Módulo de acentuación para o galego en Freeling

Miguel Anxo Solla Portela
Universidade de Vigo
miguelsolla@uvigo.es

Resumo

Descrición do módulo de acentuación para a lingua galega que se desenvolveu para a súa inclusión en vindeiras versións da biblioteca de ferramentas de análise lingüística Freeling.

1. Introdución

A biblioteca de ferramentas de análise lingüística Freeling ofrece amplas posibilidades no desenvolvemento de aplicacións lingüísticas para un conxunto de linguas cada vez máis extenso, polo momento: inglés, español, catalán, galego, italiano, portugués, asturiano e galés. Ata a versión actual, Freeling vén empregando no recoñecemento de formas resultantes da segmentación dunha afixación as mesmas funcións de restauración ou de supresión do acento gráfico para o galego que para o español. Esta característica limita as posibilidades do recoñecemento morfolóxico de numerosas raíces en lingua galega, xa que as regras de afixación en Freeling contemplan tanto a afixación léxica mediante prefixos e sufixos coma a segmentación de formas verbais con pronomes enclíticos, que é a posición habitual ou non-marcada do pronome persoal en lingua galega. A frecuencia deste tipo de secuencias xunto cun tratamento inadecuado da restauración da acentuación gráfica nas formas verbais segmentadas producen anomalías na análise. No entanto, Freeling é un proxecto de código aberto cunha atinada arquitectura modular para as especificidades de cada lingua, que permite o desenvolvemento de código para o tratamento do acento gráfico en cada lingua sen que interfira coas necesidades das demais. Co fin de evitar as interferencias do tratamento da acentuación gráfica de raíces para o castelán, desenvolveuse un novo módulo para o procesamento da acentuación gráfica que reúne un conxunto de funcións específicas que actúan sobre estas formas tras a segmentación da afixación consonte as regras de acentuación da lingua galega e remodeláronse as regras de afixación dos datos lingüísticos do galego para obter a forma illada no dicionario da aplicación.

2. O tratamento da acentuación gráfica das formas afixadas en Freeling

As regras de afixación para cada lingua da biblioteca atópanse no ficheiro afixos.dat dos datos lingüísticos correspondentes.

Freeling diferencia as regras de segmentación de elementos que anteceden na secuencia á forma que debe buscar no dicionario (*prefixes*) das regras para secuencias nas que debe segmentar un elemento ao final da secuencia (*sufixes*). Neste último grupo inclúense tanto as regras de sufixación léxica coma as de segmentación de formas verbais e pronomes enclíticos, pero os parámetros que permite establecer a aplicación en cada regra rexen comportamentos moi diferentes:

```
mente *  ^AQ0CS  RG 1  0 0  L  1  -
lle   *  ^V      *  0  1 0  L  1  $$+lle:$$+PP
```

A regra para o sufixo *-mente* vai segmentar esta terminación, activar a función de acentuación para bases de sufixos léxicos (5ª columna) que crea un candidato sen ningún acento gráfico e un candidato con acento en cada unha das vogais que conteña a raíz e, se atopa unha base adxectival, etiquetará como adverbio o derivado deadxectival; mentres que a regra para o sufixo *-lle* vai segmentar a raíz e procesala coas funcións de acentuación gráfica para as formas verbais segmentadas (6ª columna) e, se atopa no dicionario unha forma verbal, etiquetará os dous segmentos da secuencia, a forma verbal flexionada e o pronome persoal.

