-clínico do Hospital Prisional S. João de Deus` é anotado no *corpus* como um único token/lema. A solução adoptada nestes casos foi a seguinte: Os elementos marcados como locuções no Bosque foram extraídos automaticamente, e adicionados à lista de expressões multipalavra depois de serem revistos manualmente. Nos *corpora* de treino e avaliação, porém, estas formas foram divididas em tokens individuais, pelo que o treino e a avaliação foram realizadas sem locuções.

O desenvolvimento da versão para Galego foi realizado com recursos de diferente procedência, pelo que o processo de adaptação foi diferente ao realizado para PE.

O *corpus* utilizado para treinar o módulo PoS-tagger para Galego foi criado no projecto Gari-Coter (Barcala et al., 2007), e contém mais de 237.000 tokens; o *corpus*, gerado a partir de notícias jornalísticas, é especializado em economia. Uma vez que a anotação morfossintáctica deste recurso seguiu os *standards* do Grupo EAGLES, as únicas adaptações precisas para o treino foram relativas à homogeneização de alguns elementos dos tags: Modificou-se, por exemplo, o caso de alguns pronomes (de nominativo para oblíquo), ou o género dos determinantes indefinidos (de comum para neutro), de acordo com o tagset definido e utilizado no dicionário.

Em relação ao léxico, o trabalho partiu do dicionário criado pelo Seminario de Lingüística Informática da Universidade de Vigo, que fazia parte de anteriores versões de FreeLing. O dicionário foi ampliado com entradas verbais e nominais extraídas de *corpora* e flexionadas automaticamente com ajuda de flexionadores e conjugadores gerados pela equipa de trabalho. Actualmente, o dicionário contém mais de 428.000 entradas, o que se corresponde com mais de 577.000 formas se tivermos em conta aquelas entradas com mais de uma análise.

Assim mesmo, as regras de lematização verbal e nominal (do sub-módulo de tratamento de afixos), foram também ampliadas a partir das publicadas nas anteriores versões de FreeLing.

4 Avaliação do PoS-tagger

Para subsequentes tarefas de PLN, a precisão da anotação morfossintáctica é crucial, sobretudo em aquelas formas que contêm ambiguidade, e que maior índice de erros podem provocar em processamentos posteriores.

Nos últimos anos, vários trabalhos têm avaliado diferentes algoritmos de PoS-tagging, tendo em conta variáveis como o tagset utilizado, o *corpus* de treino ou a língua analisada (Megyesi, 2001; Branco e Silva, 2004).

De modo geral, considera-se que a *baseline* para esta tarefa situa-se em 90%, e que o estado-da-arte supera o 97% nos melhores resultados. Contudo, têm surgido algumas críticas à avaliação destas ferramentas, com base no tipo de texto utilizado durante o processo. Comparando diferentes avaliações de PoS-taggers sobre textos de diversas procedências (blogues, jornais digitais e outros *sites*) e tipologias (literário, científico, jornalístico, etc.), e não em texto com "condições artificiais", a precisão desce abaixo de 93%, e apresenta grandes níveis de variação em função do género textual (Giesbrecht e Evert, 2009).

A avaliação do processo de anotação morfossintáctica é realizada dividindo o número de tokens cuja etiqueta foi correctamente atribuída pelo número total de tokens do texto.

Esta tarefa, aparentemente trivial, pode apresentar problemas derivados do alinhamento entre o *gold-standard* e o texto etiquetado automaticamente. Este último pode conter um número diferente de tokens em relação ao primeiro, devido à tokenização ou à identificação de nomes próprios, locuções, etc. Assim, a forma `Presidente Mário Soares`, pode ser analisada como um único nome próprio (`Presidente_Mário_Soares`), pode ser divida em dois elementos (`Presidente / Mário_Soares`), ou em três (`Presidente / Mário / Soares`). Para tratar estes casos, o *script* de avaliação contém três parâmetros de execução, com o funcionamento seguinte:

