MANUAL DE EDICIÓN Y ANOTACIÓN EN TEITOK DE LOS MATERIALES DE P. S. POST SCRIPTUM:

- Edición modernizada
- Anotación morfosintáctica (POS)
- Anotación sintáctica (en portugués)

Gael Vaamonde y Catarina Magro

Última actualización: 22/03/2018

Tabla de contenido

1. Edición modernizada

La edición modernizada de los textos cumple un doble propósito en *P. S. Post Scriptum*:

- Facilitar la anotación morfosintáctica de carácter automático.
- Ofrecer al público lego una edición de fácil lectura.

Partiendo de la edición original del texto (i.e. edición semipaleográfica), la edición modernizada se obtiene mediante la intervención en dos, y sólo dos, aspectos lingüísticos:

- La normalización ortográfica de las formas originales, incluyendo la acentuación y la inserción de mayúsculas donde corresponda.

- La puntuación del texto de acuerdo con las normas de puntuación de la lengua contemporánea, excepción hecha de la división en párrafos, que se mantiene fiel al texto original.

Las soluciones de las conjeturas en el original (marcadas con el elemento <supplied> en XML-TEI) son integradas en el texto modernizado, mientras que los segmentos omitidos (marcados con el elemento <gap>) se marcan con puntos suspensivos entre corchetes ([...]).

Las modificaciones realizadas sobre el texto se ciñen únicamente al nivel ortográfico, por lo que no se eliminan ni añaden palabras respecto del contenido original de la carta. Tampoco se interviene sobre el nivel léxico: se conservan los regionalismos y los arcaísmos léxicos, así como cualquier otra forma léxica no estándar, si bien estos casos son tratados en un nivel independiente para facilitar su recuperación (cf. apartado 1.2.2.4.)

El trabajo de modernización del texto se realiza a través de la plataforma TEITOK y forma parte de un proceso más amplio de revisión, adición y edición de información textual y extratextual. Este proceso se puede dividir en tres partes fundamentales:

- Tareas de preedición
- Edición del texto
- Tareas de posedición

1.1. Tareas de preedición

1.1.1. Revisión general del fichero XML

En esta fase, se revisa íntegramente el contenido del fichero XML. El objetivo es depurar posibles errores en los metadatos (<teiHeader>) o en la transcripción (<text>). A modo de referencia, téngase en cuenta que el contenido del fichero debe cumplir las directrices que se recogen en la *Guía para la edición digital de textos en P. S. Post Scriptum*.

El responsable de la revisión debe constar en el siguiente apartado:

```
<respStmt><resp subcat="revision"><name></name></resp></respStmt>
```

1.1.2. Añadir cambios de línea

Es posible que en la transcripción XML del texto falten cambios de línea. Estos se pueden añadir directamente desde la plataforma TEITOK. Para añadir cambios de línea, es necesario hacer lo siguiente:

- Pinchar en el token posterior al cambio de línea que se quiere añadir
- En la ventana de edición, pinchar en el enlace que aparece destacado en la Figura 1:

Figura 1. Añadir cambio de línea.

Como resultado, en el fichero XML se creará un nuevo elemento <lb/> a continuación del token seleccionado.

Téngase en cuenta que el elemento <lb/> aparecerá por defecto pegado al token seleccionado. Sin embargo, la estrategia correcta para marcar cambios de línea que separan dos palabras es que dichos elementos aparezcas delimitado por espacio. Por tanto, es preferible añadir los cambios de línea manualmente en el editor XML Oxygen. Tampoco se recomienda utilizar el enlace de cambio de párrafo en TEITOK, puesto que este creará un elemento <sb/> en el fichero XML, cuando la estrategia correcta para cambios de párrafo es .

1.1.3. Contracciones no estándar y separación de tokens

Como regla general, en el momento de la transcripción del texto ya se normaliza la frontera de palabra. Por tanto, generalmente la división de palabras que se haya aplicado en la transcripción se corresponderá con la división de tokens generados al tokenizar al texto (cf. 1.1.4). No obstante, conviene tener en cuenta algunos casos que pueden ser problemáticos.

En el caso de los numerales y de las contracciones estándar, es necesario respetar la frontera de palabra del manuscrito original, pues lo contrario podría implicar que se falsee la historia de las contracciones. En el caso de las contracciones no estándar, se aplica la frontera de palabras contemporánea, lo que implica diferenciar tantos tokens como sean necesarios para la correcta modernización ortográfica. Por tanto:

- Numerales:

Original	Transcripción	Normalización
veinte y un	veinte y un	veinte y un
veinteyun	veinte y un	veinte y un
diez y seis	diez y seis	diez y seis
diezyseis	diezyseis	dieciséis

- Contracciones estándar:

Original	Transcripción	Normalización
de el	de el	de el
deel	de el	de el
del	del	del
a el	a el	a el
ael	a el	a el
al	al	al

- Contracciones no estándar:

Original	Transcripción	Normalización
avido tempestades	a vido tempestades	ha habido tempestades
fue alcala	fue a lcala	fue a Alcalá
le escrito cartas	le e scrito cartas	le he escrito cartas
del	d el	de él (prep. + pron.)
laspereza	l aspereza	la aspereza
deste	d este	de este
astaora	ast aora	hasta ahora
ques posible	qu es posible	que es posible
mescribio	m escribio	me escribió

Las modernizaciones de las contracciones no estándar (e.g. *a > ha; d > de; l > la; m > me; que > que;...*) se realizan dentro del atributo **@nform** y no dentro del atributo **@fform**, puesto que no se trata de abreviaturas. En ningún caso se usa apóstrofo.

1.1.4. Tokenización del texto

En esta fase, se realiza la tokenización automática del texto. Para ello, una vez importado el fichero a la plataforma TEITOK, es necesario pinchar en el enlace que aparece destacado en la Figura 2:

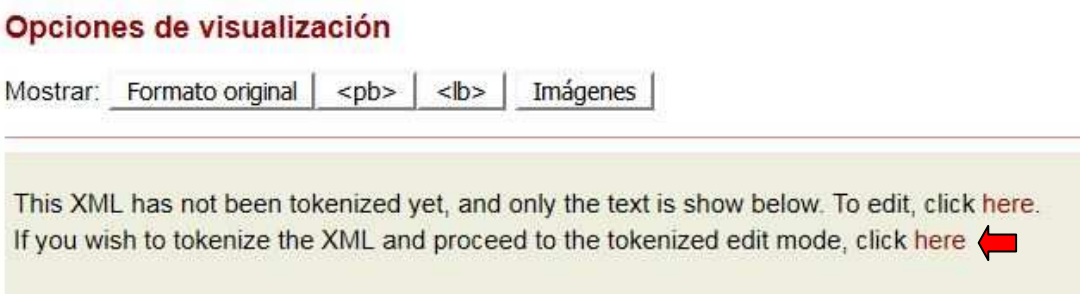


Figura 2. Tokenización automática

Como resultado, cada token original (i.e. palabras y signos de puntuación) es incluido dentro en un elemento <tok> al que se le asigna una identificación única.

```
<tok id="w-4">amigo</tok>
```

En el caso de formas que incluyen etiquetas XML, la forma original (i.e. sin etiquetas XML) aparecerá como valor del atributo **@form**:

```
<tok form="amigo" id="w-4">amig<add hand="FA3" place="supralinear">o</add></tok>
```

En el caso de abreviaturas, el token resultante aparecerá con la abreviatura original dentro del elemento <tok> y la correspondiente expansión aparecerá como valor del atributo **@fform**:

```
<tok fform="amigo" id="w-4">amo</tok>
```

Todas las etiquetas XML utilizadas en el proceso de transcripción del texto aparecerán siempre fuera del elemento <tok>

```
<supplied><tok fform="amigo" id="w-4">amo</tok></supplied>

<unclear><tok fform="amigo" id="w-4">amo</tok></unclear>

<hi><tok fform="amigo" id="w-4">amo</tok></hi>
```

Se contemplan solo cuatro excepciones a esta regla:

- Cambio de línea en interior de palabra

```
<tok form="amigo" id="w-4">a<lb subcat="false"/>migo</tok>
```

- Adición de grafías en una de palabra

```
<tok form="amigo" id="w-4">a<add hand="FR4" place="supralinear"/>migo</add></tok>
```

- Supresión de grafías en una palabra

```
<tok form="amigo" id="w-4">a<del hand="FR4"/>n</del>migo</tok>
```

- Parte de una palabra subrayada

```
<tok form="amigo" id="w-4">a<hi rend="underlined"/>migo</hi></tok>
```

1.1.5. Depuración de errores

En ocasiones, la tokenización automática puede producir inconsistencias en el texto XML si este no fue debidamente transcrito/revisado. Por ejemplo, la falta de un espacio entre el elemento de cambio de página (<pb>) y la palabra precedente producirá que dicho elemento aparezca dentro del token resultante:

```
<tok id="w-4">amigo<pb id="e-2"/></tok>
```

Para depurar este tipo de errores, una vez tokenizado el texto es necesario pinchar en el enlace que aparece destacado en la Figura 3. Como resultado, el ejemplo anterior aparecerá corregido en el XML del modo siguiente:

```
<tok id="w-4">amigo</tok><pb id="e-2"/>
```

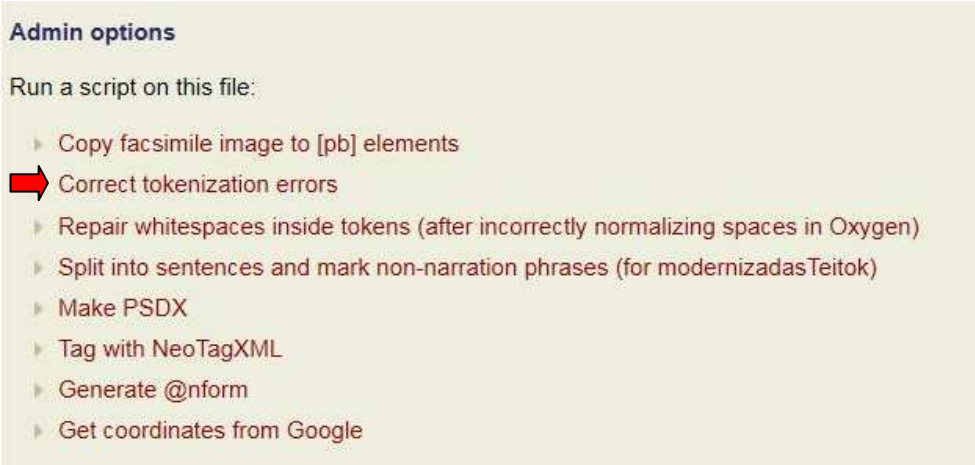


Figura 3. Depuración automática de errores de tokenización

Finalmente, el enlace que aparece destacado en la Figura 4 permite eliminar posibles espacios, tabulaciones o cambios de línea sobrantes dentro del elemento <tok>, tal como ilustramos a continuación:

```
Input:    <tok      id="w-4">amigo</tok>
Output    <tok id="w-4">amigo</tok>
```

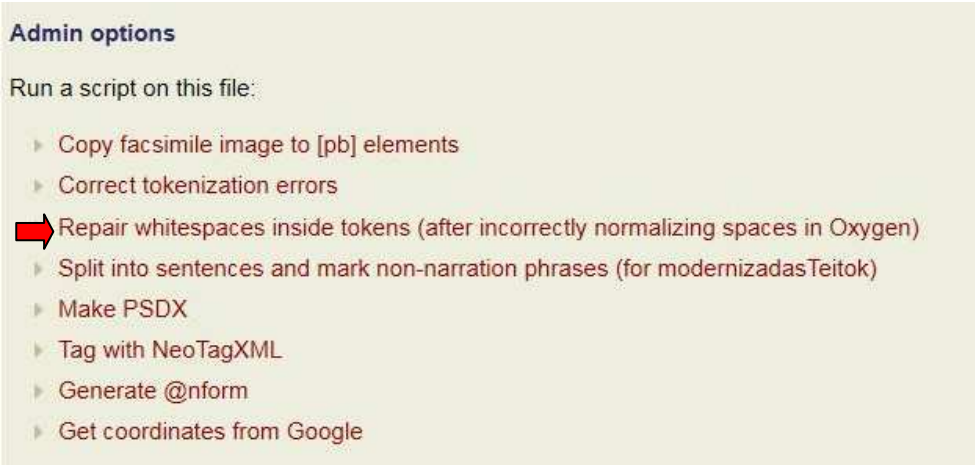


Figura 4. Eliminación automática de espacios dentro de tokens

1.1.6. Importación de imágenes

En esta fase, las imágenes correspondientes al texto que va a ser modernizado son importadas a la plataforma TEITOK. Para ello, basta con pinchar en el enlace que aparece destacado en la Figura 5:

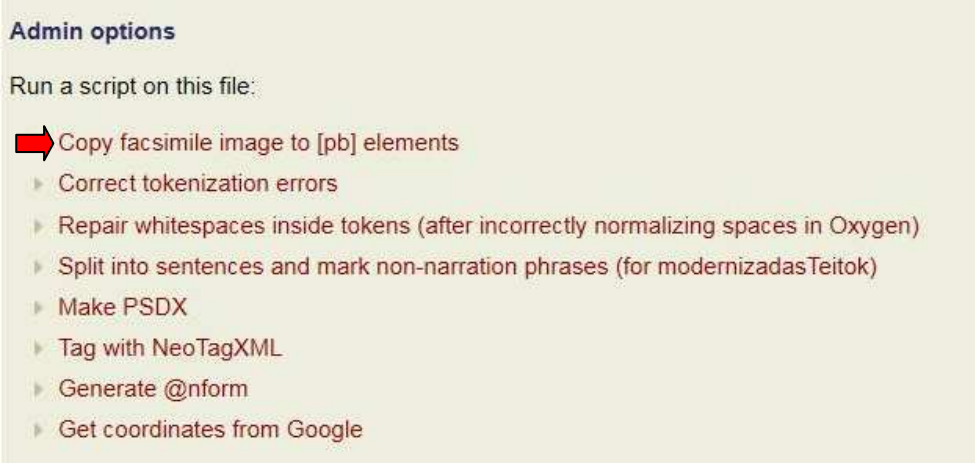


Figura 5. Importación automática de imágenes

Para esta importación automática, el script vincula dos tipos de información dentro del fichero XML:

- El número de imágenes declaradas dentro del elemento <surrogates>. Por ejemplo:

```
<surrogates>
        <p>
                <bibl>facsímile digital guardado como fichero JPEG</bibl>
                <bibl><xref subcat="PSCR6129_1.JPG"/></bibl>
                <bibl><xref subcat="PSCR6129_2.JPG"/></bibl>
                <bibl><xref subcat="PSCR6129_3.JPG"/></bibl>
                <bibl><xref subcat="PSCR6129_4.JPG"/></bibl>
        </p>
</surrogates>
```

- El número de elementos <pb> declarados en el texto.

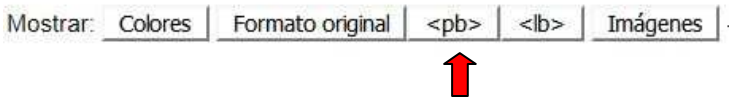
Dentro de cada elemento <pb> declarado se crea automáticamente un atributo **@facs** con la referencia de la imagen correspondiente de forma correlativa:

```
<pb n="[314]r" id="e-1" facs="PSCR6129_1"/>
<pb n="[314]v" id="e-2" facs="PSCR6129_2"/>
<pb n="[315]r" id="e-3" facs="PSCR6129_3"/>
<pb n="[315]v" id="e-4" facs="PSCR6129_4"/>
```

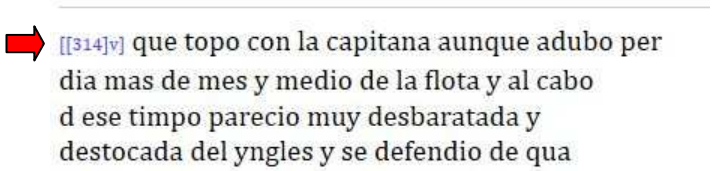
En caso de que el número de imágenes y el número de elementos <pb> no coincidan, el script desechará automáticamente la imagen número 1, que es el dígito reservado para la imagen del sobrescrito cuando este existe.

Si en el momento de la importación todavía no se cuenta con el permiso de publicación de las imágenes, se deben ocultar al público. Para ello, es necesario hacer lo siguiente:

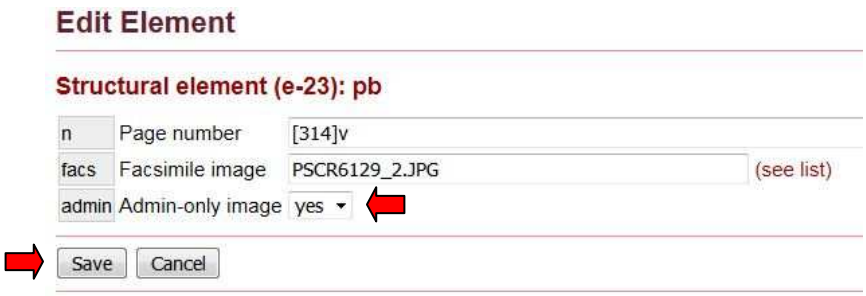
- Mostrar la visualización de cambios de página pinchando en el enlace siguiente:



- Pinchar en cada uno de los enlaces que marcan un cambio de página:



- Escoger el valor *yes* en el campo *admin*. Guardar:



Como resultado, en el fichero XML se habrá adicionado un atributo **@admin** con el valor **1** en cada elemento <pb>:

```
<pb n="[314]r" id="e-1" facs="PSCR6129_1" admin="1"/>
```

Las imágenes ocultadas al público aparecerán con el borde del facsímil en rojo.

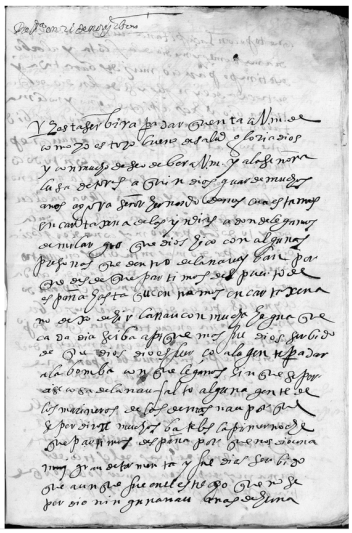


Figura 6. Imagen publicada.

