Lluís de Yzaguirre Maura
Sergi Torner Castells
Anna Matamala Ripoll

Institut de Lingüística Aplicada
Universitat Pompeu Fabra
de_yza@upf.es
www.iula.upf.es

El tratamiento automático de las ambigüedades segmentales del castellano
(XVII Congreso AESLA, Lingüística de Corpus y Computacional)

ABSTRACT

En esta comunicación se introduce el concepto de ambigüedad segmental, se analiza su especial impacto en castellano frente a otras lenguas románicas, se explica cómo se ha obtenido un listado cuasi-exhaustivo de ellas, se clasifican según criterios básicos, se propone una interpretación por defecto de las más usuales y se presenta cómo se ha resuelto la interacción entre lematizador y desambiguador para tratarlas adecuadamente.

1.- Introducción

Las ambigüedades segmentales son aquellas en que una misma palabra gráfica puede tener dos o más interpretaciones que conllevan segmentaciones distintas. El ejemplo arquetípico puede ser la palabra "verse": "los hechos sobre los que verse la prueba" (Ley del Tribunal del Jurado) vs "impedir que tal principio pueda verse falseado por prácticas desleales" (Ley de competencia desleal). Si "verse" corresponde al lema "versar" se interpreta como una única palabra, si viene del lema "ver" se interpreta como secuencia de verbo y pronombre enclítico.

Una ambigüedad como la de "verse", "pésame" o "consigo" no es a priori más dificil de resolver que la de "sobre" o "fuera". Pero en la estrategia de lematización y desambiguación que hemos aplicado al Corpus Técnico del IULA[1] nos crea una contradicción: primero lematizamos y categorizamos[2] y, luego, analizando secuencias de formas, de lemas y de etiquetas morfológicas, desambiguamos[3]; pero con las ambigüedades segmentales no podemos lematizar y categorizar hasta después de haber

[1] Véase Bach *et al.* (1997)

[2] Véase De Yzaguirre *et al.* (2000a)

[3] Véase De Yzaguirre *et al.* (2000b)

desambiguado.

2.- Ejemplos del Corpus del IULA

Para situar mejor el tema, presentamos a continuación todas las ocurrencias de “verse” de documentos jurídicos de nuestro Corpus

```
d00048 ITEM      las fincas sobre las que verse la notificación.
d00048 ITEM      las fincas sobre las que verse la notificación.
d00056 S         los hechos sobre los que verse la prueba.
d00057 ITEM  materia respectiva sobre que verse la reclamación, mediante la
d00060 ITEM      los hechos sobre los que verse la pretensión y de todos
d00060 S             el objeto del debate verse sobre preferencias atribuidas
d00060 S             el objeto del debate verse sobre preferencias atribuidas a
d00060 S    se aplicará cuando el recurso verse sobre prestaciones de la Seguridad Social
d00162 S           recaídas en juicio que verse sobre capacidad, estado civil
d00032 S  impedir que tal principio pueda verse falseado por prácticas desleales,
d00081 ITEM      a los Estados que puedan verse afectados por esos peligros o
d00091 S               Como máximo puede verse que el nuevo concurso favorece
d00107 S relación a los autonómicos puede verse, por todos, Zornoza Pérez
d00110 S   las Partes Contratantes puedan verse obligadas a adoptar por razón
d00146 S    de motivación pues como puede verse en el primer considerando de
d00146 ITEM                        Que al verse privado el apelante, por
d00146 ITEM                        Que al verse privado el apelante, por
d00158 S             fácil reparación, al verse prácticamente desposeído de un terreno
d00159 S                L.J. ), que puede verse afectado, favorable o desfavorablemente
d00159 S  por la finalidad disculpable de verse exonerado de toda responsabilidad-,
d00161 S principio fundamental que podría verse comprometido en caso de acordar
d00162 S  una mejor prestación que podría verse deteriorada con una excesiva proliferación
d00162 S  la de Cirujano-Callista , puede verse involucrado y perjudicado por mudables
d00171 ITEM      Monetario Europeo deberá verse reflejado en los niveles de
d00242 S          y sociales que pudieran verse afectados por sus actividades.
```

El d00162 (véanse las referencias al final del escrito) tiene “ver” y “versar”.

3.- Delimitación del tema

Otras lenguas románicas tienen problemas similares: el portugués “do”, el francés “des”, el catalán “cal” presentan el mismo conflicto. Pero en castellano la ambigüedad segmental tiene una magnitud mayor, siendo la mayor parte de los casos similares a “verse” (verbo+pronombre vs forma única). De las lenguas citada, sólo el castellano no separa los enclíticos del verbo. Por otro lado, la mayoría de los casos que hemos detectado tienen poca transcendencia; sólo algunos pocos merecen nuestra atención; pero si se incorpora al lematizador un mecanismo para resolver las ambigüedades segmentales interesantes, vale la pena aplicarlo de manera general.