- *NoTok*: Se são detectados erros de *split* (`Presidente_Mário_Soares NP` vs `Presidente NP / Mário NP / Soares NP`), unicamente é avaliado o tag do primeiro token, pelo que é contabilizado um acerto. Este método considera que os erros de tokenização não devem ser levados em conta na avaliação do PoS-tagger.
- *Tok*: Se houver diferenças de tokenização, são contabilizados todos os erros (no caso anterior, três). Note-se que, no caso de que a tokenização e a atribuição da etiqueta em

palavras com mais de um token sejam correctas, é marcado um único acerto.

- *NoLoc*: Este tipo de avaliação ignora todos os tokens que não estiverem alinhados; no exemplo referido, não seria contabilizado nenhum erro nem acerto.

Como foi dito na Secção 3, o Bosque apresenta alguma inconsistência na anotação das locuções e outras expressões multipalavra, facto que devemos ter em conta na consideração dos resultados destas avaliações.

Por esta razão, foi gerada uma outra versão do *corpus* de teste, na qual se realizou um *split* a todos os elementos que continham mais de um token. Assim, executando o PoS-tagger sem identificação de locuções nem de nomes próprios compostos, o *output* é um texto alinhado perfeitamente com este *gold-standard*, pelo que a avaliação resulta mais simples. Um quarto método (*OnlyTag*) tem em conta unicamente os erros e acertos, evitando diferenças de tokenização entre os *corpora* avaliados. Neste caso, na avaliação das locuções ou dos nomes próprios compostos são contabilizados todos os tokens, pelo que uma locução bem etiquetada sumará mais acertos do que com outros métodos. Esta distorção, contudo, é compensada de alguma maneira pelos casos em que a ferramenta falha, nos quais também são contabilizados um maior número de erros.

Os quatro métodos referidos avaliam a precisão do PoS-tagger com o tagset definido no Ponto 2.4. Uma vez que este contém informação muito pormenorizada, e varia notoriamente em relação aos utilizados em outros trabalhos, foi realizada também uma avaliação de cada um dos parâmetros com um tagset mais reduzido (*SingleTags*). Para este fim, unicamente se têm em conta os dois primeiros elementos das etiquetas (categoria e tipo, salvo para os verbos, que se avalia o modo), ignorando assim informação que pode ser inferida por outros meios.[4] Desta maneira, e apesar de os resultados não poderem ser directamente comparáveis (devido a diferenças não apenas no tagset, mas também nos *corpora* de treino e de teste, etc.), temos dados obtidos em condições mais próximas de outras análises, ao mesmo tempo que se verifica a importância do tagset no desempenho de um etiquetador morfossintáctico.

A Tabela 1 mostra os resultados das diferentes avaliações do PoS-tagger para Português Europeu. Os valores são a média de cinco execuções sobre extractos aleatórios de quase 10.000 tokens, com o sistema treinado nos restantes 130.000, salvo para o método *OnlyTag*, treinado sobre 90.000 tokens e avaliado sobre perto de 50.000.

Avaliação	Tag Completo	SingleTags
NoTok	94,788%	96,012%
Tok	94,470%	95,728%
NoLoc	95,044%	96,263%
OnlyTag	94,324%	95,537%

Tabela 1: Avaliação PoS-tagger PE.

Avaliação	Tag Completo	SingleTags
NoTok	97,695%	98,037%
Tok	97,191%	97,562%
NoLoc	97,724%	98,067%
OnlyTag	97,503%	97,914%

Tabela 2: Avaliação PoS-tagger GA.

Para Galego (Tabela 2), o treino foi realizado sobre o *corpus* completo (quase 238.000 tokens), e a avaliação sobre um *corpus* extraído de jornais electrónicos e etiquetado manualmente, de 6.200 tokens.[5]

Os resultados das várias avaliações dos dois sistemas indicam que, actualmente, FreeLing consegue realizar análises morfossintácticas de textos de diversa procedência com desempenhos próximos do estado-da-arte.