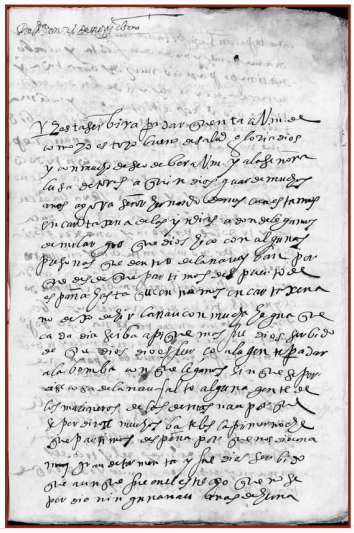


Figura 7. Imagen no publicada: **@admin**="1".

1.1.7. Importación de coordenadas geográficas

En esta fase, se importan al fichero XML las coordenadas geográficas del lugar de origen de la carta. Para ello, es necesario pinchar en el enlace que aparece destacado en la Figura 8:

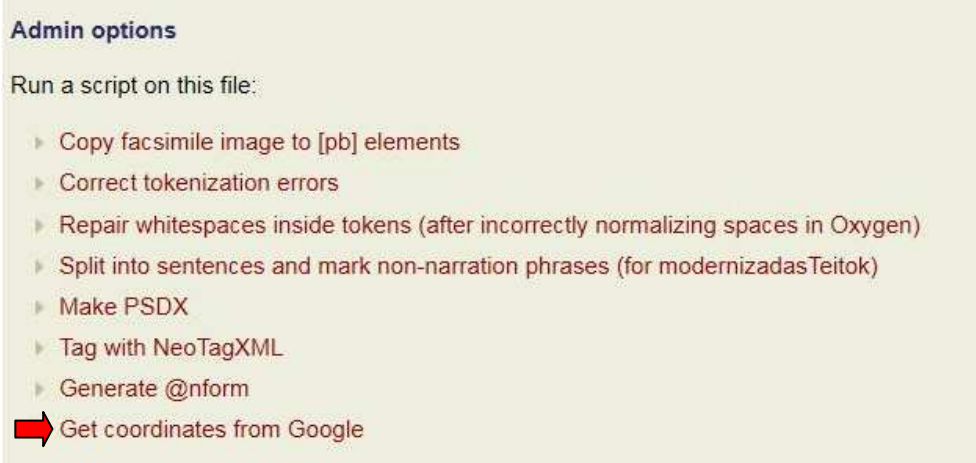


Figura 8 Importación automática de coordenadas geográficas.

Como resultado, en el elemento <placeLet> se creará un nuevo atributo **@geo**, con los valores de latitud (eje Y) y longitud (eje X) separados por una coma y un espacio en blanco:

```
<placeLet attested="yes" geo="39.874968, -4.049044">España, Toledo</placeLet>
```

1.2. Edición del texto

En esta fase, se realiza la edición modernizada del texto de acuerdo con las consideraciones señaladas anteriormente (cf. apartado 1). Este proceso se divide en dos partes: normalización automática y revisión manual

1.2.1. Normalización automática

La normalización automática del texto se realiza pinchando en el enlace que aparece destacado en la Figura 9:

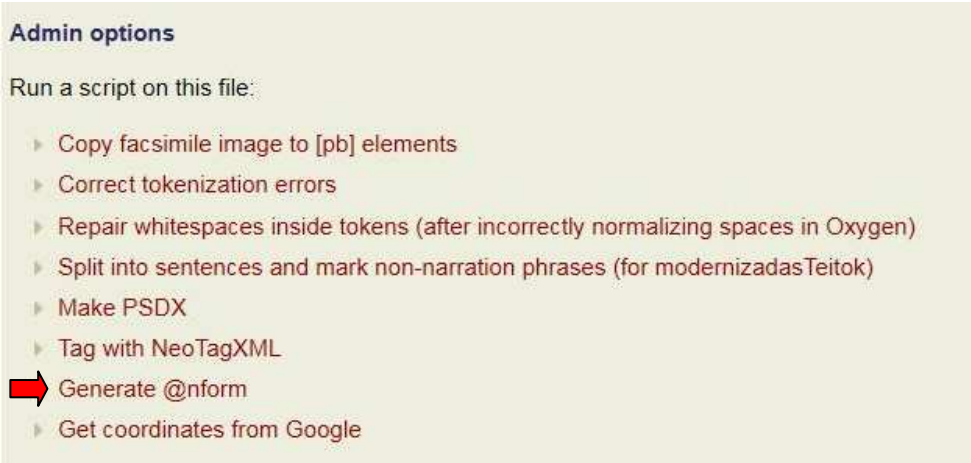


Figura 9. Normalización automática.

Como resultado, en el elemento <tok> de aquellos tokens cuya grafía haya sido automáticamente editada aparecerá un atributo **@nform** con el valor de la forma normalizada correspondiente:

```
<tok id="w-1" nform="vergüenza">berguenza</tok>
```

Téngase en cuenta que la normalización automática se ciñe a corregir la forma ortográfica de las palabras, incluyendo acentos y mayúsculas. No añade ni elimina signos de puntuación, ni opera en otros niveles de análisis como son la marcación de variedades no estándar o la aplicación de palabras clave. Esa información debe tratarse en la revisión manual.

1.2.2. Revisión manual

El resultado del script de modernización automática debe ser revisado manualmente. Esta revisión incluye las tareas siguientes:

- Editar tokens
- Añadir o eliminar puntuación
- Marcar variedades no estándar
- Marcar palabras clave de lingüística

1.2.2.1. Editar tokens

Para editar aquellos tokens que no han sido corregidos automáticamente (o que han sido tratados de forma incorrecta), es necesario acceder a la ventana de edición de tokens. Para ello, hay que pinchar en el token que se quiere editar, lo que nos devolverá una ventana como la que aparece en la Figura 10. Se explican a continuación brevemente cada una de estas opciones. Téngase en cuenta que la edición de la normalización se corresponde con la opción ***nform Standardization*** que aparece en dicha figura.

Figura 10. Niveles de edición de token.

- XML Raw XML value. La forma original tal como fue transcrita en el fichero XML (i.e. incluyendo, si fuera el caso, etiquetas XML). Es decir, equivale al contenido incluido dentro del elemento <tok>:

```
<tok form="amigo" id="w-4">amig<add hand="FA3" place="supralinear">o</add></tok>
```

- form. La forma original sin etiquetas XML. Se trata de un atributo obligatorio siempre que la forma original incluya elementos XML.

```
<tok form="amigo" id="w-4">amig<add hand="FA3" place="supralinear">o</add></tok>
```

- fform. La forma expandida de una abreviatura. Se trata de un atributo obligatorio siempre que la forma original sea una abreviatura.

  <tok fform="amigo" id="w-4">amo</tok>

- dipl. Nivel creado automáticamente como valor de un atributo oculto **@dipl**. Contiene el material gráfico no explícito de una abreviatura. Se calcula automáticamente restando el valor de **@form** al valor de **@fform**.

- dform. Variante no estándar (cf. apartado 1.2.2.4.).

  <tok dform="agora" nform="ahora" id="w-4">amo</tok>

- nform. La forma normalizada. Se trata de un atributo obligatorio siempre que exista una forma no estándar

  <tok nform="vergüenza" id="w-4">berguença</tok>

- ltags. Palabra clave de carácter lingüístico (cf. apartado 1.2.2.5.).

  <tok nform="asegurar" ltags="unstressed_vowels" id="w-4">asigurar</tok>

Sobre el resto de opciones, que están relacionadas con la anotación morfosintáctica, consúltese el apartado 2.2.2.1.

1.2.2.2. Añadir puntuación

Para añadir puntuación es necesario pinchar en el token anterior o posterior al signo que se quiere añadir. En la ventana de edición de token, habrá que pinchar en uno de los cuatro enlaces (destacados en rojo) que aparecen en la Figura 11, según corresponda:

insert tok after: attached / separate • before: attached / separate •

Figura 11. Añadir puntuación (I).

A continuación, aparecerá una nueva ventana de edición correspondiente a un nuevo token. El contenido de ese token será siempre <ee/> (i.e. *empty element*), que aparecerá automáticamente. El signo de puntuación correspondiente se añadirá en el nivel de normalización (i.e. nfrom):

XML	Raw XML value	<ee/>
form	Provisional transcription	
fform	Expanded/Free form	
dipl	Edition	
dform	Variant form	
nform	Standardization	,

Figura 12. Añadir puntuación (II).

Como resultado, en el fichero XML se habrá creado un nuevo token con la forma siguiente:

```
<tok nform="," id="w-4"><ee/></tok>
```

1.2.2.3. Eliminar información del texto

La eliminación de texto desde la plataforma TEITOK implica siempre el uso de la secuencia de caracteres **--**. Existen al menos cuatro situaciones en las que es necesario eliminar información, repartidas en dos estrategias diferentes según el nivel de edición al que se quiere aplicar la eliminación:

1.2.2.3.1. En el nivel de transcripción: @form

El uso de la secuencia **--** en el nivel de transcripción se utiliza en las dos situaciones siguientes:

- Información que está tachada en el texto original. Se trata de tokens marcados con el elemento <del> en el fichero XML. Estos casos se tratan automáticamente al tokenizar el texto en TEITOK. Sólo se visualizan en el nivel de transcripción y aparecen en el fichero XML del modo siguiente:

```
<del hand="FG4"><tok form="--" id="w-4">amigo</tok></del>
```

- Información que se conjetura como tachada en el texto original. Se trata de casos marcados con la combinación de elementos <supplied><del> en el fichero XML. Estos casos se tratan automáticamente al tokenizar el texto en TEITOK. Sólo se visualizan en el nivel de transcripción y aparecen en el fichero XML del modo siguiente:

```
<supplied resp="GV" reason="abandoned"><del hand="FG4"><tok form="--" id="w-4">amigo</tok></del></supplied>
```

1.2.2.3.2. En el nivel de normalización: @nform

El uso de la secuencia **--** en el nivel de normalización se utiliza en las dos situaciones siguientes. Ambas forman parte del proceso de revisión manual:

- Puntuación que se quiere eliminar. Es necesario pinchar en el signo de puntuación que corresponda y, en la ventana de edición del token, añadir la secuencia de caracteres **--** en el nivel de normalización:

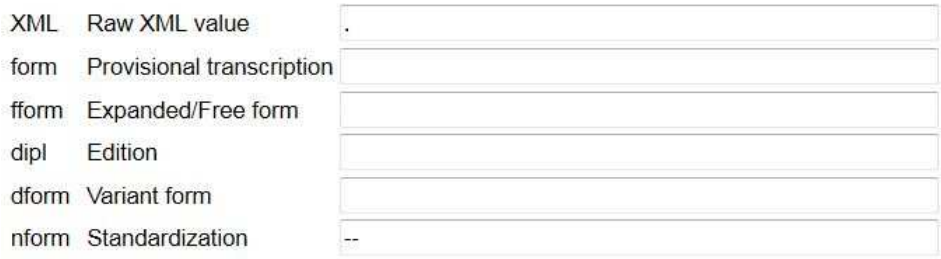


Figura 13. Eliminar puntuación.

Como resultado, el token correspondiente desaparecerá de la visualización de la edición modernizada, y en elemento correspondiente dentro del fichero XML aparecerá del modo siguiente:

```
<tok nform="--" id="w-4">.</tok>
```

- Repetición intencional de palabras. Se trata de la repetición de la misma *palabra* al final de un folio y al inicio del folio siguiente (*reclamos* o llamadas). La forma repetida que se debe eliminar es siempre la primera que aparece en el texto. Esta eliminación se realiza del mismo modo que la de la puntuación.

1.2.2.4. Marcar variedades no estándar: atributo @dform

1.2.2.4.1. Principios generales

Las formas léxicas correspondientes a variedades no estándar del idioma se marcan en el nivel **Variant form**, tal como se recoge en la Figura 14:

Figura 14. Marcación de léxico no estándar.

Como resultado, en el token correspondiente se habrá creado un nuevo atributo **@dform**, cuyo valor será la forma léxica no estándar con grafía normalizada:

```
<tok dform="agora"  nform="ahora" id="w-4">hagora</tok>
```

La marcación de léxico no estándar se rige por las normas siguientes:

- La forma no estándar se escribe con la grafía normalizada:

Original (form)	No estándar (dform)	Estándar (nform)
hagora	agora	ahora
nayde	naide	nadie
ansi	ansí	así

- Toda forma marcada como no estándar (dform) debe tener una forma estándar correspondiente (nform). En términos de XML, siempre que exista un atributo **@dform** debe existir un atributo **@nform**.

- Las formas léxicas que son marcadas como no estándar incluyen, principalmente, casos de arcaísmos y de regionalismos. En el caso del español, generalmente se trata de formas que en el diccionario de la RAE aparecen con marcas de uso de carácter diacrónico o diatópico. Respecto a las primeras, el diccionario de la RAE hace uso de cuatro marcas de uso:

ant.	Anticuado	última documentación anterior a 1500
desus.	Desusado	última documentación entre 1500 y 1900
p. us.	Poco usado	última documentación posterior a 1900
germ.	Germanía	código empleado durante el Siglo de Oro

- También se utiliza la marcación de léxico no estándar para señalar formas ya desaparecidas del léxico y que no siempre aparecen registradas en los diccionarios contemporáneos (PT: *tôdolos*, *rem*, *al*).

- La forma no estándar y la forma estándar deben compartir un mismo étimo y un mismo lema. No se marcan como léxico no estándar, por tanto, casos como los siguientes:

mercaduría

1. f. desus. **mercancía.**

Real Academia Española © Todos los derechos reservados

amicicia

Del lat. *amicitia*.

1. f. desus. **amistad** (‖ afecto).

Real Academia Española © Todos los derechos reservados

1.2.2.4.2. Ejemplos

- Simplificación de grupos consonánticos latinos:

No estándar (dform)	Estándar (nform)
acetar/aceutar	aceptar
astener	abstener
ecelente	excelente
efeto/efeuto	efecto
dino	digno
dotor	doctor
sinificar	significar
solene	solemne

- Prótesis de prefijo *a-*:

No estándar (dform)	Estándar (nform)
assuceder	suceder
arrecibir	recibir

- Formas verbales no estándar:

No estándar (dform)	Estándar (nform)
cantás	cantáis
caya	caiga
creya	crea
entenderés	entenderéis
estea	esté
habemos	hemos
habés	has
habíe	había
haiga	haya
oya	oiga
saberá	sabrá
sabíe	sabía
terné	tendré
trairés	traeréis
traya	traiga
trujo	trajo
verná	vendrá
vía	veía
vide	vi
vivíe	vivía

- Otros casos:

No estándar (dform)	Estándar (nform)
agora	ahora
aguacil	alguacil
ansí	así
comigo	conmigo
cuán	cuán
de	desde
do	donde
grand	grande
inviar	enviar
mesmo	mismo
muncho	mucho

naide	nadie
priesa	prisa
proprio	propio

1.2.2.5. Marcar palabras clave: atributo @ltags

1.2.2.5.1. Principios generales

Las palabras clave están pensadas para marcar fenómenos lingüísticos no recuperables de otra forma. Por tanto, sólo se utilizan para señalar casos que escapan a la búsqueda automática del corpus normalizado y anotado morfológica y sintácticamente. En consecuencia, la mayoría de las palabras clave indican fenómenos que tienen que ver con la fonología del español o del portugués.

Se ha optado por utilizar un número reducido de etiquetas para facilitar su anotación y evitar dudas en la aplicación. El criterio que orienta la selección de una palabra clave es doble: por un lado, se señalan los arcaísmos; por otro lado, se señalan las innovaciones. Puesto que se están clasificando textos de la Edad Moderna, este doble criterio se traduce en lo siguiente:

- Ocurrencias excepcionales de fenómenos medievales (arcaísmos).
- Ocurrencias de fenómenos innovadores que no llegaron a ser adoptados por el estándar del español y del portugués contemporáneos (innovaciones).

Por razones de objetividad, se escogió un conjunto de términos que, en lugar de describir mudanzas (e.g. armonización vocálica, disimilación, metátesis, nasalización, palatalización, …), describen estructuras. El conjunto completo de etiquetas para cada lengua es el que se recoge a continuación (en rojo las opciones exclusivas del español; en azul, las exclusivas del portugués):

Fonología	Morfología	Sintaxis	Semántica léxica	Pragmática
complex_syllable	verb_paradigm	mesoclisis	existential_ter	address_form
consonant_system		negative_concord	negative_words	
diphthong_and_hiatus			possessive_haber	
unstressed_vowels			possessive_haver	
			ser_vs_estar	

Figura 14. Palabras clave.

Cuando es necesario combinar dos palabras clave, estas se separan por un guión y se ordenan alfabéticamente. Por ejemplo:

complex_syllable-unstressed_vowels

A falta de una palabra clave apropiada, la necesidad de destacar determinada grafía idiosincrática lleva la marcación “other”.

Las palabras clave se marcan en el nivel **Linguistic notes**, tal como se recoge en la Figura 16:

XML	Raw XML value	Asiguro
form	Provisional transcription	
fform	Expanded/Free form	
dipl	Edition	
dform	Variant form	
nform	Standardization	Aseguro
pos1	Tycho POS tag	
pos	Word Class	
mfs	Detailed POS	
lemma	Lemma	
ltags	Linguistic notes	unstressed_vowels

Figura 16. Marcación de palabras clave.

Como resultado, en el token correspondiente se habrá creado un nuevo atributo **@ltags**, cuyo valor será la palabra clave correspondiente:

```
<tok nform="Aseguro"  ltags="unstressed_vowels id="w-4">Asiguro</tok>
```

Téngase en cuenta que la marcación de palabras clave se aplica siempre al nivel de <tok> y nunca de <dtok>, aun cuando el fenómeno marcado se produce en un <dtok>. Téngase en cuenta también que las palabras clave no se aplican en textos no originales (i.e. copias). Por tanto, los tokens de cartas que son copias nunca llevan el atributo **@ltags**.