Para analizar con pretensiones de exhaustividad el problema, lematizamos con PALIC las casi 900000 formas que se derivan del lemario del DALE con una configuración peculiar que puso de manifiesto las que podían tener una segunda interpretación con dos segmentos. Con ello obtuvimos una relación de 160 casos hasta entonces inadvertidos

como “tétanos”, “cáseos”, “barreros”, “velarte”, etc.

4.- Clasificación de los casos hallados

Hemos procedido a clasificar los casos de ambigüedad segmental como sigue:

4.1. Posibles:

Algunos de ellos son teóricamente posibles, aunque es poco probable que aparezcan en textos modernos, puesto que el nombre que denotan tiene un referente anticuado (por ejemplo, los nombres de oficios o *velarte*). En textos escritos, son más improbables las formas en las que el verbo está en imperativo que aquellas en las que el verbo está en infinitivo.

barreros,S051{2,barrer,barrer,VI----,os,[pr],REE626P}{1,barreros,barrero,N5-MP}
moleros,S093{2,moler,moler,VI----,os,[pr],REE626P}{1,moleros,molero,N5-MP}
placeros,S100{2,placer,placer, VI----,os,[pr],REE626P}{1, placeros,placero,N5-MP}
prenderos,S151{2,prender,prender,VI----,os,[pr],REE626P}{1,prenderos,prendero,N5-MP}
solerte,S116{2,soler,soler,VI----,te,[pr],REE626S}{1,solerte,solerte,JQ--6S}
tejeros,S120{2,tejer,tejer,VI----,os,[pr],REE626P}{1,tejeros,tejero,N5-MP}
tenderos,S121{2,tender,tender,VI----,os,[pr],REE626P}{1,tenderos,tendero,N5-MP}
trágala,S125{2,trága,tragar,V8RS6-,la,[pr],REEC3FS}{1,trágala,trágala,N5-MS}
trágalas,S126{2,trága,tragar,V8R6S-,las,[pr],REEC3FP}{1,trágalas,trágala,N5-MP}
velarte,S135{2,velar,velar,VI----,te,[pr],REE626S}{1,velarte,velarte,N5-MS}
verse,S035{2,ver,ver,VI----,se,[pr],R6-----}{1,verse,versar,V9R6S}

4.2. Posibles sólo en textos específicos:

Algunas de las palabras pueden presentar ambigüedad tan sólo en textos especializados, puesto que uno de sus dos lemas en un término.

Textos de materias científicas (biología, botánica, geología y medicina):
amaros,S001{2,amar,amar, VI----,os,[pr],REE626P}{1,amaros,amaro, N5-MP}
asómate,S050{2,asóma,asomar,V8R6S-,te,[pr],REE626S}{1,asómate,asómate,N5-MS}
buscarla,S055{2,buscar,buscar,VI----,la,[pr],REEC3FS}{1,buscarla,buscarla,N5-FS}
buscarlas,S139{2,buscar,buscar,VI----,las,[pr],REEC3FP}{1,buscarlas,buscarla,N5-FP}
cálaos,S004{2,cála,calar,V8R6S-,os,[pr],REE626P}{1,cálaos,cálao,N5-MP}
(13 más)

Otros:
a) Textos de historia:
ámala,S037{2,áma,amar, V8R6S-,la,[pr],REEC3FS}{1,ámala,ámalo,N5-FS}
ámalas,S038{2,áma,amar, V8R6S-,las,[pr],REEC3FP}{1,ámalas,ámalo,N5-FP}

ámalo,S039{2,áma,amar, V8R6S-,lo,[pr],REEC3MS}{1,ámalo,ámalo,N5-MS}
ámalos,S040{2,áma,amar, V8R6S-,los,[pr],REEC3MP}{1,ámalos,ámalo,N5-MP}

b) Derecho:
cúmplase,S071{2,cúmpla,cumplir, V9R6S-,se,[pr],R6-----}{1,cúmplase,cúmplase,N5-MS}

c) Marina:
tésalo,S030{2,tésa,tesar, V8R6S-,lo,[pr],REEC3MS}{1,tésalo,tésalo,JQ-MS}

4.3. Textos dialectales:

En algunos casos, uno de los lemas existe sólo en algunas variantes dialectales del español (generalmente, pero no únicamente, en el español de América).

