USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

INSTITUTO DA LINGUA GALEGA (ILG)

INSTRUCIÓNS PARA A ANOTACIÓN, CORRECCIÓN E LEMATIZACIÓN EN TEITOK

DOS TEXTOS DE

CORTEGAL, CORPUS DE TEXTOS GALEGOS ESCRITOS POR ESTUDANTES NO ÁMBITO ACADÉMICO

Novembro de 2022

MINISTERIO DE CIENCIA E INNOVACIÓN

UNIÓN EUROPEA
FONDO EUROPEO DE DESARROLLO REGIONAL

AGENCIA ESTATAL DE INVESTIGACIÓN

XUNTA DE GALICIA
CONSELLERÍA DE CULTURA, EDUCACIÓN, FORMACIÓN PROFESIONAL E UNIVERSIDADES

PGC2018-096069-B-I00

TÁBOA DE CONTIDOS

1. TOKENIZACIÓN E ALMACENAMENTO NA CARPETA TAGGED

Unha vez transcrito e revisado o texto, e almacenado na carpeta Revised, cómpre entrar nel e tokenizalo. Para iso, débese premer en "If you wish to tokenize the XML and proceed to the tokenized edit mode, click here". A continuación, renoméase, cambiándoo de carpeta e pasándoo a Tagged. Para iso dámoslle a Rename e poñemos 3_Tagged en vez de 7_Revised. Poñemos ademais _pr (previo) despois do número do documento, para saber que realmente aínda non está etiquetado.

Para que as palabras ou anotacións dun texto sexan buscables teñen que darse dúas circunstancias simultaneamente:

1) Que o texto estea na carpeta Tagged.
2) Que se rexenere o corpus. Se introducimos un texto en Tagged, as palabras ou anotacións dese texto non se encontrarán ata o momento en que se rexenere o corpus. O mesmo vale para calquera cambio que se faga sobre textos que xa son buscables. Mentres non se rexenere o corpus, ese cambio non será encontrado polo buscador. Para rexenerar o corpus, debemos ir a Administrar, "(re)generate the CQP corpus".

2. ANOTACIÓN DOS TEXTOS

2.1 Procedemento xeral

Unha vez tokenizados os textos e gardados en Tagged, procederemos á súa anotación. Iremos texto a texto e cando encontremos algunha diverxencia do estándar nun token debemos premer nel. Automaticamente aparece un formulario, que será o lugar en que levaremos a cabo a anotación. Para facela, seguiremos o *Manual de anotación das formas non estándares*, asignando a forma estándar e o código ou códigos que identifican o tipo e, no seu caso, orixe da desviación. Os códigos inclúense nos campo "Type of deviation of the standard" e "Source of the non-standard form" e a forma estandarizada no campo Ortographic Standard, Morphological Standard, Lexical Standard, Grammatical Standard, Semantic Standard ou Discoursive Standard, segundo corresponda.

> Así, por exemplo, se encontramos nun texto *saludables*, corrixiremos no nivel léxico (Lexical Standard), escribindo a forma estándar *saudables* e indicaremos en "Type of deviation of the standard" o tipo de desviación que presenta esta forma (substitución dunha unidade léxica estándar por unha non estándar, representada polo código L_w_su), así como en "Source of the non-standard form" a orixe (o español, feito representado por L_sp), tal e como se ve na Figura 1.

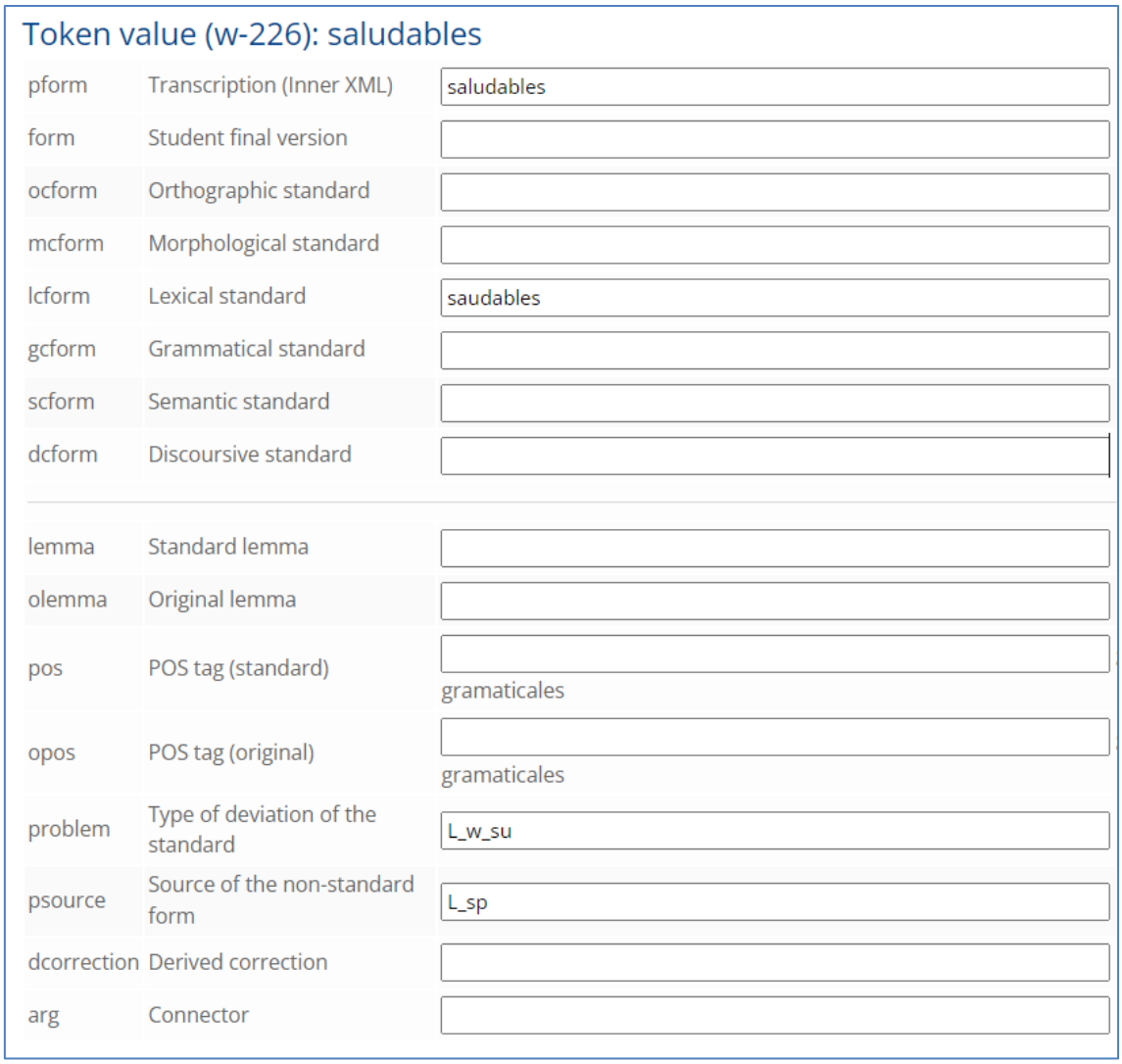


Figura 1. Corrección e asignación de código identificador dun problema en TEITOK

Debe terse en conta, en calquera caso, que **non sempre se asignará unha forma estándar**. Así, por exemplo, en "Tamén é certo que isto ten unha serie de desventaxas, xa que dentro deste está o que cada persoa poida permitirse gañar en cada momento", anotaremos un problema relativo á referencia de *este*, que é irrecuperable (D_ref_ad-su), pero resulta imposible corrixir o texto precisamente por non saber cal é o referente do demostrativo.

Antes ou despois de revisar un texto na vista normal, debe revisarse na **vista con cortes de palabra en final de liña** (ten que estar na vista Transcrición e activado Formato), para verificar se hai algún problema nos cortes de palabra en final de liña e daquela, se hai que introducir un código de O_hy_wp. A estes problemas asígnanselle código de desviación do estándar, pero non corrección.

2.2 Edición multitoken

No caso de encontrar un problema que consideremos que pode ser recorrente no corpus, e que sempre se vai corrixir da mesma maneira, podemos facer unha anotación xeral en todo o corpus. Para iso facemos a busca do token que nos interese (por exemplo *platos*) e debaixo dos resultados prememos en "use this query for multitoken edit".

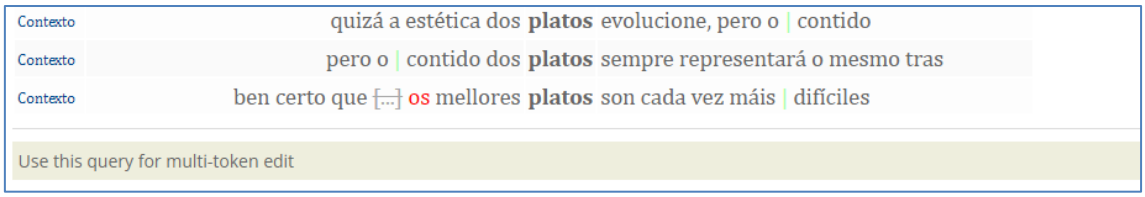


Fig. 2 Multitoken edit (I)

Ábrese un formulario onde poñemos as anotacións que nos interesan (neste caso, corrección léxica *pratos*, Tipo de desviación do estándar, L_w_su, Fonte da desviación do estándar L_sp e lema orixinal *plato*).

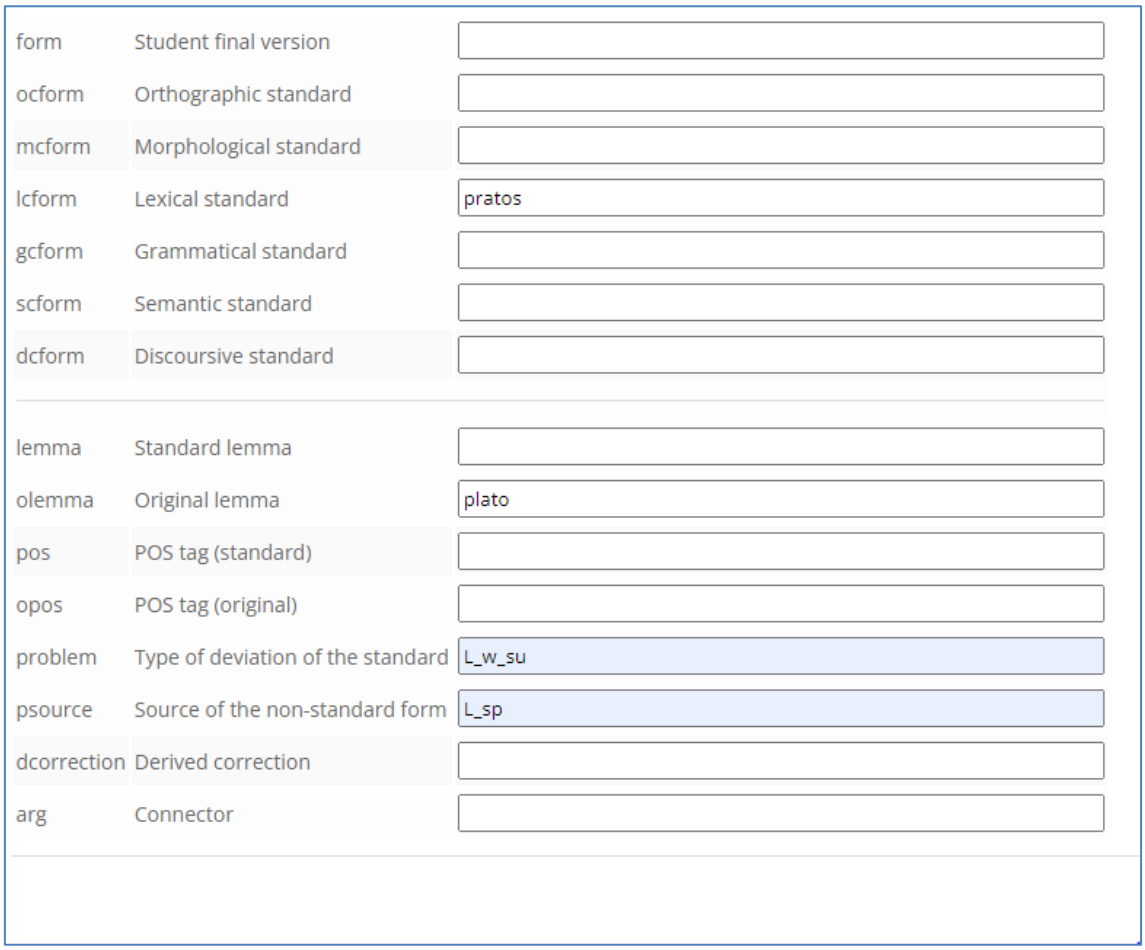


Fig. 3 Multitoken edit (II)

Embaixo do formulario figura a listaxe de textos resultado da busca anterior. Seleccionamos todos ("select all") despois de revisar que o cambio se debe aplicar a todos eles e dámoslle a "Change selected". No caso de que non haxa que seleccionar todos, vanse marcando un a un ou ben, se interesa máis, selecciónanse todos e despois desmárcanse os que non correspondan.

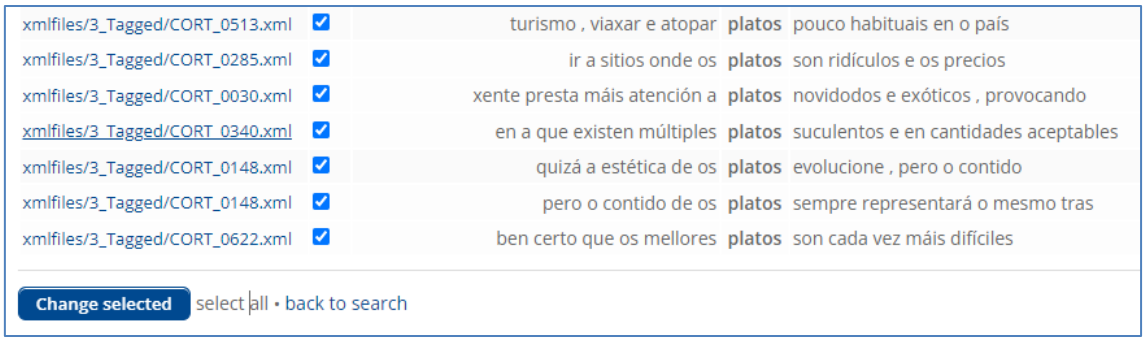


Fig. 4 Multitoken edit (III)

Con isto queda a anotación feita en todos os documentos seleccionados.

Importante!!:

1. **Antes de lematizar, non se pode poñer nada en Lema estándar e categoría gramatical estándar (por exemplo, no caso anterior non se pode poñer *prato* como lema). Se se fai iso, TEITOK di que o texto xa está lematizado e ao premer en Tag con Freeling non lematizará.**
2. **Antes de aplicar Multitoken edit, debemos rexenerar o corpus.**

2.3 Anotación de varias desviacións asociadas a unha mesma forma.

Unha mesma forma pode ter desviacións do estándar en diferentes niveis e todos os códigos correspondentes serán asignados conxuntamente no formulario de anotación. **Os diferentes códigos inclurianse en "Type of deviation of the standard" separados por coma, sen espazo**. Ordenaranse tendo en conta a orde en que figuran no *Manual de anotación das formas non estándares*.

> Así, por exemplo, a forma *hirmán* na secuencia "a hirmán" ten tanto un problema ortográfico (presenza de h, O_cons_ad) como unha diverxencia morfolóxica (M_gen_su). Ambos os códigos, tal e como se observa na Figura 5, aparecen na caixa "Type of deviation of the standard" separados por coma sen espazo. Primeiro figura o código relativo ao problema ortográfico e despois o código relativo á deviación morfolóxica, porque no manual de anotación O_cons_ad figura antes de M_gen_su.

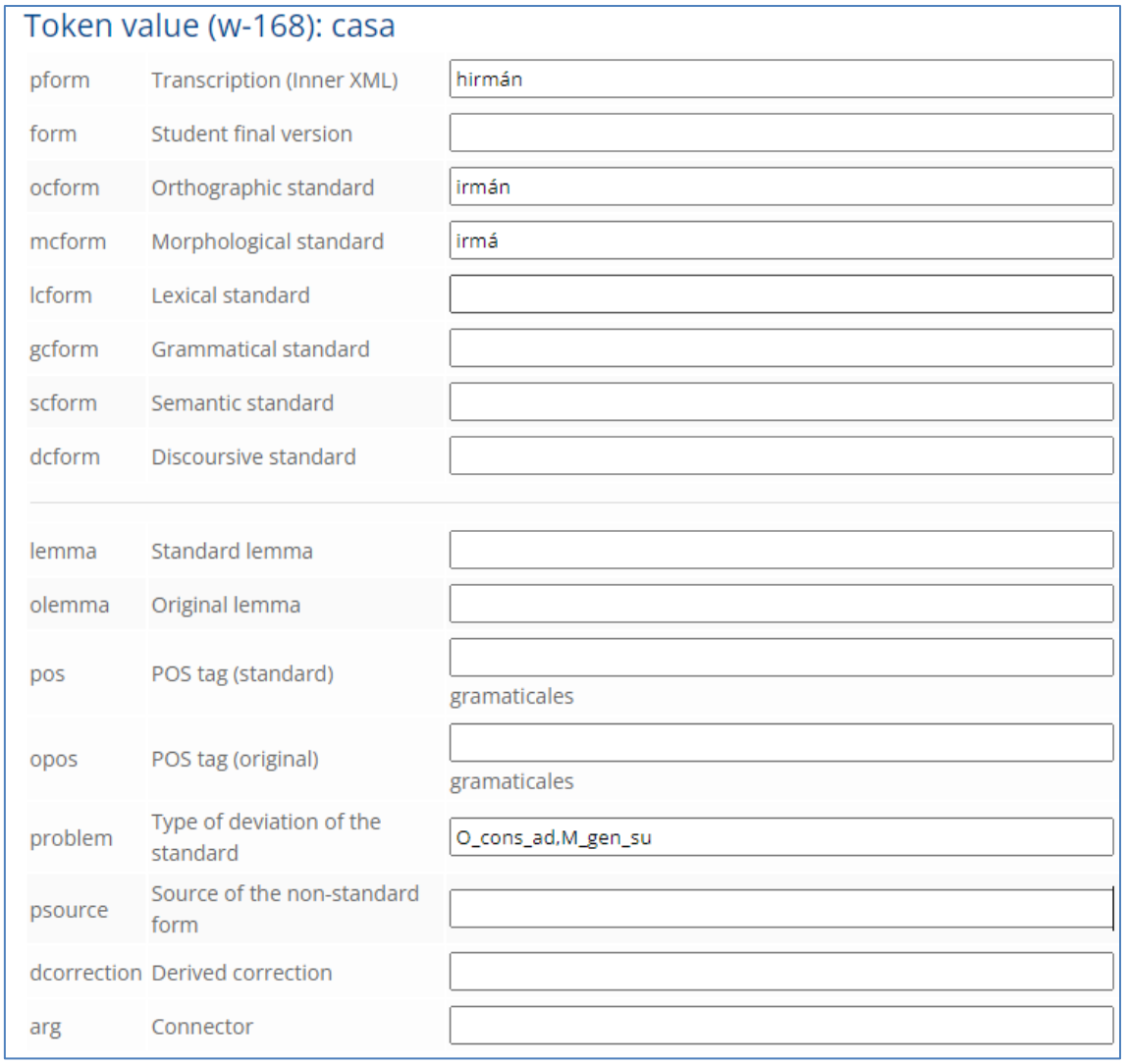


Figura 5. Estandarización e asignación de código para varias desviacións do estándar asociadas a unha mesma forma en TEITOK

Nótese que as estandarizacións que fagamos en cada nivel trasladámolas aos niveis inferiores. A corrección ortográfica realizada no nivel ortográfico (supresión do <h>) incorpórase no nivel morfolóxico: poñemos *irmá* sen <h> e non *hirmá*).