Porén, as funcións de restauración ou supresión do acento gráfico para o español non resultan adecuadas en raíces verbais galegas, xa que as regras de acentuación gráfica do español difiren das do galego na acentuación diacrítica, na silabación de certos encontros vocálicos (español *atribuimos* / galego *atribuímos, atribuíu, atribuïamos*) e na consideración das secuencias polisilábicas que rematan en ditongo decrecente ou en ditongo decrecente seguido de *-n* ou *-s* (español *comeréis, fuereis* / galego *comerei, amábeis*). Ademais, os encontros cos pronomes enclíticos presentan particularidades propias: tres alomorfos para o pronome persoal acusativo de terceira persoa en distribución complementaria segundo a terminación verbal (*la, las, lo* ou *los* tras as formas que rematan en *-r* ou *-s*: *comerala* ~ *comerás* + *a*; *na, nas, no* ou *nos* tras as formas que rematan en ditongo decrecente: *comereinas* ~ *comerei* + *as*; e *a, as, o*

ou *os* nos casos restantes: *cómoa* ~ *como* + *a*) e perda do *-s* final das formas flexionadas en primeira persoa do plural cando levan enclítico un pronome persoal que tamén sexa de primeira persoa de plural (*achegámonos* ~ *achegamos* + *nos*).

Co fin de procesar estas características, optouse polo desenvolvemento dun novo módulo de acentuación para a lingua galega, que mantén o tratamento da acentuación gráfica para a afixación léxica que operaba co módulo do español, pero que establece un novo comportamento para engadirlles, manterlles ou suprimirlles a acentuación gráfica ás formas verbais segmentadas. Asemade, revisáronse as regras de afixación para que operasen consonte o novo módulo e retomouse un ficheiro moi similar ao sufixos.dat da versión 2.0 de Freeling que resulta máis acorde, pois as versións posteriores intentaron suplir as limitacións na análise co módulo de acentuación para o español ampliando o número de regras para cada pronome mediante a inclusión da parte final da terminación verbal, mais esta ampliación non é necesaria cun comportamento máis axeitado do módulo de acentuación específico. Nesta regresión do ficheiro afixos.dat os cambios respecto do sufixos.dat da versión 2.0 son pouco significativos:

- Adaptouse o nome do ficheiro e o número de columnas ao das versións actuais da biblioteca.
- Aumentáronse as regras para o encontro do alomorfo do pronome persoal átono acusativo cando vai enclítico a unha forma verbal que termina en ditongo decrecente (ditongos *-ei* e *-íu*).
- Incluíronse novas regras para o encontro de formas verbais en 1ª persoa de plural cun pronome –ou unha secuencia de pronomes que o leve en pirmeira posición– de 1ª persoa de plural (cruz*ámonos* ~ cruz*amos* + *nos*).
- Engadíronse algunhas regras para secuencias de dous pronomes enclíticos (*cheme*, *chelle*...).
- Incluíuse unha regra regexp para tratar de evitar a segmentación da primeira forma do pronome persoal acusativo de terceira persoa nas secuencias nas que a presenza lle corresponda a algún dos seus alomorfos. Deste xeito, impídese, por exemplo, a separación de *comeras* en *comer* + *as*, pois este encontro sempre se resolve en galego co segundo alomorfo do pronome de acusativo: *comelas*):

```
Comeras comer VMIM2S0 1
Comelas comer+o VMN0000+PP3FPA00 0.330674
comer+o VMN03S0+PP3FPA00 0.330674
comer+o VMN01S0+PP3FPA00 0.330674
comer+o VMSF3S0+PP3FPA00 0.00398928
comer+o VMSF1S0+PP3FPA00 0.00398928
```