1.2.2.5.2. complex_syllable

Por regla general, se utiliza en los casos siguientes:

- Sílabas con coda añadida (implica generalmente sibilantes o líquidas).

form	nform	lengua
indozlencias	indulgências	PT
sastifazer	satisfazer	PT
fazeri	fazer	PT
comere	comer	ES / PT
secera	sequer	PT
adejunta	adjunta	ES

- Sílabas con ataque ramificado (implica generalmente sibilantes o líquidas).

form	nform	lengua
mersta	mestra	PT
obegto	objeto	ES / PT
prubilcasam	publicação	PT
Grabiel	Gabriel	ES

- Metátesis o disimilación (implica generalmente líquidas).

form	nform	lengua
Calros	Carlos	ES
altrose	artrose	PT

- Combinación de infinitivo/imperativo + clítico.

form	nform	lengua
hacello	hacerlo	ES
amallo	amarlo	ES
sello	serlo	ES
haceldo	hacedlo	ES
poneldo	ponedlo	ES

1.2.2.5.3. consonant_system

Por regla general, se utiliza en los casos siguientes:

- Grafías que reflejan seseo o ceceo en español.

form	nform	lengua
abraso	abrazo	ES
diferensia	diferencia	ES
Joze	José	ES
yzabel	Isabel	ES

- Grafías que reflejan betacismo en portugués.

form	nform	lengua
binho	vinho	PT
envora	embora	PT

- Grafías que reflejan velarización, palatalización, dentalización en español

form	nform	lengua
aguelo	abuelo	ES
algondiga	alhóndiga	ES
azgunta	adjunta	ES

- Elisión de *–d* final (generalmente, en formas de imperativo plural) en español

form	nform	lengua
ama	amad	ES
canta	cantad	ES
aministrar	administrar	ES

- Elisión de *–d–* intervocálica (generalmente en las formas de participio) en español

form	nform	lengua
benio	venido	ES
estao	estado	ES

No se aplica la etiqueta *consonant_system* a la simplificación del grupo consonántico latino *–SC–*. Casos como *deser* (*descer*) o *diciplina* (*disciplina*) no se marcan con palabra clave.

1.2.2.5.4. diphthong_and_hiatus

Grafías no canónicas que tienen que ver con diptongos o hiatos.

form	nform	lengua
diente	diante	PT
pehor	pior	PT
necesairo	necessário	PT
vigayras	vigárias	PT
audencia	audiencia	ES
bente	veinte	ES
corenta	cuarenta	ES
acredor	acreedor	ES

1.2.2.5.5. unstressed_vowels

Etiqueta que se aplica en casos variados, pero que tienen que ver siempre con la grafía no canónica de vocales pretónicas o postónicas.

form	nform	lengua
escrebieron	escribieron	ES
mijor	mejor	ES
soplico	suplico	ES
ascondida	escondida	ES
noteficar	notificar	ES
deligencia	diligencia	ES
sacraficio	sacrificio	ES

1.2.2.5.6. verb_paradigm

Esta etiqueta se usa para cuestiones morfológicas (nunca fonológicas) que afecten a las formas verbales. Generalmente, se trata de cuestiones que afectan a la desinencia verbal. Algunos de los casos más frecuentes son los siguientes:

- Regularización analógica de formas irregulares:

hacido	(hecho)
conducieron	(condujeron)
andó	(anduvo)
teniste	(tuviste)
hacería	(haría)

- Adición de –*s* final en la segunda persona del pretérito perfecto simple:

cantastes	(cantaste)
cantastes	(cantasteis)

- Desinencia *-des* en la segunda persona plural de algunos tiempos verbales:

cantásedes	(cantaseis)
cantardes	(cantareis)
cantáredes	(cantareis)
cantábades	(cantabais)
podíades	(podíais)
pudierdes	(pudiereis)
pudiéredes	(pudiereis)

- Otros casos:

quisiendo	(queriendo)
tuviendo	(teniendo)

Téngase en cuenta que el uso de la etiqueta *verb_paradigm* es incompatible con la marcación de léxico no estándar. Por tanto, ninguna de las formas arriba indicadas debe llevar el atributo **@dform**. Del mismo modo, ninguna de las formas verbales no estándar indicadas en el apartado 1.2.2.4.2. debe llevar la etiqueta *verb_paradigm*.

Todas las formas marcadas como *verb_paradigm* no se modernizan a su correspondiente forma estándar. Solo se normaliza la grafía:

form	nform	ltags
digistes	dijistes	verb_paradigm
tubiendo	tuviendo	verb_paradigm
cantabades	cantábades	verb_paradigm

1.2.2.5.7. mesoclisis

Esta etiqueta se usa para marcar clíticos insertados en mitad de una palabra:

*E meu Irmão benzer-**se**-á de tornar a fiar a ninguém*
*Quando a matéria seja tal que não consinta dilação, levar-**me**-ão nos braços*

La mesoclisis es un fenómeno característico del portugués. No obstante, en el corpus español se ha marcado como mesoclisis el siguiente caso:

*Y si comprase unas gallinas, holgar**me**ía para que a las tardes me enviara algún cuartillo.*

1.2.2.5.8. negative_concord

Esta etiqueta se usa para marcar partículas negativas superfluas y que, por tanto, no deben aparecer en la estructura estándar correspondiente:

*De ningún modo **no** enseñes ninguna carta mía.*
*Ya nada que me digan **no** me hará fuerza para dejar de cansarme.*

1.2.2.5.9. existential_ter

Esta etiqueta se reserva para el corpus portugués:

***Tem** côco na Bahia*

1.2.2.5.10. negative_words

Esta etiqueta se usa para marcar partículas que no son negativas *per se*, pero que son usadas con valor negativo en determinados contextos:

*Mas ele não tem **real***
*E se lhe ela amarga, não se me dá dele um **figo***
*E do que eles mandaram não falarei mais **palavra***

1.2.2.5.11. possessive_haber

Esta etiqueta se usa para marcar usos del verbo *haber* con valor posesivo en español.

*Solicite al señor regente si pudiéremos **haber** alguna pieza o pensión.*
*De Roma muchos días ha que no hemos **habido** letra.*

1.2.2.5.12. possessive_haver

Esta etiqueta se usa para marcar usos del verbo *haver* con valor posesivo en portugués.

> *Porque o senhor infante **havendo** misericórdia de mim e do outros muitos.*
> *E lhe digais que rogue a Pero Fernandes que **haja** esta venda por boa.*

1.2.2.5.13. ser_vs_estar

Esta etiqueta se usa para marcar usos intercambiados de las formas *ser* y *estar*.

> *Eu, a 10 ou 11 do maio à noite, aí **sou**, mas não é preciso dizer-se.*
> *E se Manuel Fernandes **é** neste mundo, eu tenho confiança nele.*
> ***Somos** a 12 de febrero de 1767.*

1.2.2.5.14. address_form

Esta etiqueta se usa para señalar referencias al destinatario de la carta mediante formas no canónicas (i.e. pronombres personales que no son de segunda persona):

> *Nosso Senhor ma traga diante de meus olhos como **ela** deseja.*

1.3. Tareas de posedición

1.3.1.1. División de frases

Una vez terminada la tarea de edición del texto (que incluye la modernización, la marcación de variedades no estándar y la marcación de palabras clave lingüísticas), el siguiente paso consiste en dividir el texto en frases. Entendemos aquí por frase toda unidad delimitada por puntuación fuerte, es decir:

- punto (.)
- punto y coma (,)
- cierre de signo de interrogación (?)
- cierre de signo de exclamación (!)

La división del texto en frases se realiza automáticamente desde la plataforma TEITOK pinchando en el enlace que aparece destacado en la Figura 17:

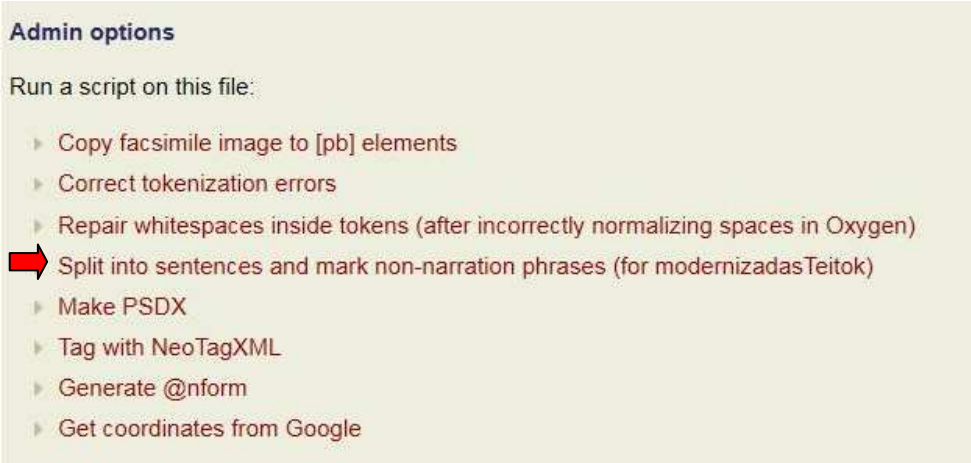


Figura 17. División automática del texto en frases.

Como resultado, en el fichero XML aparece un elemento <s> con una identificación única por cada frase que haya sido delimitada en el texto:

```
<s id="s-1”>
        <tok id="w-1">Querido</tok> <tok id="w-2">amigo</tok><tok id="w-3">.</tok>
</s>
```

Después de pasar el script para la división automática de frases, se habrá creado un nuevo enlace **Sentence view** para la carta en cuestión en la interfaz de TEITOK. Pinchando en ese enlace, que aparece destacado en la Figura 18, se puede acceder a la visualización del texto de la carta dividido por frases.

Figura 18. Visualización por frases.

1.3.1.2. Marcación de partes formulares

En esta fase se marcan las partes formulares del texto. Esta marcación se realiza en el nivel de la frase y desde la ventana de visualización por frases indicada en el apartado anterior. En realidad, esta tarea consiste en una revisión manual del resultado generado automáticamente por el script de división en frases.

Como resultado del paso anterior (cf. apartado 1.3.1.1.), el script habrá creado un atributo **@formula** dentro del elemento <s> en aquellas frases que aparezcan delimitadas por elementos XML referidos a partes formulares. Concretamente, las correspondencias que aplica el script son las siguientes:

	input	output
Fuera de <p>	frase dentro de <opener>	<s formula"opener”>
	frase dentro de <closer>	<s formula"closer”>
Dentro de <p>	frase dentro de <salute>	<s formula"salute”>
	frase dentro de <letAddress>	<s formula"letAddress”>
	frase dentro de <letDate>	<s formula"letAddress”>

Sobre este resultado automático, y en la ventana de visualización por frases, se deben corregir manualmente tres cuestiones:

- Cambiar los casos de **salute** y **letAddress** por **opener** o **closer**, según corresponda (i.e. en función de si aparecen al inicio o al final del texto).

- Incluir, si fuese necesaria, la marcación de arengas (**harangue**) y peroraciones (**peroration**).

- Incluir, si fuese necesaria, la marcación de lagunas textuales (**lacuna**). Esta indicación se utiliza para marcar aquellas frases que presentan alguna omisión textual (<gap>) o algún fragmento que, pese a haber sido transcrito como parte de la misiva, no forma parte del género epistolar propiamente dicho (e.g. oraciones, poemas, texto legislativo, etc). La identificación de ambos casos permite excluirlos del corpus sintácticamente anotado.

Para editar una frase es necesario pinchar en el enlace que muestra el código de identificación de la frase a la izquierda del texto. Por ejemplo, en la Figura 19 se observa que la frase número 2 está marcada como **salute**:

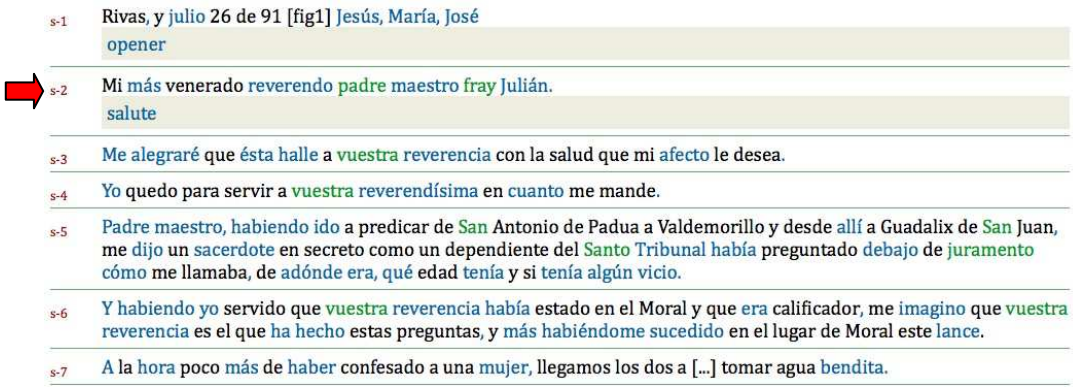


Figura 19. Marcación de partes formulares (I).

Pinchando en el enlace que aparece destacada en la figura anterior, accedemos a la ventana de edición de frase. En el espacio correspondiente al valor del atributo **@formula** cambiamos *salute* por *opener* y guardamos:

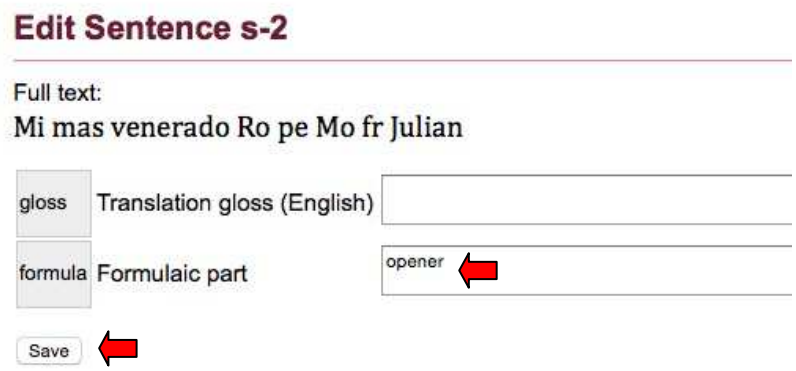


Figura 20. Marcación de partes formulares (II).

El resultado es el que aparece en la Figura 21, que también refleja la marcación de *harangue* para las frases 3 y 4:

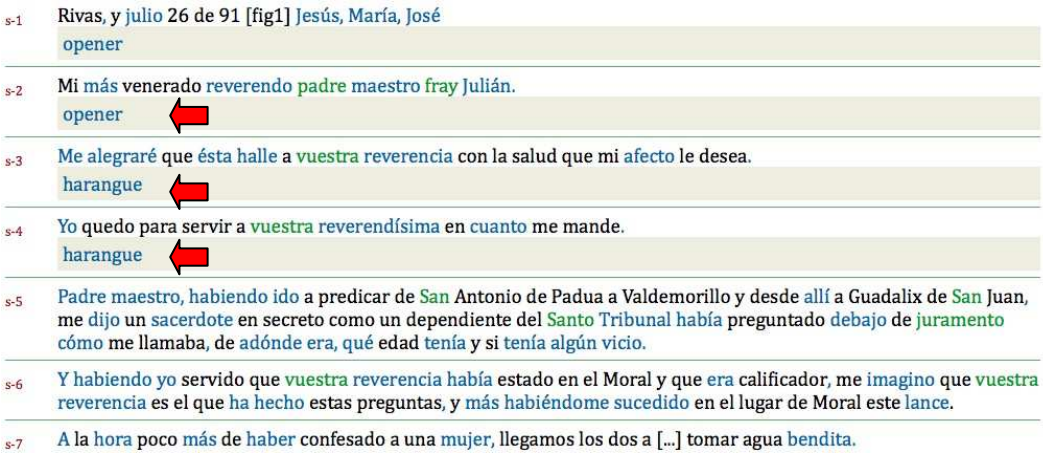


Figura 21. Marcación de partes formulares (III).

1.3.1.3. Asignación manual del subcorpus: atributo @gold

El corpus de *P. S. Post Scriptum* está formado por un número determinado de cartas escritas por un número determinado de autores, sin que exista un equilibrio entre ambos aspectos, puesto que la finalidad principal del corpus es la de recuperar el mayor número de cartas posible hasta alcanzar los 3500 documentos por lengua. La única condición que se ha aplicado a priori es la de limitar a 25 el número de cartas por autor para evitar así una desproporción excesiva entre ellos.

El número de cartas por autor es, por tanto, variable, como también lo es, obviamente, el número de palabras por carta. En suma, quiere esto decir que las características propias del corpus *P. S. Post Scriptum* imposibilitan una representatividad entre autores y aun menos entre variedades lingüísticas dentro de un mismo corpus.

No obstante, y con el objeto de minimizar este desequilibrio, se ha creado un subcorpus compuesto por una carta de cada autor. Esta selección es realizada automáticamente por un script (cf. apartado 1.4.1.2.) que elige, para cada conjunto epistolar de una misma mano, aquella carta que presente un número mayor de tipos de palabra (i.e. formas normalizadas diferentes). Esta información se registra en el propio archivo XML mediante un elemento <class> que adopta uno de dos valores posibles: el valor 0, si la carta no fue seleccionada para el subcorpus; o el valor 1, si la carta sí fue seleccionada para el subcorpus. Es decir:

Carta no seleccionada:
<class subcat="balancedSelection">0</class>

Carta sí seleccionada:
<class subcat="balancedSelection">1</class>

El subcorpus, por tanto, estará formado por todas aquellas cartas que muestren el valor 1 dentro de dicho elemento. Sin embargo, esta selección automática puede producir dos situaciones no deseables:

- Que una carta particularmente interesante (por las razones que sean) no haya sido seleccionada.

- Que una carta particularmente deteriorada sí haya sido seleccionada.

Para corregir ambas situaciones se ha contemplado el uso del atributo **@gold** dentro de un elemento <class>. Este atributo también adopta uno de dos valores, que en este caso deben ser añadidos manualmente: el valor 2, si la carta es particularmente buena; o el valor 1, si la carta es particularmente mala. En consecuencia, al ejecutar el script que selecciona automáticamente el subcorpus se mantendrá el valor 1 en aquellas cartas que fueron anotadas manualmente con gold="2"; y se buscará un mejor candidato para aquellas cartas que fueron anotadas manualmente con gold="1". Como resultado, el elemento <class> quedará del siguiente modo:

Carta particularmente buena:
<class subcat="balancedSelection" gold="2">1</class>

Carta particularmente mala:
<class subcat="balancedSelection" gold="1">0</class> (si existe otro candidato)
<class subcat="balancedSelection" gold="1">1</class> (si no existe otro candidato)

Por norma general, se anotan con gold="2" todas las cartas que hayan recibido anotación sintáctica. También reciben gold="2" aquellos casos esporádicos de cartas particularmente interesantes por razones lingüísticas y/o históricas. Por su parte, se anotan con gold="1" casos esporádicos de cartas muy deterioradas y que, por tanto, presentan numerosas lagunas de transcripción.