Pode haber **varias desviacións no mesmo nivel** e os códigos aparecerán tamén nese caso separados por coma sen espazo e na orde en que figuran no *Manual de anotación das formas non estándares*. Se unha mesma desviación se repite dúas veces (por exemplo, dous casos de O_cons_su na mesma palabra), o código tamén se repite.

> Por exemplo, tal e como se observa na figura 6, no caso de *vevian*, ofrécese no nivel ortográfico a forma correcta *bebían* e en Type of deviation of the standard, os tres códigos, correspondentes todos eles ao citado nivel: O_ac_om,O_cons_su,O_cons_su. A orde corresponde á que figura no manual de anotación.

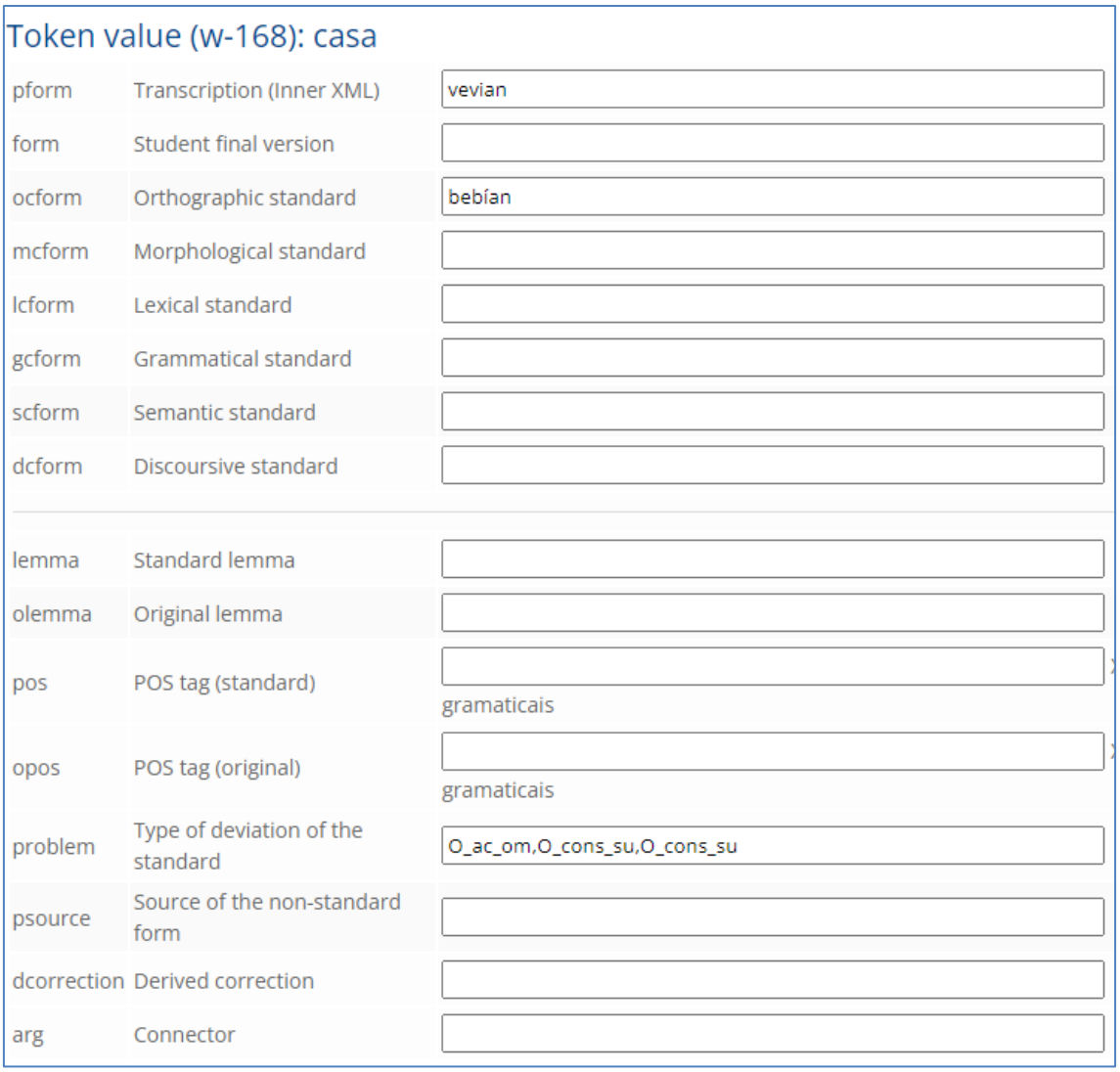


Figura 6. Anotación de problemas pertencentes ao mesmo nivel lingüístico

2.4 Anotación da fonte de desviación do estándar

Nalgúns casos é necesario indicar a **orixe** da desviación (véxase o *Manual de anotación das formas non estándares* para saber en que problemas hai que ofrecer esta información e os códigos correspondentes). Neste caso, inclúense os códigos no campo Source of the non-standard form, como no seguinte exemplo, en que lle asignamos a marca L_sp (Lexical_Spanish) ao castelanimo *receitas*.

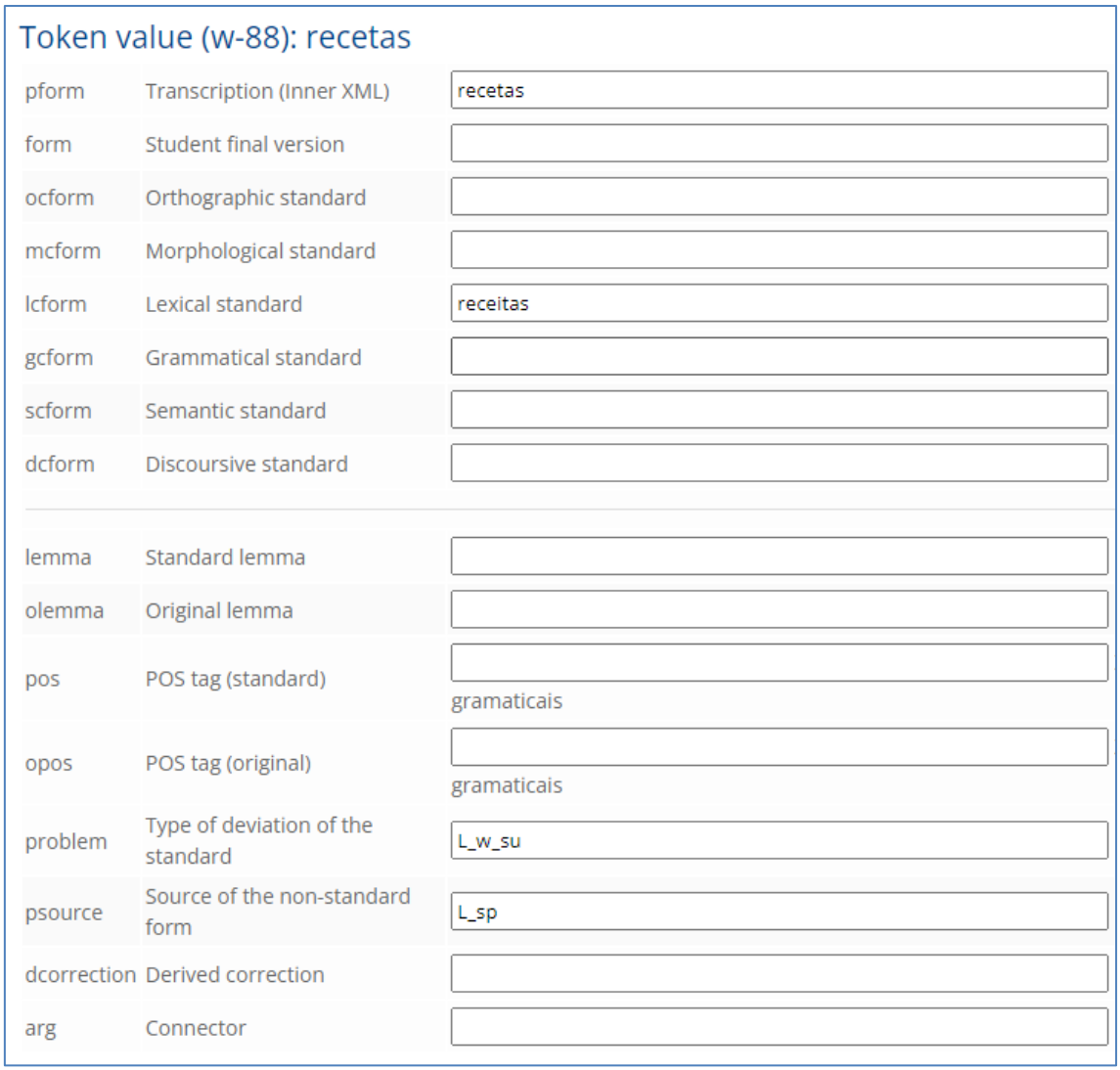


Figura 7. Asignación de códigos no campo fonte da forma non estándar

Dado que unha mesma forma pode ter varias desviacións pertencentes a diferentes niveis e dado que tales diverxencias poden ter orixes distintas, antes da marca identificadora da orixe (sp, gal...) identificamos o nivel ao que corresponde tal marca, mediante as mesmas letras que empregamos nos códigos de tipo de desviación do estándar, separadas por guión baixo (O, M, L, G, SP e D). Así teremos M_gal, L_sp, G_or etc. Se debemos atribuír diferentes orixes a unha mesma palabra (ben sexa, dúas explicacións posibles para unha mesma orixe, ben sexa dúas explicacións para dúas desviacións diferentes) estas separanse por coma sen espazo de separación.

> Así, por exemplo, se algúen escribe *salíamos* (forma que irá etiquetada con M_ac_su e con L_w_su) no campo Source escribiremos M_gal,M_sp,L_sp. As dúas primeiras etiquetas corresponden ás dúas posibles orixes do uso non estándar da acentuación e a terceira ao emprego do castelanismo *salir* por *saír*.

Os criterios de ordenación serán en primeiro lugar o nivel, na orde en que figuran no *Manual de anotación das formas non estándares* (O, M, L, G, SP e D) e en segundo lugar, para dúas explicacións asociadas ao mesmo nivel, a orde en que figuran expostas as diferentes marcas no citado manual, correspondente á ordenación alfabética (Táboa 1).

2.5 Correccións derivadas

Por outro lado, tal e como se indica no *Manual de anotación das formas non estándares* (apartado 6 do punto 1.5), as diverxencias do estándar que derivan dunha corrección previa, pero que non supoñen un problema no dominio do estándar por parte da/do estudante, corrixiranse, pero non se anotarán. Así, por exemplo, se estandarizamos *pobo* por *vila* no nivel semántico na frase "o pobo de Iago Aspas", a frase que resulta ao estandarizar será "o vila de Iago Aspas". Para evitar esta falta de concordancia, substituiremos tamén *o* por *a* **no mesmo nivel en que se fixo a estandarización** que deu lugar ao novo problema (neste caso no nivel semántico), pero non asignaremos, claro está, código de desviación do estándar. Para ter identificados estes casos, no formulario existe unha caixa chamada dcorrection ("Derived Correction"), que se marcará con DC nos casos en que se produza unha corrección deste tipo. No resto dos casos, quedará baleira.

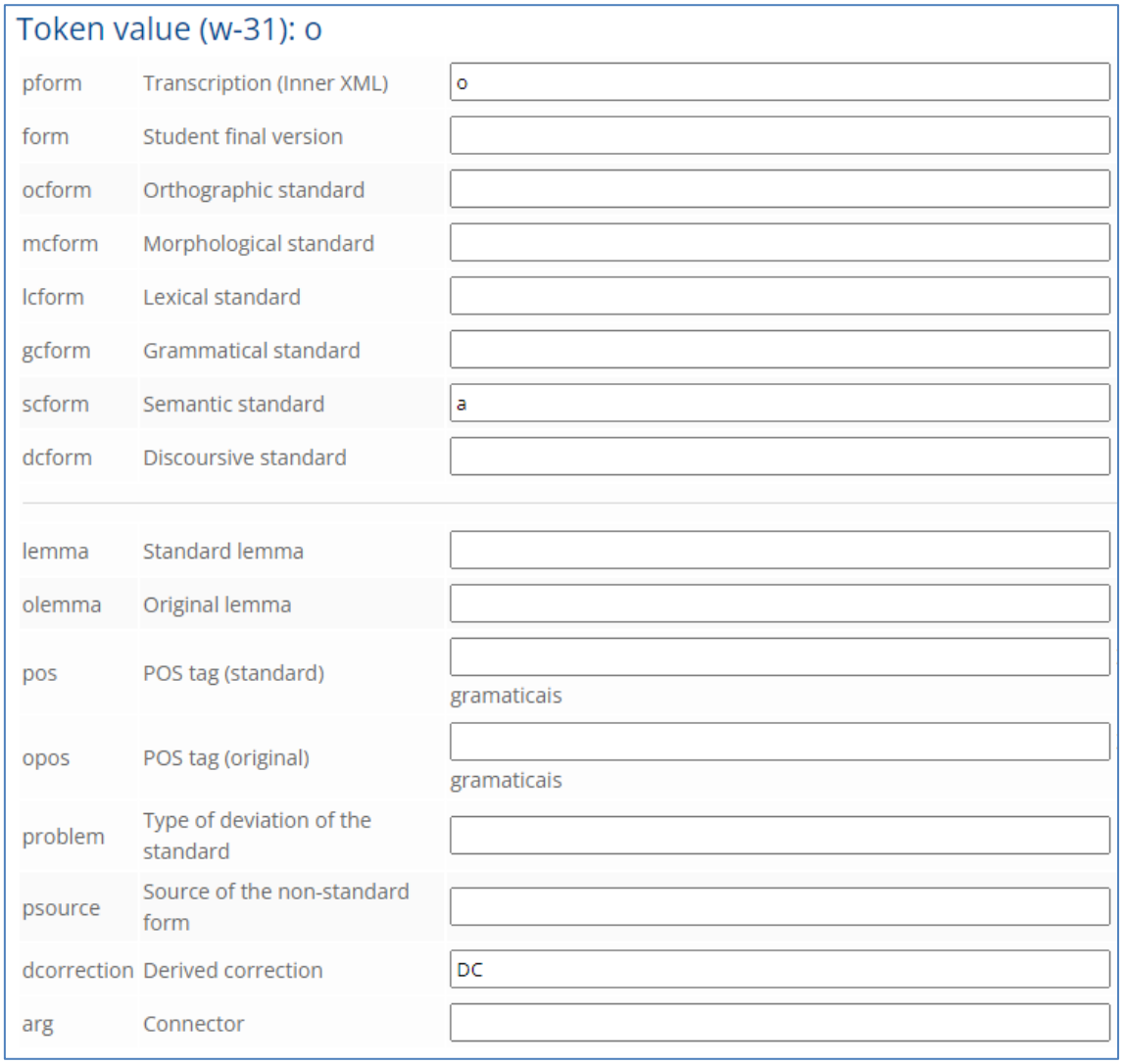


Figura 8. Anotación dunha corrección derivada

Empregaremos esta etiqueta tamén cando a/o estudante omite signos de puntuación "dobres" (por exemplo para separar un inciso). Así, en "Ana que vive aquí ten quince anos" é preciso

poñer unha coma despois de *Ana* e despois de *aquí* (“Ana, que vive aquí, ten quince anos”). Etitquetamos a primeira coma que engadimos con D_pm_om, pero na segunda poñemos DC.

Tamén usamos a etiqueta “Corrección derivada” para desviacións que afectan a varias palabras, pero que son manifestación dun único problema. Por exemplo:

En “tanto o consumismo como a produtividade segue e seguirá sendo no futuro imprescindible para a sociedade” hai un problema de concordancia, que afecta a *segue*, *seguirá* e *imprescindibles*. Se anotamos o problema nas tres formas estamos multiplicando por tres o que realmente é un único problema de concordancia. Por tal motivo, só colocamos a etiqueta G_num_su no **primeiro** elemento, *segue*, mentres que en *seguirá* e *imprescindible* poñemos a etiqueta DC. Agora ben, no caso de que situacións deste tipo afecten a unha frase cunha palabra léxica e palabras gramaticais, márcase só a léxica e as gramaticais van con DC, aínda que aparezan antes no texto. Por exemplo, en “A última moda no sector estudantil e de traballo ven sendo o chamado gastronomía”, corriximos por “a chamada”, poñemos en *chamada* G_gen_su e en *a* DC.

Non se anotarán como correccións derivadas as que se fan no marco da mesma palabra que sofre a estandarización.

