Esta presentación é parte do proxecto de I+D+i PGC2018-096069-B-I00, financiado/a por MCIN/AEI/10.13039/501100011033/ e FEDER “Una manera de hacer Europa”

V XORNADAS DE LINGUA E LITERATURA GALEGAS NO ENSINO

CORTEGAL: unha nova ferramenta para traballar a competencia discursiva

MARÍA ÁLVAREZ DE LA GRANJA

INSTITUTO DA LINGUA GALEGA-UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

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

CONTIDOS

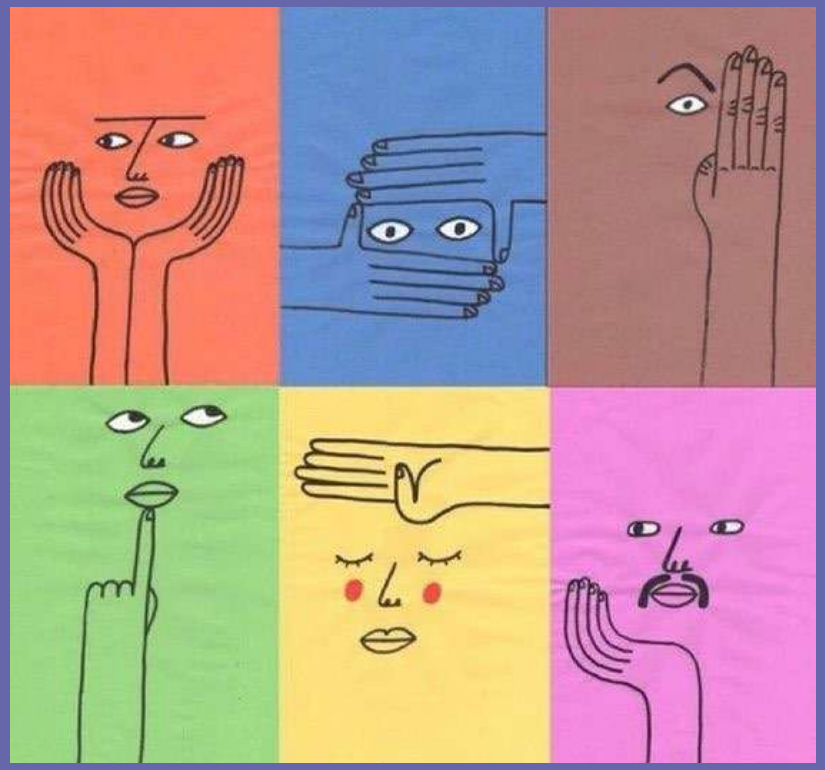
1. Os textos
2. A mostra
3. A plataforma
4. A anotación dos textos
5. A visualización dos textos
6. As buscas
7. Algúns resultados

Corpus para a análise de dificultades na produción escrita en lingua galega

Convenio do Instituto da Lingua Galega coa Secretaría Xeral de Política Lingüística

A SELECCIÓN DOS TEXTOS

Que textos conformarán o corpus?



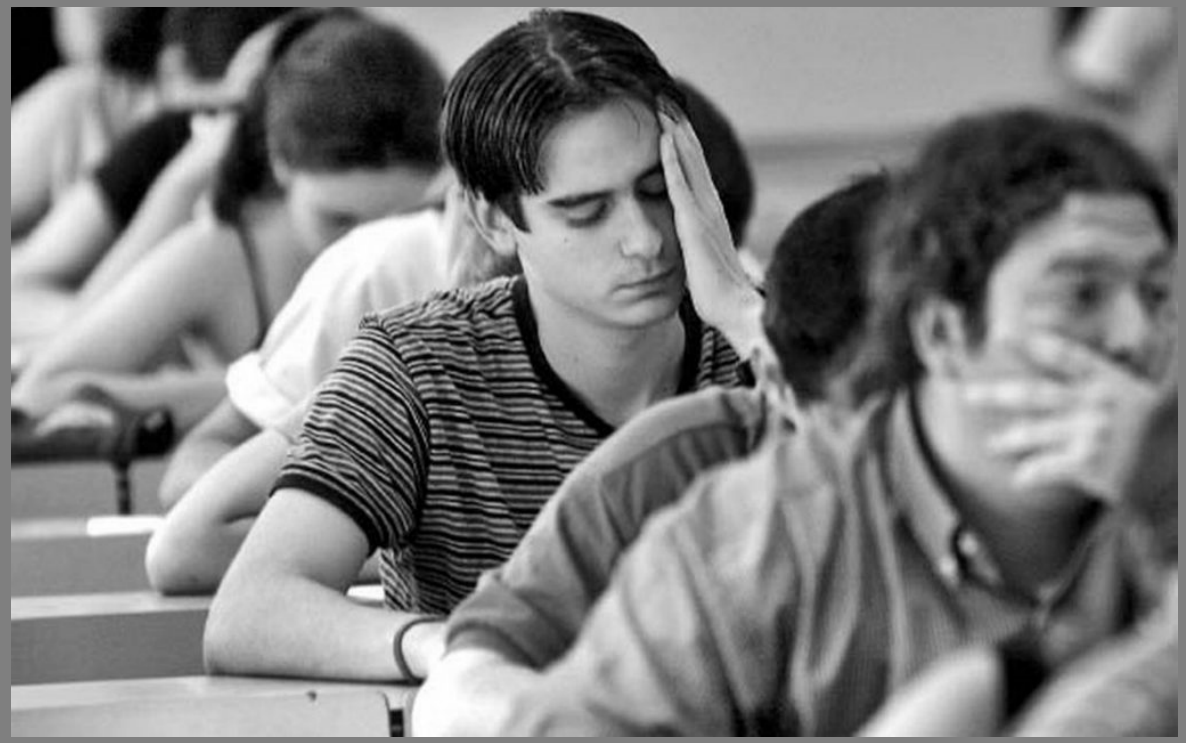
RESPOSTA Á PREGUNTA DE COMENTARIO DOS EXAMES DE LINGUA E LITERATURA GALEGA DAS PROBAS ABAU (CURSO 2016-2017)

OS TEXTOS

Textos de carácter argumentativo, de entre 200 e 250 palabras, sobre un tema concreto, vinculado cun texto previo

ASPECTOS NEGATIVOS

- ✓ Transcrición manual dos textos
- ✓ Ausencia de información sociolingüística
- ✓ Tensión e restricións temporais



ASPECTOS POSITIVOS

- ✓ Dispoñibilidade inmediata dos textos
- ✓ Homoxeneidade
- ✓ Garantía de seriedade
- ✓ Emprego da variedade estándar
- ✓ Seguimento do proceso de composición textual
- ✓ Relevancia e representatividade dos textos

OS TEXTOS

XUÑO 2017

Opción A

Nos últimos anos a gastronomía e a cociña acadaron moita popularidade. Redacta un texto expoñendo a túa opinión sobre este fenómeno: as súas causas, o que ten de moda pasaxeira ou de cambio cultural máis duradeiro...