1.3.1.4. Palabras clave sociohistóricas: atributo @key

Además de las palabras clave de carácter lingüístico (cf. apartado 1.2.2.5.), que se anotan directamente en el texto al nivel del token, cada carta está asociada a un conjunto de palabras clave de carácter sociohistórico. En este apartado se explica el proceso llevado a cabo para la marcación de estas palabras clave y su importación al sistema de búsqueda en TEITOK.

1.3.1.4.1. Creación de la lista general

En un primer momento se creó una lista trilingüe (inglés, español y portugués) de palabras clave. Cada palabra clave está asociada a un código formado por la letra K más un número arbitrario (por ejemplo: K126). Esta lista se puede consultar en el enlace siguiente: http://ps.clul.ul.pt/index.php?action=keywords.

Para editar la lista, es necesario registrarse en TEITOK, ir al enlace anterior y pinchar en el enlace **edit** correspondiente a la palabra clave que se quiera modificar:

edit K1	Aborto	Aborto	Abortion
edit K3	Accidente	Acidente	Accident
edit K5	Acusación	Acusação	Accusation
edit K2	Absolución	Absolvição	Acquittal
edit K6	Administración	Administração	Administration
edit K7	Adulterio	Adultério	Adultery
edit K53	Consejos	Conselhos	Advice
edit K8	Agricultura	Agricultura	Agriculture
edit K28	Ayuda	Ajuda	Aid
edit K138	Limosna	Esmola	Alms

A continuación, aparecerá una ventana como la de la Figura 22, desde la que es posible añadir/editar el nombre de la palabra en cuestión o su descripción.

Figura 22. Edición de palabras clave sociohistóricas.

Aunque la lista está prácticamente cerrada, también es posible añadir nuevas palabras si fuese necesario. Para ello, hay que ir al final de la página y pinchar en el botón que aparece destacada en la Figura 23. El sistema asignará automáticamente un nuevo código a la palabra creada:

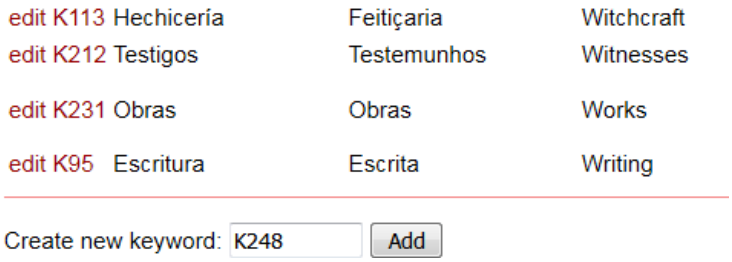


Figura 23. Adición de palabras clave sociohistóricas.

Sobre la lista general de palabras clave, ténganse en cuenta las dos normas siguientes:

- Todas las palabras deben tener su correspondiente término en las tres lenguas (inglés, español, portugués).

- No puede haber términos repetidos en inglés (sí puede haberlos en español o en portugués, aunque conviene evitarlos). Por ejemplo, los términos *destierro* y *exilio* no pueden ser ambos traducidos a *exile* (actualmente, *destierro* está traducido a *banishment*, y *exilio* a *exile*)

1.3.1.4.2. Creación de listas particulares

Por otro lado, cada fichero XML cuenta con un elemento destinado a recoger la lista de palabras clave asociadas a cada carta en particular:

```
<class subcat="socioHistoricalSource"></class>
```

Como parte del trabajo de transcripción y codificación XML (y, por tanto, previamente a la importación del XML a la plataforma), los historiadores habrán dejado constancia de una lista provisional de palabras clave dentro de dicho elemento:

Ejemplo tomado de la carta PSCR1435:

```
<class subcat="socioHistoricalSource">fraude, justiça</class>
```

Como método alternativo, existe también una hoja de cálculo en Excel que fue utilizada por el grupo de historiadores de *P. S. Post Scriptum* para recoger lista provisionales de palabras clave para un número importante de cartas. Tanto el propio XML como dicho Excel son archivos de entrada válidos para el siguiente paso, que es la importación automática de los datos a la plataforma TEITOK.

1.3.1.4.3. Importación al sistema de búsqueda

Esta última fase constituye la tarea de posedición propiamente dicha, pues las dos anteriores son tareas que ya vienen dadas en el momento de trabajar en TEITOK. Se trata de importar al sistema de búsqueda de la plataforma las palabras clave de historia que previamente hayan sido indicadas en el archivo XML de la carta (o en el citado Excel). Para ello, y una vez visualizado en la interfaz el texto de la carta de la que se quiere importar la lista de palabras, es necesario pinchar en el enlace **Edit keywords** que aparece destacado en la Figura 24:

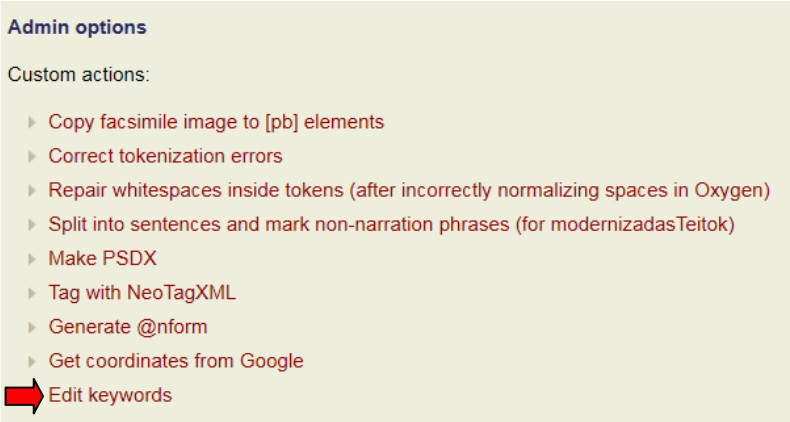


Figura 24. Importación de palabras clave sociohistóricas (I).

A continuación, aparecerá el resultado de una operación automática que compara la lista provisional (particular) de la carta en cuestión con la lista definitiva (general) del sistema. En otras palabras, el sistema verifica si las palabras propuestas están registradas en el conjunto general, trilingüe, de palabras clave elegidas como ideales por los historiadores. Tan solo es necesario guardar el resultado pinchando en **Save**:

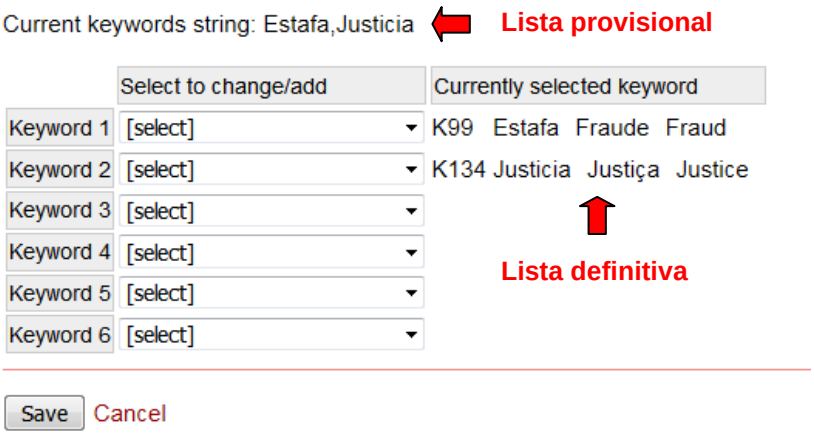


Figura 25. Importación de palabras clave sociohistóricas (II).

Como resultado, en el fichero XML aparecerá, por cada palabra clave, un elemento <term> con los atributos **@n** y **@key**. El atributo **@n** sirve para numerar cada término que se ha utilizado; el atributo **@key** sirve para asociar cada término con un código único que remite al archivo externo KW.xml, en donde está guardada la lista general de palabras clave. Es el valor de este atributo **@key** y su referencia al archivo externo KW.xml el que posibilita la búsqueda automática trilingüe de las palabras clave sociohistóricas:

```
<class subcat="socioHistoricalSource">
        <term n="1" key="KW.xml#K99"/>
        <term n="2" key="KW.xml#K134"/>
        Estafa, Justicia
</class>
```

En el ejemplo de la Figura 25, todas las opciones de la lista provisional están ya recogidas en la lista definitiva. No obstante, en este proceso de importación pueden darse dos situaciones más:

- Que una o más opciones de la lista provisional no estén registradas en la lista definitiva.

- Que no exista una lista provisional previa.

El primer caso es debido a que para la carta en cuestión se indicaron palabras clave que finalmente no fueron seleccionadas en la lista definitiva. El sistema alertará de estos casos indicando aquellas palabras que no aparecen registradas en la lista definitiva, tal como se aprecia en la Figura 26. Al pinchar en **Save** las palabras no registradas se desecharán de la importación al sistema:

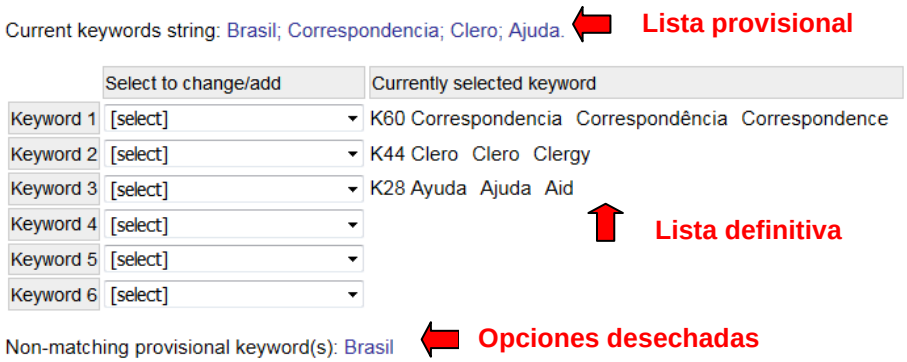


Figura 26. Importación de palabras clave sociohistóricas (III).

El segundo caso puede ser debido a que el contenido de esa carta no sugiere ninguna marcación de tipo sociohistórico o a que todavía no se realizó dicha marcación. En cualquier caso, téngase en cuenta que desde la propia ventana de edición de palabras clave siempre es posible seleccionar nuevas palabras o eliminarlas, si fuese necesario. Para ello, basta con utilizar el menú desplegable que aparece en la Figura 27:

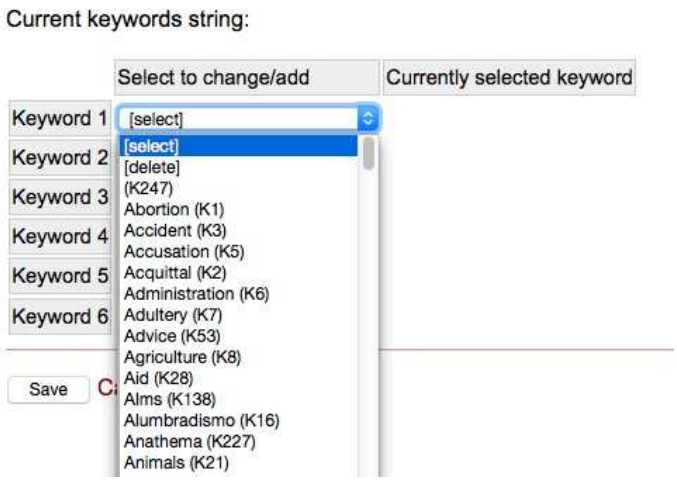


Figura 27. Añadir/eliminar palabras clave sociohistóricas.

1.3.1.5. Responsable de edición

En esta fase, se deja constancia del responsable de la edición modernizada del texto. Este registro debe realizarse tanto en la plataforma TEITOK como en el fichero XML de la carta.

En TEITOK, este registro de realiza en el repositorio de datos, al que se accede pinchando en el enlace que aparece destacado en la Figura 28:

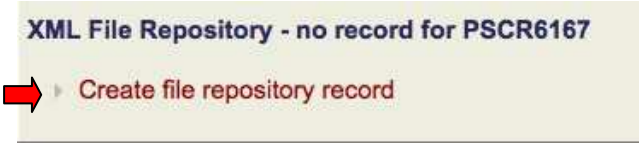


Figura 28. Crear nuevo registro en el repositorio de datos.

En la ventana que se abre a continuación, se deben cubrir los campos **Original Code** y **Modernization**. En el primero, se indica el código de la carta que ha sido importada a TEITOK para ser posteriormente editada; en el segundo, se indican las siglas del modernizador y la fecha de modernización, tal como aparece en la Figura 29:

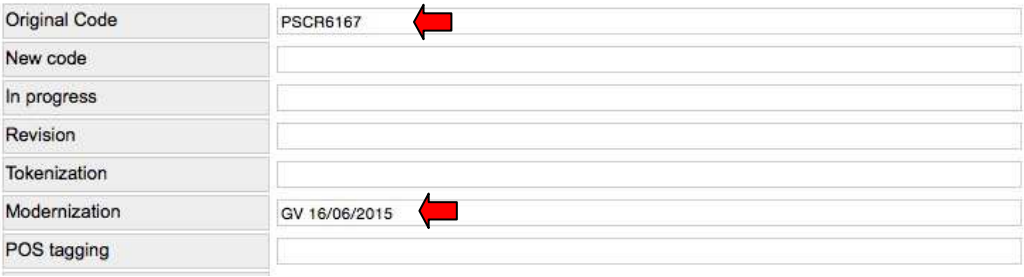


Figura 29. Cubrir nuevo registro en el repositorio de datos.

En el fichero XML, el responsable de la modernización debe constar en el siguiente apartado:

```
<respStmt><resp subcat="modernizedEdition"><name></name></resp></respStmt>
```

1.4. Tareas periódicas

Existen dos tareas adicionales que deben realizarse periódicamente con el objetivo de actualizar los datos del corpus y/o de los ficheros XML. En ambos casos, se trata de pasar un script que realiza dicha actualización automáticamente. Son los casos siguientes:

1.4.1.1. Actualización del script de normalización automática

El script que realiza la normalización automática de los textos incorpora los datos del corpus ya normalizado y los aplica estadísticamente sobre las nuevas formas aún no normalizadas. Este conjunto de datos (i.e. corpus de entrenamiento) es almacenado dentro del servidor en un fichero llamado **corpus.vrt**, del que existen dos copias sincronizadas: una guardada en la carpeta **freeling-es** y otra guardada en la carpeta **freeling-pt**. Este fichero se debe actualizar periódicamente con el fin de mejorar el comportamiento estadístico del script. Para ello, es necesario hacer lo siguiente:

- Abrir la línea de comandos y entrar en el servidor de postscriptum:

  postscriptum@cards.clul.ul.pt

- Dirigirse a una de estas dos carpetas:

  [postscriptum@cards ~]$ cd **freeling-es**
  [postscriptum@cards ~]$ cd **freeling-pt**

- Escribir en la línea de comandos la orden siguiente[1]:

find /home/postscriptum/cards/ -type f -exec perl makevrt.pl {} withlang \; > /home/postscriptum/freeling-pt/corpus.vrt

1.4.1.2. Actualización del subcorpus

De acuerdo con lo explicado en el apartado 1.3.1.3, la selección de cartas que forman parte del subcorpus debe ser actualizada periódicamente. Para ello, es necesario ejecutar algunos scripts, todos ellos guardados en la carpeta **balancer** dentro del servidor. El proceso de actualización es el siguiente:

- Abrir la línea de comandos y entrar en el servidor de postscriptum:

  postscriptum@cards.clul.ul.pt

- Dirigirse a la siguiente carpeta:

  [postscriptum@cards ~]$ cd **balancer**

- Escribir en la línea de comandos la orden siguiente:

[1] Esta orden utiliza comandos básicos de UNIX, por lo que sólo funciona en sistemas operativos basados en UNIX: Linux o Macintosh. Para hacerlo funcionar en Windows, es necesario instalar una herramienta que proporcione un ambiente de trabajo UNIX (e.g. Cygwin).

```
sh todo.txt
```

Los tres script se ejecutarán automáticamente y la asignación de valores (1 o 0) dentro de elemento <class subcat="balancedSelection"> será actualizada.

2. Anotación morfosintáctica (POS)

2.1. Conjunto de etiquetas

Para la anotación morfosintáctica del corpus *P.S. Post Scriptum,* se utiliza una versión ligeramente modificada de las etiquetas EAGLES propuestas para el español. Esta versión modificada facilita la conversión a las etiquetas CLAWS usadas en el corpus Tycho Brahe.

El conjunto de etiquetas EAGLES se rige por un sistema de posiciones: cada etiqueta consta de una secuencia de letras y números, donde cada letra o número representa un rasgo morfosintáctico determinado dependiendo de su posición dentro de la secuencia. El significado de cada posición está asociado a la categoría principal, representada por la primera letra de la secuencia. Por ejemplo, la forma *bonita* lleva la etiqueta AQ0FS0, donde la primera A indica que se trata de un adjetivo, y la S en la quinta posición indica el número, en este caso singular.

Para una descripción completa de las etiquetas utilizadas, con su posición, significado y valores posibles, véase el etiquetario de *P. S. Post Scriptum*.

2.2. Proceso de anotación en TEITOK

La anotación morfosintáctica del corpus se lleva a cabo en dos fases: anotación automática y revisión manual. Además, conviene actualizar el anotador después de cada nueva carta anotada y revisada, con el fin de incorporar los datos al anotador automático y mejorar así su rendimiento.