Porén, esta limitación non resulta suficiente en todos os casos, particularmente co terceiro alomorfo do pronome acusativo *-na, -nas, -no* ou *-nos* que aparece tras ditongo decrecente, xa que a restrición da segmentación imponse tras a etiquetación de Freeling sobre as formas da flexión verbal. As formas flexivas que rematan en *-r* ou *-s* son constantes na flexión de todos os verbos, pero algunhas formas flexivas que rematan en ditongo decrecente non presentan unha recorrencia tan constante: a terceira persoa de singular do pretérito de indicativo de todos o verbos regulares remata en ditongo decrecente (*andou, varreu, partiu*), mais moitos irregulares adoptan outras terminacións (*dixo, houbo*); na primeira persoa deste mesmo tempo, os verbos regulares da primeira conxugación rematan en ditongo decrecente (*andei*) fronte aos da segunda e da terceira conxugación (*varrín, partín*); algúns verbos irregulares teñen formas que rematan en ditongo decrecente na primeira e na terceira persoa do singular do presente de indicativo (*dou, vou, hei, sei, fai, vai*)... Como non se atopou unha restrición satisfactoria para evitar que a regra da primeira forma do acusativo se aplique tras un ditongo decrecente, nalgunhas secuencias que presentan esta ditongación seguida de *-a, -as, -o* ou *-os* prodúcese unha segmentación incorrecta*:*

```
Vaia ir VMSP3S0 0.477778 ir VMSP1S0 0.477778 ir
VMM03S0 0.0111111 vaia I 0.0111111 ir+o
VMIP3S0+PP3FSA00 0.0111111 ir+o VMM02S0+PP3FSA00
0.0111111
```

```
Faios faio NCMP000 0.758333 faios NP00000
0.210714 facer+o VMIP3S0+PP3MPA00 0.0297619
facer+o VMM02S0+PP3MPA00 0.00119048
```

Nos dous exemplos, as dúas últimas análises parten dunha segmentación incorrecta, pois a forma *vaia* non admite unha segmentación en *vai* + *a*, xa que esta secuencia se resolve co terceiro alomorfo do pronome, *vaina* e, do mesmo modo, *faino* é o resultado de *fai* + *o*:

```
Vaina ir+o VMIP3S0+PP3FSA00 0.997312 ir+o
VMM02S0+PP3FSA00 0.00268817
```

```
Faino facer+o VMIP3S0+PP3MSA00 0.997312 facer+o
VMM02S0+PP3MSA00 0.00268817
```

- Engadiuse tamén outra regra regexp que evita a segmentación do pronome enclítico *se* (ou de secuencias de pronomes que comecen por *se*) con formas verbais flexionadas en primeira ou en segunda persoa.

3. Comportamento da acentuación gráfica no novo módulo

Cada regra de segmentación dun afixo establece o tratamento da acentuación gráfica que requiren os candidatos a formas de dicionario. A candidatura vai recibir un procesamento moi diferente cando provén da segmentación de pronomes persoais enclíticos de cando provén dunha afixación léxica.