> Por exemplo, se corriximos *convinar* por *combinar*, a presenza do <m> vén derivado da modificación de <v> por <b>, pero non poñeremos DC na corrección e unicamente poñeremos unha etiqueta de O_cons_su en Type of deviation of the standard, correspondente ao cambio de <b> por <v>, non ao de <m> por <n>. Do mesmo xeito, se ao corrixir cómpre colocar un pronome enclítico e como consecuencia hai que modificar a acentuación da palabra (“me dixeron”, “dixéronme”), neste caso tampouco poñemos DC en Derived correction.

2.6 Anotación dalgúns problemas particulares

A seguir, ofrécense algunhas indicacións máis concretas sobre o proceso de anotación.

2.6.1 Corrección mediante a supresión de tokens

En certos casos , é preciso estandarizar mediante a supresión duha palabra ou signo de puntuación. Por exemplo, se alguén escribe "porque se comeu todo", é necesario indicar que hai un problema de adición inadecuada de pronome e que a forma estándar sería "porque comeu todo". Nestes casos poñemos -- no nivel de corrección que corresponda, neste caso o gramatical. TEITOK interpreta esta marca como omisión á hora de xerar a capa de estandarización correspondente e a palabra substitúese por nada.

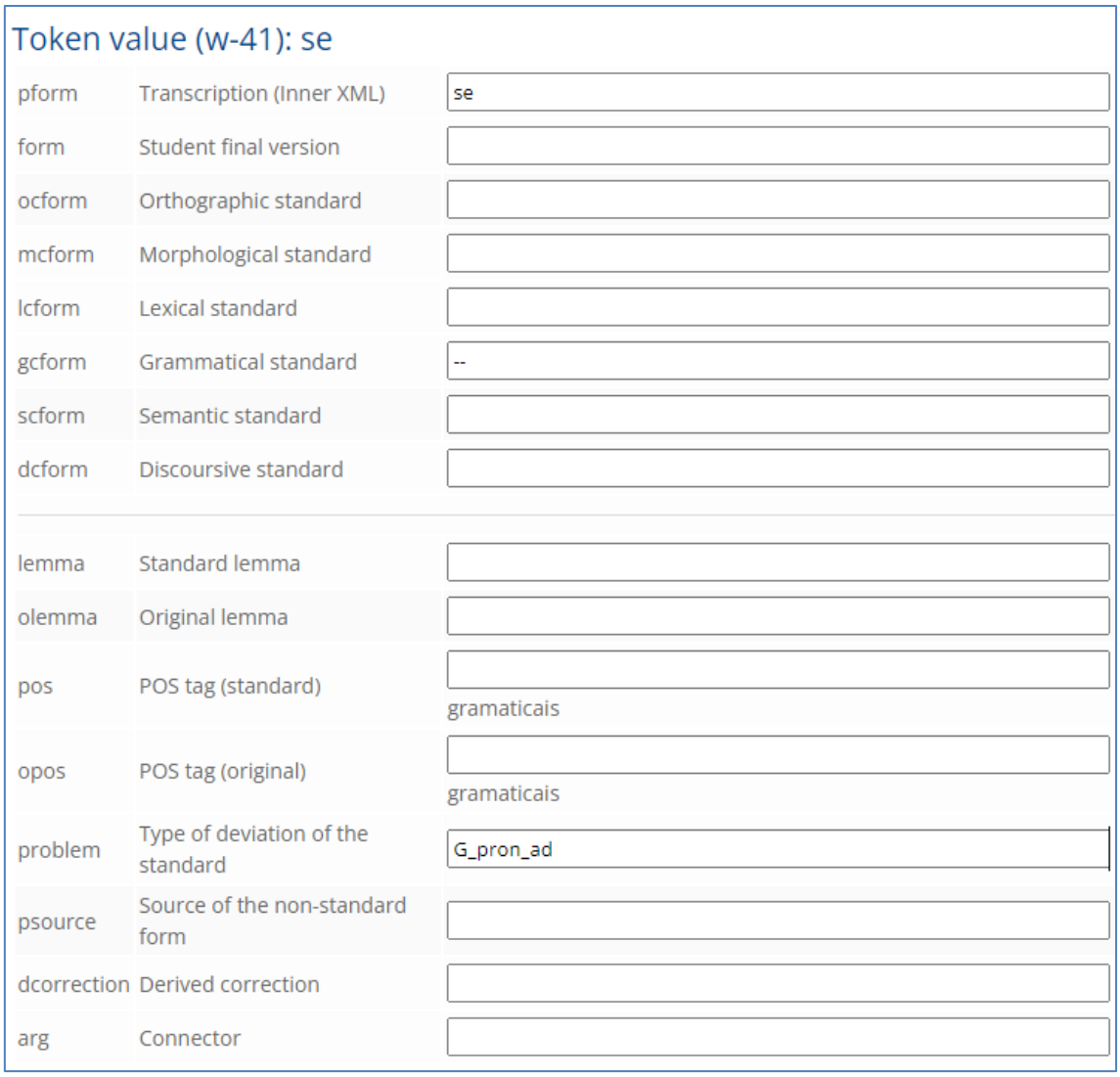


Figura 9. Anotación de casos de adición inadecuada de palabras ou signos de puntuación

2.6.2 Corrección mediante o engadido de tokens

Nalgúns casos, a corrección supón o engadido dun ou varios elementos (palabras ou signos). En tal caso, a non ser que a anotación deba facerse en standoff (vid. 2.7), debe procederse do seguinte xeito:

a) Cómpre premer na forma inmediatamente anterior ao lugar en que se debe inserir a palabra ou o signo.
b) A seguir, en “Insert tok after”, prémese en “attached” se o elemento debe ir pegado á forma (por exemplo, se é un signo de puntuación como un punto ou unha coma) e en “separated”, se debe ir separado (que é o que sucederá normalmente se se insire unha ou máis palabras).
c) No novo formulario que se abre, onde se xera <ee/> automaticamente en *pform*, insírese o elemento ou elementos na capa de estandarización que corresponda, así como o código ou códigos que corresponda(n). Así, no seguinte exemplo inseriuse a frase *do programa* despois de *caso* (na secuencia "Porén, algúns programas televisivos

non se apartan da tradición e simplemente incitan e dan máis popularidade a cociñar (é o caso de Carlos Sovera)").

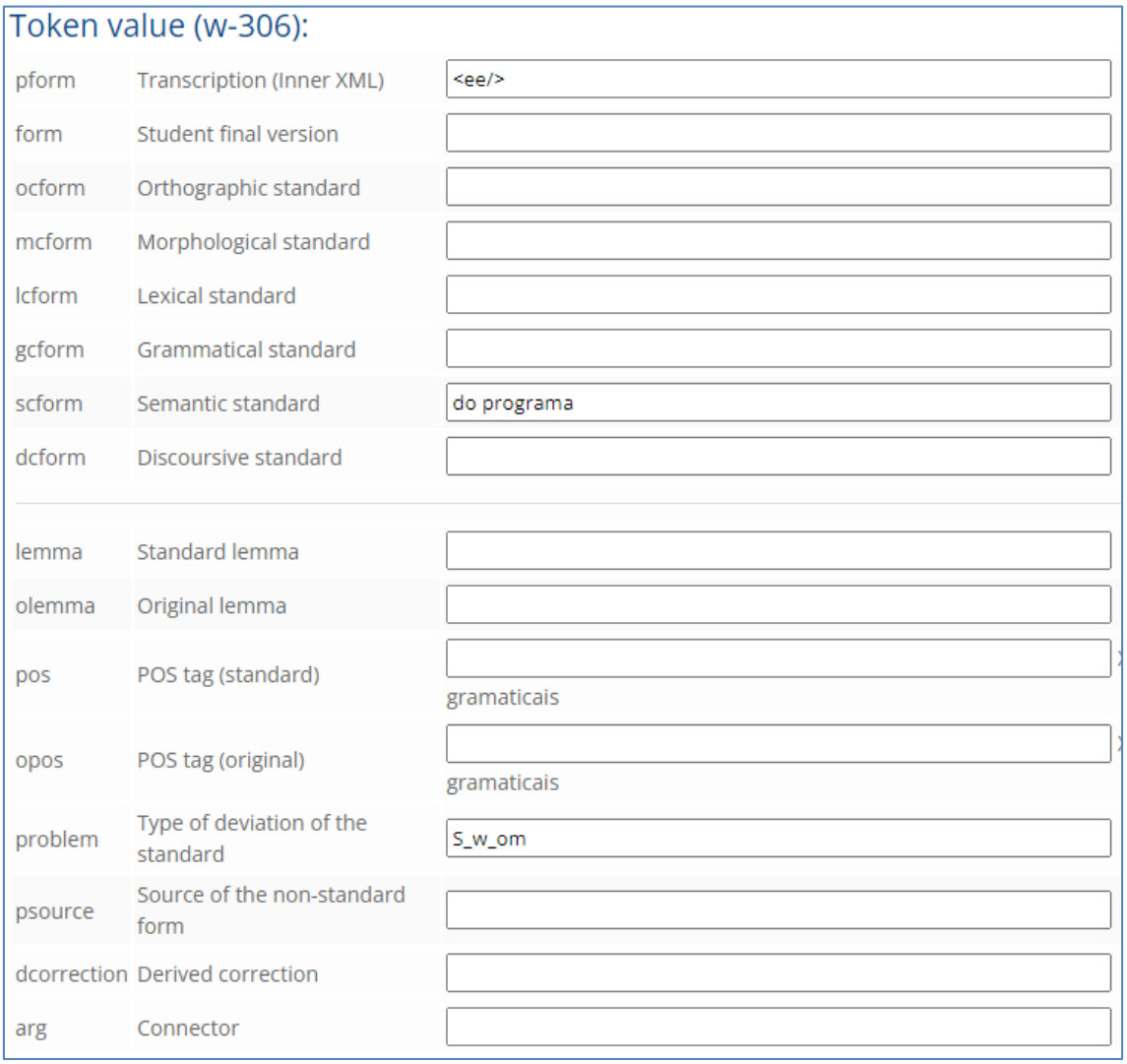


Figura 10. Anotación de casos en que cómpre engadir palabras ou signos de puntuación na estandarización

d) No caso de que o elemento que se deba inserir deba ir pegado á palabra seguinte (por exemplo, unha paréntese inicial), clicarase en “Insert tok before attached”.

2.6.3 Anotación de elementos constituídos por varias palabras

Este apartado trata sobre a anotación de unidades lexicalizadas constituídas por varios elementos (tipo *sen embargo*), así como de contraccións, asimilacións e grupos de verbo e clítico. As anotacións sobre outros grupos de palabras (por exemplo, unha secuencia que debe suprimirse: os libros necesarios que fan falta) faranse en standoff e deben consultarse as indicacións ao respecto (vid. 2.7).

Ante un token complexo, TEITOK permite crear dous ou máis d-token para cada un dos seus constituíntes. Se TEITOK identifica ese token complexo (faino en principio nas contraccións e combinacións de verbo e pronome enclítico estándares), crea automaticamente os d-tokens correspondentes no proceso de lematización con Freeling[1]. Así, por exemplo, para *coa* crea o d-token *con* e *o.* Se hai unha unidade complexa con varios d-tokens, o sistema só permite realizar buscas sobre estes e sobre as súas anotacións (e non sobre a unidade complexa e o que está anotado nela), de modo que debemos ter isto en conta á hora de levar a cabo a anotación e corrección.

2.6.3.1 Contraccións, asimilacións e grupos de verbo e clítico

No caso de que teñamos unha **desviación relativa a un dos compoñentes da contracción**, anotaremos o problema sobre o compoñente que corresponda. Así, por exemplo, se alguén escribe "interese coa rapaza" en vez de "interese pola rapaza", onde se selecciona a preposición *con* en vez de *por*, no d-token *con,* e no nivel semántico, asignaremos a forma estándar *por* e en "Type of deviation of the standard" indicaremos que hai un caso de S_w_su. Con todo, ademais diso, no Token completo (*coa*) asignaremos como forma correcta, tamén no nivel semántico, *pola*, porque se non, a estandarización non se visualizaría na capa semántica. Agora ben, aquí non asignamos código, tal e como se pode comprobar na Figura 11.

[1] Con todo, hai un problema con esta cuestión, posto que o segundo d-Token das contraccións e dos grupos de verbo e clítico créao en masculino singular, cando debería facelo mantendo a flexión (por exemplo, no caso de *coa* xera *con* e *o* e debería xerar *con* e *a*). Temos o mesmo problema no caso dos verbos, pois debería xerar a forma concreta que apareza (cólleo= colle + o), pero xera o infinitivo por sistema (*coller* + *o*). Por tal motivo, todos estes casos deben corrixirse manualmente (si asigna axeitadamente a categoría gramatical). Véxase o apartado sobre lematización.

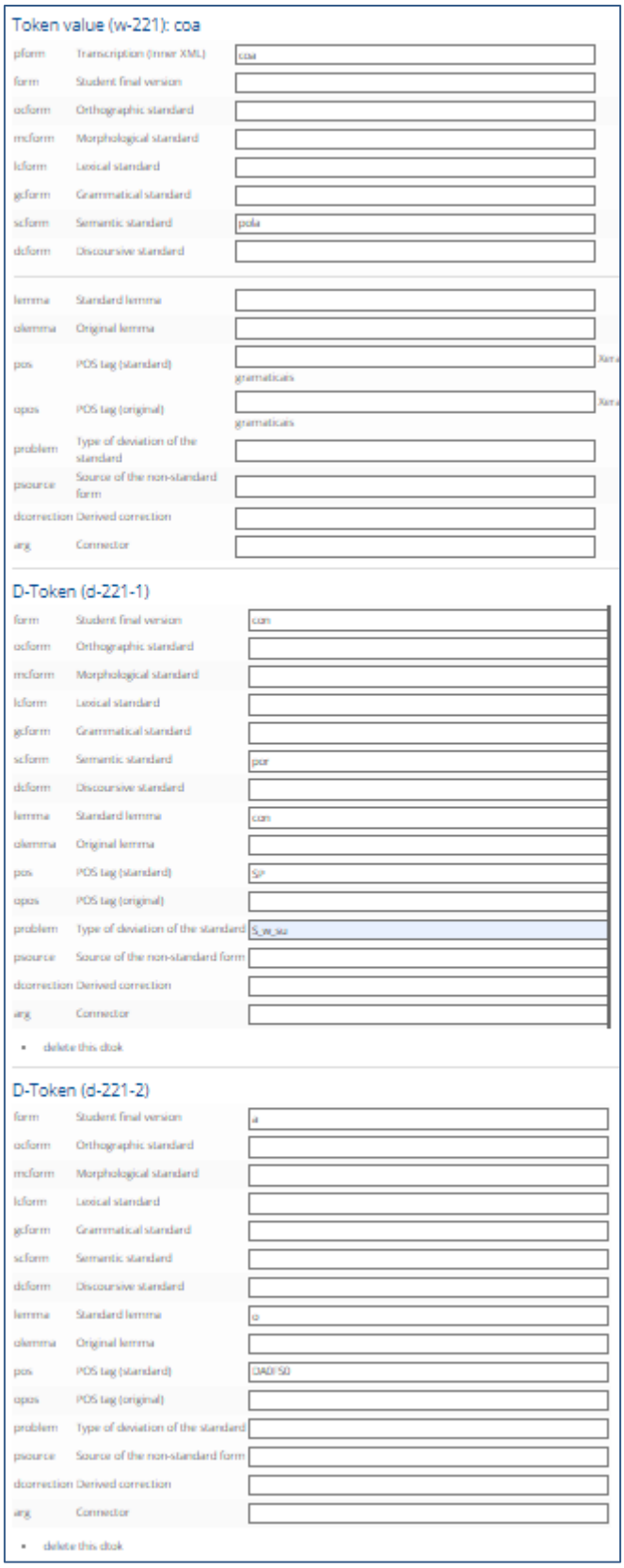


Figura 11. Anotación dos problemas que afectan a un dos elementos das contraccións

O mesmo é aplicable aos casos en que teñamos un verbo e un clítico. Así, por exemplo, "vinche onte" será corrixido do seguinte xeito.

1) En *vinche*, poñemos *vinte* na capa de estandarización gramatical, pero non asignamos código de desviación do estándar.
2) En *che*, estandarizamos con *te* na capa de estandarización gramatical e asignamos código G_pron_su.

Nos casos anteriores e nos que veñen a seguir, **a anotación pode facerse despois da lematización, unha vez que TEITOK xere os d-tokens** (neste caso tomarase nota para non esquecer esta corrección) **ou ben poden xerarse os d-tokens manualmente**, premedo debaixo do formulario no número que acompaña a "Split in dtoks", (2), escribindo os d-tokens manualmente en *form* e asignando a anotación e corrección nese momento.

O principais problemas xorden **cando a desviación afecta a toda a unidade**, isto é, cando temos unha diverxencia na propia conformación da contracción (ou na unión de verbo e pronome)[2]. Así, se unha persoa escribe *ca* en vez de *coa*, a anotación sería así (vid. Figura 12):

1. Estandarizamos sobre o token completo para que na capa de visualización ortográfica se vexa a estandarización.
2. Estandarizamos tamén sobre o primeiro d-token para que a persoa usuaria poida encontrar todos os casos en que se corrixiu para *coa*.
3. Asignamos código de desviación do estándar no primeiro d-token para que a persoa usuaria poida buscar todos os casos de O_cont_su.

Unha anotación do mesmo tipo valería para casos en que se asigna unha contracción non estándar (por exemplo, *dalgo*), por suposto, cambiando o código (O_cont_ad).

[2] Se existen tanto a unidade conxunta como d-tokens TEITOK non permite facer buscas sobre o token completo, nin sobre as súas anotacións (só sobre os d-tokens). Así pois, se incluímos na unidade completa a corrección e o código de desviación do estándar estes non se poderán encontrar. Por tal motivo, nestes casos imos anotar a desviación no primeiro d-token da contracción / asimilación.