2.2.1. Tratamiento automático

La anotación automática se realiza con el anotador morfosintáctico NeoTag. Para aplicar este anotador al texto de una carta, hay que pinchar en el enlace que aparece destacado en la Figura 30:

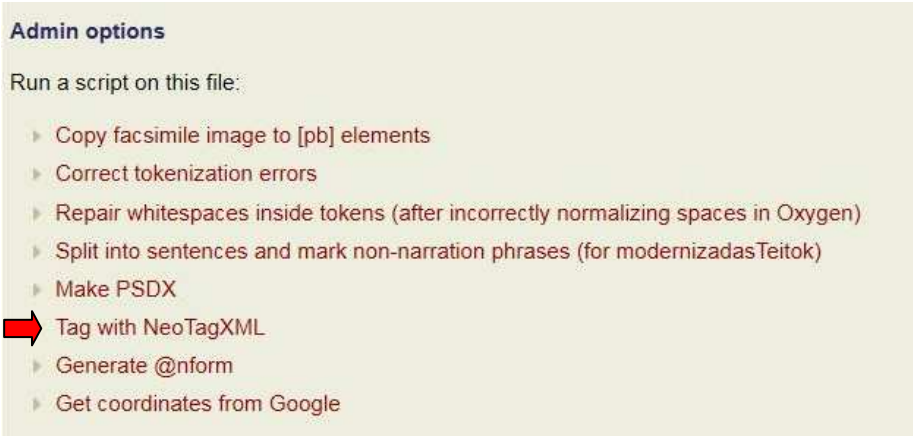


Figura 30. Anotación automática con NeoTag.

Como resultado, en el elemento <tok> de aquellos tokens que hayan sido anotados automáticamente aparecerá un atributo **@mfs** con la etiqueta morfosintáctica correspondiente, y un atributo **@lemma** con el lema correspondiente:

```
<tok id="w-1" nform="vergüenza" mfs="NCFS000 lemma="vergüenza>berguenza</tok>
```

2.2.2. Revisión manual

2.2.2.1. Token simple

El resultado del script de anotación automática debe ser revisado manualmente. La edición de la anotación se realiza en la ventana de edición de token. Dentro de esta ventana, la etiqueta morfosintáctica se edita en el nivel **Detailed POS**; la edición del lema se edita en el nivel **Lemma**.

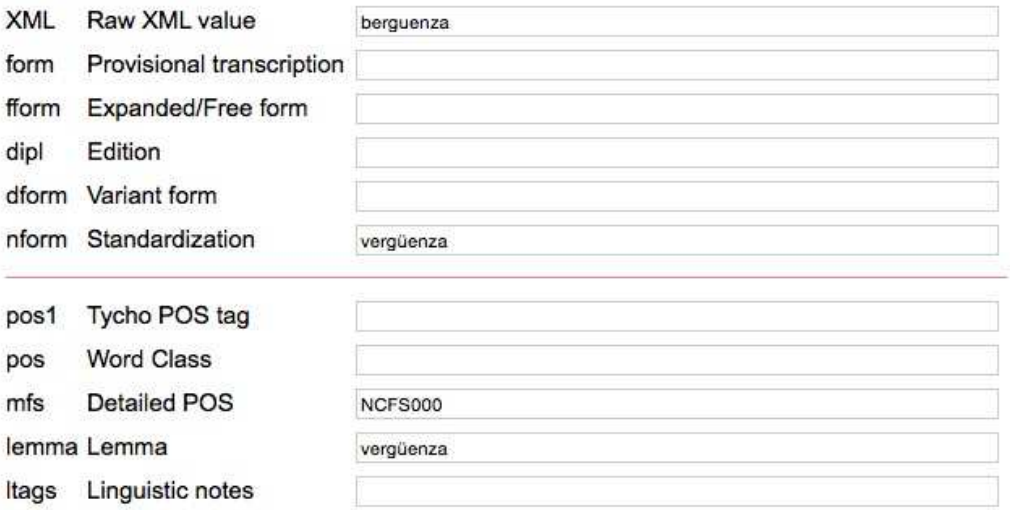


Figura 31. Información morfosintáctica del token.

Como se aprecia en la Figura 31, existen otros dos niveles de información relacionados con la anotación morfosintáctica de cada token. El contenido de ambos niveles es posible obtenerlo automáticamente y, por tanto, no es necesario cubrirlo de forma manual. Concretamente, están reservados para lo siguiente:

- pos1. Correspondencia con las etiquetas morfosintácticas CLAWS utilizadas en el corpus Tycho Brahe.

- pos. Etiqueta general sobre la clase de palabra (nombre, adjetivo, verbo, adverbio, ...).

2.2.2.2. Token complejo: <dtok>

En caso de que el token anotado sea complejo (i.e. un token formado por dos o más tokens), el anotador automático creará un elemento <dtok> por cada unidad que conforma el token complejo. En el XML, este elemento <dtok> aparecerá dentro del elemento <tok> y con un identificador único propio. Por ejemplo:

```
<tok id="w-1">mandarme
        <dtok id="d-1-1" form="mandar" mfs="VMN0000 lemma="mandar/>
        <dtok id="d-1-2" form="me" mfs="PP1CS000 lemma="me/>
</tok>
```

En la revisión manual de estos casos, ténganse en cuenta las normas siguientes:

- La información lingüística (etiqueta y lema) de los tokens complejos se incluye siempre dentro del elemento <dtok>, y no dentro del elemento <tok>. Por tanto, en estos casos los atributos @**mfs** y @**lemma** son atributos de <dtok>.

- La información metalingüística (de palabras clave) de los tokens complejos se incluye siempre dentro del elemento <tok>, y no dentro del elemento <dtok>. Por tanto, el atributo @**ltags** es siempre un elemento de <tok>.

- Dentro del elemento <dtok>, el atributo @**form** es obligatorio y corresponde a la forma libre de la palabra en cuestión tal como aparecería escrita por el autor:

Ejemplo 1:

```
<tok id="w-1">mandarme
        <dtok id="d-1-1" form="mandar" mfs="VMN0000” lemma="mandar”/>
        <dtok id="d-1-2" form="me" mfs="PP1CS000” lemma="me”/>
</tok>
```

Ejemplo 2:

```
<tok id="w-1">del
        <dtok id="d-1-1" form=”de” mfs="SPS00 lemma="de”/>
        <dtok id="d-1-2" form="el" mfs="DA0MS0” lemma="el”/>
</tok>
```

Ejemplo 3:

```
<tok id="w-1">al
        <dtok id="d-1-1" form="a" mfs="SPS00 lemma="a”/>
        <dtok id="d-1-2" form="el" mfs="DA0MS0” lemma="el”/>
</tok>
```

Ejemplo 4:

```
<tok id="w-1" nform="dizer-lho" >dizerlho
        <dtok id="d-1-1" form="dizer" mfs="VMN0000 lemma="dizer”/>
        <dtok id="d-1-2" form="lhe" mfs="PP3CSD00” lemma="lhe”/>
        <dtok id="d-1-3" form="o" mfs="PP3MSA00” lemma="o”/>
</tok>
```

- Dentro del elemento <dtok>, el atributo **@nform** se utiliza para modernizar el valor de **@form** cuando su contenido no presenta ortografía contemporánea. Si se tratase de una contracción o de una enclisis abreviadas, la expansión de la abreviatura se hace dentro del atributo **@fform de los niveles <tok> y <dtok>.** No es necesario que la suma de todos los valores del atributo **@nform** dé como resultado el token complejo en su forma modernizada:

Ejemplo 5:

```
<tok id="w-1" nform="habiéndome">aviendome
        <dtok id="d-1-1" form="aviendo" nform="habiendo" mfs="VAG0000” lemma="haber”/>
        <dtok id="d-1-2" form="me" mfs="PP1CS000” lemma="me”/>
</tok>
```

Ejemplo 6 (portugués):

```
<tok id="w-1" nform="daquela" fform="daquella">daqlla
        <dtok id="d-1-1" form="de" mfs="SPS00” lemma="de”/>
        <dtok id="d-1-2" form="aqlla" fform="aquella" nform="aquela" mfs="DD0FS0”
        lemma="aquele”/>
</tok>
```

- Los nombres compuestos no generan <dtok>. Se tratan dentro de <tok>, respetando la frontera de palabra que aparezca en el original (i.e. se respeta, si fuere el caso, el espacio en blanco entre las dos partes del nombre compuesto):
-

  Ejemplo 7:
  <tok id="w-1" nform="capitão-mor">capitão mor</tok>

  Ejemplo 8:
  <tok id="w-1" nform="capitão-mor">capitãomor</tok>

- Los numerales complejos no generan <dtok>. Se crean tantos <tok> como fuese necesario en función de la frontera de palabra que aparezca en el original (cf. apartado 1.1.3.):

  Ejemplo 9:
  <tok id="w-1" nform="dieciséis">diezyseis</tok>

  Ejemplo 10:
  <tok id="w-1">diez</tok> <tok id="w-1">y</tok> <tok id="w-1">seis</tok>

Para añadir manualmente un elemento <dtok> dentro de un token, es necesario seleccionar ese token y, en la ventana de edición de token, pinchar en el enlace que aparece destacado en la Figura 32:

Figura 32. Añadir dtok en TEITOK.

Para eliminar manualmente un elemento <dtok> dentro de un token, es necesario seleccionar ese token y, en la ventana de edición de token, pinchar en el enlace que aparece destacado en la Figura 33:

Figura 33. Eliminar dtok en TEITOK.

2.2.3. Actualización del anotador

El script que realiza la anotación automática incorpora todos los datos del corpus ya anotado (i.e. corpus de entrenamiento) y los aplica estadísticamente. Estos datos son almacenados en un fichero llamado **es.xml** (en el caso del español) y **pt.xml** (en el caso del portugués), ambos guardados en la carpeta **neotag** del servidor del

proyecto. Estos ficheros se deben actualizar periódicamente (idealmente, tras revisar manualmente la anotación de cada carta) con el fin de mejorar el comportamiento estadístico del script. Para ello, es necesario hacer lo siguiente:

- Dirigirse a la siguiente ubicación dentro de la plataforma TEITOK:

  Admin / Check or update the NeoTag tagset(s)

- Seleccionar la lengua (ES o PT):

- Pinchar en el enlace que aparece destacado en la Figura 34:

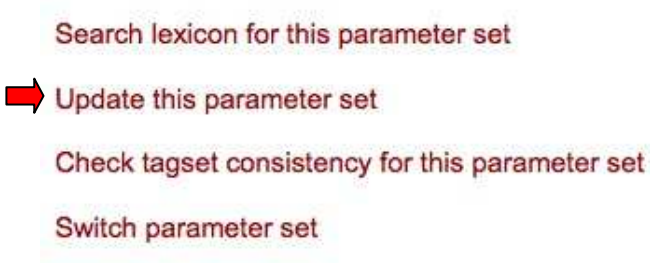


Figura 34. Actualización del anotador NeoTag.

Como resultado, el corpus de entrenamiento incorporará los datos que fueron anotados desde la última actualización. El estado actual de cada corpus de entrenamiento (ES y PT) se puede consultar al seleccionar el conjunto de parámetros correspondiente:

pid	es
restriction	//language[@subcat="ES"]
params	/home/postscriptum/neotag/es.xml
folder	./neotag/pt
training	xmlfiles/Revistas/
tagform	nform
tagpos	mfs
formtags	lemma,mfs
lemmatize	nform
last update	Tue Apr 19 14:48:45 2016
training size	85074
lexicon size	8492
tagset size	327

Training size se refiere al número de tokens del corpus de entrenamiento.

Lexicon size se refiere al número de lemas del corpus de entrenamiento.
Tagset size se refiere al número de etiquetas del corpus de entrenamiento.

2.2.4. Depuración de errores en la anotación

TEITOK permite comprobar automáticamente la presencia de errores en las etiquetas morfosintácticas. El sistema realiza esta comprobación comparando el conjunto de etiquetas registrado en el corpus de entrenamiento (i.e. todas las etiquetas utilizadas en el proceso de anotación) con el conjunto de etiquetas previamente definido en el etiquetario de *P. S. Post Scriptum* (i.e. todas las etiquetas esperables).

Para realizar esta comprobación, es necesario hacer lo siguiente:

- Dirigirse a la siguiente ubicación dentro de la plataforma TEITOK:

  Admin / Check or update the NeoTag tagset(s)

- Seleccionar la lengua (ES o PT):

- Pinchar en el enlace que aparece destacado en la Figura 35:

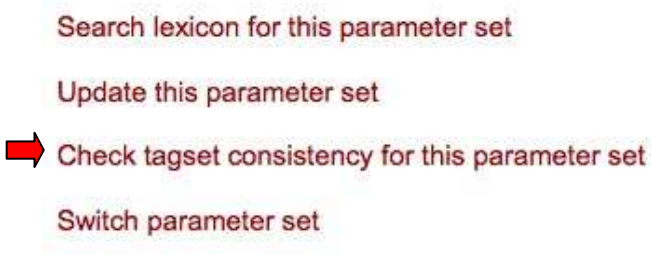


Figura 35. Comprobación de errores en la aplicación de etiquetas EAGLES.

Como resultado, aparecerá una tabla con cuatro columnas (ver Figura 36). La primera columna recoge todas las combinaciones de etiquetas registradas en el corpus de entrenamiento; la segunda columna recoge la frecuencia de aparición de cada etiqueta; la tercera columna recoge la descripción de cada etiqueta registrada; la cuarta y última columna es la que informa de posibles errores. En la Figura 36, se informa de un error en la etiqueta AW0FP0 (no es esperable una W en la primera posición después de A) :

AQ0MS0	558	Adjective;Qualitative;masculine;singular	(ok)
AQ0MSP	9	Adjective;Qualitative;masculine;singular;participle	(ok)
AQSFP0	5	Adjective;Qualitative;superlative;feminine;plural	(ok)
AQSFS0	31	Adjective;Qualitative;superlative;feminine;singular	(ok)
AQSFSP	1	Adjective;Qualitative;superlative;feminine;singular;participle	(ok)
AQSMS0	25	Adjective;Qualitative;superlative;masculine;singular	(ok)
AQSMSP	5	Adjective;Qualitative;superlative;masculine;singular;participle	(ok)
AW0FP0	1	Adjective;?;feminine;plural	Invalid W in position 1 for A;
CC	3919	Conjunction;Coordinate	(ok)
CCN	21	Conjunction;Coordinate;negative	(ok)
CS	3915	Conjunction;Subordinate	(ok)
DA0FP0	400	Determiner;Article;feminine;plural	(ok)
DA0FS0	1354	Determiner;Article;feminine;singular	(ok)

Figura 36. Visualización de errores en la aplicación de etiquetas EAGLES.

Para corregir este error, es necesario localizar en el corpus la combinación no esperable. Esto se puede realizar mediante el modo CQP del sistema de búsqueda. Para el ejemplo anterior, la consulta sería la siguiente:

Text Search

Search method: CQP Word Search

CQP Query:

[pos="AW0FP0"]

Una vez localizados y corregidos todos los errores, se recomienda actualizar de nuevo el anotador (cf. apartado 2.2.3.) y comprobar que no existen más errores.

2.3. Casos particulares

2.3.1. Participios y adjetivos

Los participios que forman parte de tiempos compuestos (i.e. HABER + participio) no incluyen información de género y número, puesto que no se flexionan:

habíamos ***cantado***	**VMP0000**

Los participios que no forman parte de tiempos compuestos incluyen información de género y número:

fuimos ***rechazados***	**VMP00PM**
fue ***rechazada***	**VMP00SF**

Para los participios que funcionan como adjetivo se contemplan dos soluciones. Si la forma del adjetivo coincide con la forma regular del participio, se anota como participio. Si la forma del adjetivo no coincide con la forma regular del participio, se anota como adjetivo y se marca una P en la última posición:

*mi **querido** amigo*	**VMP00SM**	
*el año **pasado***	**VMP00SM**	
*la carta **incluida***	**VMP00SF**	
*el niño está bien **peinado***	**VMP00SM**	
*la carta **adjunta***	**AQ0FSP**	(cf. *adjuntado*) (lema: *adjunto*)
*las cartas **inclusas***	**AQ0FPP**	(cf. *incluido*) (lema: *incluso*)
*el niño **desnudo***	**AQ0MSP**	(cf. *desnudado)* (lema: *desnudo*)
*su **afecta** servidora*	**AQ0FSP**	(cf. *afectado*) (lema: *afecto*)
***fecha** en octubre de 1674*	**AQ0FSP**	(cf. *hecho*) (lema: *fecho*)
***bendito** sea Dios*	**AQ0MSP**	(cf. *bendecido*) (lema: *bendito*)

Téngase en cuenta que esta distinción entre formas regulares e irregulares sólo se aplica cuando el participio funciona como adjetivo. Por tanto, no se aplica cuando se trata de tiempos compuestos, en cuyo caso el participio siempre se anota como VMP0000:

*el papel **imprimido***	**VMP00SM**	(lema: *imprimir*)
*el papel **impreso***	**AQ0MSP**	(lema: *impreso*)
*habían **impreso** un papel*	**VMP0000**	(lema: *imprimir*)
*habían **imprimido** un papel*	**VMP0000**	(lema: *imprimir*)

Si el participio está léxicamente sustantivado o no es posible recuperar del contexto un núcleo nominal, se anota como nombre (cf. apartado 2.3.21):

*el **peinado** del niño*	**NCMS000**
*los **nominados** de este año*	**NCMP000**
*el **pasado** no tiene remedio*	**NCMS000**

Téngase en cuenta que las formas *fecha*/*hecha* deben tratarse del modo siguiente:

***fecha** en Valladolid a 7 de junio*	**AQ0FSP**	(lema: *fecho*)
***hecha** en Valladolid a 7 de junio*	**VMP00SF**	(lema: *hacer*)
*día de la **fecha***	**NCFS000**	(lema: *fecha*)

2.3.2. Los verbos HABER y SER

Para la anotación de las formas del verbo HABER se contemplan dos soluciones. Si funciona como verbo auxiliar (i.e. en los tiempos compuesto), se anota con una A en la segunda posición. En el resto de casos, se anota con una M en la segunda posición:

habíamos *cantado*	**VAII1P0**
hayan *cantado*	**VASP3P0**
hay *muchas desgracias*	**VMIP3S0**
días ***ha*** *que llegué*	**VMIP3S0**

Las formas del verbo SER se anotan, en todas sus ocurrencias, con una S en la segunda posición:

fue *rechazada*	**VSIS3S0**
ellos ***son*** *hermanos*	**VSIP3P0**

El resto de formas verbales se anotan, en todas sus ocurrencias, con una M en la segunda posición.