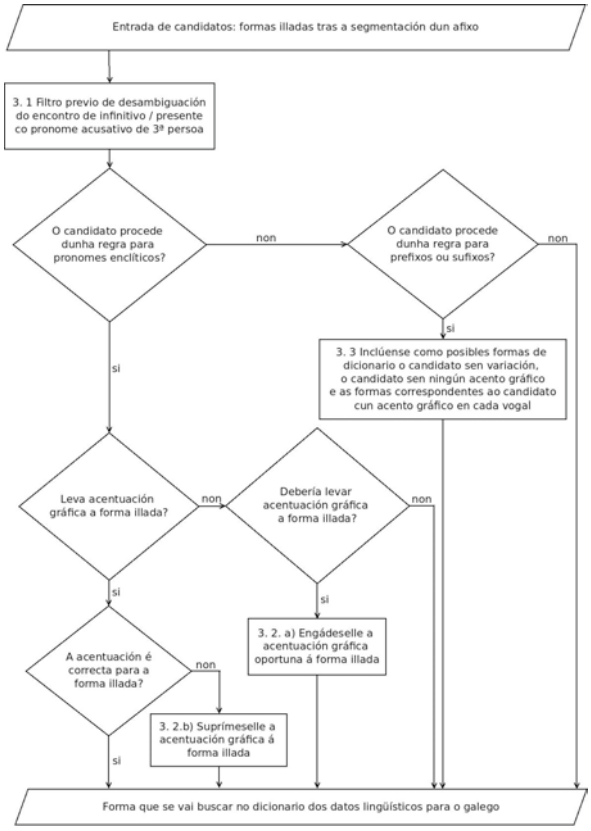

Ilustración 1. Diagrama de fluxo do módulo

3.1 O filtro de desambiguación

Incluíuse unha función de desambiguación de formas resultantes de segmentar o alomorfo do pronome persoal átono de terceira persoa *-la, -las, -lo, -los*. A regra para este enclítico establece que se lle debe engadir un *-s* ou un *-r* á raíz verbal para recuperar a forma verbal no dicionario da aplicación. A función de desambiguación compara ambas as dúas candidaturas e examina se o acento que tiña co enclítico ten función fonolóxica antes de que se apliquen as demais funcións que determinarán se lle corresponde ou non levar acento gráfico á forma verbal resultante; isto é, diferencia exemplos como *comela* (~ *comer* + *a*) de *cómela* (~ *comes* + *a*) e mantén a dobre posibilidade de análise con determinados presentes de indicativo polisilábicos oxítonos (*prevelos ~ prevés + os / prever + os*). O motivo de establecer este filtro con anterioridade ao tratamento da acentuación gráfica é que, cando a ambas as dúas formas illadas lles corresponde eliminar a acentuación gráfica, a desambiguación a posteriori xa non sería posible, pois calquera das dúas constitúe unha entrada no dicionario.

3.2 Formas verbais de secuencias con pronomes persoais enclíticos

O tratamento que recibe o acento gráfico cando unha regra activa o módulo de acentuación para formas verbais resultantes da segmentación de pronomes enclíticos é, a grandes trazos, o seguinte:

a) Cando da segmentación se obtén unha forma verbal que non ten acento gráfico, compróbase que non se trate dunha forma polisílabica, que termine nas vogais *-a, -e, -o,* e *-i* cando non forma parte dun ditongo decrecente nin en ningunha das terminacións anteriores e mais un *-n* ou un *-s* final, que levase enclítico un pronome persoal monosílabo. Se se trata dun destes casos, incorpórase o acento gráfico da forma verbal oxítona para validar a forma no dicionario (*prevese ~ prevé + se*, *darasme ~ darás + me*) e, se non, valídase sen o acento gráfico (*deille ~ dei + lle, enviounos ~ enviou + nos*).

b) Cando da segmentación resulta unha raíz con acento gráfico, compróbase que o acento sexa correcto:

- Acentos diacríticos (*dálle ~ dá + lle*).
- Segunda persoa do singular ou terceira persoa, singular ou plural, do futuro de indicativo de todos os verbos e segunda ou terceira persoa de singular ou terceira de plural de formas oxítonas de certos verbos en presente de indicativo, que levasen enclítica unha secuencia polisílaba de pronomes persoais (*fáranche*~*me ~ farán + che + me, estánvola ~ están + vos + a*).
- Primeira ou segunda persoa do plural do pretérito de subxuntivo (*cantásemoslles ~ cantásemos + lles*).
- Acentuación dunha vogal pechada que marca un hiato (*sabíao ~ sabía + o, saínlles ~ saín + lles, constituíuna ~ constituíu + a*).

Se non se trata de ningún dos casos anteriores, elimínase o acento gráfico da forma resultante (*quíxeno ~ quixen + o, perseguíndoas ~ perseguindo + as, cáelles ~ cae + lles, caéralle ~ caera + lle, tróuxoma ~ trouxo + me + a, atéivolas ~ atei + vos + as, cantáballe ~ cantaba + lle*).

3.3 Afixación léxica

As regras dun prefixo ou dun sufixo seguen contando, coma no caso do español, cunha posibilidade diferente no módulo de acentuación, unha función específica coa que a forma candidata vaise reconstruír en varias: unha forma sen ningún acento gráfico e esa mesma forma con acento gráfico en cada unha das vogais que conteña. Cada unha destas formas vaise procurar no dicionario. Deste xeito, aínda que *calidamente* non figura no dicionario, Freeling identifica que se trata dunha

derivación de *cálido* grazas á regra do sufixo *-mente* que segmenta e reconstrúe a forma candidata, e activa este tratamento da acentuación no ficheiro afixos.dat dos datos lingüísticos para o galego.

4. O código do módulo

O código do módulo de acentuación incorporouse ao repositorio de subversion da versión en desenvolvemento de Freeling e pódese obter mediante a instrución `svn checkout http://devel.cpl.upc.edu/freeling/svn/latest/freeling`.