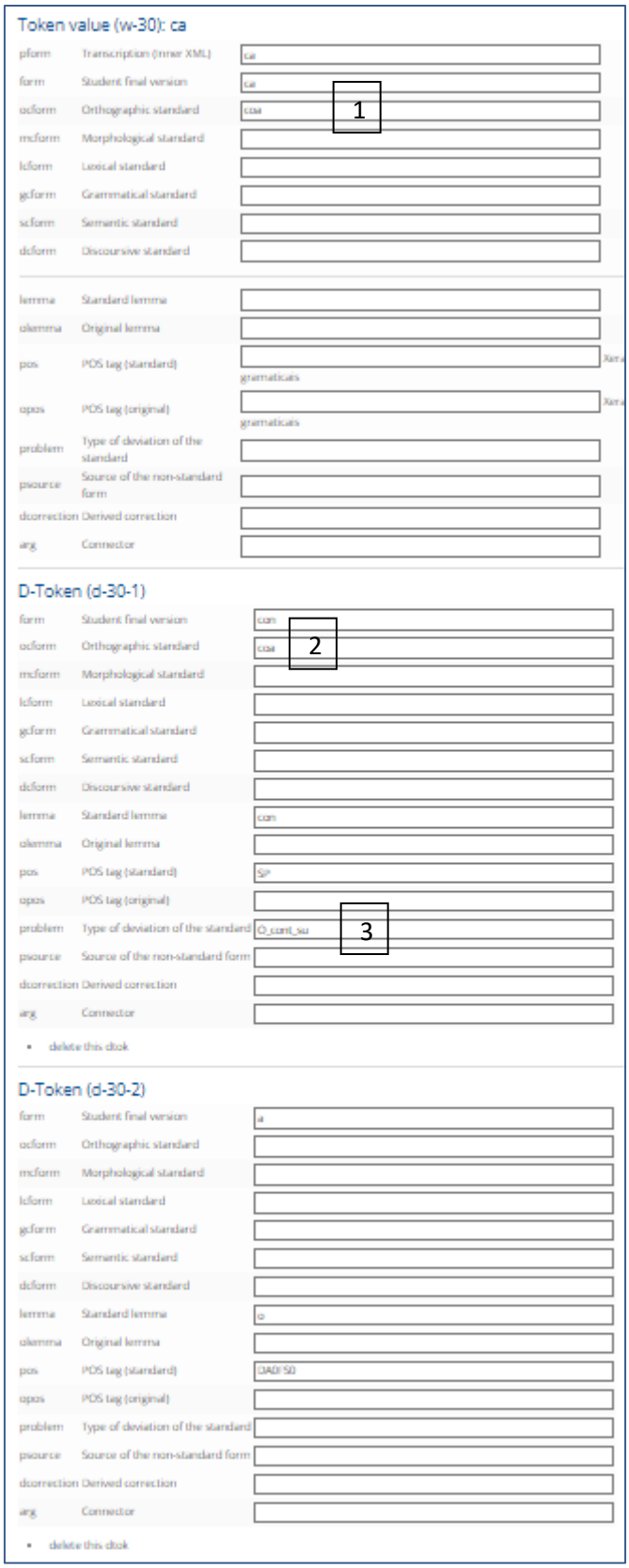


Figura 12. Anotación das desviacións que afectan a toda a contracción

O indicado vale tamén para unións inadecuadas de verbo e pronome, do tipo de *irémosnos* (O_as_om) ou de *permítenos* (por *permítennos*) (O_as_ad), aínda que nestes casos no primeiro d-token non poñeremos nada na corrección (só no código de desviación do estándar). Así pois,

1. Estandarizamos sobre o Token completo na capa de estandarización ortográfica (*irémonos*) sen asignarmos código de desviación do estándar.
2. No primeiro d-token, asignamos código de desviación do estándar (O_as_om).

No caso de termos un caso de **ausencia de contracción** (por exemplo, *de unha*), debemos actuar do seguinte xeito:

1. Prememos sobre o segundo elemento (neste caso *unha*).
2. No formulario que se crea, prememos en "merge left to w_nº".
3. Gardamos
4. No token completo, poñemos *dunha* na capa de estandarización ortográfica
5. No primeiro d-token (que se xera tras lematizar ou que creamós nós), poñemos *de* en *form*, asignamos o código de desviación do estándar O_cont_om en Type of deviation of the standard e repetimos *dunha* na capa de estandarizacion ortográfica.
6. No segundo d-token poñemos *unha* (ou corriximos *un* por *unha* se os dtokens son xerados automaticamente).

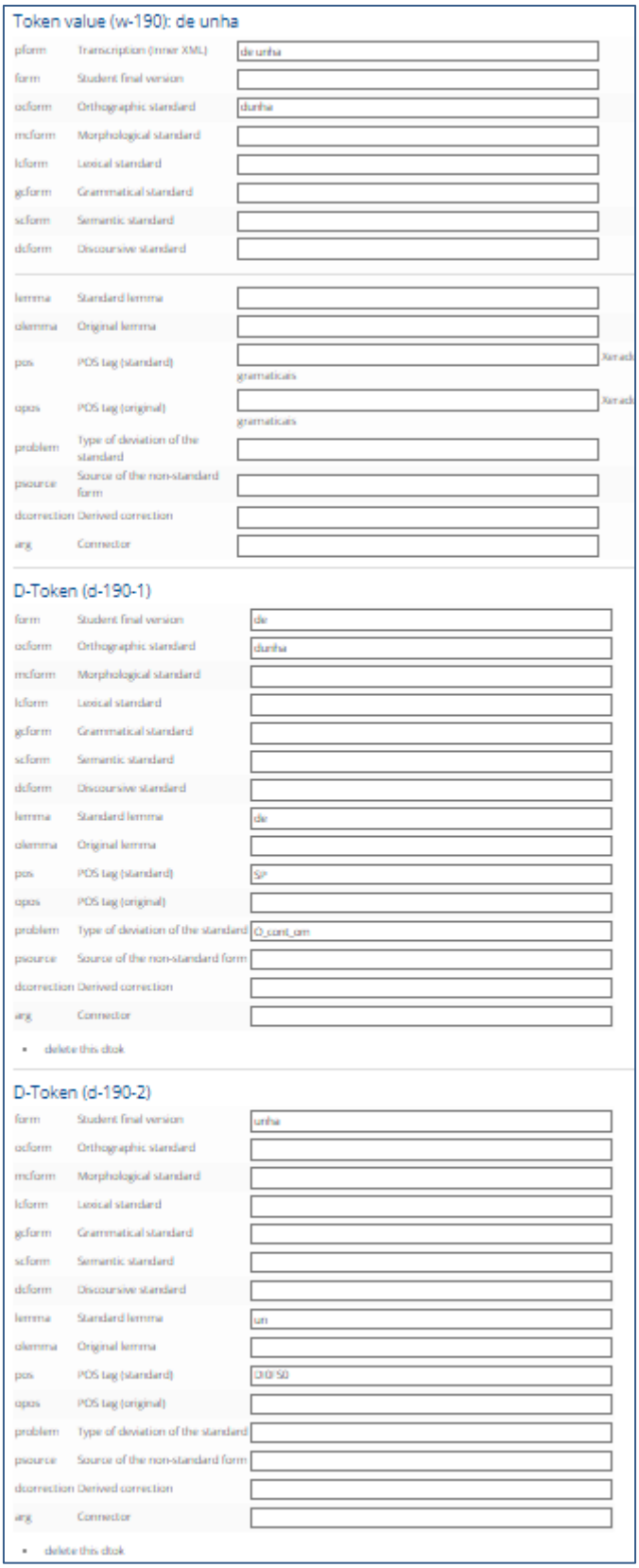


Figura 13. Anotación dos problemas de omision dunha contracción

2.6.3.2 Expresións complexas

Outro problema para anotación supóñeno as unidades lexicalizadas, do tipo de *sin embargo*, *sen embargo* ou *a pescadilla que se morde a cola*. Neste caso, dado que toda a expresión é unha unidade léxica, imos crear unha unidade conxunta a través de *merge* e non imos crear d-tokens. Ao non crear d-tokens, a anotación feita sobre o conxunto si é localizable a través do buscador.

Freeling non funciona se debe lematizar unha unidade constituída por varias palabras. Asi pois, **imos facer esta fusión despois de lematizar** (**no proceso de estandarización de cada texto, deben apuntarse estas formas para recordar que hai que estandarizalas tras a lematización**). A única excepción son os casos en que asignamos unha estandarización no nivel ortográfico, morfolóxico ou léxico que está constituída por unha soa palabra. Neste caso Freeling funcionará, porque simplemente terá que lematizar esa palabra.

> Debe terse en conta en primeiro lugar que Freeling asigna o lema estándar a partir da corrección do nivel léxico. Se este está baleiro, acudirá ao nivel morfolóxico. Se aquí tampouco hai nada irá ao nivel ortográfico, se neste non hai tampouco nada escrito irá á forma escrita polo estudante. *Pescadilla que se morde a cola* debe corrixirse por *carioca co rabo na boca* no nivel léxico. Se fusionamos previamente e asignamos a estandarización, o lematizador mirará esta última forma para lematizar, e como está constituida por varias palabras, non funcionará. Tampouco funcionará se corriximos no nivel semántico *a cambio de* por *fronte a*. Neste caso, dado que o lematizador non mira o nivel semántico para lematizar, e non hai nada no nivel léxico, morfolóxico nin ortográfico, mirará a unidade escrita polo estudante, *a cambio de*, que tamén é unha unidade complexa, de modo que Freeling non funcionará. Unicamente se poderá fusionar antes de levar a cabo a lematización cando a forma da que parta o lematizador sexa unha única palabra. Por exemplo, se corriximos *sin embargo* no nivel léxico por *porén*, Freeling lematizará *porén* e, ao tratarse dunha única palabra, non dará problema.

Para crear unidades pluriverbais cómpre colocarse no último elemento da unidade (por exemplo, *cola* en *pescadilla que se morde a cola*), darlle a “merge left to nº” e posteriormente a gardar. Isto vai unindo cada elemento co inmediatamente anterior. Repítese a operación (unindo cada pequeno grupo formado coa palabra anterior) ata conformar a unidade completa[3]. A continuación, codificaremos o problema que afecta a toda a unidade (por exemplo, no caso de a *pescadilla que se morde a cola*, L_w_su e L_spadapt) e asignaremos a forma estándar no nivel que corresponda (neste caso, no nivel léxico, *carioca co rabo na boca*). No caso de que no interior da expresión haxa outras desviacións (por exemplo, un pronome inadecuado, *se*, ou un castelanismo, *pescadilla*) estes problemas tamén se anotarán con código (pero non terán estandarización específica asociada).

> Así, *pescadilla que se morde a cola* anotaríase como L_w_su,L_w_su,G_pron_ad (o primeiro código está pola unidade completa, o segundo pola presenza de *pescadilla* e o terceiro pola presenza de *se*). Do mesmo xeito, unha forma como *sin embargo*, levará dúas marcas L_w_su, unha pola

[3] No caso de que o último elemento en unirse sexa unha contracción ("o problema **da** pescadilla que se morde a cola"), non se inclúe o artigo na unidade conformada (que neste caso sería entón "pescadilla que se morde a cola").

unidade no seu conxunto (que se corrixirá con *porén*) e outra pola presenza de *sin*. En cambio, *sen embargo* só levará un código L_w_su correspondente á unidade no seu conxunto.

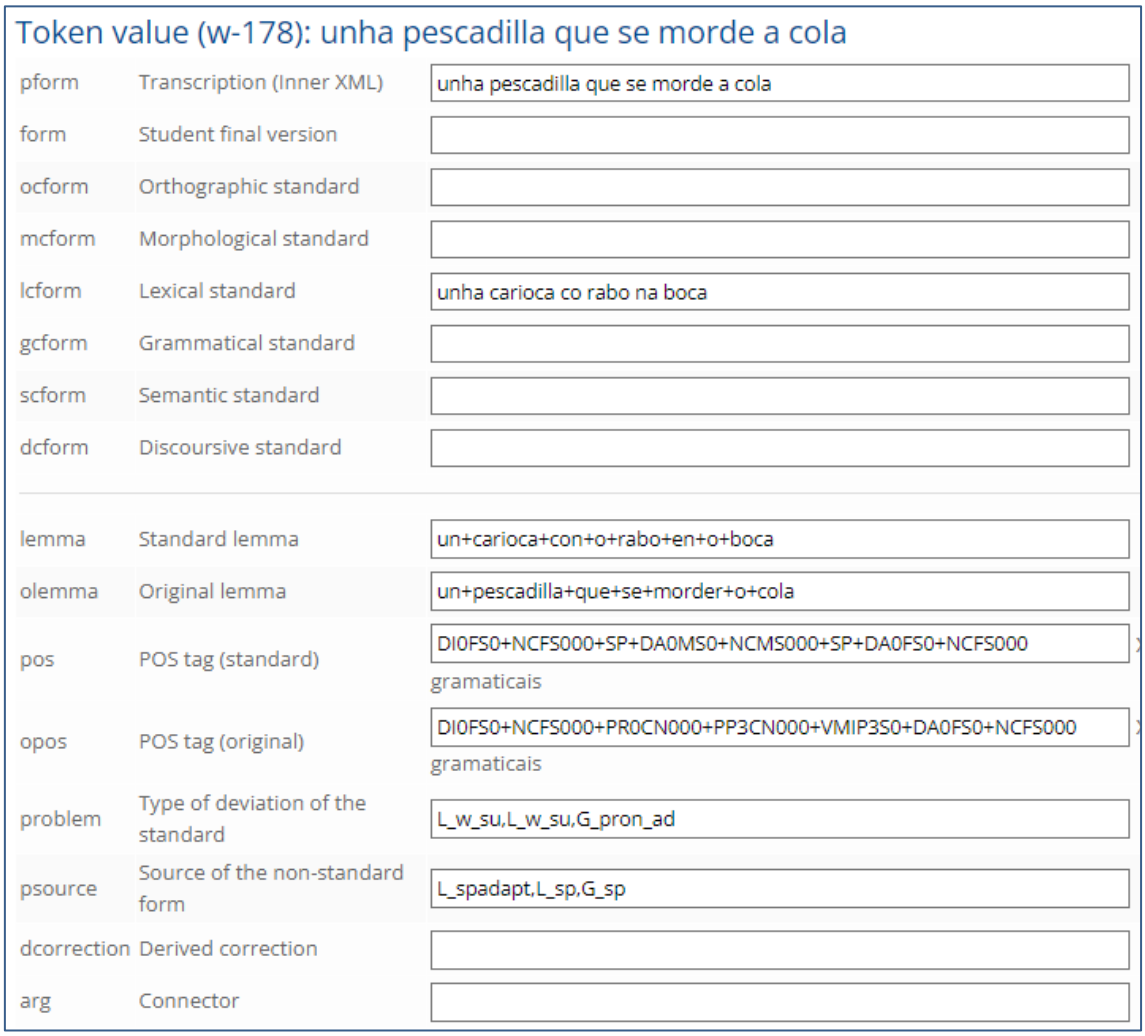


Figura 14. Anotación de unidades pluriverbais con formas non estándares no seu interior

As expresións que se crean poden ser unidades lexicalizadas (tipo *a pescadilla que se morde a cola*), pero tamén outras unidades que realizan conxuntamente unha mesma función ou un verbo acompañado por preposición (por exemplo, *optar por*) e que deba substituírse por unha única unidade nalgún nivel (por exemplo no semántico).

2.6.4 Estandarizacións no nivel ortográfico, morfolóxico ou léxico dunha palabra mediante dúas ou máis palabras

De acordo co que se acaba de indicar, cando unha palabra se estandariza no nivel ortográfico, morfolóxico ou léxico con varias palabras (por exemplo, algún *senon* que deba corrixirse con *se non*), **se levamos a cabo a estandarización antes de lematizar, Freeling non funcionará**. Por tal motivo, estes casos deben anotarse e corrixirse unha vez levada a cabo a lematización. Isto só afecta a estandarizacións que teñen lugar nos niveis ortográfico, morfolóxico ou léxico posto que, tal e como indicamos, o lematizador "mira" as correccións deses niveis para lematizar. Se as correccións afectan ao nivel semántico, gramatical ou discursivo Freeling funcionará

correctamente, de modo que en tales casos a estandarización pode facerse previamente á lematización.

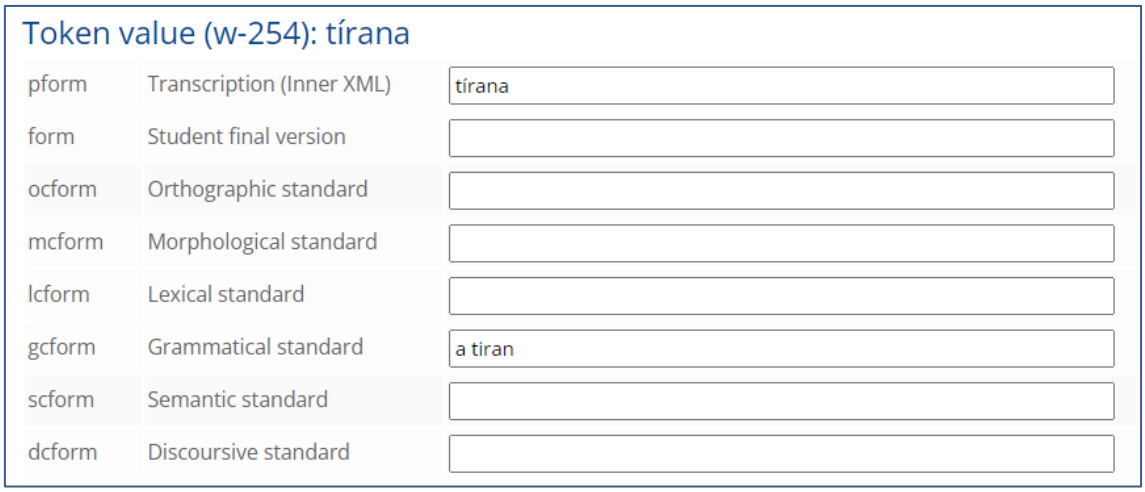
2.6.5 Varias opcións de estandarización

Nalgúns casos, as opcións de estandarización están claramente limitadas (en "vinche onte", só cabe "vinte onte", corrección de *che* por *te*) ou, habendo varias posiblidades, hai unha que se presenta claramente como a máis natural (se algúen escribe *bello*, o máis lóxico é substituír por *belo* e non por *fermoso*). Con todo, outras veces a forma que se debe incorporar no formulario de corrección non é tan evidente. Así, por exemplo, no caso de *sin embargo*, podería corrixirse por *non obstante*, *porén*, *no entanto*... Como o conveniente é que haxa a maior homoxeneidade posible na corrección, escolleremos unha forma común para todos os casos. Como criterio xeral, cando se trata de castelanismos, acudiremos ao *Dicionario Castelán-Galego* da Real Academia Galega e poñeremos o primeiro dos equivalentes que figuran nel. Non obstante, este criterio non é vinculante e pode optarse por outras solucións se así se considera conveniente.