2.3.3. La forma SE

Para la anotación de las forma SE en español se contemplan dos soluciones. Si equivale al pronombre “le” (objeto indirecto), se etiqueta como pronombre de tercera persona (PP3CN000). En el resto de casos (reflexivo, recíproco, media, mediopasiva, impersonal, …) se etiqueta P00CN000:

le di el regalo = ***se*** *lo di*	**PP3CN000**	
*me atreví a decír****se****lo*	**PP3CN000**	
Dios ***se*** *lo perdone*	**PP3CN000**	
se *vio a sí mismo*	**P00CN000**	*(reflexivo)*
se *escriben cartas*	**P00CN000**	*(recíproco)*
se *sentó*	**P00CN000**	*(media)*
se *realizaron encuestas*	**P00CN000**	*(mediopasiva)*
se *habló de muchas cosas*	**P00CN000**	*(impersonal)*

2.3.4. Leísmo, laísmo, loísmo

En español, la variación en el empleo de los pronombres átonos no reflexivos de tercera persona permite establecer una diferencia entre los llamados usos canónicos o etimológicos y los llamados usos innovadores o anti-etimológicos. Los empleos

canónicos son aquellos empleos que se ajustan al canon heredado del latín, es decir, formas de acusativo *la*(*s*), *lo*(*s*) para el objeto directo (OD) y formas de dativo *le*(*s*) para el objeto indirecto (OI). Los segundos son aquellos usos que no se atienen a la función sintáctica del referente y dan lugar a los fenómenos de leísmo, laísmo y loísmo.

Concretamente, el leísmo es el empleo de las formas *le(s)* en lugar de *la(s), lo(s)* para referirse a un OD. El laísmo es el empleo de las formas *la(s)* en lugar de *le(s)* para referirse a un OI. El loísmo es el empleo de las formas *lo(s)* en lugar de *le(s)* para referirse a un OI:

le *di un regalo al niño*	**PP3CSD00**	(uso canónico)
le *ayudé con el problema*	**PP3CSA00**	(uso innovador: leísmo)
la *llamé por teléfono*	**PP3FSA00**	(uso canónico)
la *compré unos pantalones*	**PP3FSD00**	(uso innovador: laísmo)
lo *convencí de que viniera*	**PP3MSA00**	(uso canónico)
lo *dije que no viniera*	**PP3MSD00**	(uso innovador: loísmo)

La sexta posición en la etiqueta de los pronombres permite marcar, en el caso de los pronombres de tercera persona, si se trata de un empleo canónico o innovador. La tabla siguiente recoge todas las posibilidades para los pronombres átonos no reflexivos de tercera persona:

	Uso canónico	Uso innovador	Lema
le	PP3CSD00	PP3CSA00	le
les	PP3CPD00	PP3CPA00	le
la	PP3FSA00	PP3FSD00	lo
las	PP3FPA00	PP3FPD00	lo
lo	PP3MSA00	PP3MSD00	lo
los	PP3MPA00	PP3MPD00	lo

Finalmente, téngase en cuenta que la forma *lo* también puede referirse al pronombre neutro:

eso no ***lo*** *puedo remediar*	**PP3CNA00**	(uso canónico)

2.3.5. El imperativo

Solo se marcan como imperativo las segundas personas (tú, vosotros). El resto de formas del paradigma se marcan como presente de subjuntivo:

cante *yo*	**VMSP1S0**	
canta *tú*	**VMM02S0**	(imperativo)
cante *él / usted*	**VMSP3S0**	
cantemos *nosotros*	**VMSP1P0**	
cantad *vosotros*	**VMM02P0**	(imperativo)
canten *ellos /ustedes*	**VMSP3P0**	

2.3.6. Los posesivos

En la etiqueta de los posesivos, conviene distinguir la información referida al poseedor (persona y número) de la información referida a la cosa poseída (género y número). En el caso de los determinantes, esta información se distribuye del modo siguiente:

Posición	Atributo	
0	categoría	
1	tipo	
2	persona	(poseedor)
3	género	(poseído)
4	número	(poseído)
5	número	(poseedor)

Por ejemplo:

mi *casa*	**DP1CSS**
mis *casa*	**DP1CPS**
tu *casa*	**DP2CSS**
nuestra *casa*	**DP1FSP**
su *casa*	**DP3CS0**

En el caso de los pronombres, esta información se distribuye del modo siguiente:

Posición	Atributo	
0	categoría	
1	tipo	
2	persona	(poseedor)
3	género	(poseído)
4	número	(poseído)
5	caso	
6	número	(poseedor)
7	polaridad	

Por ejemplo:

la casa es ***mía***	**PX1FS0S0**
las casas son ***mías***	**PX1FP0S0**
la casa es ***tuya***	**PX2FS0S0**
la casa es ***nuestra***	**PX1FS0P0**
la casa es ***suya***	**PX3FS000**

Por regla general, las formas *mi*, *tu*, *su, …* se etiquetan como determinante posesivo, mientras que las formas *mío*, *tuyo*, *suyo, …* se etiquetan como pronombre posesivo. No obstante, ténganse en cuenta estructuras como las siguientes:

muy señor ***mío***	**DP1MSS**	(lema: mío)
muy señores ***míos***	**DP1MPS**	(lema: mío)
muy señor ***nuestro***	**DP1MSP**	(lema: nuestro)

2.3.7. Las formas CONMIGO, CONTIGO, CONSIGO

Estas formas se consideran un token complejo y, por tanto, deben ser tratadas con dos elementos <dtok> del modo siguiente:

```
Ejemplo 1:
<tok id="w-1">conmigo
        <dtok id="d-1-1" form="con" mfs="SPS00” lemma="con”/>
        <dtok id="d-1-2" form="mi" mfs="PP1CSO00” lemma="mí”/>
</tok>

Ejemplo 2:
<tok id="w-1">contigo
        <dtok id="d-1-1" form="con" mfs="SPS00” lemma="con”/>
        <dtok id="d-1-2" form="ti" mfs="PP2CSO00” lemma="ti”/>
</tok>

Ejemplo 3:
<tok id="w-1">consigo
        <dtok id="d-1-1" form="con" mfs="SPS00” lemma="con”/>
        <dtok id="d-1-2" form="si" mfs="PP3CNO00” lemma="sí”/>
</tok>
```

En el caso de la forma no estándar *comigo*, el análisis sería el siguiente:

```
<tok id="w-1" dform"comigo” nform="conmigo”>comigo
        <dtok id="d-1-1" form="con" mfs="SPS00” lemma="con”/>
        <dtok id="d-1-2" form="mi" mfs="PP1CSO00” lemma="mí”/>
```

</tok>

2.3.8. Las formas (A)DONDE, COMO, CUAL y CUANDO

Se reserva la etiqueta PA000000 para los pronombres relativos o interrogativos con valor adverbial e invariables *donde*, *adonde*, *como, cual* y *cuando* (y algunos usos de *cuanto*: ver apartado siguiente):

*la casa **donde** vives*	**PA000000**	(relativo con antecedente expreso)
*trabajo **donde** puedo*	**PA000000**	(relativo sin antecedente expreso)
*no sé **dónde** vives*	**PA000000**	(interrogativa indirecta)
*¿**dónde** vives?*	**PA000000**	(interrogativa directa)
*la manera **como** cantas*	**PA000000**	(relativo con antecedente expreso)
*me ayudó **como** pudo*	**PA000000**	(relativo sin antecedente expreso)
*no sé **cómo** cantas*	**PA000000**	(interrogativa indirecta)
*¿**cómo** cantas?*	**PA000000**	(interrogativa directa)
*vivieron **cual** reyes*	**PA000000**	(relativo con antecedente expreso)
*el momento **cuando** vino*	**PA000000**	(relativo con antecedente expreso)
*iré **cuando** me digas*	**PA000000**	(relativo sin antecedente expreso)
*no sé **cuándo** vino*	**PA000000**	(interrogativa indirecta)
*¿**cuándo** vino?*	**PA000000**	(interrogativa directa)

En el caso de la forma *como*, téngase en cuenta que se contemplan otras posibilidades de anotación: conjunción (subordinada o coordinada), preposición y adverbio:

*es (tan) rubio **como** su padre*	**CS**	(conjunción comparativa)
*lo aceptaba tal **como** venía*	**CS**	(conjunción comparativa)
*no hay **como** correr para sudar*	**CS**	(conjunción comparativa)
***como** no estudié, me suspendieron*	**CS**	(conjunción causal)
***como** no estudies, suspenderás*	**CS**	(conjunción condicional)
*ya verás **como** no llega a tiempo*	**CS**	(conjunción completiva)
*invitaron **tanto** a Juan **como** a Pedo*	**CC**	(conjunción copulativa)
*invitaron así a Juan **como** a Pedro*	**CC**	(conjunción copulativa)
*pusieron su casa **como** aval*	**SPS00**	(preposición: “en calidad de”)
***como** presidente, no lo hizo mal*	**SPS00**	(preposición: “en calidad de”)
*llevaba **como** dos mil reales*	**RG**	(adverbio:"aproximadamente")

En el caso de la forma *cual*, téngase en cuenta que se contemplan otras posibilidades de anotación: relativo (pronombre) o interrogativo (determinante o pronombre):

*sean **cuales** sean sus razones*	**PR0CP000**
*se muestra siempre tal **cual** es*	**PR0CS000**
*perdió su dinero, el **cual** no era mucho*	**PR0CS000**
*¿**cuál** quieres?*	**PT0CS000**
*¿**cuál** camisa quieres?*	**DT0CS0**

En la anotación del corpus portugués, téngase en cuenta que la forma *donde* puede adoptar dos soluciones, dependiendo de si se trata de una contracción o de una variante de *onde*.

Las dos posibles soluciones son:

- Como token simple marcado como variante de *onde*.

*Ejemplo: A terra **donde** nasci*

```
<tok id="w-1" dform="donde" nform="onde" mfs="PA000000" lemma="onde">donde</tok>
```

- Como token complejo formado por dos dtoks (contracción).

*Ejemplo: Venho **donde** nasci* (= Venho do sítio onde nasci)

```
<tok id="w-1">donde
        <dtok id="d-1-1" form="de" mfs="SPS00" lemma="de"/>
        <dtok id="d-1-2" form="onde" mfs="PA000000" lemma="onde"/>
</tok>
```

Nótese, finalmente, que las formas portuguesas *aonde* y *adonde* son siempre tratadas como variantes de *onde*, nunca como contracciones. Por tanto:

*Ejemplo: A terra **aonde** nasci*

```
<tok id="w-1" dform="aonde" nform="onde" mfs="PA000000" lemma="onde">aonde</tok>
```

*Ejemplo: A terra **adonde** nasci*

```
<tok id="w-1" dform="adonde" nform="onde" mfs="PA000000" lemma="onde">adonde</tok>
```

2.3.9. La forma CUANTO

Para la forma *cuanto* se contemplan varias posibilidades. Si funciona como determinante, se anota como interrogativo (T), tanto en su valor propiamente interrogativo como en su valor relativo:

*¿**cuánto** dinero tienes?*	**DT0MS0**
*le entregué **cuantas** monedas tenía*	**DT0FP0**

Si funciona como pronombre, se anota como interrogativo (T), exclamativo (E) o relativo (R), según corresponda:

*¿**cuánto** tienes?*	**PT0MS000**
*¡**cuántas** tienes!*	**PE0FP000**
*miré **cuanto** me rodeaba*	**PR0MS000**

Si funciona como adverbio, es invariable y se anota igual que *donde*, *adonde*, *cuando* y *como* (ver apartado anterior):

*en **cuanto** a eso, no tengo opinión*	**PA000000**
*vendré **cuanto** antes*	**PA000000**
*por **cuanto** sé de ese tema, se conseguirá*	**PA000000**

Finalmente, la forma *cuanto* se anota como conjunción coordinada (CC) en correlación con *tanto*, en estructuras como la siguiente:

***tanto** Juan **cuanto** Pedro se retrasaron*	**CC**

2.3.10. La forma AUNQUE

La forma *aunque* se anota como conjunción coordinada (CC) cuando tiene valor adversativo, y como conjunción subordinada (CS) cuando tiene valor concesivo:

*es listo, **aunque** un poco vago*	**CC**
*iré a trabajar **aunque** llueva*	**CS**

2.3.11. La forma SI

Para la forma *si* se contemplan dos posibilidades dependiendo de si encabeza una condicional o una interrogativa indirecta:

***Si** llueve, no iremos al cine*	**CS**
*Me preguntó **si** íbamos al cine*	**CSI**

2.3.12. La forma DEMÁS

Para la forma *demás* se contemplan varias posibilidades. Si funciona como determinante o pronombre, se incluye dentro de los indefinidos (I):

*Los **demás** amigos no vinieron*	**DI0CP0**
*Los **demás** no vinieron*	**PI0CP000**

En la construcción *lo demás* equivale a “lo restante, el resto” y, por tanto, se etiqueta como nombre:

*Lo **demás** no me interesa*	**NCMS000**

Finalmente, la forma *además* se anota como adverbio en todas sus ocurrencias:

***Además** de listo, es honrado*	**RG**
***Además**, es listo y honrado*	**RG**

2.3.13. Las formas negativas NO, NI, NADA, NADIE, NINGUNO y NUNCA

La anotación de estas formas negativas se realiza del modo siguiente. Las formas *no* y *nunca* se anotan como adverbio de negación en todas sus ocurrencias:

***No** escribo más cartas*	**RN**
***Nunca** escribe cartas*	**RN**

La forma *nadie* se anota como pronombre indefinido en todas sus ocurrencias:

***Nadie** lo convenció*	**PI0NN00N**

Para la forma *ni* se contemplan dos posibilidades. En la estructura *ni...ni* se anota como conjunción copulativa. En el resto de ocurrencias, se anota como conjunción copulativa negativa:

***Ni** tengo dinero **ni** lo tendré*	**CC**
***Ni** come **ni** deja comer*	**CC**
*Nunca lo hizo **ni** lo hará*	**CCN**
*No lo hizo **ni** lo hará*	**CCN**
Ni siquiera sé si lo hizo	**CCN**

La forma *nada* puede ser pronombre indefinido (equivalente a “ninguna cosa”) o adverbio de cantidad (equivalente a “en modo alguno”, “de ninguna manera”):

Nada *es para siempre*	**PI0NN00N**
No vio ***nada*** *extraño*	**PI0NN00N**
No hubo ***nada*** *particular*	**PI0NN00N**
No hubo ***nada*** *de particular*	**PI0NN00N**
No tengo ***nada***	**PI0NN00N**
No tengo ***nada*** *más*	**PI0NN00N**
No estoy ***nada*** *contento*	**RQ**
El licor no me sienta ***nada*** *bien*	**RQ**

Las forma *ninguno(s), ninguna(s)* pueden ser pronombre indefinido o determinante indefinido:

No hizo ***ningún*** *comentario*	**DI0MS0**
No hizo comentario ***ninguno***	**DI0MS0**
No tuvo ***ninguna*** *respuesta*	**DI0FS0**
No tuvo ***ninguna***	**PI0FS00N**
No hizo ***ninguno***	**PI0MS00N**

2.3.14. Las formas TAN y TANTO

Para la forma *tanto* se contemplan varias posibilidades. Como determinante o pronombre, presenta variación de género y número y se incluye dentro de los indefinidos (I).

Tengo ***tantos*** *amigos como tú*	**DI0MP0**
Tengo ***tantos*** *como tú*	**PI0MP000**
Pasó ***tanta*** *hambre que casi muere*	**DI0FS0**
Pasé ***tanta*** *que casi me muero*	**PI0FS000**

Como adverbio, es invariable y se marca como adverbio comparativo (RC):

Uno de esos amigos que ***tanto*** *quiero*	**RC**
Es ***tanto*** *lo que queda por hacer*	**RC**
Corrió ***tanto*** *como su hermano*	**RC**

Ante adjetivos y adverbios, el adverbio *tanto* se apocopa en *tan*. No obstante, no se apocopa cuando va precedido de *más*, *menos*, *mayor*, *menor*, *mejor*, *peor* con valor comparativo. Se marca igualmente como adverbio comparativo:

Es ***tan*** *grande como su hermano*	**RC**
Es ***tan*** *grande que no cabe allí*	**RC**
Jamás oí frases ***tan*** *inteligentes*	**RC**
Nunca se fue ***tan*** *lejos de la ciudad*	**RC**
Cuanto más tengo, ***tanto*** *más quiero*	**RC**

La forma *tanto* se anota como sustantivo en la construcción *otro tanto*:

*Otro **tanto** sucede con los emigrantes*	**NCMS000**

En correlación con *como* o *cuanto* forma una conjunción compuesta discontinua de valor copulativo:

***Tanto** yo **como** mi hijo iremos a casa*	**CC**
***Tanto** yo **cuanto** mi hijo iremos a casa*	**CC**

Finalmente, la forma *tanto* se anota como adverbio general (RG) cuando forma parte de locuciones adverbiales: *por tanto, con tanto, hasta tanto, al tanto, entre tanto, tan siquiera*, ... (ver apartado 2.3.22.)