caberos,S059{2,caber,caber,VI----,os,[pr],REE626P}{1,caberos,cabero,N5-MP}
chócola,S062{2,chóco,chocar,VDR1S-,la,[pr],REEC3FS}{1,chócola,chócolo,N5-FS}
chócolas,S063{2,chóco,chocar,VDR1S-,las,[pr],REEC3FP}{1,chócolas,chócolo,N5-FP}
chócolo,S064{2,chóco,chocar,VDR1S,lo,[pr],REEC3MS}{1,chócolo,chócolo,N5-MS}
(15 más)

4.4. Textos coloquiales:

Bajo este epígrafe se recogen los casos en los que uno de los dos lemas (el verbo) pertenece a niveles o registros orales:

dormirlas,S141{1,dormirlas,dormirlas,N5-MS}{2,dormir,dormir,VI----,las,[pr],REEC3FP}
lárgalo,S088{1,lárgalo,lárgalo,N5-MS}{2,lárga,largar, V8R6S-,lo,[pr],REEC3MS}
lárgalos,S089{1,lárgalos,lárgalo,N5-MP}{2,lárga,largar, V8R6S-,los,[pr],REEC3MP}
pétalo,S014{1,pétalo,pétalo,N5-MS}{2,péta,petar, V8R6S-,lo,[pr],REEC3MS}

4.5. Anticuados:

Los verbos de las siguientes formas son anticuados, y no es probable que aparezcan en textos modernos.

caleros,S060{1,caleros,calero,N5-MP}{2,caler,caler,VI----,os,[pr],REE626P}
óvalo,S043{1,óvalo,óvalo,N5-MS}{2,óva,ovar, V8R6S-,lo,[pr],REEC3MS}
óvalos,S044{1,óvalos,óvalo,N5-MP}{2,óva,ovar, V8R6S-,los,[pr],REEC3MP}

4.6. El orden de los pronombres los hace anticuados:

En el español moderno, las combinaciones de forma verbal + pronombre enclítico se producen únicamente con el infinitivo o con las formas del imperativo. Por ello, la aparición de pronombres enclíticos con formas verbales distintas de éstas se puede producir tan sólo en textos antiguos o con un lenguaje arcaizante.

acabóse,S047{1,acabóse,acabóse,N5-MS}{2,acabó,acabar,VDP3S-,se,[pr],R6}
acúleos,S048{1,acúleos,acúleo,N5-MP}{2,acúle,acular,V9R6S-,os,[pr],REE626P}
apóstoles,S137{1,apóstoles,apóstol,N5-MP}{2,apósto,apostar,VDR1S-
,les,[pr],REE636P}
argénteos,S138{1,argénteos,argénteo,N5-MP}{2,argénte,argentar,V9R6S-
,os,[pr],REE626P}
átonos,S042{1,átonos,átono,JQ-MP}{2,áto,atar,VDR1S-,nos,[pr],REE616P}
brótanos,S052{1,brótanos,brótano,N5-MP}{2,bróta,brotar,V8R6S-,nos,[pr],REE616P}
(57 más)

4.7. No son posibles:

En algunos casos, el analizador detecta como posibles algunas formas que, en rigor, no son posibles, puesto que el verbo no rige los complementos representados por el pronombre (verbos intransitivos, etc.).

alagarte,S049{2,alagar,alagar,VI----,te,[pr],REE626S}{1,alagarte,alagarte,N5-MS}
bálanos,S056{2,bála,balar,V8R6S-,nos,[pr],REE616P}{1,bálanos,bálano,N5-MP}
brótola,S053{2,bróto,brotar,VDR1S-,la,[pr],REEC3FS}{1,brótola,brótola,N5-FS}
brótolas,S054{2,bróto,brotar,VDR1S-,las,[pr],REEC3FP}{1,brótolas,brótola,N5-FP}
búfalo,S002{2,búfa,bufar,V8R6S-,lo,[pr],REEC3MS}{1,búfalo,búfalo,N5-MS}
(14 más)

4.8. Otros:

Se han de eliminar de la lista que teníamos (los analiza mal: sobra el acento):
salme,S026{2,sal,salir,VRR2S-,me,[pr],REE616S}{1,salme,salme,N5-MS}
salte,S027{2,sal,salir,VRR2S-,te,[pr],REE626S}{1,salte,salte,N5-MS}
tésala,S029{2,tésa,tesar,V8R6S-,la,[pr],REEC3FS}{1,tésala,tésala,JQ-FS}
ásaros,S041{2,ásar,ásar,VI----,os,[pr],REE626P}{1,ásaros,ásaros,N5-MP}
cesáreos,S061{cesáre,cesar,VJU6S-,os,[pr],REE626P}{1,cesáreos,cesáreos,JQ-MP}
calcáreos,S140{calcáre,calcar,VJU6S-,os,[pr],REE626P}{1,calcáreos,calcáreos,JQ-MP}
cóconos,S070{cóco,cocar,VDR1S-,nos,[pr],REE616P}{1,cóconos,cóconos,N5-MP}
folíolo,S079{folío,foliar,VDR1S-,lo,[pr],REEC3MS}{1,folíolo,folíolo,N5-MS}
folíolos,S080{folío,folío,VDR1S-,los,[pr],REEC3MP}{1,folíolos,folíolos,N5-MP}
sobresalte,S153{sobresal,sobresalir,VRR2S-,te,[pr],REE626S}
{1,sobresalte,sobresalte,N5-MS}