XUÑO 2017

Opción B

Redacta un texto sobre a importancia que teñen o consumo e a produción (ou o consumismo e a produtividade) no noso modo de vida actual.

SETEMBRO 2017

Opción A

Expón, de maneira argumentada, a túa opinión persoal sobre o problema que reflicte o texto e, en xeral, sobre este tipo de conflitos familiares entre pais e fillos adolescentes.

SETEMBRO 2017

Opción B

A autora móstrase crítica co feito de que a infancia e a mocidade soñe con ser futbolista ou modelo moi maioritariamente (liña 10). Redacta un texto expoñendo de maneira argumentada o teu acordo ou desacordo co seu punto de vista.

A MOSTRA DE TEXTOS

Cantos e que textos?

A MOSTRA

- ✓ 1000 textos (9,87% do total de exames presentados)
- ✓ Reparto proporcional á cifra total de exames por comisión delegada e convocatoria (xuño e setembro)

A MOSTRA. Temas e convocatorias

Convocatoria	Tema	Nº de textos	% sobre o total
Xuño	A gastronomía	449	44,9%
	Consumo e produción	449	44,9%
Total xuño		*898*	*89,8%*
Setembro	Conflitos familiares	51	5,1%
	Os referentes da mocidade	51	5,1%
Total setembro		*102*	*10,2%*
Total		***1000***	***100%***

A MOSTRA. Cualificacións

Convocatoria	Exame galego	Desviación estándar	Comentario	Desviación estándar	ABAU galego
Xuño	6,09	1,65	6,93	1,87	5,77
Setembro	4,26	1,43	5,21	1,75	4,17
Xuño e setembro	5,90	1,72	6,75	1,93	5,58

A MOSTRA. Cualificacións

Convocatoria	Exame galego	Desviación estándar	Comentario	Desviación estándar	ABAU galego
Xuño	6,09	1,65	**6,93**	1,87	5,77
Setembro	4,26	1,43	**5,21**	1,75	4,17
Xuño e setembro	5,90	1,72	**6,75**	1,93	5,58

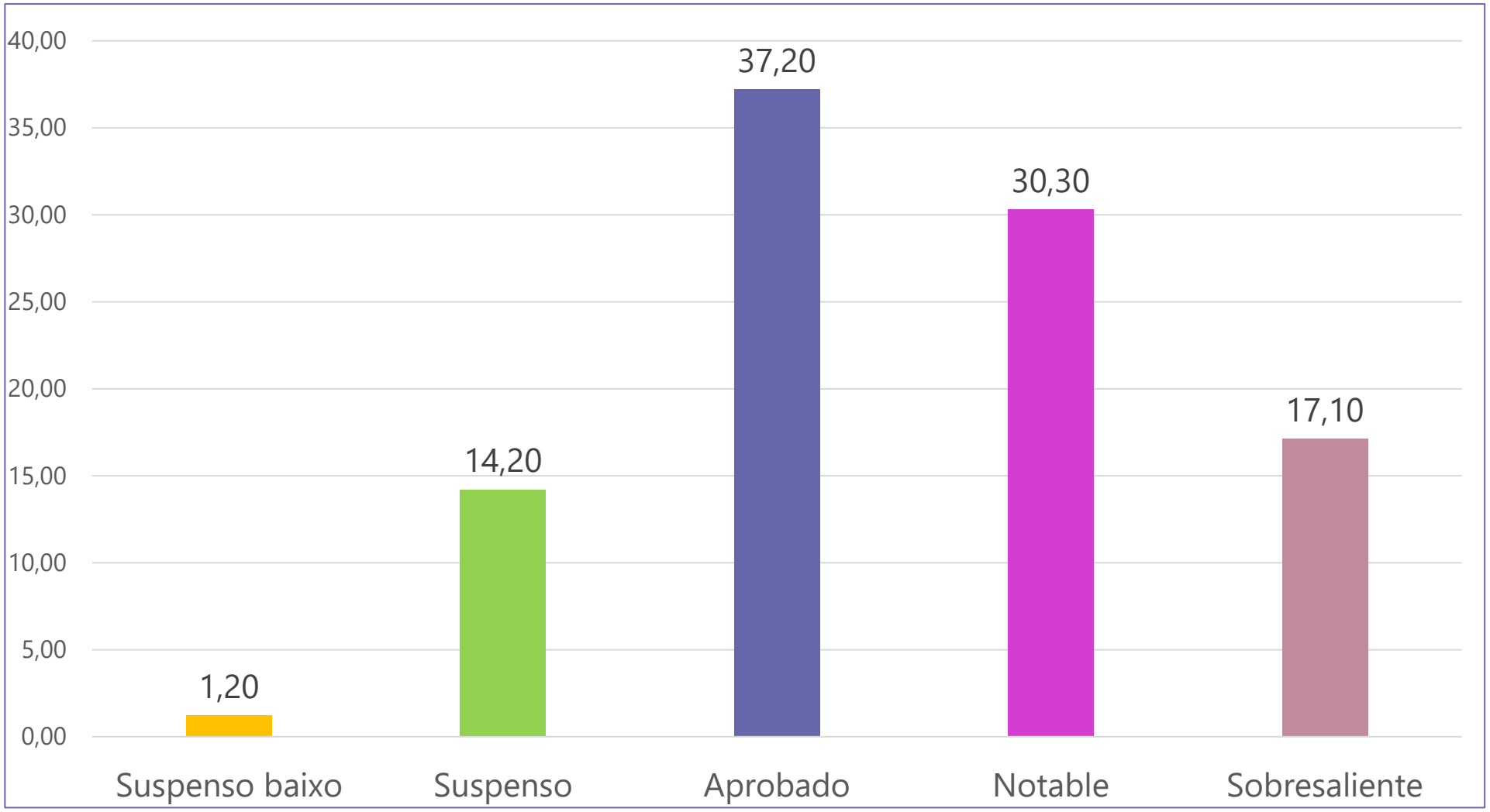
A MOSTRA. Cualificacións

Convocatoria	Exame galego	Desviación estándar	Comentario	Desviación estándar	ABAU galego
Xuño	6,09	1,65	6,93	1,87	5,77
Setembro	4,26	1,43	5,21	1,75	4,17
Xuño e setembro	5,90	1,72	6,75	1,93	5,58

A MOSTRA. Cualificacións

Convocatoria	Exame galego	Desviación estándar	Comentario	Desviación estándar	ABAU galego
Xuño	6,09	1,65	6,93	1,87	**5,77**
Setembro	4,26	1,43	5,21	1,75	**4,17**
Xuño e setembro	5,90	1,72	6,75	1,93	**5,58**