2.3.15. La forma TAL

Para la forma *tal* se contemplan varias posibilidades. Acompañando a un sustantivo con el que concuerda en número, puede ser determinante demostrativo (D) o determinante indefinido (I), dependiendo de si su significado equivale a "ese, de esa clase" o a "tanto, tan grande":

***Tal** conducta provocó su expulsión*	**DD0CS0**
*Esperaron con paciencia las **tales** armas*	**DD0CP0**
*Eso me lo dijo un **tal** Alfonso*	**DD0CS0**
*Desató **tal** polémica que fue expulsado*	**DI0CS0**
*Tenía unos dolores **tales** que eran insoportables*	**DI0CP0**

Cuando sustituye a un nombre se anota como pronombre demostrativo:

*Sus amigas dejaron pronto de ser **tales***	**PD0CP000**

Finalmente, como forma invariable *tal* se anota como adverbio comparativo (RC) equivalente a "así, de esta manera". En estos casos aparece en correlación con la conjunción *como* (CS) o con el pronombre relativo *cual* (PR0CS000):

*Lo hizo **tal** (y) como dijo*	**RC**
*Lo hizo **tal** cual dijo*	**RC**

2.3.16. La forma MISMO

Para la forma *mismo* se contemplan varias posibilidades. Como adjetivo, presenta variación de género y número:

*Llegaré el **mismo** día que tú*	**AQ0MS0**
*Era la **misma** persona de siempre*	**AQ0FS0**

También se anota como adjetivo cuando forma parte de sintagmas nominales cuyo núcleo nominal es recuperable en el contexto (cf. apartado 2.3.21)

*Sus ideas seguían siendo las **mismas***	**AQ0FP0**

Como adverbio, la forma *mismo* (sin flexión) se utiliza pospuesta a otros adverbios o locuciones adverbiales como mero refuerzo enfático:

*Se durmió al lado **mismo** de la carretera*	**RG**
*Mañana **mismo** resolveremos tu problema*	**RG**

En estructuras comparativas y precedida del artículo *lo*, la forma *mismo* puede adquirir dos valores diferentes. Con valor nominal equivale a "la misma cosa" y se anota como adjetivo; con valor adverbial equivale a "de la misma manera" y se anota como adverbio comparativo:

*Eso no es lo **mismo** que aquello*	**AQ0MS0**
*Entiende lo **mismo** inglés que francés*	**RC**

2.3.17. Las formas MÁS, MENOS, MAYOR, MENOR, MEJOR y PEOR

Las formas *más* y *menos* se anotan como adverbio comparativo (C) en todas sus ocurrencias:

*No vi hombre **más** paciente*	**RC**
*Cada vez va **más** gente al teatro*	**RC**
*Ya no hay **más** que decir*	**RC**
*Corre **más** que yo*	**RC**
*Corre **más***	**RC**

Las formas *mayor* y *menor* se anotan como adjetivo en todas sus ocurrencias:

*Amigo de mi **mayor** veneración*	**AQ0CS0**
*Lo hizo sin **mayores** dificultades*	**AQ0CP0**

Para las formas *mejor* y *peor* se contemplan dos posibilidades: adjetivo (comparativo de bueno/malo) o adverbio (comparativo de bien/mal):

*Tiene un coche **mejor** que el mío*	**AQ0CS0**
*La tortilla está **mejor** con cebolla*	**AQ0CS0**
*Las **mejores** aceitunas son estas*	**AQ0CP0**
*Cierra **mejor** la ventana*	**RC**
*Desde que hago ejercicio estoy **mejor***	**RC**
*Dormí **peor** la otra noche*	**RC**

Precedidos del artículo *lo*, las formas *mejor* y *peor*, aunque aparecen fosilizadas en masculino y singular, se anotan como adjetivos (y no como adverbios), ya que en tales contextos equivalen al comparativo de *bueno/malo* y no al de *bien/mal* (cf. apartado 2.3.21):

*Lo **peor** de no dormir es el cansancio*	**AQ0CS0**

2.3.18. Las formas MUY, MUCHO y POCO

Para las formas *mucho/poco* se contemplan fundamentalmente dos posibilidades. Como determinante o pronombre, presentan variación de género y número y se incluyen dentro de los indefinidos (I):

*Había tragado **mucha/poca** agua*	**DI0FS0**
*Había tragado **mucha/poca***	**PI0FS000**

Como adverbio, ambas formas son invariables y se marcan como adverbio de cantidad (RQ). Precedidas del artículo *lo*, *mucho* y *poco* también se marcan como adverbio de cantidad (cf. apartado 2.3.21):

*Hablamos **mucho/poco** aquella tarde*	**RQ**
*Es **mucho/poco** lo que queda por hacer*	**RQ**
Lo mucho/poco que tengo te lo doy	**RQ**

Ante adjetivos y adverbios, el adverbio *mucho* se apocopa en *muy*. No obstante, no se apocopa cuando va precedido de *más*, *menos*, *mayor*, *menor*, *mejor*, *peor, antes, después* con valor comparativo

*Él estaba **muy** rojo por el sol*	**RQ**
*Ya se encontraba **muy** lejos de aquí*	**RQ**
*Tú estás **mucho** peor que yo*	**RQ**

En la construcción *mucho más/menos + sustantivo* la forma *mucho* es determinante que concuerda en género y número con el sustantivo:

*Necesitan **muchas** más cosas* **DI0FP0**

En la construcción *mucho mayor/menor + sustantivo* la forma *mucho* es adverbio que modifica a los adjetivos *mayor* y *menor*: No obstante, por influjo de la construcción anterior puede aparecer concordando con el sustantivo. Siempre es invariable cuando el sustantivo va antepuesto:

*Lo hizo con **mucho** mayor facilidad* **RQ**
*Lo hizo con **mucha** mayor facilidad* **RQ**
*Lo hizo con una facilidad **mucho** mayor* **RQ**

Precedido del indefinido *un*, la forma *poco* se interpreta como un sustantivo que significa "cantidad pequeña":

*Necesitan un **poco** de agua* **NCMS000**

2.3.19. Casos dudosos de lematización

- Pronombres personales (español).

forma	lema
yo	yo
me	me
mí	mí
nos	nos
nosotras	nosotros
nosotros	nosotros
te	te
ti	ti
tú	tú
os	os
vos	os
vosotras	vosotros
vosotros	vosotros
él	él
ella	él
ello[2]	ello
ellas	ellos

[2] La forma *ello* no se etiqueta como pronombre personal, sino como pronombre demostrativo neutro (PD0NN000), al igual que *esto*, *eso*, *aquello*.

ellos	ellos
la, las, lo, lo, los	lo
le, les	le
se	se
sí	sí

- Pronombres personales (portugués).

forma	lema
eu	eu
tu	tu
ele	ele
ela	ele
nós	nós
vós	vós
eles	ele
elas	ele
você	você
vocês	você
mim	mim
(co)migo	mim
ti	ti
(con)tigo	ti
si	si
(con)sigo	si
(con)nosco	nós
(con)vosco	vós
o, lo, no	o
a, la, na	o
os, los, nos	o
as, las, nas	o
lhe	lhe
lhes	lhe
te	te
vos	vos
me	me
nos	nos
se	se

- Adjetivos ordinales: lema por extenso incluso en formas apocopadas.

forma	lema
primero	primero
primer	primero
tercero	tercero
tercer	tercero

- Adverbios coordinados: lema por extenso.

pura y claramente

forma	lema
pura	puramente
clara	claramente

- Formas apocopadas o reducidas: lema propio

forma	lema
tan	tan
tanto	tanto
gran	gran
grande	grande
muy	muy
mucho	mucho
cualquier	cualquier
cualesquier	cualquier
cualquiera	cualquiera
cualesquiera	cualquiera

Excepciones: *algún*, *ningún*, *un, buen*

forma	lema
algún	alguno
alguno	alguno
ningún	ninguno
ninguno	ninguno
un	uno
uno	uno
buen	bueno

2.3.20. La anotación de nombres propios

La etiqueta utilizada para etiquetar nombres propios es NP00000. Por regla general, esta etiqueta se aplica a nombres de personas (antropónimos) y lugares (topónimos). No obstante, su uso se extiende también a otro tipo de entidades que no tienen por qué aparecer escritas con mayúscula. Para la anotación de nombres propios y otro tipo de entidades nombradas, ténganse en cuenta las normas siguientes:

- Los diferentes nombres referidos a *Dios* se anotan del modo siguiente:

Dios	NP00000	(lema: dios)
Jesús	NP00000	(lema: jesús)
Ihesu	NP00000	(lema: ihesu)
Jesucrito	NP00000	(lema: jesucristo)
XPO	NP00000	(lema: xpo)

- En las expresiones multipalabra denominativas, es decir, aquellas que se refieren a una persona, a un lugar o a cualquier otra entidad análoga a un nombre propio (instituciones, organismos, etc.), se anotan como NP00000 los sustantivos que forman parte de dicha expresión. El resto de elementos llevan la etiqueta que les corresponda gramaticalmente:

Su Divina Majestad	DP3CS0 + AQ0FS0 + NP00000
Espíritu Santo	NP00000 + AQ0MS0
Nuestro Señor	DP1MSP + NP00000
Nuestra Bendita Madre	DP1FSP + AQ0FS0 + NP00000

- Las abreviaturas de las fórmulas de tratamiento *vuestra merced* (VM), *vuestra señoría* (VSa) y sus correspondientes plurales (VMs, VSas) se anotan siempre como nombre propio:

VM, VSa, VMs, VSas	NP00000

- El resto de fórmulas de tratamiento se rigen por la norma general indicada para las expresiones multipalabra:

vuestra señoría	DP2FSP + NP00000
vuestra merced	DP2FSP + NP00000
vuestra ilustrísima	DP2FSP + AQSFS0
vuestra ilustrísima señoría	DP2FSP + AQSFS0 + NP00000
vuestra reverencia	DP2FSP + NP00000
vuestra magnífica reverencia	DP2FSP + AQ0FS0 + NP00000

- Los nombres de persona formados por dos o más palabras (i.e. nombre y apellidos) se anotan con NP00000 para todas las formas que lo componen. Esto incluye los apodos, pero no los determinantes que puedan aparecer en el interior del nombre:

Manuel García	NP00000 + NP00000
Manuel García el Viejo	NP00000 + NP00000 + DA0MS0 + NP00000

- Respecto al tratamiento de los honoríficos, ténganse en cuenta las normas siguientes:

a) Si el honorífico acompaña a un nombre propio, se anota como NP00000:

don Manuel García	NP00000 + NP00000 + NP00000
señor Manuel García	NP00000 + NP00000 + NP00000
señor don Manuel García	NP00000 + NP00000 + NP00000 + NP00000
san Bartolomé	NP00000 + NP00000
fray Bartolomé	NP00000 + NP00000
padre Bartolomé	NP00000 + NP00000
reverendo padre fray Bartolomé	AQ0MS0 + NP00000 + NP00000 + NP00000
hermana Vicenta	NP00000 + NP00000
licenciado abogado García	NP00000 + NP00000 + NP00000
amigo García	NP00000 + NP00000

b) Si el honorífico acompaña a un nombre común, se anotan ambos como NP00000:

señor alcalde	NP00000 + NP00000
señor escribano	NP00000 + NP00000

c) Si no hay honorífico o no funciona como tal, no se aplica la etiqueta NP00000:

el alcalde	DA0MS0 + NCMS000
mi amigo	DP1CSS + NCMS000

La siguiente clasificación tomada de Alina Villalva puede servir como referencia en la delimitación de honoríficos que son anotados como nombre propio en el corpus de *P.S. Post Scriptum*:

> *As expressões nominais classificadoras do interlocutor podem ser de muito diversa ordem, das relações familiares (exs. Tia Alice, Prima Teresa) à expressão dos afectos, frequentemente reforçada pela presença de um possessivo (exs. Amigo João, Minha Querida Lúcia), do reconhecimento do grau académico (exs. Mestre António Pereira, Professor Doutor Sebastião Araújo), à expressão de um título profissional (exs. Arquitecto Brás de Almeida, Engenheiro Costa Dias, Juíza Teresa Antunes), que popularizou de forma arrasadora o título académico de Doutor/ Doutora atribuindo-o à generalidade dos licenciados, ou da identificação de um título nobiliárquico (exs. Visconde de Paredes, Rainha de Inglaterra) a uma mera expressão de respeito (exs. Dona Isabel, Senhor Henrique). Sem esquecer as expressões que identificam cargos de poder, quer se trate do poder temporal (exs. Ministro Manuel da Silva Gonçalves, Presidente Santos Queirós), quer do espiritual (exs. Padre Videira, Senhor Bispo António Rodrigues)*

(Alina Villalva: http://64.71.144.19/nad/PrintVersion.php?aid=1878)

- Los vocativos, las salutaciones (<salute>), las firmas (<signed>) y los destinatarios del cierre de la carta (<salute subcat="addressee") se tratan como nombres propios y se rigen por las normas ya indicadas:

mi querido padre Bartolomé	DP1CSS + VMP00SM + NP00000 + NP00000
mi querido padre	DP1CSS + VMP00SM + NP00000
mi querido amigo	DP1CSS + VMP00SM + NP00000

2.3.21. Elipsis nominal y nominalizaciones

2.3.21.1. Normas generales

El artículo determinado puede combinarse con categorías diferentes del nombre, dando lugar a sintagmas nominales en los que no aparece un núcleo nominal. En el nivel sintáctico, este tipo de estructuras se anotan sistemáticamente como NP (*Noun Phrase*). En el nivel morfosintáctico, la anotación se rige por las pautas siguientes:

- Como regla general, cada palabra incluida en el sintagma nominal se anota con la etiqueta que le corresponda como forma aislada.

- En el caso de los adjetivos o participios, aquellos cuyo núcleo nominal sea recuperable en el contexto se anotan como adjetivos o participios; aquellos cuyo núcleo nominal no sea recuperable se anotan como nombre. Se incluyen entre estos últimos los adjetivos y participios léxicamente sustantivados. El lema de los participios que son anotados como nombre será la forma del participio en masculino singular.

- Los adjetivos o participios o adverbios precedidos de la forma invariable "lo" son siempre etiquetados con la categoría que les corresponda, pese a que el núcleo nominal no sea recuperable.

2.3.21.2. Tipología y ejemplos

<u>1. *el/la/los/las* + forma no nominal</u>

1.1. Adjetivos y participios:

1.1.1. Núcleo nominal recuperable

*Las personas **ricas** tienen dietas; las **pobres**, hambre* **AQ0FP0**
*Las personas **enriquecidas** tienen dietas; las **explotadas**, hambre* **VMP00PF**
(lema: enriquecer; lema: explotar)

1.1.2. Núcleo nominal no recuperable

*Los **ricos** tienen dietas; los **pobres**, hambre*	**NCMP000**
*Los **enriquecidos** tienen dietas; los **explotados**, hambre*	**NCMP000**
(lema: enriquececido; lema: explotado)	

1.2. Sintagmas preposicionales

*Los **de arriba** no paran de hacer ruido*	**SPS00 + RG**
*El **del coche** nos hizo una señal*	**SPS00 + NCMS000**

1.3. Oraciones de relativo

*Los **que están arriba** no paran de hacer ruido*	**PR0CN000 + VMIP3P0 + RG**
*El **que conduce** nos hizo una señal*	**PR0CN000 + VMIP3S0**

2. *lo* + forma no nominal

2.1. Construcciones enfáticas (las formas flexionables concuerdan con el nombre y el verbo de la oración subordinada):

*Me impresionó lo **blanca/dura/alta** que era aquella montaña*	**AQ0FS0**
*Me impresionó lo **cansada/mareada/enfadada** que estabas*	**VMP00SF**
*Me impresionó lo **lejos/cerca** que estaba la montaña*	**RG**

2.2. Construcciones no enfáticas (las formas flexionables en género aparecen en masculino y las formas flexionables en número aparecen en singular)

*Me impresionó lo **blanco/duro/alto** de la montaña*	**AQ0MS0**
*El niño duerme lo **justo/necesario/suficiente***	**AQ0MS0**
*Eso es lo **bueno/malo** de ser padre*	**AQ0MS0**
*Eso es lo **mejor/peor** de ser padre*	**AQ0CS0**
*Me impresionó lo **mucho/poco** que se quejaron*	**RQ**

2.3.22. Locuciones

El hecho de no considerar las locuciones y otras expresiones pluriverbales como una sola unidad en el nivel morfosintáctico obliga a reservar una etiqueta por cada token que forma parte de este tipo de estructuras. Para mantener una sistematización en el análisis, se ha optado por las siguientes soluciones de anotación:

No obstante	RN + RG
Sin embargo	SPS00 + NCMS000
Tal vez	DD0CS0 + NCFS000
A solas	SPS00 + RG
Por tanto	SPS00 + RG
Entre tanto	SPS00 + RG
En efecto	SPS00 + NCMS000
En seguida	SPS00 + VMP00SF (lema: seguir)

Con todo		SPS00 + PI0MS000
Acerca de		RG + SPS00
Tocante a		RG + SPS00
A pesar de		SPS00 + NCMS000 + SPS00
Apesar de	(PT)	RG + SPS00
Respecto a/de		NCMS000 + SPS00
Respeitante a	(PT)	RG + SPS00

2.3.23. Palabras en otras lenguas

No se anotan ni se lematizan.

2.3.24. Palabras indescifrables: <unclear>

Se anotan y se lematizan con el símbolo X

3. Anotação sintática (em português)

A anotação sintática do *corpus P.S Post Scriptum* segue o sistema originalmente concebido para a anotação dos *Penn Corpora of Historical English* (Santorini, 2010). A adaptação do sistema original à anotação de dados do português foi desenvolvida em colaboração com as equipas dos projetos Tycho Brahe, WOChWEL e Cordial-Sin e é apresentada no Manual de Anotação Sintática do Português.

Nas secções abaixo, descrevem-se as diferentes etapas do processo de anotação sintática do *corpus P.S. Post Scriptum*.

3.1. Conversão de etiquetas EAGLES em etiquetas CLAWS

A anotação sintática é introduzida sobre uma versão do *corpus* anotada morfossintaticamente com o sistema de etiquetas CLAWS. Assim, numa primeira fase, é necessário converter as etiquetas EAGLES (cf. Secção 2 do presente manual) em etiquetas CLAWS[3]. Esta operação é realizada automaticamente através da aplicação sequencial de dois *scripts*: o primeiro *script* substitui as etiquetas EAGLES por etiquetas CLAWS; o segundo desagrega as formas contraídas e enclíticas ou mesoclíticas. Uma forma como *dá-lho*, por exemplo, que no sistema EAGLES é anotada conforme ilustrado em (1), será transformada em (2) por ação do primeiro *script* e em (3) por ação do segundo:

(1) dá-lho VMIP3S0+PP3CSD00+PP3MSA00
(2) dá-lho/VB-P-3S+CL+CL
(3) dá@/ VB-P-3S @lhe@/CL @o/CL

Uma vez que muitas das operações de conversão e desagregação são sensíveis a aspetos lexicais, existe um conjunto de dois *scripts* para cada língua representada no corpus.