Entre as duas variedades linguísticas, as diferenças de precisão são notórias, mas os resultados não devem ser directamente cofrontados. A este respeito, devemos notar, por um lado, os *corpora* de treino utilizados; enquanto o PoS-tagger de PE foi treinado sobre extractos de 130.000 tokens, para GA usou-se um texto mais de 100.000 tokens superior, pelo que é esperável que o desempenho seja maior (Banko e Brill, 2001). Em relação a isto, note-se que os resultados mais baixos da avaliação do PE foram com o método *OnlyTag*, treinado sobre um *corpus* menor. Por outro lado, é importante destacar também as características dos *corpora* de teste utilizados para a avaliação. O PE foi avaliado sobre extractos do próprio Bosque 8.0, com mais ruído —e de maior tamanho— do que o *corpus* de avaliação do GA (mais consistente e com menos ruído). Estas diferenças de desempenho sugerem, como (Giesbrecht e Evert, 2009), que as características do texto a etiquetar influenciam decisivamente a qualidade da etiquetagem.

[4] Este tagset pode ver-se na Tabela 3 (também sem os símbolos de pontuação).

[5] A velocidade de etiquetação de FreeLing executado numa máquina com um processador Core 2 Quad a 2.8 GHz num sistema GNU/Linux foi de aproximadamente 12.000 tokens por segundo. O tempo que o *software* precisou para treinar o PoS-tagger na mesma máquina foi de 10 e de 15 segundos sobre os *corpora* de Português Europeu e Galego, respectivamente.

5 Conclusões

Este trabalho apresenta o desenvolvimento de diversos módulos de tratamento morfossintáctico de FreeLing para Português Europeu e Galego. Os primeiros módulos criados, o tokenizador e o segmentador de frases, realizam tarefas relativamente simples, pelo que a adaptação desde outras línguas românicas não apresentou dificuldades. Os sub-módulos de pesquisa em dicionário e de tratamento de afixos foram desenvolvidos fundamentalmente com base em recursos existentes, bem como os ficheiros de treino do PoS-tagger, gerados com base em *corpora* já anotados. Outros módulos de FreeLing, como os de numerais, datas ou expressões multipalavra (ainda em versões não definitivas), são também disponibilizados.

A avaliação do etiquetador morfossintáctico mostrou que, treinado com recursos já disponíveis para Português Europeu e Galego, um PoS-tagger pode atingir valores de precisão próximos do estado-da-arte. Assim, os diferentes testes realizados confirmam a importância do tagset no desempenho do etiquetador (com diferenças de mais de 1% devidas aos dois tagsets utilizados), bem como dos *corpora* de treino e teste.

O principal objectivo do presente trabalho foi mostrar o processo de adaptação de FreeLing para Português Europeu e Galego, indicando as soluções adoptadas nos casos problemáticos, e realizando diferentes avaliações dos módulos de PoS-tagging. A accessibilidade ao *software* e a alguns dos recursos linguísticos permitiu desenvolver ferramentas de alta precisão, bem como disponibilizar recursos de análise morfossintáctica para Português Europeu e Galego com licenças livres.

Agradecimentos

Este trabalho recebeu apoio do Governo Galego através dos projectos com referências PGIDIT07PXIB204015PR e 2008/101.

Referências

Banko, Michele e Eric Brill. 2001. Mitigating the Paucity-of-Data Problem: Exploring the Effect of Training Corpus Size on Classifier Performance for Natural Language Processing. Em *Proceedings of the Conference on Human Language Technology*.