Cualificacións dos textos de CORTEGAL



A PLATAFORMA

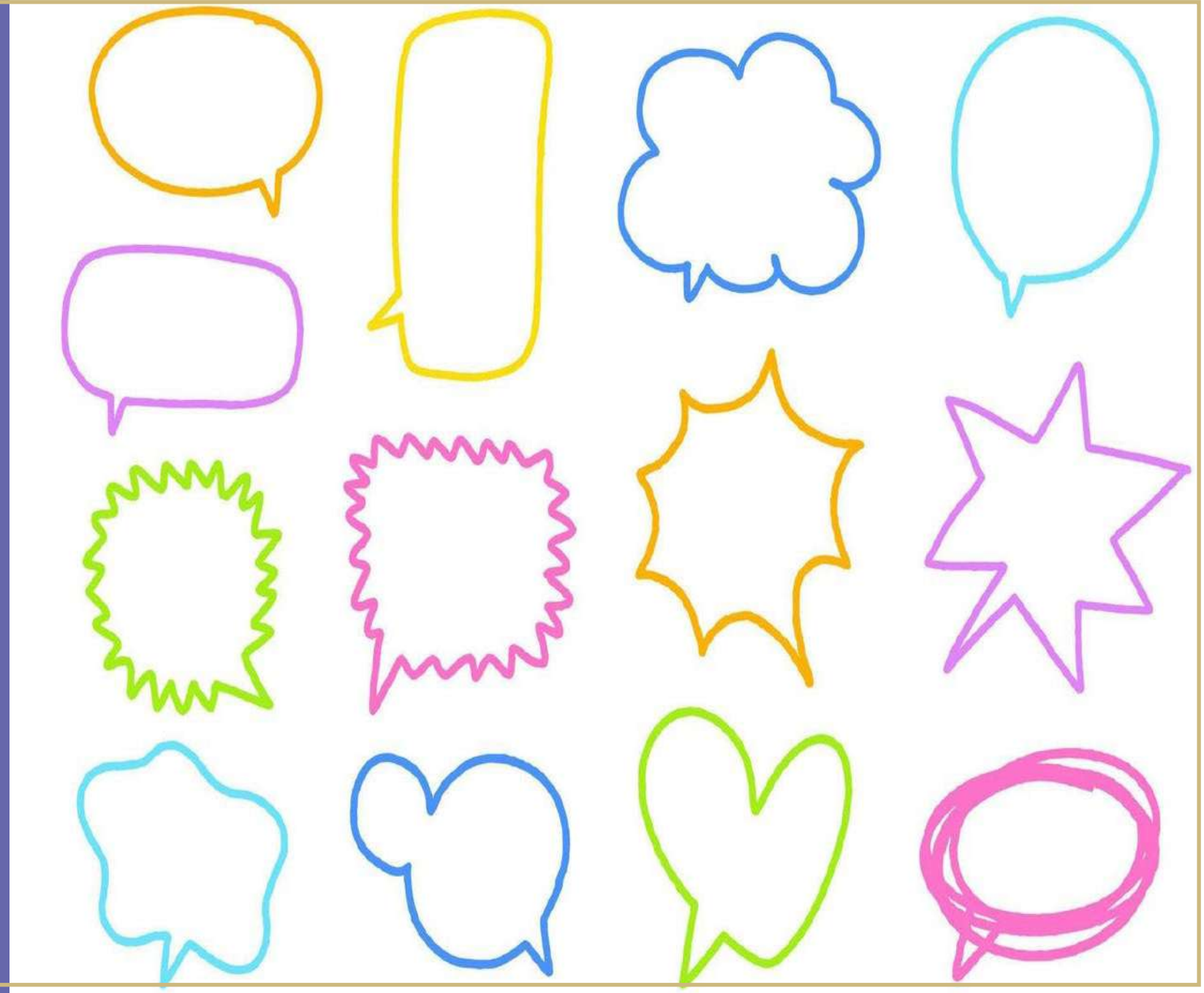
Onde elaborar o corpus?

A PLATAFORMA

- ✓ Transcrición e anotación dos textos en TEITOK (Janssen, 2016), unha plataforma para crear, editar, visualizar e consultar corpus.
- ✓ Posibilidade de visualizar os textos en diferentes capas.
- ✓ Experiencia previa en TEITOK (corpus Gondomar).
- ✓ Plataforma gratuíta e relativamente doada de usar

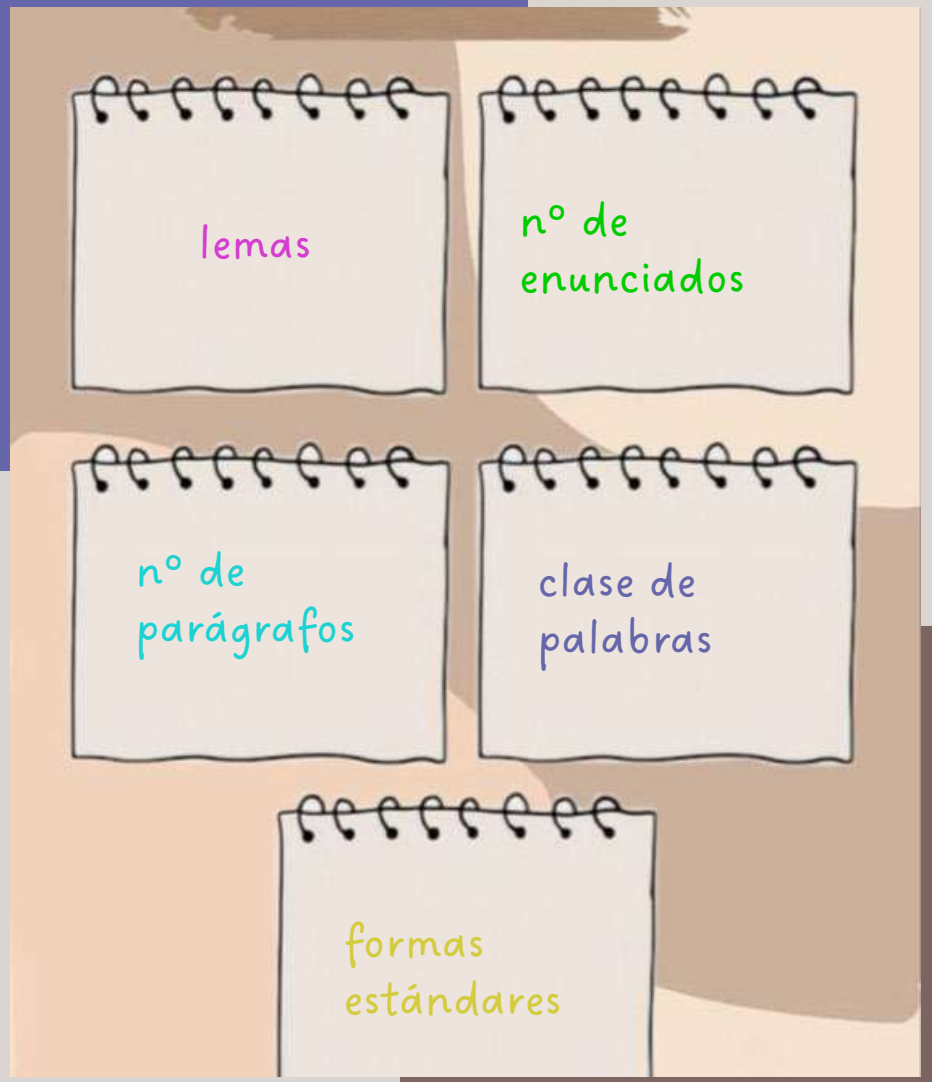
A ANOTACIÓN DOS TEXTOS

Que información ofrecer sobre os textos?