Ademais das modificacións que xa se viron para o ficheiro dos datos lingüísticos coas regras de afixación, o ficheiro accents.cc modificouse para que a análise en lingua galega deixe de utilizar o módulo de acentuación para o español e pase a utilizar o módulo novo. No ficheiro accents_modules.h decláranse as clases e as funcións que se definen no ficheiro accents_modules.cc, no que figuran diferentes funcionalidades de adecuación da acentuación gráfica para as linguas que as precisan.

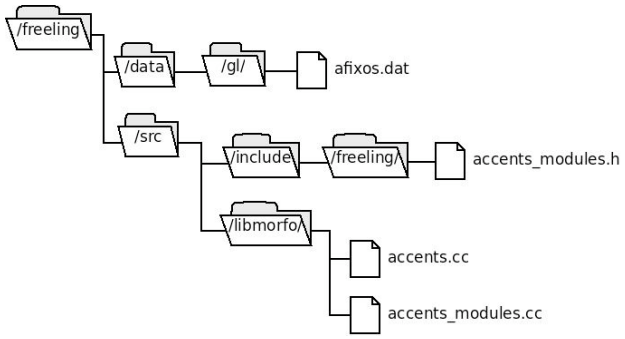

Ilustración 2. Rutas dos ficheiros que se modificaron na versión en desenvolvemento.

5. Avaliación dos resultados

Para a avaliación dos resultados analizouse o ficheiro corpus_economia_prensa_oracions.txt do *Corpus xiada*, versión 2.4, que distribúe o Centro Ramón Piñeiro para a Investigación en Humanidades baixo os termos da licenza Lesser General Public License for Linguistic Resources. O corpus analizouse primeiro coa versión estable de Freeling 2.2 e despois coa versión en desenvolvemento. O ficheiro contén, consonte o cómputo do editor de textos, 205.370 palabras.

Nesta análise sobre o mesmo corpus, a cantidade de secuencias lingüísticas que a versión en desenvolvemento segmenta como formas verbais con pronomes enclíticos (3.213) practicamente duplica as secuencias que atopa a versión 2.2 de Freeling (1.636). A continuación figuran algúns exemplos deste comportamento:

Resultados con Freeling 2.2	**Resultados coa versión en desenvolvemento**
faise faise NCFS000 0.894226 faise VMSI3S0 0.0769279 faise VMSP3S0 0.0288462	faise facer+se VMIP3S0+PP3CN000 1
reclamándollo reclamándollo NCMS000 1	reclamándollo reclamar+lle+o VMG0000+PP3CSD00+PP3MSA00 1
adoptárense adoptárense NP00000 1	adoptárense adoptar+se VMN03P0+PP3CN000 1
encontrámonos encontrámonos NCMP000 0.962947 encontrámonos AQ0MP0 0.0370529	encontrámonos encontrar+nos VMIP1P0+PP1CP000 0.5 encontrar+nos VMIS1P0+PP1CP000 0.5
vaise vaise NCFS000 0.894226 vaise VMSI3S0 0.0769279 vaise VMSP3S0 0.0288462	vaise ir+se VMIP3S0+PP3CN000 1
Báixansenos báixansenos NP00000 1	Báixansenos baixar+se+nos VMIP3P0+PP3CN000+PP1CP000 1
déixenos déixenos RG 0.893127 déixenos AQ0MP0 0.0733362 déixenos NCMP000 0.0335372	déixenos deixar+o VMSP3P0+PP3MPA00 0.333269 deixar+nos VMSP3S0+PP1CP000 0.333269 deixar+nos VMSP1S0+PP1CP000 0.333269 deixar+o VMM03P0+PP3MPA00 9.62927e-05 deixar+nos VMM03S0+PP1CP000 9.62927e-05
mantela mantela NCFS000 1	mantela manter+o VMIP2S0+PP3FSA00 0.6 mantela NCFS000 0.1 manter+o VMN0000+PP3FSA00 0.1 manter+o VMN03S0+PP3FSA00 0.1 manter+o VMN01S0+PP3FSA00 0.1
subilo subilo NCMS000 0.470691 subilo NP00000 0.382288 subilo AQ0MS0 0.14702	subilo subir+o VMN0000+PP3MSA00 0.330674 subir+o VMN03S0+PP3MSA00 0.330674 subir+o VMN01S0+PP3MSA00 0.330674 subir+o VMSF3S0+PP3MSA00 0.00398928 subir+o VMSF1S0+PP3MSA00 0.00398928
faino faino AQ0MS0 0.409727 faino NCMS000 0.360778 faino NP00000 0.213106 faino VMIP1S0 0.0163881	faino facer+o VMIP3S0+PP3MSA00 0.997312 facer+o VMM02S0+PP3MSA00 0.00268817
Dío díos NP00000 1	Dío dicir+o VMM02S0+PP3MSA00 1
podérense podérense NP00000 1	podérense poder+se VMN03P0+PP3CN000 1
foise foise NCFS000 0.894226 foise VMSI3S0 0.0769279 foise VMSP3S0 0.0288462	foise ir+se VMIS3S0+PP3CN000 1
Dise dise NP00000 1	Dise dicir+se VMIP3S0+PP3CN000 1