Acudiremos cando sexa posible á edición multipalabra (vid. 2.2) para asegurar a maior coherencia posible, na anotación, pero isto non sempre é posible, como sucede con *sen embargo*. Para tales casos, crearase unha listaxe en word cunha columna para a forma corrixida e outra para a corrección seleccionada (así, se aparecen outros casos da mesma expresión errada, buscarase o indicado no documento e corrixirase do mesmo xeito).

2.6.6 Anotación de clíticos mal colocados

No caso de que o problema sexa de énclise en vez de próclise, os códigos (G_pron-en_wp en tipo de desviación e G_hc, seguido ou non de G_or, segundo corresponda) anotaranse no pronome unha vez xerados os d-tokens e a forma estándar figurará asociada ao Token completo para que se visualice na capa de corrección gramatical. Aquí pode verse un exemplo:



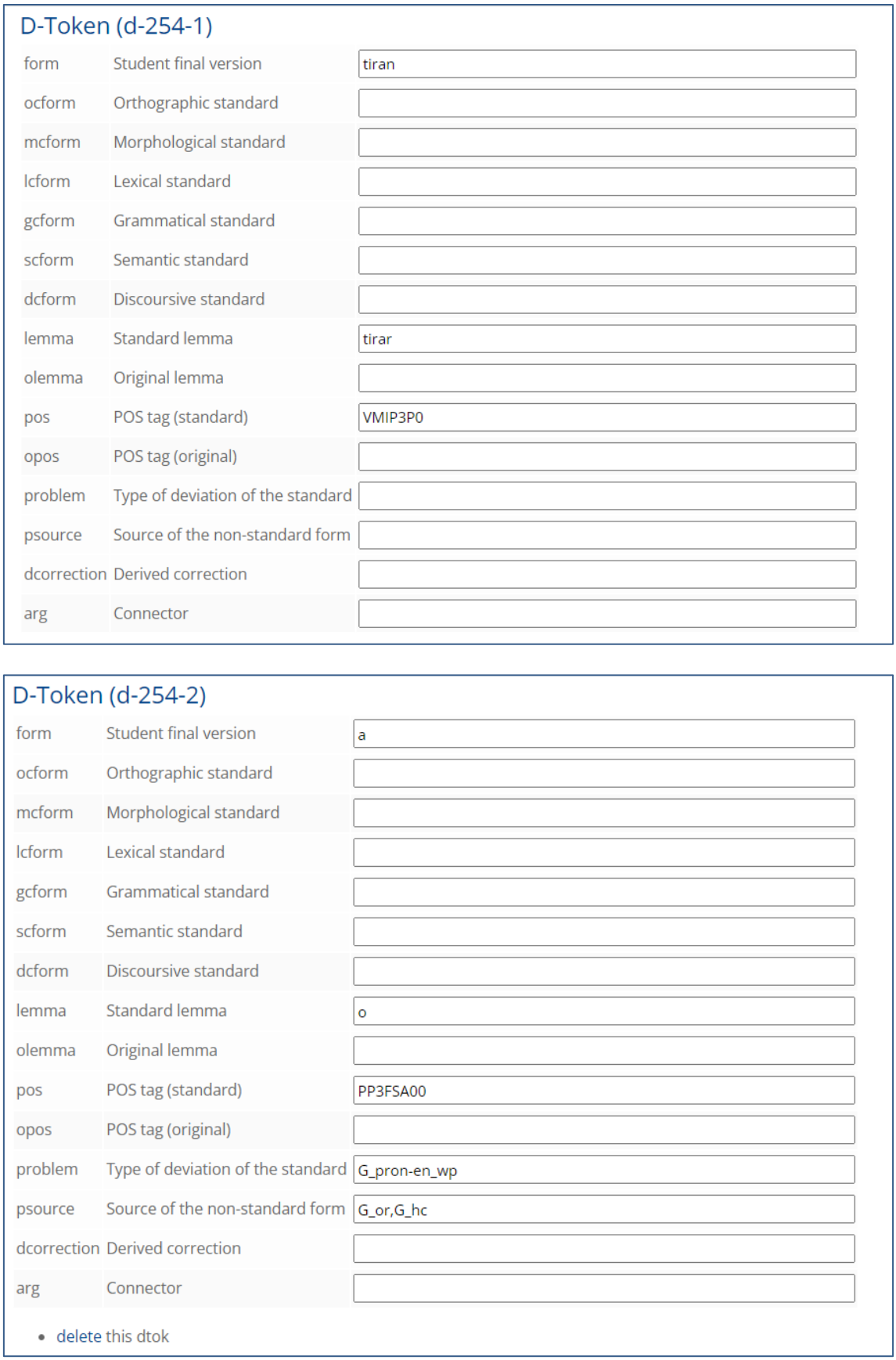
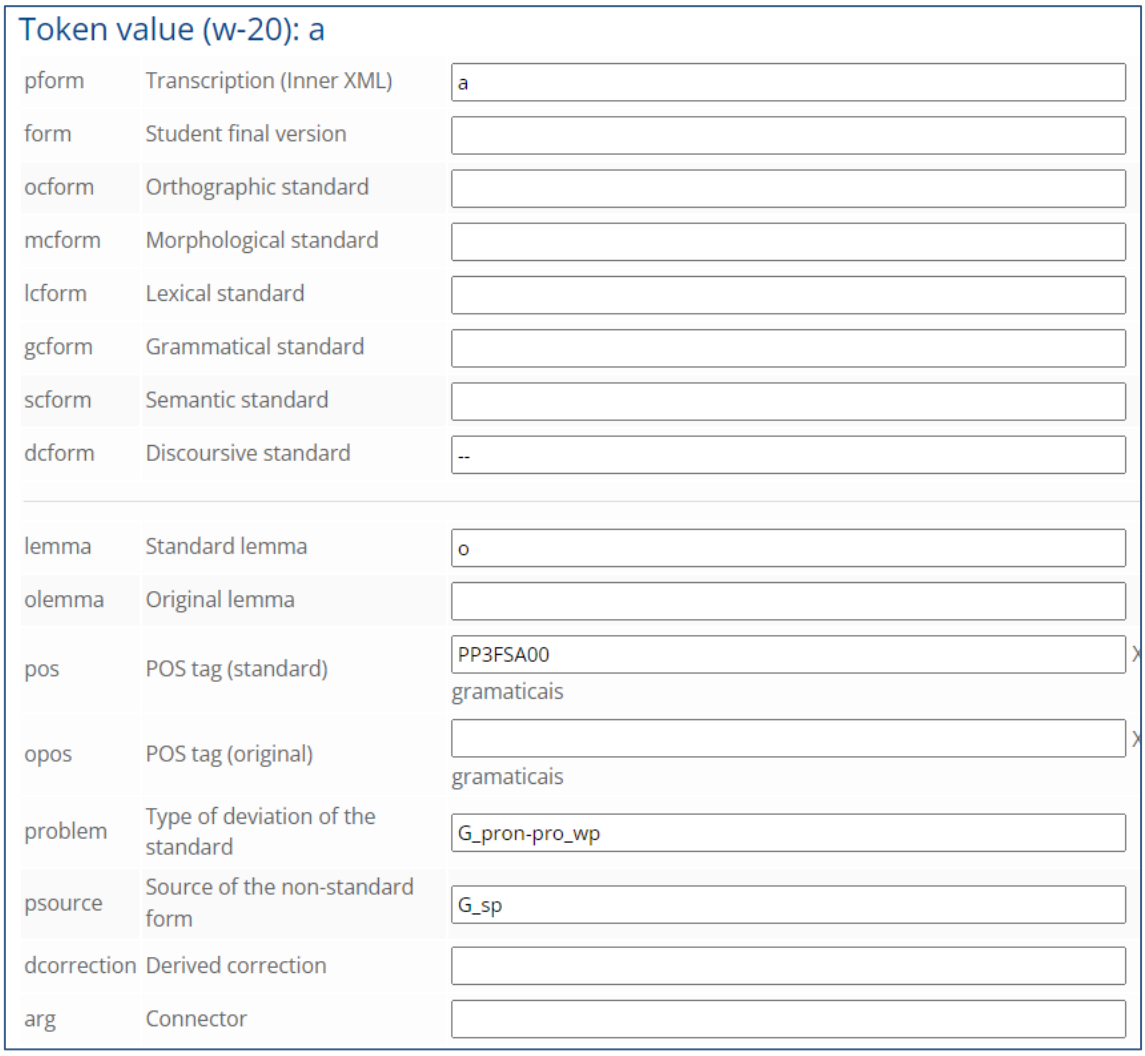


Figura 15. Corrección dun pronome enclítico que debería aparecer proclítico

No caso de que o problema sexa de próclise por énclise, prememos no clítico e asignámoslle os códigos correspondentes (G_pron-pro_wp en tipo de desviación e G_sp no código de orixe). Na capa de corrección gramatical poñemos --. Despois prememos no verbo e poñemos a forma co pronome enclítico na capa de corrección gramatical. En corrección derivada marcamos DC. Así, por exemplo, para "Sobre a metade da nosa vida a pasamos" teriamos:



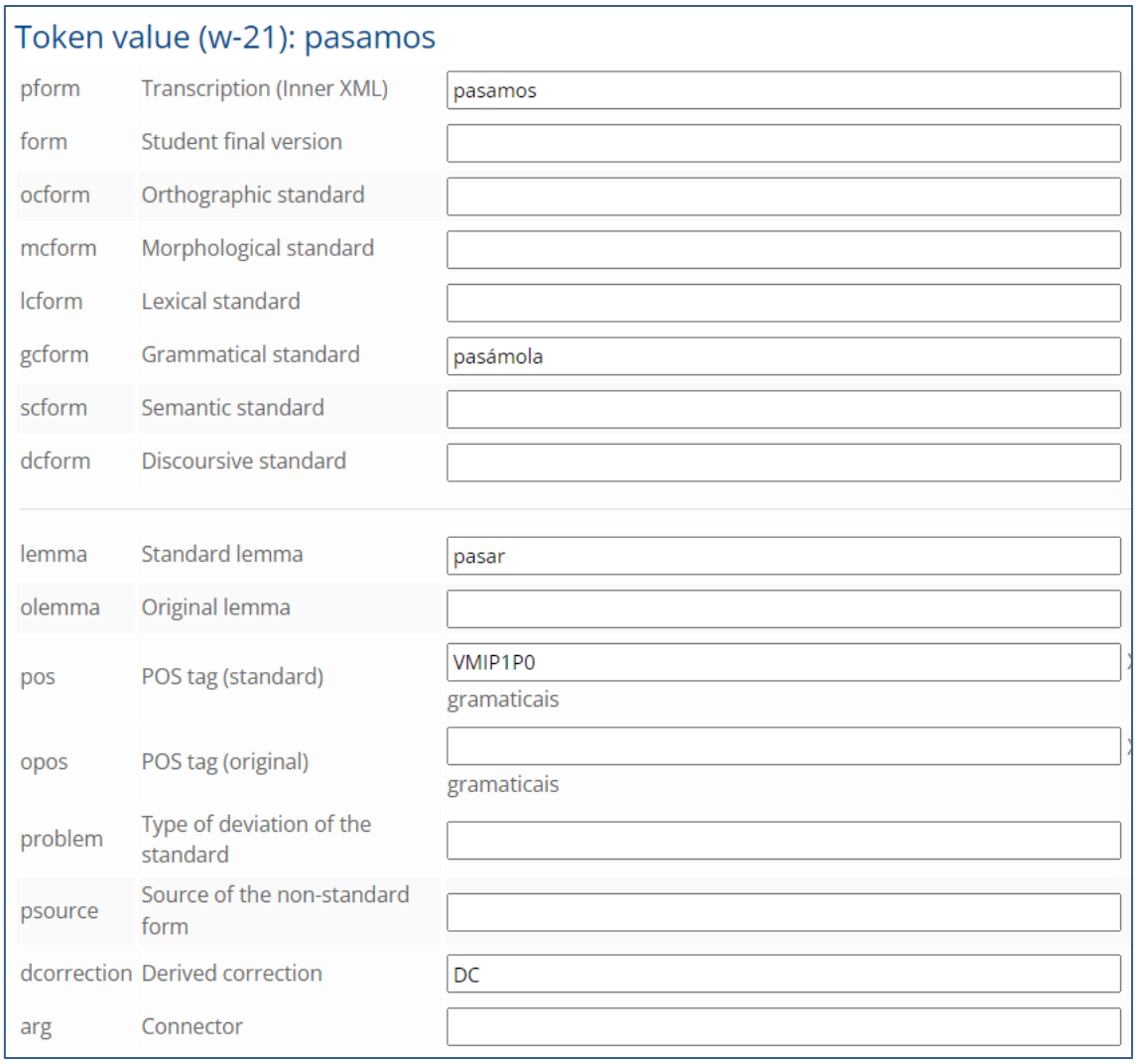


Figura 16. Corrección dun pronome proclítico que debería aparecer enclítico

2.6.7 Anotación dos signos de puntuación

Tal e como se indicou en 2.6.2, no caso de que haxa que **engadir un punto, coma, dous puntos ou punto e coma**, prémese na palabra inmediatamente anterior ao lugar en que debe aparecer o signo e despois, embaixo, prémese en "Insert tok after: attached". Abrirase un formulario con <ee/> en *pform*. En "Discoursive standardization" engádese o signo que corresponda e en "Type of deviation" D_pm_om. Así, por exemplo:

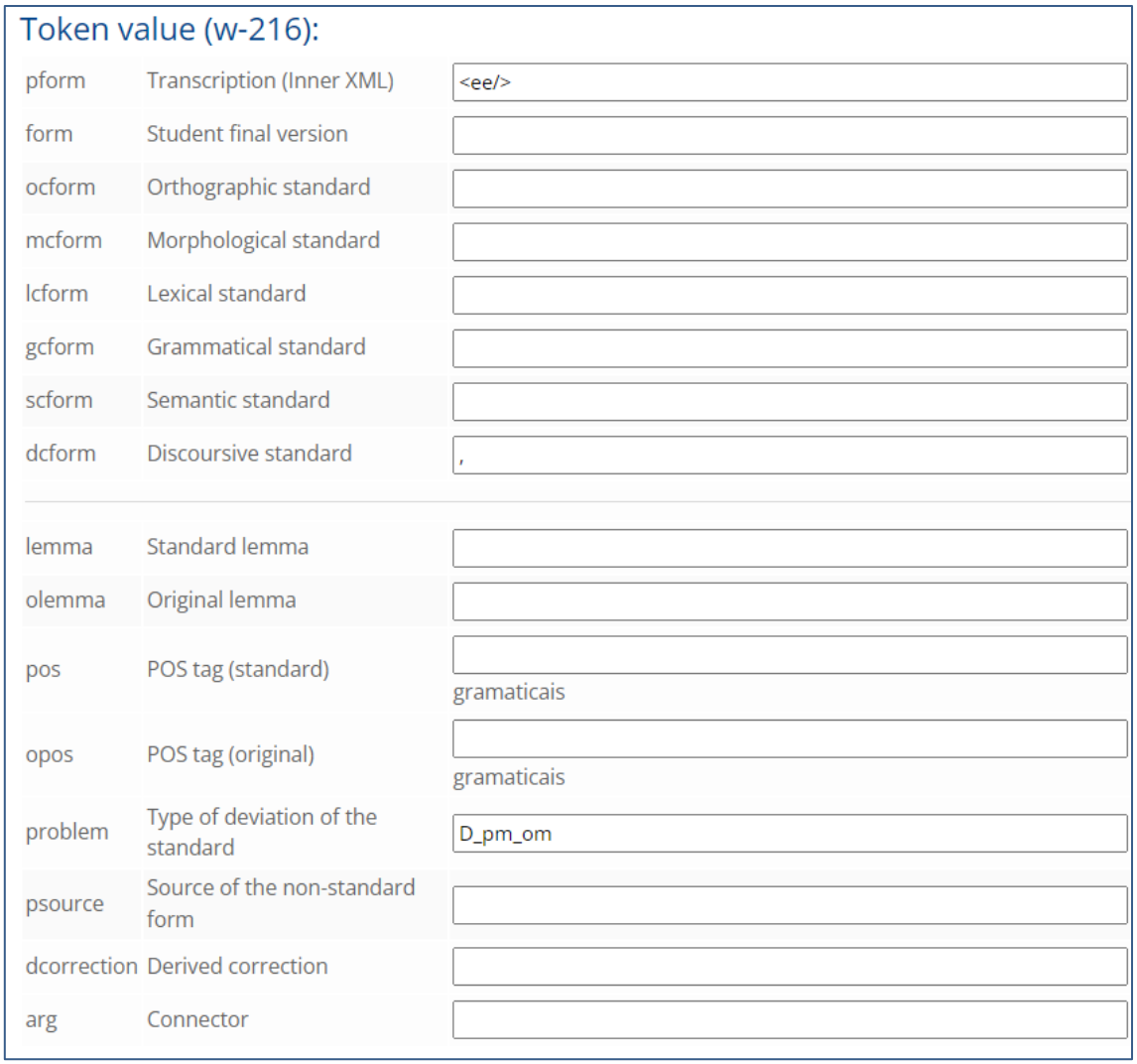


Figura 17. Anotación dos casos de omisión dun signo de puntuacion

Como xa indicamos, no caso de que haxa que **engadir dous signos de puntuación** (para delimitar un inciso ou similar), só anotamos o código de desviación no primeiro signo e o segundo etiquetámolo con DC. Así o faremos, por exemplo, en "Porque aínda que non o pensemos inflúe en máis cousas das que pensabamos".

No caso de que haxa que **engadir paréntese, raia ou comiñas**, prémese na palabra inmediatamente posterior ao lugar en que debe aparecer o signo de apertura e clícase embaixo, en "Insert tok before: attached". Abrirase un formulario con <ee/> en *pform*. En "Discoursive standardization" engádese o signo que corresponda e en "Type of deviation of the standard" ponse D_pm_om. Con respecto ao signo de peche, clícase na palabra inmediatamente anterior ao lugar en que debe aparecer o signo e despois en "Insert tok after: attached". Colócase o signo de peche en "Discoursive Standarization". Neste caso, non se inclúe nada en "Type of deviation of the standard", para evitar computar o problema (que é só un) dúas veces e colócase DC en "Derived Correction".