Barcala, Fco. Mario, Eva Mª Domínguez Noya, Pablo Gamallo Otero, Marisol López Martínez, Eduardo Miguel Moscoso Mato, Guillermo Rojo, María Paula Santalla del Río, e Susana Sotelo Docío. 2007. A corpus and Lexical Resources for Multi-word Terminology Extraction in the Field of Economy in a in a Minority Language. Em Zygmunt Vetulani, editor, *Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 3rd Language & Technology Conference*, pp. 359–363, Poznan. Wydawnictwo Poznaskie Sp. z o.o.

Bick, Eckhard. 2000. *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Tese de doutoramento, University of Aarhus, Denmark.

Bosque. Uma floresta integralmente revista por linguistas. `http://www.linguateca.pt/Floresta/corpus.html#bosque`.

Branco, António e João Silva. 2003. Contractions: breaking the tokenization-tagging circularity. Em *Lecture Notes in Artificial Intelligence*, volume 2721, pp. 167–170, Berlin. Spinger.

Branco, António e João Silva. 2004. Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, e Raquel Silva, editores, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 507–510, Paris. ELRA.

Brants, Thorsten. 2000. TnT – A Statistical Part-of-Speech Tagger. Em *Proceedings of the 6th Conference on Applied Natural Language Processing, ANLP*. ACL.

Carreras, Xavier, Isaac Chao, Lluís Padró, e Muntsa Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. Em *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*.

DepPattern. An Open Source Dependency-Based Analyzer. `http://gramatica.usc.es/pln/tools/deppattern.html`.

Eleutério, Samuel, Elisabete Ranchhod, Cristina Mota, e Paula Carvalho. 2003. Dicionários Electrónicos do Português. Características e Aplicações. Em *Actas del VIII Simposio Internacional de Comunicación Social*, pp. 636–642, Santiago de Cuba.

FreeLing. An Open Source Suite of Language Analyzers. `http://www.lsi.upc.edu/~nlp/freeling/`.

Gamallo, Pablo e Isaac González. 2009. Una gramática de dependencias basada en patrones

de etiquetas. Em *XXV Congresso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural*, Donostia.

Giesbrecht, Eugenie e Stefan Evert. 2009. Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. Em *Proceedings of the 5th Web as Corpus Workshop (WAC5)*, Donostia.

Graña, Jorge, Fco. Mario Barcala, e Jesús Vilares, 2002. *Formal Methods of Tokenization for Part-of-Speech Tagging*, volume 2276/2002, pp. 123–144.

Leach, Geoffrey e Andrew Wilson. 1996. Recommendations for the Morphosyntactic Annotation of Corpora. Relatório técnico, Expert Advisory Group on Language Engineering Standard (EAGLES).

Marques, Nuno e Gabriel Lopes. 2001. Tagging with Small Training Corpora. Em *Proceedings of the International Conference on Intelligent Data Analysis*, volume 2189 of *Lecture Notes on Artificial Intelligente (LNAI)*, pp. 63–72. Springer-Verlag.

Megyesi, Beáta. 2001. Comparing Data-Driven Learning Algorithms for PoS Tagging of Swedish. Em *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing*, pp. 151–158.

Padró, Lluís. 1998. *A Hybrid Environment for Syntax-Semantic Tagging.* Tese de doutoramento, Dept. Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya.

Ribeiro, Ricardo, Luís C. Oliveira, e Isabel Trancoso. 2003. Using Morphossyntactic Information in TTS Systems: Comparing Strategies for European Portuguese. Em *Proceedings of the 6th Workshop on Computational Processing on the Portuguese Language (PROPOR 2003)*, pp. 143–150, Faro. Springer-Verlag.