ANOTACIÓNS

- ✓ Anotacións sobre os textos
- ✓ Anotacións sobre as palabras e secuencias de palabras



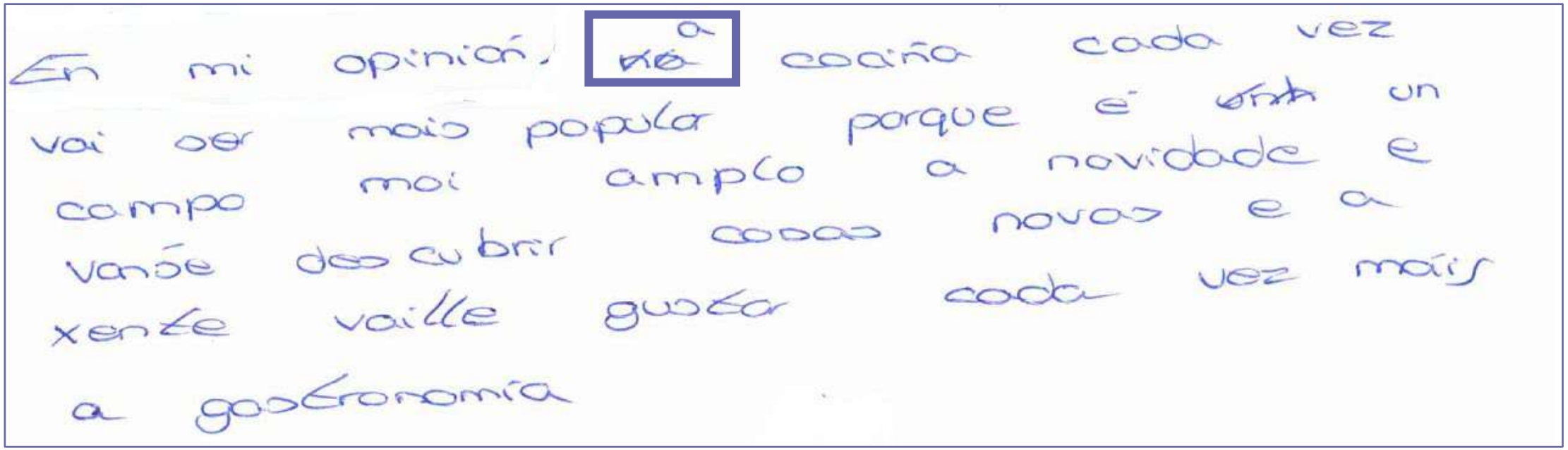
Anotación sobre os textos

DATOS CUALITATIVOS	DATOS CUANTITATIVOS
Tema	Número de palabras
Convocatoria	Número de lemas
Comisión delegada	Densidade léxica (nº de lemas / nº de palabras)
	Número de enunciados
	Media de palabras por enunciado
	Número de palabras do enunciado máis longo
	Número de palabras do enunciado máis curto
	Número de parágrafos
	Media de enunciados por parágrafo

Anotación sobre as palabras

- ✓ Formas riscadas, engadidas a posteriori e de lectura dubidosa
- ✓Formas non estándares
- ✓ Lemas e categorías gramaticais
- ✓Conectores entre enunciados

Transcrición dos textos



En mi opinión, ~~na~~ <add>a</add> cociña cada vez <lb/>vai ser mais popular porque é ~~unh~~ un <lb/>campo moi amplo a novidade e <lb/>vánse descubrir cosas novas e a <lb/>xente vaille gustar cada vez máis <lb/>a gastronomía

Formas non estándares

CODIFICACIÓN	ESTANDARIZACIÓN
Tipo de diverxencia	Nivel ortográfico
Orixe da diverxencia	Nivel morfolóxico
	Nivel léxico
	Nivel gramatical (sintáctico)
	Nivel semántico
	Nivel discursivo

abundáncia

Nivel ortográfico

1. Emprego de signos diacríticos (acentuación e diérese)
2. Maiúsculas e minúsculas
3. Escritura conxunta ou separada das palabras
4. Confusión, adición ou omisión de letras ou dígrafos
5. Escritura de estranxeirismos, abreviaturas e siglas
6. Representación de contraccións e asimilacións.

pform	Transcription (Inner XML)	abundáncia
form	Student final version	
ocform	Orthographic standard	abundancia
mcform	Morphological standard	
lcform	Lexical standard	
gcform	Grammatical standard	
scform	Semantic standard	
dcform	Discoursive standard	
lemma	Standard lemma	abundancia
olemma	Original lemma	
pos	POS tag (standard)	NCFS000
opos	POS tag (original)	
problem	Type of deviation of the standard	O_ac_ad

Nivel morfolóxico

1. Flexión: verbos, adxectivos, substantivos e palabras gramaticais.
2. Xénero, número, conxugación verbal e morfemas avaliativos.