No caso de que haxa que **substituír un signo simple (punto, coma...) por un dobre**, cuxo signo de apertura vai pegado á seguinte palabra, na estandarización discursiva poñemos un espazo e

o signo (por exemplo, espazo e paréntese de apertura). (Quedará un espazo entre a paréntese e a seguinte palabra).

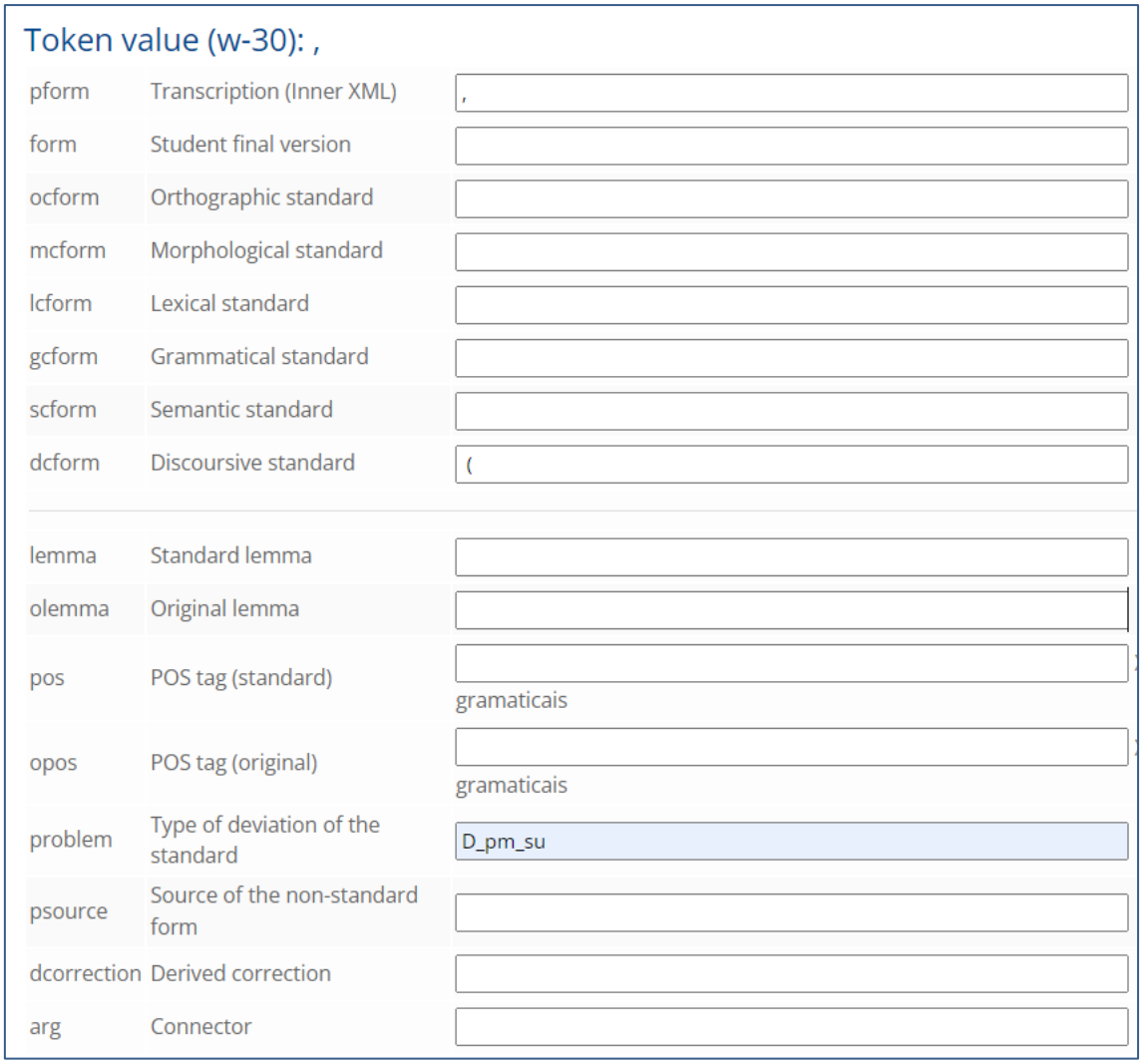


Figura 18. Anotación dos casos de substitución dun signo de puntuacion simple por un dobre

No caso de que haxa que propoñer **a supresión dun signo de puntuación**, actúase tal e como se indicou no apartado 2.6.1.

Se estamos ante un problema de **mala colocación dun signo de puntuación** (por exemplo, colocar coma despois de *pero* en vez de poñela antes), prememos no signo mal colocado e colocamos -- na capa de estandarización discursiva, asignando o código de desviación (D_pm_wp). Despois situámonos na palabra inmediatamente anterior ao lugar en que debería aparecer o signo de puntuación, inserimos token e colocamos o signo na capa de estandarización discursiva, con DC en correción derivada.

Token value (w-249): ,

pform	Transcription (Inner XML)	,
form	Student final version	
ocform	Orthographic standard	
mcform	Morphological standard	
lcform	Lexical standard	
gcform	Grammatical standard	
scform	Semantic standard	
dcform	Discoursive standard	--
lemma	Standard lemma	
olemma	Original lemma	
pos	POS tag (standard)	gramaticais
opos	POS tag (original)	gramaticais
problem	Type of deviation of the standard	D_pm_wp
psource	Source of the non-standard form	
dcorrection	Derived correction	
arg	Connector	

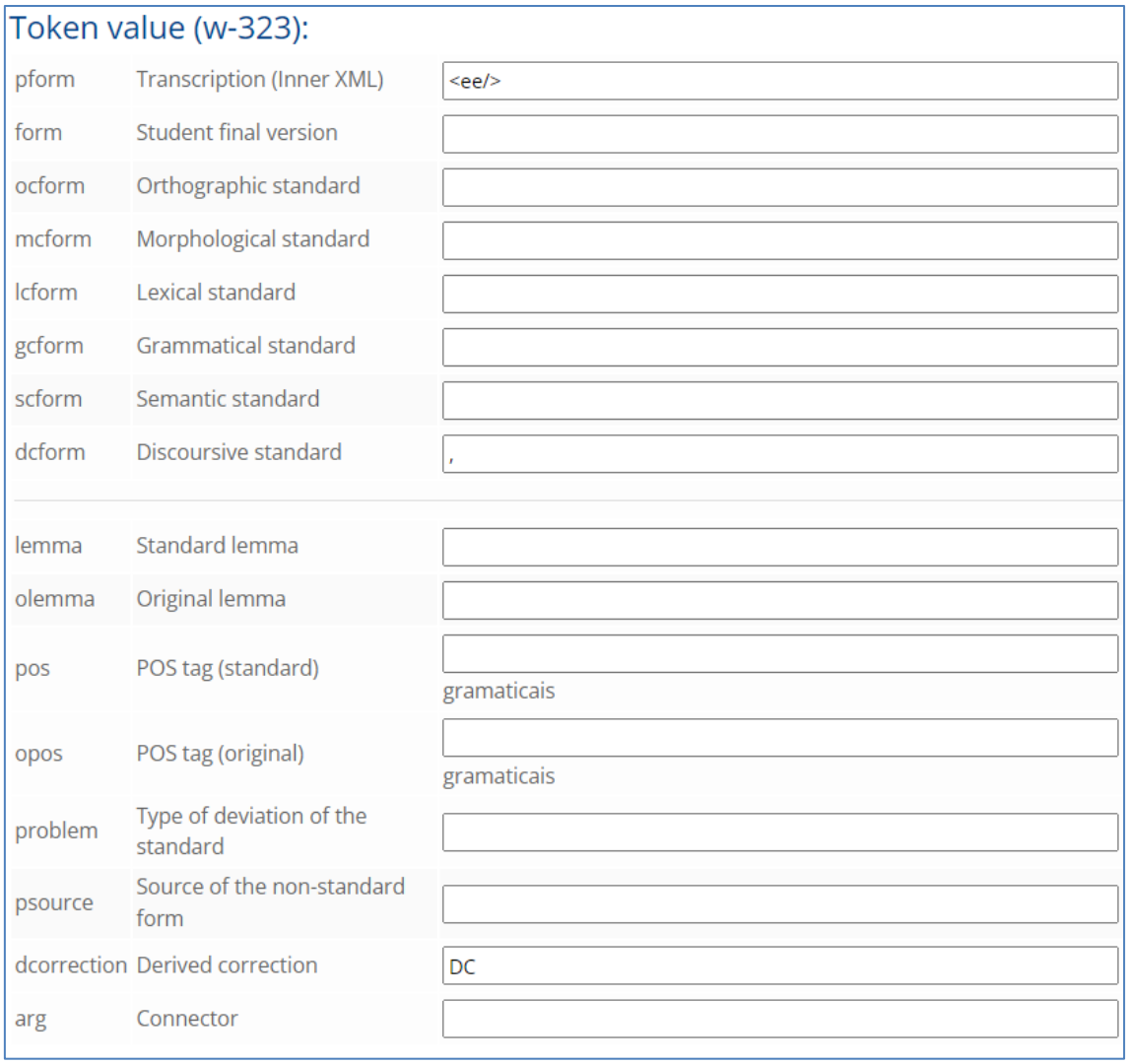


Figura 19. Anotación dos casos de colocación inadecuada dun signo

Con respecto ás estandarizacións que supoñen a creación de novos parágrafos ou a súa supresión (porque a/o estudante omitiu un parágrafo necesario ou engadiu un onde non debía), anotaranse co código correspondente, respectivamente D_np_om e D_np_ad, pero resulta inviable a súa corrección na capa correspondente (discursiva). A anotación farase no signo de puntuación que exista entre os enunciados que se deben unir / separar.

No caso de que falte un punto nunha abreviatura (por exemplo *etc*), dado que realmente non estamos ante un problema discursivo, imos consideralo como un problema ortográfico (O_ab_su). Para corrixir, poñemos esa etiqueta en "Type of deviation of the standard" e **despois de lematizar**, poñemos na capa de estandarización ortográfica a abreviatura seguida de punto (*etc.*)

2.6.8 Anotación de problemas de concordancia en cadea

Tal e como xa indicamos, no caso de que encontremos varios problemas gramaticais de concordancia encadeados, en que se poida supoñer que os segundos, terceiros... van arrastrados polo primeiro, só se marcará este último con código de desviación do estándar. Nas outras formas que non concordan levarase a cabo a corrección (no nivel gramatical), pero non

se asignará código de desviación do estándar. Así, por exemplo, se alguén escribe "o pobo, concienciado da situación, pensaban emocionados e contentos que...", só asignaremos a etiqueta G_num_su en *pensaban*, que ademais se estandarizará con *pensaba*. Os adxectivos *emocionados* e *contentos* estandarizaranse na capa gramatical, pero non se lles asignará código de desviación do estándar e irán ademais etiquetados con DC en Derived correction. No caso de estaren implicadas unha palabra léxica e palabras gramaticais, anotarase co código de desviación do estándar a palabra léxica, mentres que as palabras gramaticais serán corrixidas e levarán DC.

2.6.9 Anotación de problemas de cambio de xénero

Encontramos algúns casos en que a/o estudante escribe un palabra cuxa forma é estándar, pero que presenta un xénero non estándar (por exemplo, "a costume"), normalmente por influencia do español (vid. a observación do código L_gen_su no *Manual de anotación das formas non estándares*). Os determinantes e outros modificadores que acompañen estas formas estandarizaranse (no nivel léxico tamén), pero non levarán código de desviación do estándar; unicamente DC en Derived correction. **Nestas formas etiquetadas con DC, a categoría gramatical debe corresponder coa da forma que escribiu a/o estudante (se escriben *a costume*, a categoría gramatical de *a* debe ser artigo feminino). Por tal motivo, dado que ao corrixir na capa léxica vai cambiarlle o xénero, cómpre corrixir manualmente estes casos para asignar a forma e o xénero orixinal.** Con respecto aos substantivos, levarán a etiqueta L_gen_su, pero non haberá forma estándarizada no nivel léxico, pois esta coincide coa forma estándar. Ofrecerase, iso si, a categoría non estándar en POS orixinal.

2.7 Anotación en standoff

Algúns códigos de erro, que afectan ou poden afectar a varias palabras (sen constituíren expresións complexas), deberán anotarse mediante un sistema diferente ao visto ata de agora, chamado standoff. Os códigos que se anotan en standoff son os que se indican a continuación:

Grammar_structure_substitution G_str_su
Grammar_structure-reformulation-topicalization_substitution G_str_ref-top_su
Grammar_word_addition G_w_ad
Semantics_word_addtion S_w_ad
Discourse_word(s)_wrong place D_w_wp
Discourse_utterance_complexity D_ut_com
Discourse_utterance_unintelligibility D_ut_unin

Aínda que algún exemplo concreto destes problemas puidese marcarse de maneira doada a través do sistema de anotación habitual, **todos os problemas que leven estas etiquetas deben marcarse mediante standoff**, posto que deben poder recuperarse conxuntamente (e o sistema de buscas mediante standoff é diferente do sistema de buscas normal).

Para marcar un problema en standoff, debemos premer en *Anotación multipalabra* (embaixo do texto).

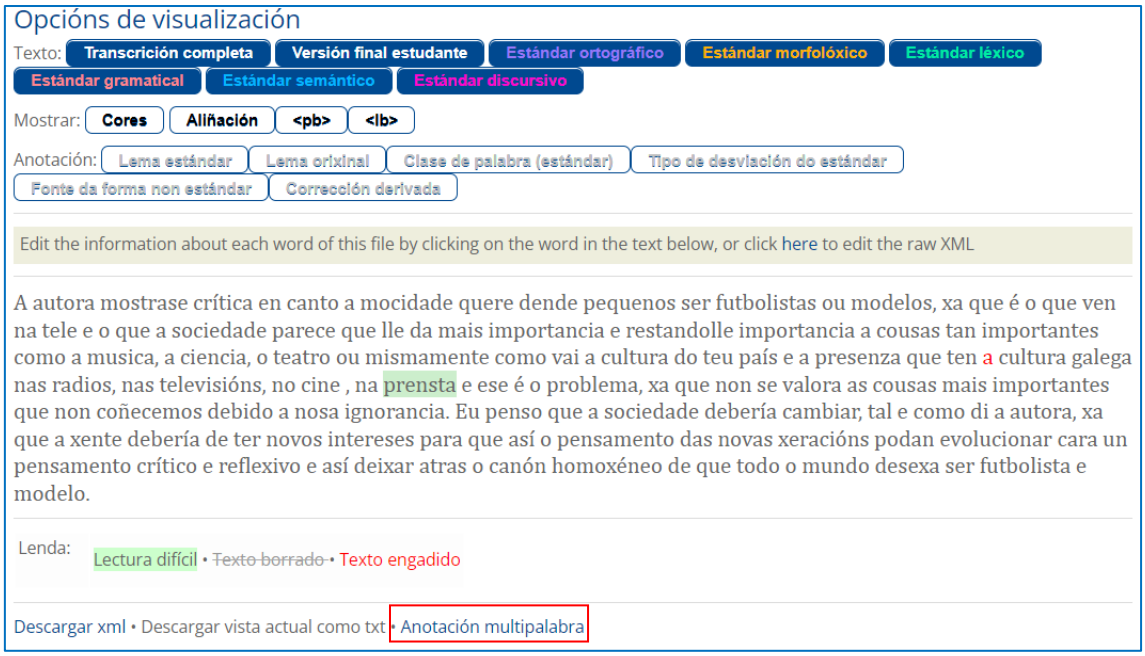


Figura 20. Anotación en standoff (I)

A seguir ábrese unha pantalla como a seguinte:

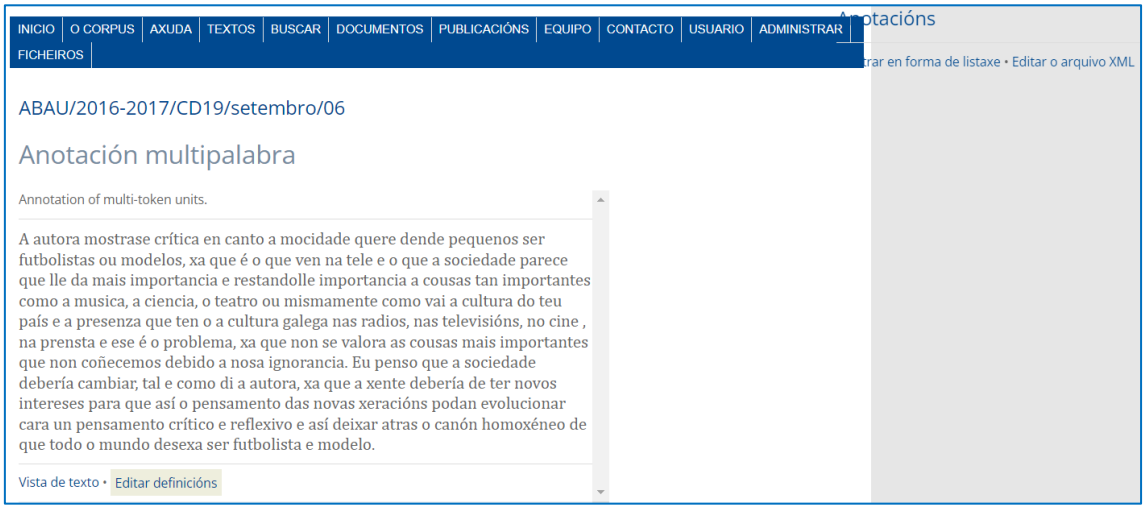


Figura 21. Anotación en standoff (II)

Seleccionamos entón o fragmento de texto que queremos marcar. Por exemplo, neste caso seleccionamos o primeiro parágrafo para marcar que hai un problema de D_ut_com. Colocámonos na primeira palabra e arrastramos o rato para seleccionar todo o fragmento. Ao facer isto, ábrese un formulario á dereita:

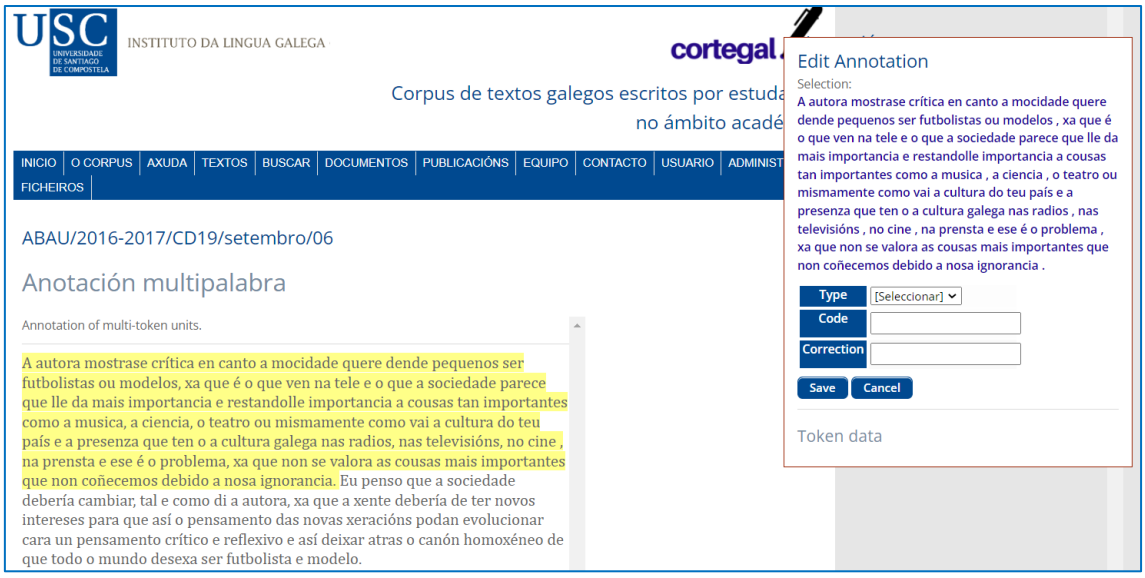


Figura 22. Anotación en standoff (III)

Prememos en “Type” e seleccionamos “Discourse”, “Semantics” ou “Grammar” segundo o nivel ao que corresponda o problema. En “Code” poñemos o código de desviación (neste caso D_ut_com) e en corrección poñemos o texto modificado, pero atendendo exclusivamente ao aspecto gramatical, semántico ou discursivo a que se refire o código (manteríanse castelanismos, problemas ortográficos etc.). Unha vez cuberto o formulario dámoslle a “Gardar” e nese momento aparece unha pantalla como esta, que nos indica que a anotación xa está gardada:

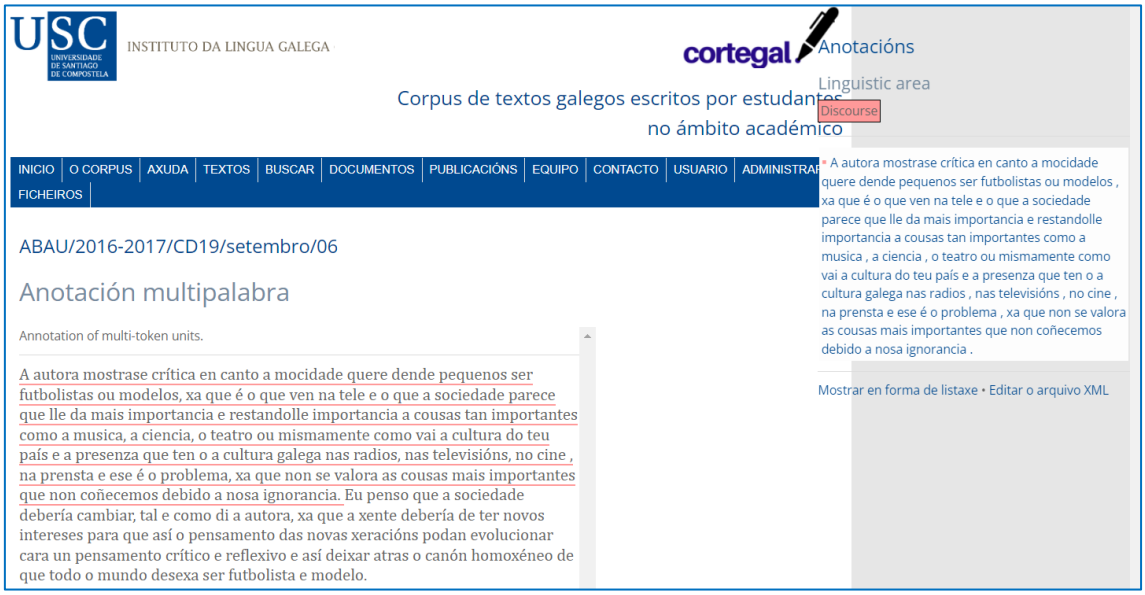


Figura 23. Anotación en standoff (IV)

Pódese incluír unha anotación dentro dun fragmento xa anotado en standoff (para iso seleccionamos o fragmento de texto dentro do fragmento xa marcado e anotamos seguindo o mesmo procedemento que acabamos de indicar). Con todo o buscador só localiza a primeira

anotación que se introduza. Por tal motivo, sempre que nos encontremos con esta necesidade de marcar unha anotación dentro doutra, debemos comentalo para ver como proceder.

No caso de que se queira borrar unha anotación en standoff, prémese no fragmento que figura á dereita (baixo “Linguistic area”) e na pantalla que se abre dáselle a *delete segment*.

Debe terse en conta que as correccións que se fagan aquí non se van trasladar ás capas de visualización correspondentes, de tal modo que nestas faremos as estandarizacións oportunas para poder ver o texto corrixido, pero sen asignar códigos. Con todo, no caso de que sexa doado reducir a enunciados máis simples un enunciado moi complexo marcado con D_ut_com, cambiando ou engadindo puntuación, anótase en standoff o conxunto, coa posible corrección, e despois individualmente cada problema de puntuación no sistema normal de anotación, con códigos incluídos.

Por outra banda, debe seleccionarse para anotar o fragmento ao que afecta o problema, c**o contexto que xustifica que se trata dunha desviación do estándar**. Por exemplo, no caso de “Ademais, por outro lado cómpre dicir que o constante fomento…” marcamos “Ademais, por outro lado”, e na corrección propoñemos “Por outro lado”, eliminando “Ademais”.

IMPORTANTE:

Se hai varias anotacións standoff nun texto, é preciso anotalas en orde (primeiro a que está antes no texto, despois a seguinte e así consecutivamente).

As anotacións en standoff desaparecen cando se cambia o nome do documento!!

3. LEMATIZACIÓN E ASIGNACIÓN DE CATEGORÍA GRAMATICAL

3.1 Lema estándar e clase de palabra estándar

Unha vez estea feita a estandarización e codificación dos problemas, levarase a cabo a lematización automática con Freeling (debaixo do texto, en “Admin options”, “Tag with Freeling)”. Freeling asigna unha forma na caixa "Lema estándar" e unha categoría gramatical na caixa "Clase de palabra (estándar)". Como xa indicamos, o lema asígnase sobre a forma corrixida no nivel léxico. Se este está baleiro, acudirá ao nivel morfolóxico. Se aquí tampouco hai nada irá ao nivel ortográfico, se neste non hai tampouco nada escrito irá á forma escrita pola/o estudante.

Unha vez lematizado o texto, deberemos revisar se a asignación de lema e categoría gramatical se fixo adecuadamente. Para iso, na vista da transcrición provisional iremos vendo cada forma para ver cal foi o lema e a categoría asignados. No caso de que haxa algún erro, debe premerse na forma e corrixir o que corresponda.

Para corrixir a categoría gramatical, bórrase a categoría mal asignada e ponse o código ou ben bórrase e dáselle a tag builder (xerador de etiquetas gramaticais), que abre embaixo unha caixa, asígnase a categoría que corresponda e finalmente dáselle a *insert*. Con todo, hai que usar con coidado o *tag builder*, porque non sempre é evidente que valor hai que asignar neste xerador de etiquetas. É preferible buscar outros exemplos da mesma clase de palabras que queremos

asignar, copiar a etiqueta e pegala (ou ben facer un documento cunha listaxe de etiquetas e consultalo cando sexa preciso).

Sinálanse aquí algunhas particularidades sobre o proceso de revisión:

1. Deben revisarse sobre todo as conxuncións, determinantes e pronomes. Freeling confunde con frecuencia a preposición *a* co artigo, non diferencia axeitadamente entre os diferentes valores de *que* e de *como* etc.
2. No caso dos nomes propios cómpre corrixir o lema e poñelo con maiúscula inicial.
3. Algúns numerais son identificados como nomes. Hai que poñer Z na categoría gramatical.
4. No caso dos adverbios en *-mente*, quítalle a algúns a terminación *–mente* no lema e cómpre repoñela.
5. No caso dos adxectivos coincidentes con participios e que Freeling identifica como tales, non se modifica a asignación categorial, aínda que no contexto estean funcionando como adxectivos e non como verbos.
6. No caso das locucións conformadas por varios elementos, mantemos a categoría gramatical dos seus constituíntes. Por exemplo, en *grazas a* Freeling identifica cada un dos elementos por separado (un substantivo plural e unha preposición). Non se modifica esta atribución.

Ademais, debe terse en conta que é posible que quedase algunha cuestión pendente despois da anotación. A este respecto, cómpre:

1) Corrixir agora as formas cuxa corrección requiría incluír varios elementos no nivel ortográfico, morfolóxico ou léxico.
2) Se non foron anotados antes, anotar os erros nos d-tokens creados automaticamente que o requiran (contraccións e verbos con pronomes enclíticos). Ademais, nestes casos debe revisarse o resultado de Freeling, porque a "form" do segundo elemento de certas contraccións e grupos de verbo e clítico vai aparecer sempre en masculino singular e o verbo destes en infinitivo.
3) Crear as unidades pluriverbais que sexa preciso e anotalas.

Con relación ás contraccións, en vez de corrixir unha a unha, imos empregar o sistema Multitoken edit para corrixir de vez o mesmo problema en varios documentos. Por exemplo, se queremos corrixir todos os casos en que o segundo elemento da contracción é o artigo masculino plural (*os*) (e nos que, como se indicou, Freeling pon *o* na "form"), faremos o seguinte. En primeiro lugar, facemos unha busca de casos en que teñamos *o* na "form" e artigo masculino plural na clase de palabras estándar (DA0MP0) (podemos engadir *o* no lema, aínda que non sería necesario):

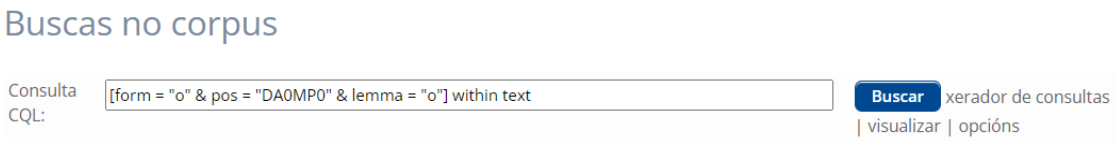


Figura 24. Mult-token edit nas contraccións(I)

A continuación obtemos os resultados e baixo eles prememos en "use this query for multi-token edit".

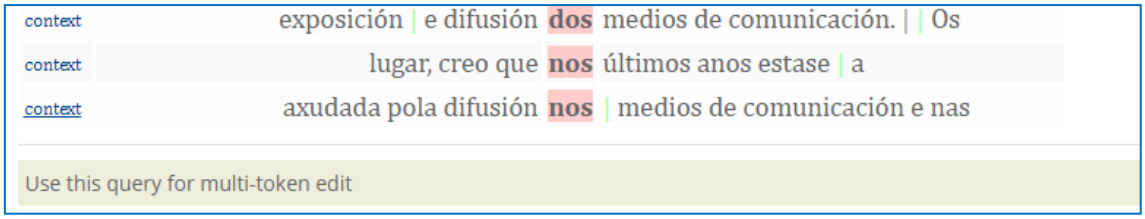


Figura 25. Mult-token edit nas contraccións (II)

Ábrese un formulario que, como se indicou, nos permite facer conxuntamente, en todos os textos que queiramos, un cambio. Neste caso o que nos interesa é que cambie *o* por *os* na "form", e por tal motivo escribimos *os* na caixa de "form".

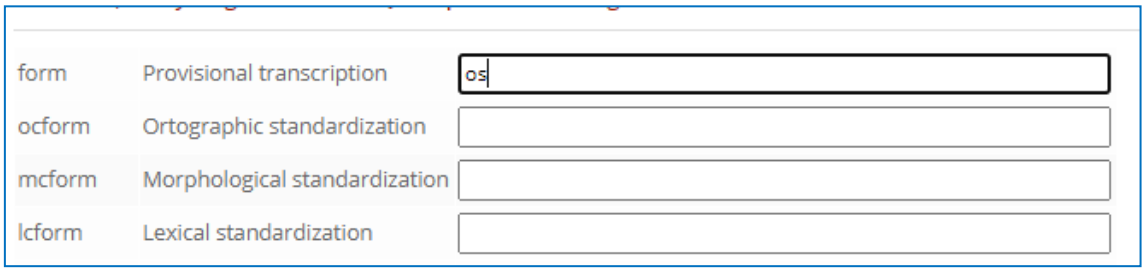


Figura 26. Multi-token edit nas contraccións (III)

Embaixo do formulario aparecen todos os exemplos que conteñen a busca realizada. Seleccionamos os exemplos (con “select all”) e dámoslle a "change selected". E xa queda o cambio feito en todos os textos.

Ao respecto das unidades pluriverbais, debe terse en conta que ao crear unha expresión pluriverbal despois de que o proceso de lematización tivera lugar, TEITOK o que fai é unir os diferentes lemas e categorías gramaticais mediante o signo +. Imos conservar este sistema, revisando sempre que as categorías asignadas foron correctas.

> Por exemplo, se temos a expresión pluriverbal *a cambio de*[4], ao lematizar Freeling asignará unha etiqueta SP (Preposición), outra NCMS000 (Substantivo común masculino singular) e outra SP e respectivamente os lemas *a*, *cambio* e *de*. Ao crear a expresión pluriverbal, Freeling poñerá na caixa do lema estándar *a+cambio+de* e na caixa da clase de palabra estándar SP+NCMS000+SP.

Agora ben, se despois de lematizar, unha vez creada a expresión pluriverbal, incluimos unha estandarización no nivel léxico (ou se non, no morfolóxico ou se non, no ortográfico), e dado que os lemas deben construírse sempre sobre estas formas corrixidas en tales niveis, debemos modificar o lema e a categoría gramatical asignado por Freeling.

Por exemplo, se o que temos é a forma *pescadilla que se morde a cola*, Freeling, despois de crear a expresión complexa, ofrecerá isto.

[4] Creouse a expresión pluriverbal para poder corrixila no nivel semantico por *fronte a*, pois aparece nun contexto en que o seu uso é inadecuado.

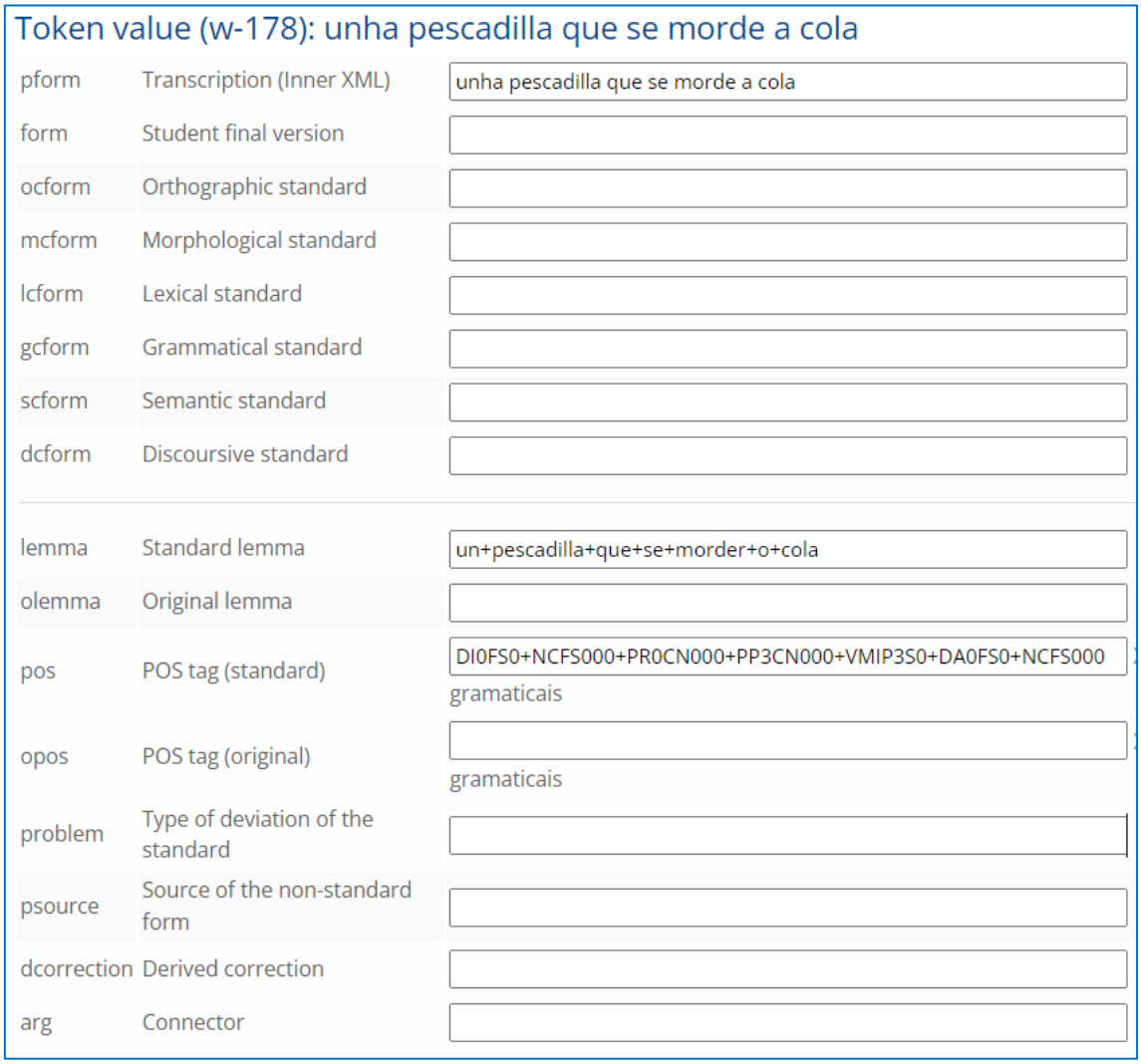


Figura 27. Lematización de expresións complexas (I)