Estes *scripts* tomam como *input* os ficheiros com anotação POS descarregados a partir da página de *downloads* do TEITOK e são executados num Terminal através da seguinte sequência de comandos:

[3] Este passo não se aplica ao caso dos 350 ficheiros portugueses que, na fase inicial do projeto Post Scriptum, foram modernizados e morfossintaticamente anotados com a ferramenta eDictor, que atribuía etiquetas CLAWS. Tal processo implicava exportar, mediante a seleção da opção "tagged sentences" de eDictor, os ficheiros de texto que continham a versão normalizada da edição Post Scriptum (incluindo eventuais variantes regionais e sociais) e respetiva marcação POS. Eram ficheiros de texto que exigiam, antes de serem processados pelo parser de Dan Bikel, algumas operações de edição automática: limpeza de cabeçalhos de identificação, marcação de clíticos e de contrações com o sinal @, codificação em UTF-8 e conversão dos finais de linha para o formato UNIX.

cartas portuguesas perl input_syn_PT.pl nomedoficheiro.txt > nomedoficheiro_output.txt
perl fix_dtoks_PT.pl nomedoficheiro_output.txt

cartas espanholas perl input_syn_ES.pl nomedoficheiro.txt > nomedoficheiro_output.txt
perl fix_dtoks_ES.pl nomedoficheiro_output.txt

O ficheiro resultante desta operação apresenta os dados com o formato adequado à prossecução do processo de anotação, o qual envolve tarefas e ferramentas diferenciadas nas duas línguas. Assim, no caso das cartas portuguesas, os dados são dispostos no formato "one sentence per line" (cf. exemplo (4)); no caso das cartas espanholas, os dados são dispostos no formato "one word per line" (cf. exemplo (5)); em ambas as línguas, as frases são separadas por uma linha em branco.

```
(4)   Em@/P  @o/D  Monte/NPR  de@/P  @o/D  Caroço/NPR  ,/,  pergunta/VB-I-2S
      onde/WADV é/SR-P-3S o/D Monte/NPR de@/P @a/D-F Misericórdia/NPR  ,/,
que/WPRO é/SR-P-3S um/D-UM monte/N que/WPRO está/ET-P-3S   mesmo/FP    a@/P
@o/D pé/N de@/P @o/D Monte/NPR de@/P @o/D      Caroço/NPR ./.
```

```
(5)   Hara@/VB-R-3S
      @se/SE
      cuanto/WADV
      se/SE
      pudiere/VB-SR-3S
      para/P
      que/C
      Francavila/NPR
      no/NEG
      apele/VB-SP-3S
      más/ADV-R
      ./.
```

Para os utilizadores de sistemas operativos UNIX (incluindo Mac OSX) ou Linux, este ficheiro está preparado para receber anotação sintática. Os utilizadores de Windows, antes de passarem à etapa seguinte do processo, terão de converter este ficheiro em formato UNIX, codificando os carateres em UTF8 e traduzindo os finais de linha de "carriage return" e "line feed" ("\r\n") para apenas "line feed" ("\n"), o que poderá ser feito num editor de texto como o *Notepad*++, por exemplo.

Uma vez que o sucesso do parseamento depende da correção da anotação morfossintática, o resultado da conversão automática das etiquetas POS deve ser sempre cuidadosamente revisto.

3.2. Implementação da anotação sintática

3.2.1. Cartas portuguesas

A anotação sintática das cartas portuguesas é implementada com recurso a um *parser* de base estatística, criado por Collins (1999) e Bikel (2004), modificado por Seth Kulick para a construção do *Penn Treebank* e adaptado ao português por Pablo Faria.

Este *parser* corre sobre os ficheiros .txt produzidos pela operação descrita acima e é executado através da seguinte linha de comandos no Terminal:

tb-backgroundParsing.pl 500 8 10 nomedoficheiro.txt

O ficheiro gerado pelo parser recebe o mesmo nome que o ficheiro de *input* e tem uma extensão .psd (de "parsed"). Este ficheiro apresenta os dados anotados sob a forma parentetização etiquetada, conforme ilustrado em (6):

```
(6)

(IP-IMP (PP (P Em@)
          (NP (D @o)
              (NPR (NPR Monte) (P de@) (D @o) (NPR Caroço))))
        (, ,)
        (VB-I pergunta)
        (CP-QUE (WADVP-1 (WADV onde))
            (IP-SUB (ADVP *T*-1)
                (SR-P é)
                (NP-SBJ (D o)
                        (NPR (NPR Monte) (P de@) (D-F @a) (NPR Misericórdia))
                        (, ,)
                        (CP-REL (WNP-2 (WPRO que))
                            (IP-SUB (NP-SBJ *T*-2)
                                (SR-P é)
                                (NP-ACC (D-UM um)
                                    (N monte)
                                    (CP-REL (WNP-3 (WPRO que))
                                            (IP-SUB (NP-SBJ *T*-3)
                                                (ET-P está)
                                                (FP mesmo)
                                                (PP (P a@)
                                                    (NP (D @o)
                                                     (N pé)
                                                     (PP (P de@)
                                                         (NP (D @o)
                                                          (NPR (NPR Monte)
                                                               (P de@)
                                                               (D @o)
                                                               (NPR Caroço)))))))))))))
         (. .))
```

3.2.2. Cartas espanholas

Para a anotação sintática das cartas espanholas, o projeto *Post Scriptum* usa o programa *CorpusSearch 2* (Randall, 2005-15). Originalmente concebido como um motor de busca para *corpus* anotado, este programa integra atualmente a funcionalidade de revisão automática de *corpus*. Esta componente opera por meio de *queries* que combinam instruções de pesquisa com instruções de alteração, permitindo introduzir modificações num *corpus* anotado quer ao nível das etiquetas sintáticas, quer ao nível da estrutura sintática.

O aproveitamento desta funcionalidade do *CorpusSearch* para a criação da anotação sintática tem sido uma estratégia explorada no âmbito de vários projetos de construção de *corpora* com o formato do *Penn Treebank*. Adotando esta estratégia, a equipa do *Post Scriptum* escreveu 135 *queries* que, quando executadas em sequência, atribuem estrutura sintática aos dados do espanhol sob a forma de parentetização etiquetada.

A aplicação ordenada das *queries* do *P.S. Post Scriptum* é assegurada por um conjunto de 13 *scripts*. O primeiro destes *scripts* corre sobre o ficheiro .txt que resulta da operação descrita na Secção 3.1. e gera um ficheiro .psd; os restantes correm sequencialmente sobre o ficheiro .psd. *Queries* e *scripts* estão reunidos numa mesma diretoria, irmã da diretoria que contém o *CorpusSearch*.

A atribuição de estrutura sintática aos dados do espanhol implica, então, os seguintes passos:

- abrir o Terminal;
- ir para a diretoria que contém as *queries*/*scripts*;
- dar a seguinte sequência de comandos:

```
./make-flat-parse nomedoficheiro.txt
./PARS0 nomedoficheiro.psd
./PARS1 nomedoficheiro.psd
./PARS2 nomedoficheiro.psd
./PARS3 nomedoficheiro.psd
./PARS4 nomedoficheiro.psd
./PARS5 nomedoficheiro.psd
./PARS6 nomedoficheiro.psd
./PARS7 nomedoficheiro.psd
./PARS8 nomedoficheiro.psd
./PARS9 nomedoficheiro.psd
./PARS10 nomedoficheiro.psd
./PARS11 nomedoficheiro.psd
```

./PARS12 nomedoficheiro.psd

Os *outputs* das diferentes etapas são numerados sequencialmente. O *output* final da operação tem o mesmo nome do *input* de PARS0 (ficheiro .psd inicial).

3.3. Edição da anotação sintática

3.3.1. Correção da anotação

Quer o resultado do *parser* automático (cf. Secção 3.2.1.), quer o resultado das *queries* de anotação (cf. Secção 3.2.2.) têm de ser objeto de correção/revisão manual. Esta tarefa é desenvolvida com o apoio do *CorpusDraw*, uma interface gráfica de edição da anotação que é parte integrante do *CorpusSearch.* Nesta etapa do processo, procede-se à correção de eventuais erros de etiquetagem e estrutura, à introdução de categorias vazias em falta e à coindexação de constituintes interdependentes.

O *CorpusDraw* é invocado no Terminal, a partir da diretoria acima daquela que contém o *CorpusSearch* (diretoria CS, no exemplo abaixo), através do seguinte comando:

java -classpath CS/CS_2.003.03.jar drawtree/CorpusDraw nomedoficheiro.psd

Conforme se pode ver na Figura 37, esta *interface* apresenta a anotação sintática sob a forma de diagrama em árvore, estando dispostos no topo da janela os botões com as opções de edição da árvore:

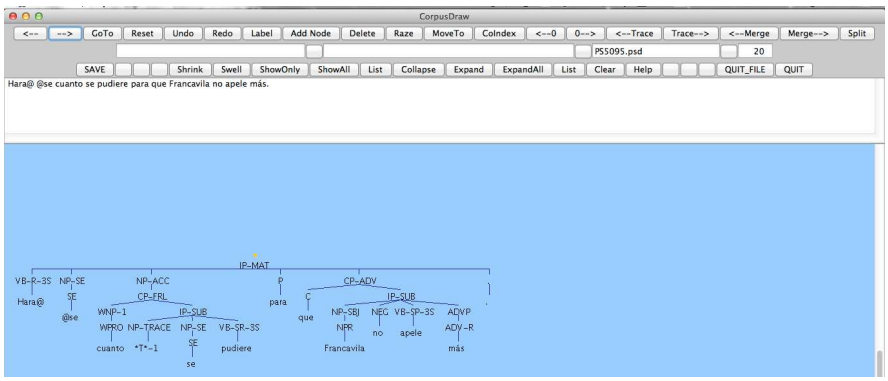


Figura 37. Aparência do *CorpusDraw*

De forma a reduzir a possibilidade de introduzir erros durante o processo de edição manual da anotação, a diretoria que contém o *CorpusSearch* é irmã de um documento intitulado LEGALTAGS.tag, que lista todas as etiquetas sintáticas

previstas no sistema de anotação do português/espanhol.

Um ficheiro editado no *CorpusDraw* é automaticamente gravado como .psd.new, mantendo-se o ficheiro de *input* com o nome original. No final de cada sessão de edição de uma carta, é conveniente correr um script de atualização dos nomes dos ficheiros. Este *script*, que atribui a extensão .old ao ficheiro de *input* e apaga a extensão .new do ficheiro de *output*, é executado através do seguinte comando:

```
update-new.bat nomedoficheiro
```

3.3.2. Atribuição de ID

Uma vez concluído o processo de correção/edição da anotação, é atribuído a cada frase um código identificador (ID), composto pelo código da carta e pelo número da frase. Para introduzir estes ID, é necessário dar o seguinte comando no Terminal:

```
tb-addID.pl nomedoficheiro.psd >  nomedoficheiro_id.psd
```

Se modificações à anotação, posteriores à atribuição dos ID, implicarem a renumeração de frases, será preciso remover os ID antes de os voltar a atribuir. Para executar o *script* de remoção dos ID, dá-se o seguinte comando:

```
tb-rmID.pl  nomedoficheiro_id.psd
```

3.4. Disponibilização da anotação sintática

3.4.1. Geração do ficheiro psdx

O alojamento da anotação sintática do *corpus* no TEITOK implica codificar as árvores sintáticas numa linguagem com sintaxe xml. Assim, numa primeira fase, o *script* psd2psdx.pl transforma a informação codificada sob a forma de parentetização etiquetada (ficheiro .psd) em informação codificada de acordo com os princípios estruturais do xml (ficheiro psdx). Os ficheiros .psdx são, pois, a versão xml dos ficheiros .psd.

No formato .psdx, a estrutura sintática é representada pela hierarquização de elementos xml, caracterizados por atributos que captam a restante informação da anotação. Apresenta-se abaixo o conjunto de elementos e atributos utilizados em psdx (cf. Tabela 1) e o exemplo de uma árvore sintática convertida neste formato (cf. (7)).

Elementos	Atributos	Descrição
forest		nó raiz de uma árvore sintática
eTree		nó sintático ou morfossintático (não terminal)
	Label	etiqueta sintática ou morfossintática
	index	índice numérico que codifica dependência sintática (emparelha com o índice de outro elemento da mesma árvore)
eLeaf		nó terminal (vazio ou com conteúdo lexical)
	Text	conteúdo lexical de um eLeaf
	Notext	conteúdo nulo de um eLeaf (categoria vazia)
	index	índice numérico que codifica dependência sintática (emparelha com o índice de outro elemento da mesma árvore)

Tabela 1. Elementos e atributos dos ficheiros PSDX

(7)
```
<forest forestId="18" File="PS5095" Location=".18" sentid="">
<eTree Label="IP-MAT">
<eTree Label="NP-SBJ">
<eLeaf Notext="*pro*-1" index="1"/>
</eTree>
<eTree Label="VB-R-3S">
<eLeaf Text="Hara@"/>
</eTree>
<eTree Label="NP-SE-1" index="1">
<eTree Label="CL">
<eLeaf Text="@se"/>
</eTree>
</eTree>
<eTree Label="NP-ACC">
<eTree Label="CP-FRL">
<eTree Label="WADVP-2" index="2">
<eTree Label="WADV">
<eLeaf Text="cuanto"/>
</eTree>
</eTree>
<eTree Label="IP-SUB">
<eTree Label="ADVP">
<eLeaf Notext="*T*-2" index="2"/>
</eTree>
<eTree Label="NP-SBJ-3" index="3">
<eLeaf Notext="*pro*"/>
</eTree>
<eTree Label="NP-SE-3" index="3">
<eTree Label="CL">
<eLeaf Text="se"/>
</eTree>
</eTree>
<eTree Label="VB-SR-3S">
<eLeaf Text="pudiere"/>
```

```
</eTree>
</eTree>
</eTree>
</eTree>
<eTree Label="PP">
<eTree Label="P">
<eLeaf Text="para"/>
</eTree>
<eTree Label="CP-ADV">
<eTree Label="C">
<eLeaf Text="que"/>
</eTree>
<eTree Label="IP-SUB">
<eTree Label="NP-SBJ">
<eTree Label="NPR">
<eLeaf Text="Francavila"/>
</eTree>
</eTree>
<eTree Label="NEG">
<eLeaf Text="no"/>
</eTree>
<eTree Label="VB-SP-3S">
<eLeaf Text="apele"/>
</eTree>
<eTree Label="ADVP">
<eTree Label="ADV-R">
<eLeaf Text="mÃ¡s"/>
</eTree>
</eTree>
</eTree>
</eTree>
</eTree>
<eTree Label=".">
<eLeaf Text="."/>
</eTree>
</eTree>
</forest>
```

Para gerar o ficheiro .psdx a partir do .psd correspondente, devem dar-se os seguintes passos:

- abrir o Terminal e entrar no servidor do *Post Scriptum*:

  ssh postscriptum@cards.clul.ul.pt

- ir para a diretoria PSDX, na qual deverá estar o ficheiro psd a converter;
- escrever a seguinte linha de comando:

  perl psd2psdx.pl Annotations/nomedoficheiro.psd --encoding=utf8 > Annotations/nomedoficheiro.psdx

3.4.2. Alinhamento do ficheiro .psdx com o ficheiro .xml

O último passo do processo de disponibilização da anotação sintática consiste no alinhamento do ficheiro .psdx, que contém a informação relativa à anotação sintática, e do ficheiro .xml, que contém a restante informação linguística e filológica. Esta operação é assegurada pelo *script* mergepsdx.pl, que é executado através do seguinte comando:

perl mergepsdx.pl nomedoficheiro

Este *script* alinha as frases anotadas com as frases de igual numeração do .xml e importa para o .psdx os <id> de cada *token*. O resultado tem a seguinte aparência:

```
(8) <forest forestId="18" File="PS5095" Location=".18" sentid="s- 19"
    id="tree-18">
    <eTree Label="IP-MAT" id="node-1093">
    <eTree Label="NP-SBJ" id="node-1094">
    <eLeaf Notext="*pro*-1" index="1" id="node-1095"/>
    </eTree>
    <eTree Label="VB-R-3S" id="node-1096">
    <eLeaf Text="Hara@" tokid="d-365-1" id="node-1097"/>
    </eTree>
    <eTree Label="NP-SE-1" index="1" id="node-1098">
    <eTree Label="CL" id="node-1099">
    <eLeaf Text="@se" tokid="d-365-2" id="node-1100"/>
    </eTree>
    </eTree>
    <eTree Label="NP-ACC" id="node-1101">
    <eTree Label="CP-FRL" id="node-1102">
    <eTree Label="WADVP-2" index="2" id="node-1103">
    <eTree Label="WADV" id="node-1104">
    <eLeaf Text="cuanto" tokid="w-366" id="node-1105"/>
    </eTree>
    </eTree>
    <eTree Label="IP-SUB" id="node-1106">
    <eTree Label="ADVP" id="node-1107">
    <eLeaf Notext="*T*-2" index="2" id="node-1108"/>
    </eTree>
    <eTree Label="NP-SBJ-3" index="3" id="node-1109">
    <eLeaf Notext="*pro*" id="node-1110"/>
    </eTree>
    <eTree Label="NP-SE-3" index="3" id="node-1111">
    <eTree Label="CL" id="node-1112">
    <eLeaf Text="se" tokid="w-367" id="node-1113"/>
    </eTree>
    </eTree>
    <eTree Label="VB-SR-3S" id="node-1114">
    <eLeaf Text="pudiere" tokid="w-368" id="node-1115"/>
    </eTree>
    </eTree>
    </eTree>
```

```
</eTree>
<eTree Label="PP" id="node-1116">
<eTree Label="P" id="node-1117">
<eLeaf Text="para" tokid="w-369" id="node-1118"/>
</eTree>
<eTree Label="CP-ADV" id="node-1119">
<eTree Label="C" id="node-1120">
<eLeaf Text="que" tokid="w-370" id="node-1121"/>
</eTree>
<eTree Label="IP-SUB" id="node-1122">
<eTree Label="NP-SBJ" id="node-1123">
<eTree Label="NPR" id="node-1124">
<eLeaf Text="Francavila" tokid="w-371" id="node-1125"/>
</eTree>
</eTree>
<eTree Label="NEG" id="node-1126">
<eLeaf Text="no" tokid="w-372" id="node-1127"/>
</eTree>
<eTree Label="VB-SP-3S" id="node-1128">
<eLeaf Text="apele" tokid="w-373" id="node-1129"/>
</eTree>
<eTree Label="ADVP" id="node-1130">
<eTree Label="ADV-R" id="node-1131">
<eLeaf Text="más" tokid="w-374" id="node-1132"/>
</eTree>
</eTree>
</eTree>
</eTree>
</eTree>
<eTree Label="." id="node-1133">
<eLeaf Text="." tokid="w-375" id="node-1134"/>
</eTree>
</eTree>
</forest>
```


Concluído este processo, a anotação sintática fica disponível no TEITOK para visualização, pesquisa *online* e *download*. O formato psdx, bem como o correspondente psd de cada ficheiro singular, é descarregável selecionando sucessivamente as opções *Syntactic annotation* e *Download file* na margem inferior da janela.