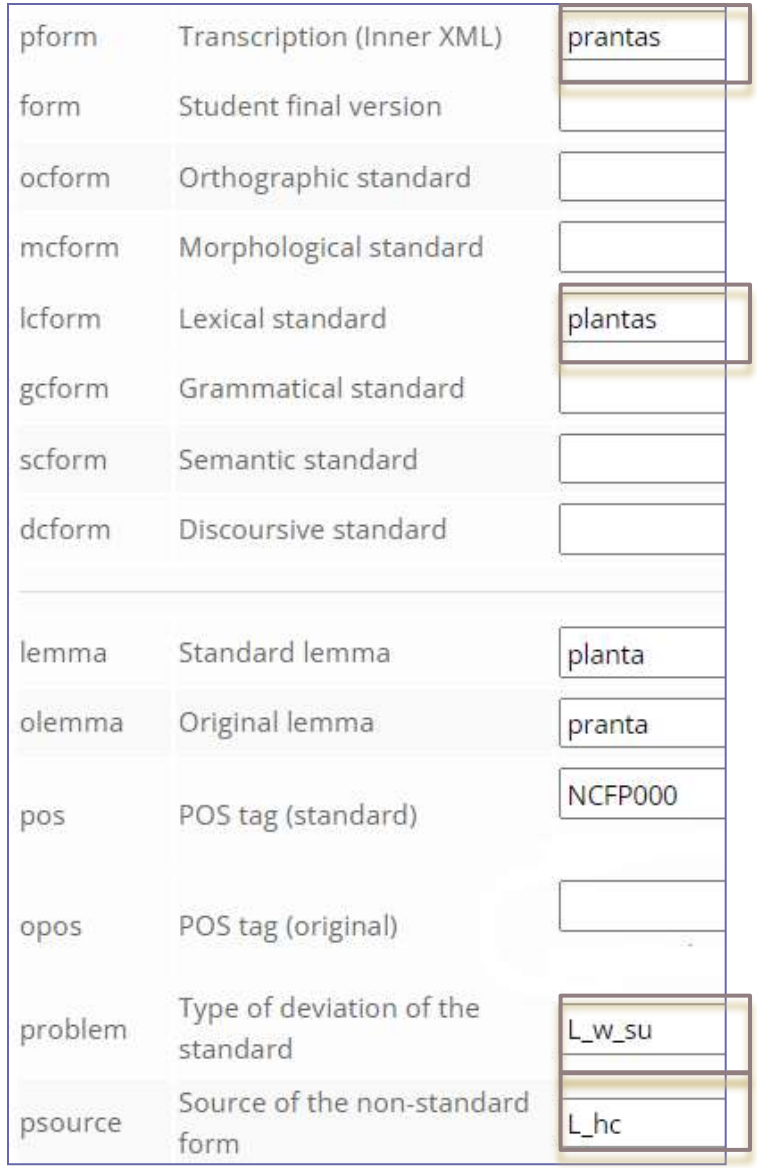
razones

pform	Transcription (Inner XML)	razones
form	Student final version	
ocform	Orthographic standard	
mcform	Morphological standard	razóns
lcform	Lexical standard	
gcform	Grammatical standard	
scform	Semantic standard	
dcform	Discoursive standard	
lemma	Standard lemma	razón
olemma	Original lemma	
pos	POS tag (standard)	NCFP000
opos	POS tag (original)	
problem	Type of deviation of the standard	M_num_su
psource	Source of the non-standard form	M_sp

Nivel léxico

1. Emprego dunha unidade léxica con diferenzas de xénero con respecto ao galego estándar (*a leite*)
2. Emprego dunha unidade léxica con diferenzas na acentuación de intensidade con respecto ao galego estándar (*élite*)
3. Outras formas léxicas en que as diferenzas co estándar van máis alá do xénero ou da acentuación (*receta*, *abandoar*, *todavía*...)

prantas

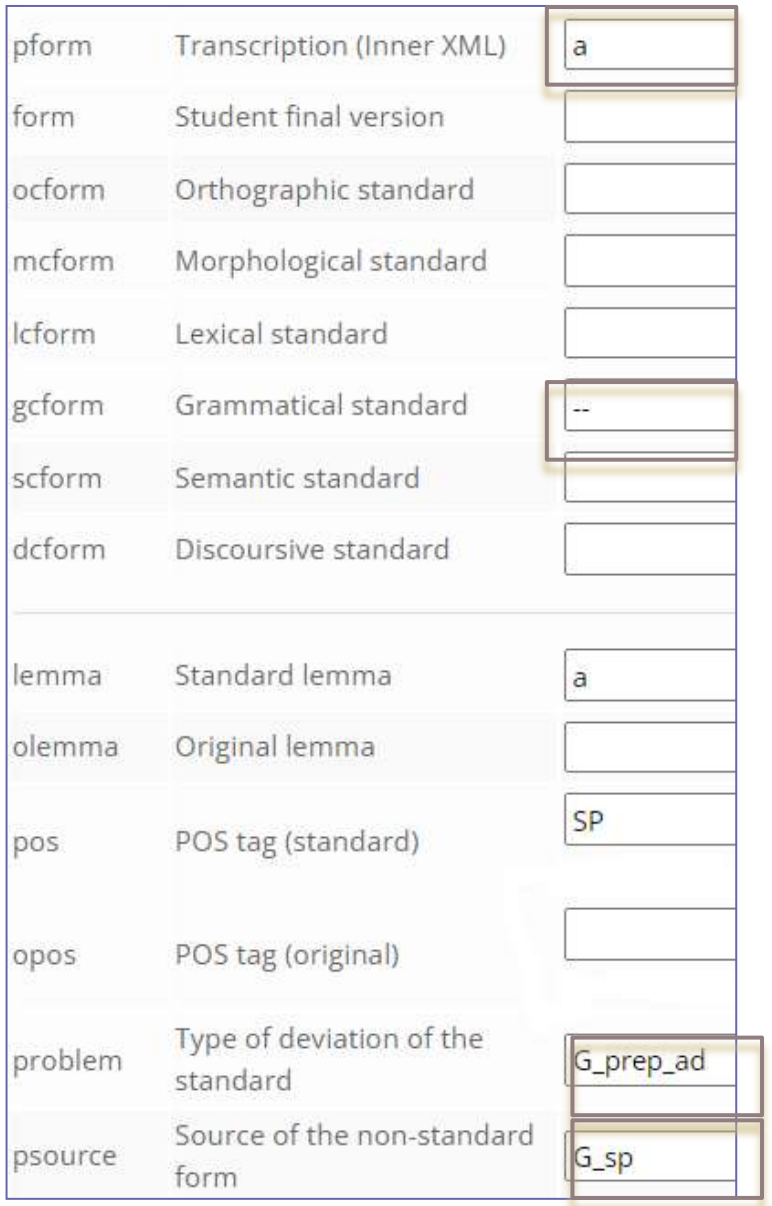


pform	Transcription (Inner XML)	prantas
form	Student final version	
ocform	Orthographic standard	
mcform	Morphological standard	
lcform	Lexical standard	plantas
gcform	Grammatical standard	
scform	Semantic standard	
dcform	Discoursive standard	
lemma	Standard lemma	planta
olemma	Original lemma	pranta
pos	POS tag (standard)	NCFP000
opos	POS tag (original)	
problem	Type of deviation of the standard	L_w_su
psource	Source of the non-standard form	L_hc

a xente sempre vai **a** ter que comer

Nivel gramatical

1. Concordancia
2. Omisión e adición de preposicións
3. Emprego de determinantes, pronomes e conxuncións
4. Selección de tempo, modo ou voz verbal
5. Esquemas verbais
6. Escolla de clase de palabra
7. Estruturas sintácticas en xeral