5.- Interpretación

El formalismo de los casos precedentes se interpreta así:

velarte,S135{2,velar,velar,VI,te,[pr],REE626S}{1,velarte,velarte,N5MS}

La forma “velarte”, código S135, tiene dos interpretaciones:
– la primera {2,velar,velar,VI,te,[pr],REE626S} está formada por dos segmentos
– la segunda {1,velarte,velarte,N5MS} por un solo segmento

Para cada segmento se da forma, lema y etiqueta[4].

6.- Tratamiento

Nuestra manera de resolver las ambigüedades segmentales es la siguiente:

1.- PALIC detecta estos casos como primer paso de la lematización (para evitar que sean tratados como el resto de las palabras) y les asigna como lema el código de ambigüedad segmental (v.g. S135).

2.- AMBILIC posee un paquete de reglas especiales para los casos resolubles (v.g. artículo + que + verse => un solo segmento; verbo + verse => dos segmentos). Si una de sus reglas cumple las condiciones, AMBILIC lematiza y desambigua con la información pertinente.

3.- Una de las últimas rutinas resuelve ciegamente los casos que no se han resuelto por regla, aplicando la segmentación que nuestros expertos han preferido, que es la que aparece en primera posición.

Las reglas de desambiguación són idénticas a las descritas en AMBILIC excepto en dos puntos:
- pueden tener como lema el código, como S135
- la parte ejecutiva de la regla es “segment=X” donde X es 1 o 2 y se refiere a la primera o a la segunda interpretación

7.- Bibliografía

BACH, C. ; SAURÍ, R.; VIVALDI, J. y M.T. CABRÉ (1997) "El Corpus de l'IULA: descripció", *Papers de l'IULA*, sèrie informes, 17, Barcelona: IULA, Universitat Pompeu Fabra.

[4] Para el etiquetario, véase Morel *et al.* (1997)

DE YZAGUIRRE, L. , A. MATAMALA y T. CABRÉ (2000a): El lematizador "PALIC" del IULA (UPF), comunicación presentada al XVII Congreso AESLA, panel sobre Lingüística de Corpus y Computacional.

DE YZAGUIRRE, L., A. MATAMALA, C. BACH, N. CASTILLO y E. USTRELL (2000b): AMBILIC, el desambiguador lingüístico del corpus del IULA (UPF), comunicación presentada al XVII Congreso AESLA, panel sobre Lingüística de Corpus y Computacional.

MOREL, J.; TORNER, S; VIVALDI, J; DE YZAGUIRRE, Ll. y M.T. CABRÉ (1997) "El corpus de l'IULA: etiquetaris", *Papers de l'IULA*, sèrie informes, 18, Barcelona: IULA, Universitat Pompeu Fabra.

8.- Referencia del Corpus Técnico:

# d00032 (1)	Ley de competencia desleal
# d00048 (2)	Real Decreto 1684/1990 de recaudación de tributos
# d00056 (1)	Ley del Tribunal del Jurado
# d00057 (1)	Ley 230/1963. Ley general tributaria
# d00060 (4)	Procedimiento laboral. Texto refundido de la ley
# d00081 (1)	Convenio del 5-6-1992 sobre diversidad biológica
# d00091 (1)	La organización administrativa del Estado
# d00107 (1)	Autonomies. Revista Catalana de Derecho Público
# d00110 (1)	Convenio internacional sobre la amortización de los controles de mercancías en las fronteras (BOE 48-25-2-86)
# d00146 (3)	Jurisprudencia TS-Contencioso Administrativo. Primer Semestre 1993 (Parte 2)
# d00158 (1)	Jurisprudencia TS-Contencioso Administrativo. Segundo semestre 1992 (Parte 1)
# d00159 (2)	Jurisprudencia TS-Contencioso Administrativo. Segundo semestre 1992 (Parte 2)
# d00161 (1)	Jurisprudencia TS-Constencioso Administrativo. Primer semestre 1992 (Parte 1)
# d00162 (3)	Jurisprudencia del TS-Contencioso Administrativo. Primer semestre 1992 (Parte 2)
# d00171 (1)	Tratado de Maastricht
# d00242 (1)	Ley 21/1992 de 16 de Julio "Ley de industria"