Para ilustrar os resultados da aplicación do novo módulo, contrastouse o número de análises destas secuencias coa cantidade de veces en que a etiquetación era certeira.

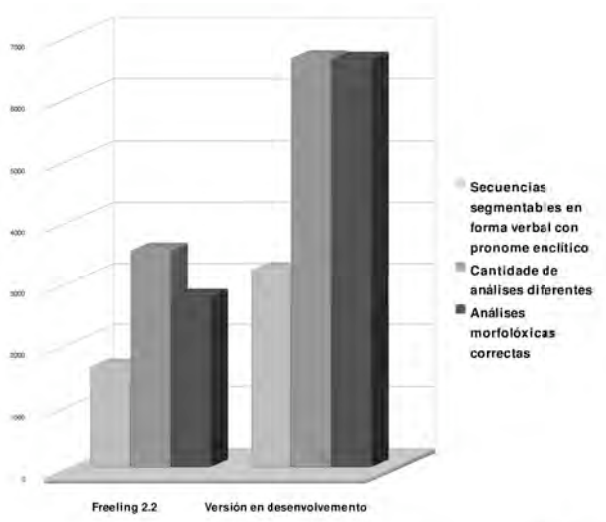

Ilustración 3. Gráfico comparativo das secuencias de forma verbal con pronomes enclíticos nas etiquetacións do *Corpus xiada*.

Consideráronse análises incorrectas as que contiñan descricións morfolóxicas da secuencia inadecuadas e as que etiquetaron o pronome persoal *se* en énclise cunha forma verbal flexionada en 1ª ou 2ª persoa. Deste xeito, constatouse que a marxe de erro da versión 2.2 de Freeling era bastante superior (21,41%, 756 etiquetacións erróneas nas 3.531 análises das 1.636 secuencias segmentadas) respecto da versión en desenvolvemento (0,30%, 20 análises erradas nas 6.647 etiquetacións de formas verbais con pronomes enclíticos nas 3.213 secuencias que se detectaron).

As etiquetacións que se obtiveron coa versión en desenvolvemento resultaron máis axeitadas ca as da versión estable e ofrecen, en xeral, unha descrición máis precisa de secuencias homógrafas:

Resultados con Freeling 2.2	**Resultados coa versión en desenvolvemento**
mantela mantela NCFS000 1	mantela manter+o VMIP2S0+PP3FSA00 0.6 mantela NCFS000 0.1 manter+o VMN0000+PP3FSA00 0.1 manter+o VMN03S0+PP3FSA00 0.1 manter+o VMN01S0+PP3FSA00 0.1
ilusionante ilusionar+te VMIP3P0+PP2CSA00 1	ilusionante ilusionante AQ0CS0 0.5 ilusionante AQ0MS0 0.5
convertela converter+o VMIP2S0+PP3FSA00 0.880952 converter+o VMN0000+PP3FSA00 0.0238095 converter+o VMN03S0+PP3FSA00 0.0238095 converter+o VMN01S0+PP3FSA00 0.0238095 converter+o VMSF3S0+PP3FSA00 0.0238095 converter+o VMSF1S0+PP3FSA00 0.0238095	convertela converter+o VMN0000+PP3FSA00 0.330674 converter+o VMN03S0+PP3FSA00 0.330674 converter+o VMN01S0+PP3FSA00 0.330674 converter+o VMSF3S0+PP3FSA00 0.00398928 converter+o VMSF1S0+PP3FSA00 0.00398928
serrano serrar+o VMIP3P0+PP3MSA00 1	serrano serrano AQ0MS0 1
préstamos préstamo NCMP000 1	préstamos préstamo NCMP000 0.857639 prestar+me+o VMIP3S0+PP1CS000+PP3MPA00 0.140556 prestar+me+o VMM02S0+PP1CS000+PP3MPA00 0.00180556
importe importe NCMS000 0.988095 importar VMM03S0 0.00396825 importar VMSP3S0 0.00396825 importar VMSP1S0 0.00396825	importe importe NCMS000 0.986395 importar VMM03S0 0.00226757 importar VMSP3S0 0.00226757 importar VMSP1S0 0.00226757 impor+te VMN0000+PP2CSA00 0.00226757 impor+te VMN03S0+PP2CSA00 0.00226757 impor+te VMN01S0+PP2CSA00 0.00226757
verse versar VMSP3S0 0.49988 versar VMSP1S0 0.49988 versar VMM03S0 0.000240674	verse ver+se VMN0000+PP3CN000 0.391657 ver+se VMN03S0+PP3CN000 0.391657 versar VMSP3S0 0.107597 versar VMSP1S0 0.107597 versar VMM03S0 0.00149195
dálle dar+lle VMM02S0+PP3CSD00 1	dálle dar+lle VMIP3S0+PP3CSD00 0.997312 dar+lle VMM02S0+PP3CSD00 0.00268817
querela querela NCFS000 1	querela querela NCFS000 0.321672 querer+o VMN0000+PP3FSA00 0.226109 querer+o VMN03S0+PP3FSA00 0.226109 querer+o VMN01S0+PP3FSA00 0.226109
días día NCMP000 1	días día NCMP000 0.991968 dicir+o VMIP3S0+PP3FPA00 0.00401606 dicir+o VMM02S0+PP3FPA00 0.00401606
vaise vaise NCFS000 0.894226 vaise VMSI3S0 0.0769279 vaise VMSP3S0 0.0288462	vaise ir+se VMIP3S0+PP3CN000 1
Tense tensar VMSP3S0 0.49988 tensar VMSP1S0 0.49988 tensar VMM03S0 0.000240674	Tense ter+se VMIP3S0+PP3CN000 0.86083 tensar VMSP3S0 0.069332 tensar VMSP1S0 0.069332 tensar VMM03S0 0.000506073
quedarmos quedar VMN01P0 0.75 quedar VMSF1P0 0.25	quedarmos quedar VMN01P0 0.228571 quedar+me+o VMN0000+PP1CS000+PP3MPA00 0.228571 quedar+me+o VMN03S0+PP1CS000+PP3MPA00 0.228571 quedar+me+o VMN01S0+PP1CS000+PP3MPA00 0.228571 quedar VMSF1P0 0.0285714 quedar+me+o VMSF3S0+PP1CS000+PP3MPA00 0.0285714 quedar+me+o VMSF1S0+PP1CS000+PP3MPA00 0.0285714
afaste afastar VMSP3S0 0.444444 afastar VMSP1S0 0.444444 afastar VMM03S0 0.111111	afaste afastar VMSP3S0 0.416667 afastar VMSP1S0 0.416667 afastar VMM03S0 0.0833333 afacer+te VMIP2S0+PP2CSA00 0.0833333