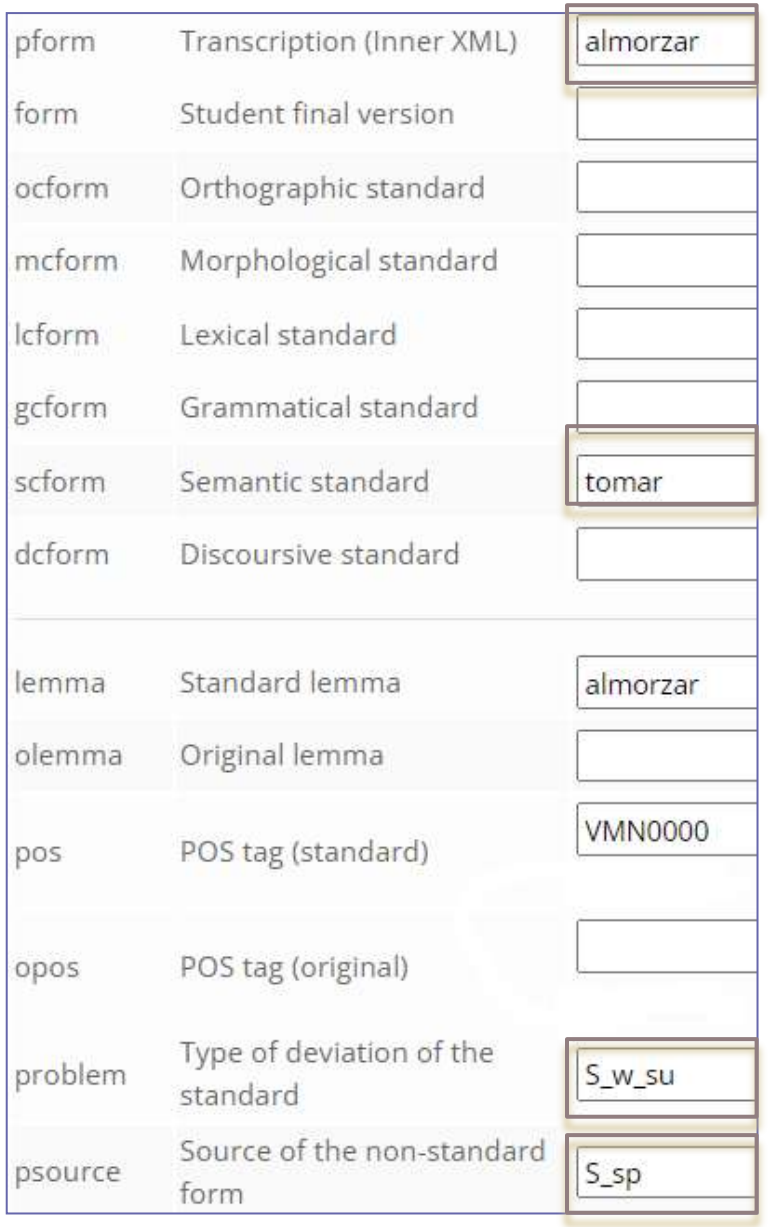


pform	Transcription (Inner XML)	a
form	Student final version	
ocform	Orthographic standard	
mcform	Morphological standard	
lcform	Lexical standard	
gcform	Grammatical standard	--
scform	Semantic standard	
dcform	Discoursive standard	
lemma	Standard lemma	a
olemma	Original lemma	
pos	POS tag (standard)	SP
opos	POS tag (original)	
problem	Type of deviation of the standard	G_prep_ad
psource	Source of the non-standard form	G_sp

lugares onde poder **almorzar** a comida típica da zona

Nivel semántico

1. Usos ou significados dunha palabra non estándares ou non adecuados ao contexto ou ao sentido que se quere transmitir
2. Omisión de voces necesarias para o sentido do texto
3. Adición de voces innecesarias

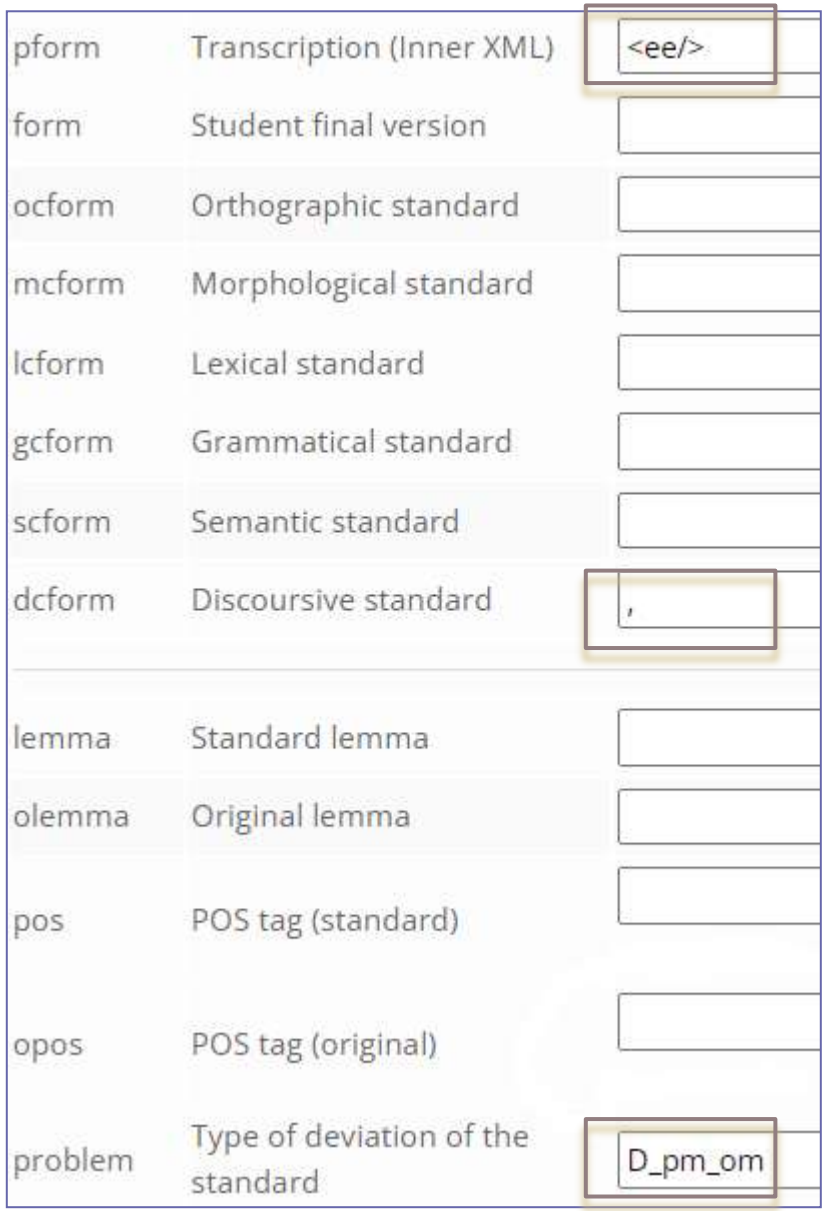


pform	Transcription (Inner XML)	almorzar
form	Student final version	
ocform	Orthographic standard	
mcform	Morphological standard	
lcform	Lexical standard	
gcform	Grammatical standard	
scform	Semantic standard	tomar
dcform	Discoursive standard	
lemma	Standard lemma	almorzar
olemma	Original lemma	
pos	POS tag (standard)	VMN0000
opos	POS tag (original)	
problem	Type of deviation of the standard	S_w_su
psource	Source of the non-standard form	S_sp

Neste caso fala dese **conflicto pero** existen moitos máis

Nivel discursivo

1. Puntuación
2. División do texto en parágrafos
3. Emprego de conectores e de partículas referenciais
4. Elipses inadecuadas
5. Orde inapropiada
6. Rexistro das palabras
7. Enunciados complexos
8. Enunciados inintelixibles