Nós debemos corrixir na capa léxica con *a carioca co rabo na boca* e o lema debe corresponder a esta expresión. Polo tanto, modificamos manualmente o lema e a categoría gramatical estándar, pasando o lema e a categoría gramatical orixinal respectivamente a *olemma* e *opos*, tal e como se explicará a seguir:

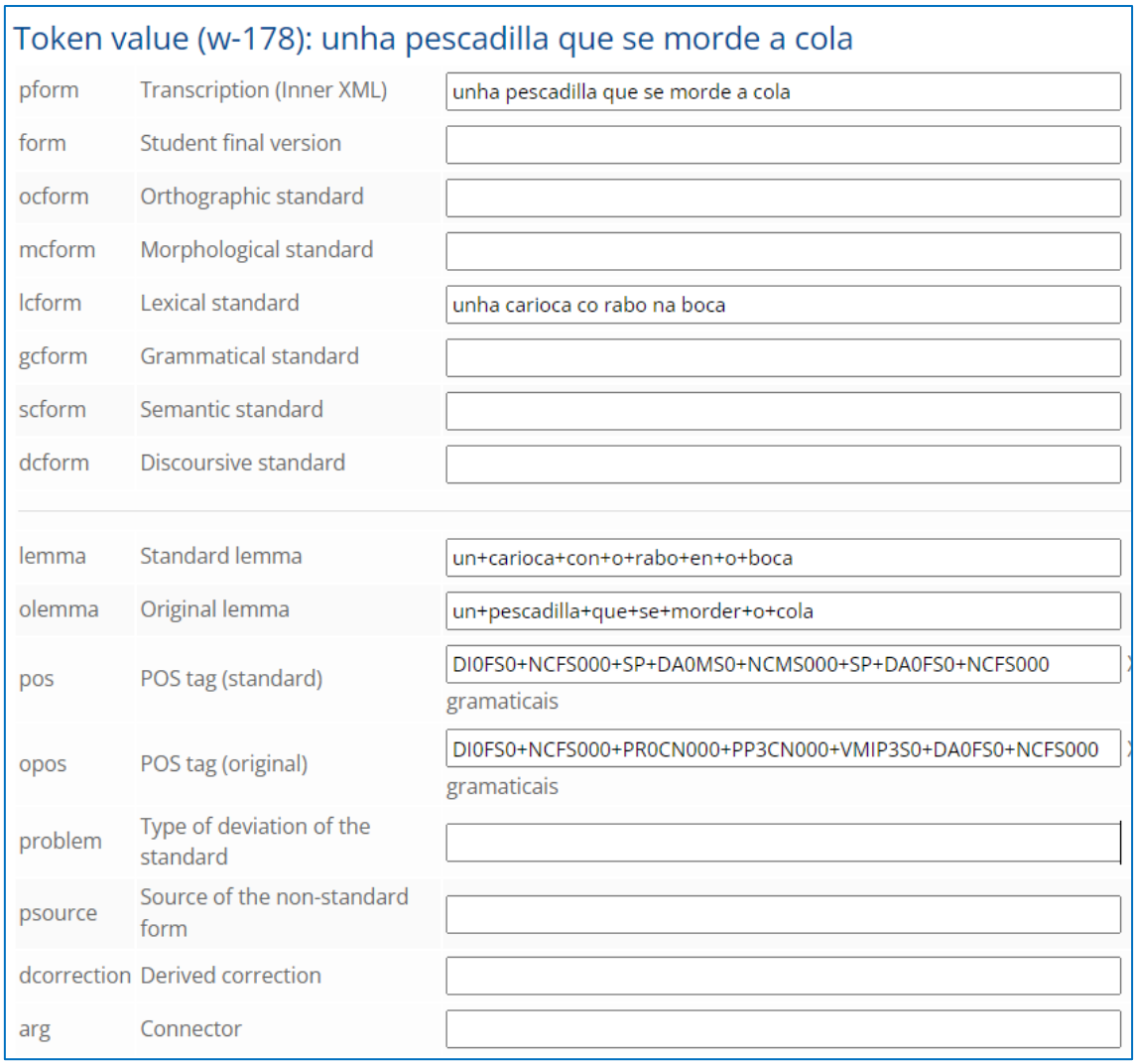


Figura 28. Lematización de expresións complexas (II)

3.2 Lema orixinal e clase de palabra orixinal

As formas estandarizadas no nivel léxico tamén serán anotadas cun lema orixinal e cunha clase de palabras orixinal (que, por exemplo, permitirá encontrar conxuntamente todos os exemplos do verbo *salir*). O lema orixinal corresponde ao lema da forma que escribiu o estudante. Se escribiu *salimos*, o lema será *salir*. O mesmo vale para a clase de palabras. Se o estudante escribe *leche* a clase de palabras orixinal será substantivo feminino. Só nos casos en que exista unha corrección no nivel léxico debemos poñer manualmente o lema orixinal (e, no seu caso, a categoría gramatical orixinal).

Se non se escribe nada nas caixas correspondentes, o lema e a clase de palabras orixinal hérdase automaticamente do lema e clase de palabras estándar, aínda que no formulario non aparezan explicitamente.

> Así, no caso de *salimos* escribiremos *salir* no lema orixinal. A categoría gramatical non cambia, porque é a mesma da palabra que foi considerada para a lematización, *saimos*, que é a forma que figura na caixa de estandarización léxica. Por tal motivo, e dado que xa se herda desta, non escribimos nada na caixa POS Tag (original). A asignación de lema e, no seu caso, categoría orixinal

pode facerse antes ou despois da lematización con Freeling (pode facerse asociada á asignación de código de desviación no nivel léxico e á correspondente corrección).

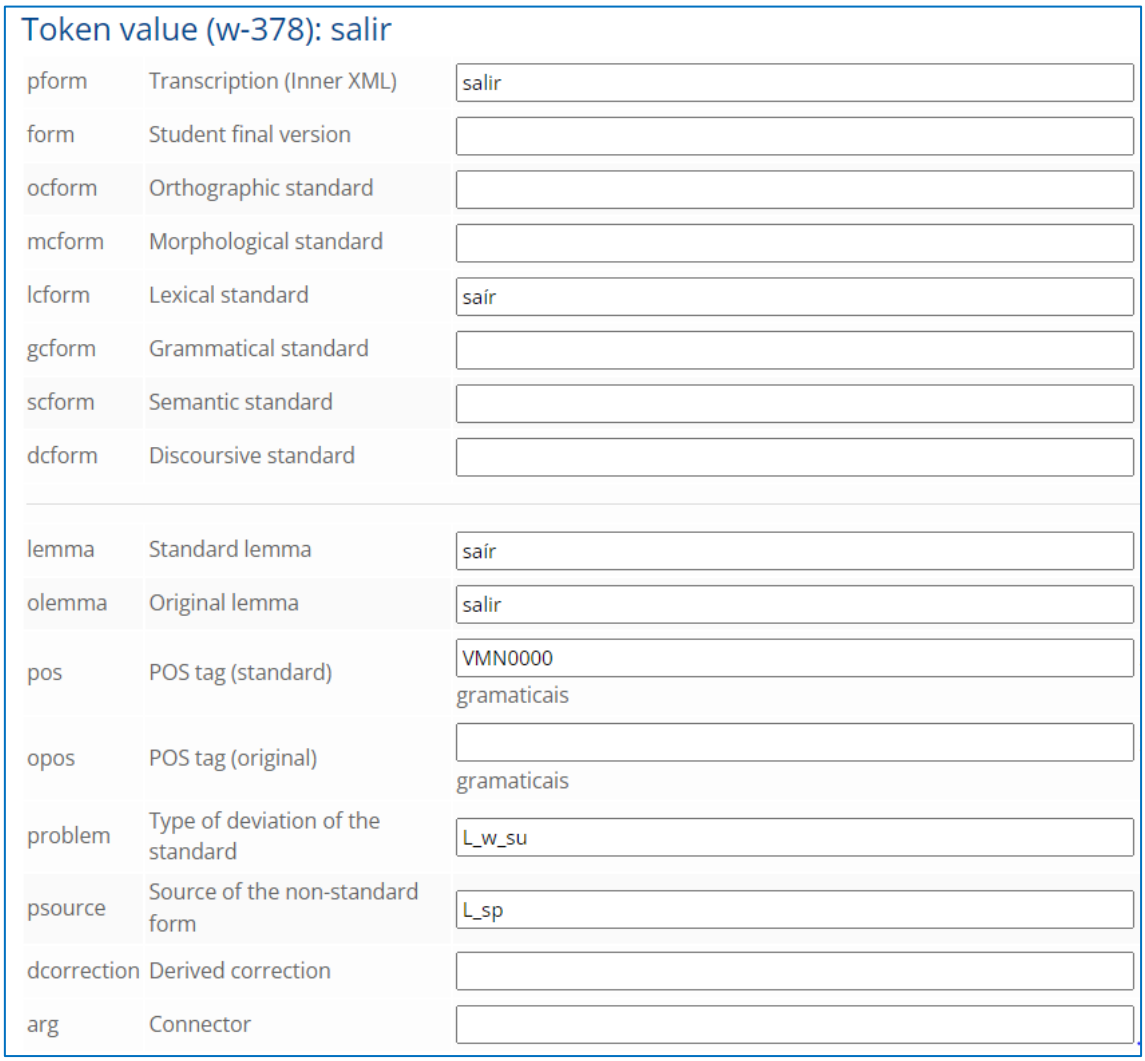


Figura 29. Asignación de lema orixinal

No caso de *leche*, tamén deberiamos asignar unha categoría gramatical diferente á estándar na etiqueta gramatical orixinal:

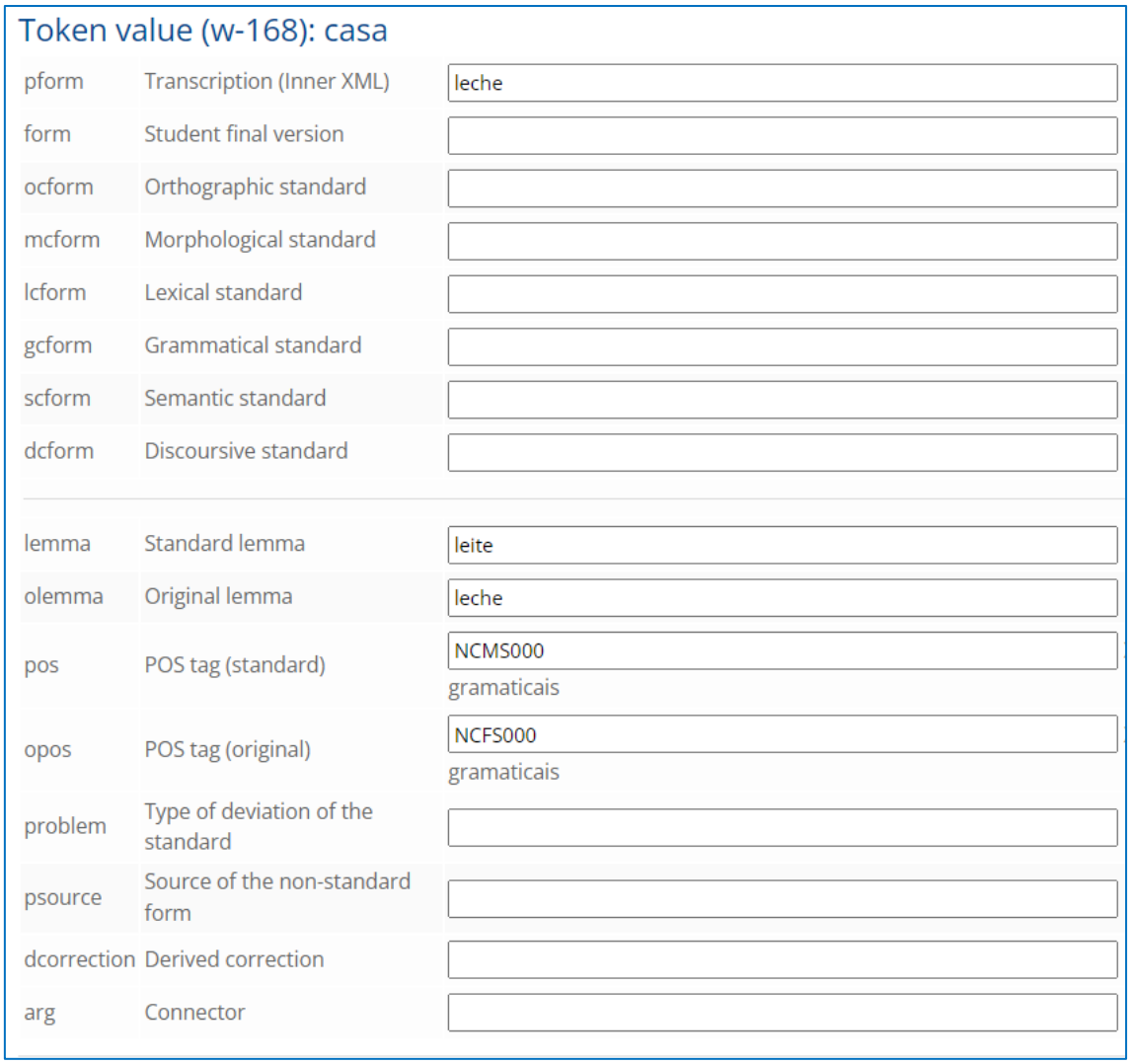


Figura 30. Asignación de lema e categoria gramatical orixinais

No caso das expresións pluriverbais, o lema orixinal debe seguir o mesmo esquema do lema estándar (diferentes lemas e categorías gramaticais separados polo signo +). Así, por exemplo, o lema orixinal de *unha pescadilla que se morde a cola* será *un + pescadilla+que+se+morder+o+cola* e a clase de palabras orixinal DI0FS0 + NCM00000+PR0CN000+PP3CN000+VMIP3S0+DA0FS0+NCFS000. Se a expresión pluriverbal se creou despois da lematización, pódese copiar o lema e clase de palabras estándar que xera TEITOK e pegalos nas caixas de lema e clase de palabras orixinal, procedendo despois á súa revisión.

4. INTRODUCIÓN DE INFORMACIÓN NA CABECEIRA DOS TEXTOS

Seguimos os seguintes pasos:

1) Unha vez o textos están revisados prememos en **estandarización léxica** e copiamos o texto.
2) Vamos a https://ilg.usc.gal/dcontado/.
3) Dámoslle a "Introduce texto", pegamos o texto e dámoslle a enviar (en principio, por defecto, a lingua seleccionada é o galego, pero cómpre asegurarse.

4) Abrimos a folla de cálculo que se xera e incluímos os seguinte datos na cabeceira. Os que están en negriña non se extraen da folla de cálculo, senón que hai que calculalos manualmente:
    a) Número de palabras do texto. Isto encontrámolo na liña 25 da primeira páxina da folla de cálculo. Por exemplo, words: 252 tokens 160 types (63,49%). A cifra subliñada é a que hai que rexistrar.
    b) Número de lemas do texto. Isto encontrámolo na liña 26 da primeira páxina da folla de cálculo. Por exemplo, lemmas: 252 tokens 137 types (54.37%) . A cifra subliñada é a que hai que rexistrar.
    c) **Densidade léxica**: Dividimos o número de lemas polo número de palabras. Sempre será un número comprendido entre 0 e 1. Co exemplo, a densidade léxica sería 137/252 =0,54. Poñemos dous decimais, separados con coma e con redondeo: se o terceiro decimal é ≥ 5 sumamos un decimal máis ao segundo. No caso de que o segundo decimal sexa 0, incluímolo (por exemplo, 0,70). Multiplicámolo por 100 para permitir consultas a través deste dato no buscador.
    d) Número de enunciados: Esa información figura na liña 36 da primeira páxina. Conta de punto a punto (pero non conta os puntos de abreviaturas como etc. ou Sr.). No caso de que falte un punto no texto, cómpre repoñelo ao pegar en Dcontado, porque se non, conta un enunciado menos.
    e) **Media de palabras por enunciado**: Esta información figura na liña 37 da primeira páxina da folla de cálculo: Con todo, Dcontado non redondea, de modo que cómpre facela manualmente. Poñemos dous decimais, seguindo o mesmo sistema indicado en c) (pero sen multiplicar).
    f) Número de palabras no enunciado máis longo: Esta información figura na liña 37 da primeira páxina da folla de cálculo: words/sentence: avg: 21 min: 9 max: 43. A cifra subliñada é a que hai que rexistrar.
    g) Número de palabras no enunciado máis curto. Esta información figura na liña 37 da primeira páxina da folla de cálculo: words/sentence: avg: 21 min: 9 max: 43. A cifra subliñada é a que hai que rexistrar.
    h) **Número de parágrafos**: Hai que contalos manualmente.
    i) **Media de enunciados por parágrafo:** Cómpre dividir o número de enunciados entre o número de parágrafos. Neste caso son 12/3 = 4. No caso de que non dea xusto, poñemos dous decimais, separados con coma e co mesmo sistema de redondeo proposto para a densidade léxica (sen multiplicar por 100). No caso de que o segundo decimal sexa 0 incluímolo (por exemplo, 2,50).
5) Incorporamos eses datos na cabeceira clicando en Edit teiHeader.

5. ALMACENAMENTO DO TEXTO EN TEITOK E WORD

Unha vez introducida a cabeceira, renomeamos o texto manténdoo na carpeta Tagged, e quitándolle simplemente "_pr". A seguir, introducimos as anotacións en standoff e unha vez feito isto copiamos todo o xml (incluída a cabeceira) e creamos un documento en word, co nome do texto, que arquivamos na carpeta Tagged en word.

6. PROBAS

Para facer probas con textos, acódese a “Ficheiros”, “Test”; embaixo de todo vaise a “Create New xml file”, ponse un nome (que non estea repetido) e rematado en .xml e logo “Create XML File”. Pódese escribir caquera texto, gardalo, en “This XML does not (yet) have a text content. To edit the raw XML of the file, click here”. Dáselle a *here* e escríbese o texto que se desexe entre <text xml:space="preserve"> e </text>, dáselle a gardar e despois en “If you wish to tokenize the XML and proceed to the tokenized edit mode, click here”, dáselle a *here* para tokenizar o texto. A partir dese momento xa se pode anotar.


cortegal