Con todo, na versión en desenvolvemento aínda se xeran análises que parten de segmentacións incorrectas nalgúns casos. A forma *vaia* aparece 7 veces no corpus e produce 14 análises que parten dunha segmentación errónea, pois, como xa se viu, as dúas últimas análises do exemplo correspóndenlle á secuencia *vainas*:

```
vaia ir VMSP3S0 0.477778 ir VMSP1S0 0.477778 ir VMM03S0
0.0111111 vaia I 0.0111111 ir+o VMIP3S0+PP3FSA00 0.0111111 ir+o
VMM02S0+PP3FSA00 0.0111111
```

As restantes análises erróneas non aparecen con tanta frecuencia no corpus:

```
explícaselles explicar+lle VMSI3S0+PP3CPD00 0.492234
explicar+lle VMSI1S0+PP3CPD00 0.492234 explicar+se+lle
VMIP3S0+PP3CN000+PP3CPD00 0.0155317

débeselle deber+lle VMSI3S0+PP3CSD00 0.492234 deber+lle
VMSI1S0+PP3CSD00 0.492234 deber+se+lle VMIP3S0+PP3CN000+PP3CSD00
0.0155317

predios predio NCMP000 0.857639 predicir+o VMIP3S0+PP3MPA00
0.140556 predicir+o VMM02S0+PP3MPA00 0.00180556

dias dicir+o VMIP3S0+PP3FPA00 0.997312 dicir+o VMM02S0+PP3FPA00
0.00268817
```

As dúas primeiras análises das secuencias *explícaselles* e *débeselle* correspóndense, en realidade, coa análise atinada das secuencias *explicáselles* e *debéselle* respectivamente:

```
Explicáselles explicar+lle VMSI3S0+PP3CPD00 0.5 explicar+lle
VMSI1S0+PP3CPD00 0.5

Debéselles deber+lle VMSI3S0+PP3CPD00 0.5 deber+lle
VMSI1S0+PP3CPD00 0.5
```

A segmentación da secuencia *predios* nunha forma verbal cun pronome enclítico tamén é errónea, xa que esta análise correspóndelle á secuencia *predíos*:

```
Predíos predicir+o VMIP3S0+PP3MPA00 0.997312 predicir+o
VMM02S0+PP3MPA00 0.00268817
```

No caso da secuencia **dia,* trátase dun erro ortográfico da forma *día*:

```
Día día NCMS000 0.758333 día NP00000 0.210714 dicir+o
VMIP3S0+PP3FSA00 0.0297619 dicir+o VMM02S0+PP3FSA00 0.00119048
```

6. Conclusións e traballo futuro

A etiquetación morfolóxica das secuencias de formas verbais con pronomes enclíticos mellorou sensiblemente co desenvolvemento do novo módulo de acentuación para a lingua galega. Unha identificación máis adecuada dos núcleos verbais debera mellorar tamén outras análises da biblioteca, como a análise de dependencias, así como o funcionamento xeral doutras aplicacións que utilicen Freeling.

O módulo que se desenvolveu estase adaptando á formalización das futuras versións de Freeling coa pretensión de acadar os mesmos resultados que na versión de desenvolvemento do Freeling 2.2 que se vén de describir, e tamén coa intención de tratar de deseñar estratexias que eviten as etiquetacións erróneas que se detectaron ata o momento.

Referencias

Atserias, Jordi, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluis Padró, e Muntsa Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source NLP library. En *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, páxinas 48-55.

Real Academia Galega / Instituto da Lingua Galega, *Normas ortográficas e morfolóxicas do idioma galego*, 18ª edición, 2003.

Freeling user manual, 2.2, September 2010 `http://nlp.lsi.upc.edu/freeling/doc/userman/userman.pdf`.

Technical reference manual, 2.2, `http://nlp.lsi.upc.edu/freeling/doc/refman`.

Centro Ramón Piñeiro para a Investigación en Humanidades, *Etiquetador/Lematizador do Galego Actual (XIADA),* versión 2.4, *`http://corpus.cirp.es/xiada`,* [Consultado o: 20/10/2010].