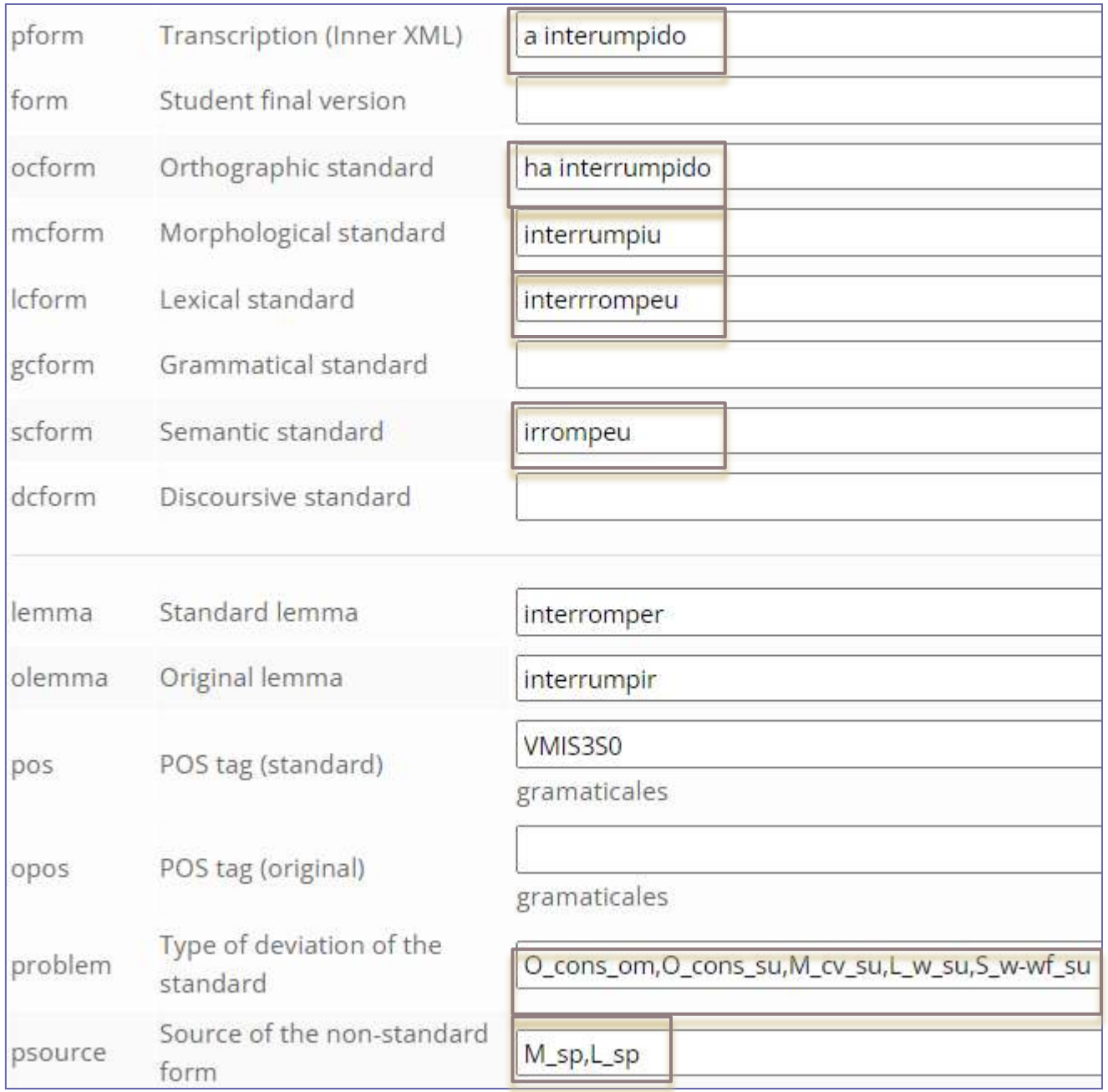


pform	Transcription (Inner XML)	<ee/>
form	Student final version	
ocform	Orthographic standard	
mcform	Morphological standard	
lcform	Lexical standard	
gcform	Grammatical standard	
scform	Semantic standard	
dcform	Discoursive standard	,
lemma	Standard lemma	
olemma	Original lemma	
pos	POS tag (standard)	
opos	POS tag (original)	
problem	Type of deviation of the standard	D_pm_om

Internet **a interumpido** de forma férrea no noso mercado

Anotación multinivel

Posibilidade de asignar a un mesmo elemento varias formas estándares diferentes pertencentes a distintos niveis lingüísticos



Anotación de secuencias

Algúns códigos que poden afectar a secuencias asígnanse mediante un sistema diferente (*stand-off*)

Co aumento deste consumismo, é lóxico pensar que a productividade ten que ir ao mesmo nivel, e, polo tanto, as ganancias das empresas son esaxeradas, e todo gracias a que, dalgunha maneira, a época na que nacimos e os avances que aparecen constantemente inculcannos o desexo de consumir, de ter cada vez máis cousas só polo feito de que a xente sepa que compras e que tanto nivel económico tes, pois agora, por exemplo, cos avances progresivos dos móbiles, comprar cada vez un mobil máis moderno fainos sentir mellor con nós mesmos, porque nos sentimos ao día dos avances e iso fainos creer que somos alguén. Ao final, non é máis ca unha maneira de sentirnos importantes.

a xente sempre vai **a** ter que comer

Dobre codificación

Códigos que describen o tipo de desviación do estándar:

G_prep_ad

Códigos que sinalan a orixe da forma desviante do estándar:

G_sp (hc, gal, anal, for, spadapt, mix, or)

pform	Transcription (Inner XML)	a
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	a
olemma	Original lemma	
pos	POS tag (standard)	SP
opos	POS tag (original)	
problem	Type of deviation of the standard	G_prep_ad
psource	Source of the non-standard form	G_sp

Lema e categoría gramatical

Asignado automaticamente mediante Freeling

hai

pform	Transcription (Inner XML)	hai
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	haber
olemma	Original lemma	
pos	POS tag (standard)	VMIP3S0

abuela

Dobre lema

Nas palabras corrixidas no nivel léxico

Lema estándar: avó
Lema orixinal: abuelo

pform	Transcription (Inner XML)	abuela
form	Student final version	
ocform	Orthographic standard	
mcform	Morphological standard	
lcform	Lexical standard	avoa
gcform	Grammatical standard	
scform	Semantic standard	
dcform	Discoursive standard	
lemma	Standard lemma	avó
olemma	Original lemma	abuelo

CORPUS DE TEXTOS ESCRITOS POR ESTUDANTES NO ÁMBITO ACADÉMICO

http://ilg.usc.gal/cortegal

OS TEXTOS E A SÚA VISUALIZACIÓN

Onde atopar os textos e que opcións existen para a súa visualización?

AS BUSCAS

Que consultas se poden facer e como se realizan?