Tag	**Valor**
AO	Adjectivo Ordinal
AQ	Adjectivo Qualificativo
CS	Conjunção Subordinativa
CC	Conjunção Coordenativa
DA	Determinante Artigo
DD	Determinante Demonstrativo
DI	Determinante Indefinido
DP	Determinante Possessivo
I	Interjeição
NC	Nome Comum
NP	Nome Próprio
PD	Pronome Demonstrativo
PE	Pronome Exclamativo
PI	Pronome Indefinido
PP	Pronome Pessoal
PR	Pronome Relativo
PT	Pronome Interrogativo
PX	Pronome Possessivo
RG	Advérbio Geral
RN	Advérbio Negativo
SP	Preposição
VG	Verbo: Gerúndio
VI	Verbo: Modo Indicativo
VM	Verbo: Modo Imperativo
VN	Verbo: Infinitivo
VP	Verbo: Particípio
VS	Verbo: Modo Conjuntivo
Z	Numeral

Tabela 3: Tagset Largo (*SingleTags*).

Adjectivos			
Elemento	**Atributo**	**Valor**	**Tag**
1	Categoria	Adjectivo	A
2	Tipo	Qualificativo	C
		Ordinal	O
3	Grau	Aumentativo	A
		Diminutivo	D
		Superlativo	S
4	Género	Masculino	M
		Feminino	F
		Comum	C
5	Número	Singular	S
		Plural	P
		Invariável	N

Advérbios			
Elemento	**Atributo**	**Valor**	**Tag**
1	Categoria	Advérbio	R
2	Tipo	Geral	G
		Negativo	N

Determinantes			
Elemento	**Atributo**	**Valor**	**Tag**
1	Categoria	Determinante	D
2	Tipo	Artigo	A
		Demonstrativo	D
		Indefinido	I
		Possessivo	P
3	Pessoa	1ª/2ª/3ª	1/2/3
4	Género	Masculino	M
		Feminino	F
		Comum	C
		Neutro	N
5	Número	Singular	S
		Plural	P
		Invariável	N
6	Possuidor	Singular	S
		Plural	P

Verbos			
Elemento	**Atributo**	**Valor**	**Tag**
1	Categoria	Verbo	V
2	Tipo	Principal	M
3	Modo	Gerúndio	G
		Indicativo	I
		Imperativo	M
		Infinitivo	N
		Particípio	P
		Conjuntivo	S
4	Tempo	Futuro do Pretérito	C
		Mais-que-Perfeito	M
		Futuro	F
		Imperfeito	I
		Presente	P
		Perfeito	S
5	Pessoa	1ª/2ª/3ª	1/2/3
6	Número	Singular	S
		Plural	P

Conjunções			
Elemento	**Atributo**	**Valor**	**Tag**
1	Categoria	Conjunção	C
2	Tipo	Coordinativa	C
3		Subordinativa	S

Preposições			
Elemento	**Atributo**	**Valor**	**Tag**
1	Categoria	Aposição	S
2	Tipo	Preposição	P
3	Forma	Simples	S

Nomes			
Elemento	**Atributo**	**Valor**	**Tag**
1	Categoria	Nome	N
2	Tipo	Comum	C
		Próprio	P
3	Género	Masculino	M
		Feminino	F
		Comum	C
4	Número	Singular	S
		Plural	P
		Invariável	N
7	Grau	Aumentativo	A
		Diminutivo	D

Pronomes			
Elemento	**Atributo**	**Valor**	**Tag**
1	Categoria	Pronome	P
2	Tipo	Demonstrativo	D
		Exclamativo	E
		Indefinido	I
		Pessoal	P
		Relativo	R
		Interrogativo	T
		Possessivo	X
3	Pessoa	1ª/2ª/3ª	1/2/3
4	Género	Masculino	M
		Feminino	F
		Comum	C
		Neutro	N
5	Número	Singular	S
		Plural	P
		Invariável	N
6	Caso	Nominativo	N
		Acusativo	A
		Dativo	D
		Oblíquo	O
7	Possuidor	Singular	S
		Plural	P

Numerais			
Elemento	**Atributo**	**Valor**	**Tag**
1	Categoria	Numeral	Z

Interjeições			
Elemento	**Atributo**	**Valor**	**Tag**
1	Categoria	Interjeição	I

Tabela 4: Formato do Tagset Estreito.