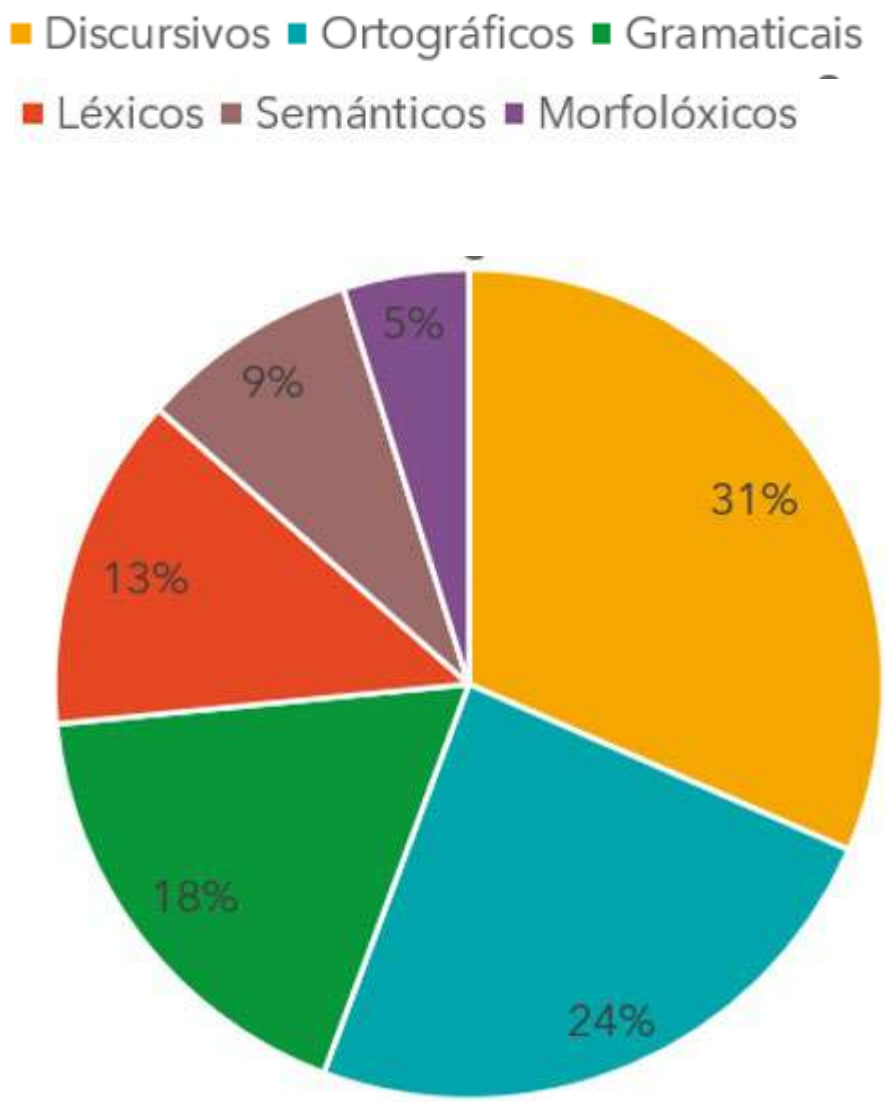
ALGÚNS RESULTADOS

Que nos mostra o corpus?

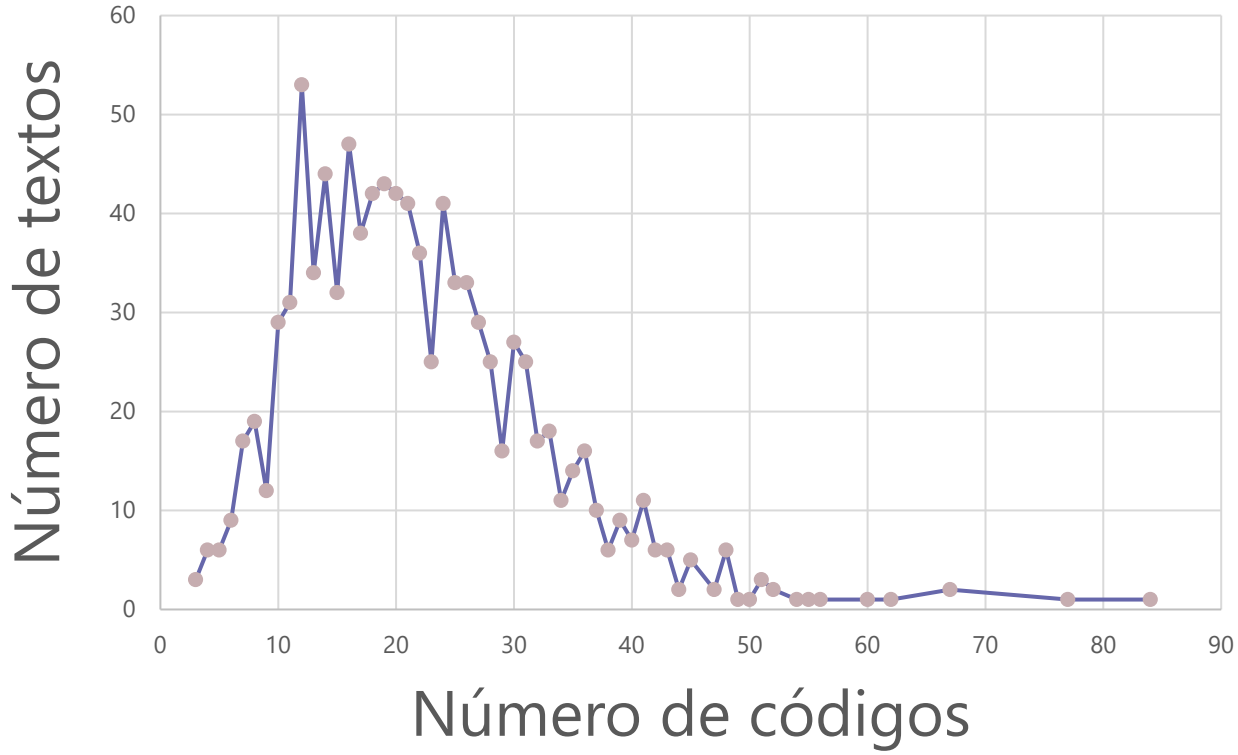
Media total	Media xuño	Media setembro	Desviación estándar	Rango
Palabras				
224,42	225,66	213,52	52,75	51-613
Enunciados				
9,03	9,15	8,06	3,23	2-29
Palabras por enunciado				
12,28	11,94	15,32	6,49	1-93
Parágrafos				
4,30	4,32	4,07	1,60	1-12
Enunciados por parágrafo				
2,41	2,42	2,33	1,39	1-11

Nivel	Número de códigos
Discursivo	7366
Ortográfico	5648
Gramatical	3958
Léxico	3049
Semántico	2004
Morfolóxico	1144
Total	23.169

23,17 códigos de desviación por texto



Rango: 84 códigos (1 texto)- 3 códigos (3 textos)



CÓDIGO	VALOR	NÚMERO DE CÓDIGOS ASIGNADOS
D_pm_om	Omisión dun signo de puntuación	3662
L_w_su	Selección dunha unidade léxica non estándar con diferenzas que van máis alá do xénero ou da acentuación	2887
O_ac_om	Omisión de acento gráfico	2362
S_w_su	Selección dunha unidade léxica inadecuada desde o punto de vista semántico ou combinatorio	1391
D_pm_su	Selección inadecuada dun signo de puntuación	1229
D_pm_ad	Adición dun signo de puntuación	919
G_num_su	Selección de número non estándar	854
O_cons_ad	Adición de consoante	798
O_ac_ad	Adición de acento gráfico	672
M_v_su	Selección dunha forma flexiva verbal non estándar	520

EQUIPO

María Álvarez de la Granja

Francisco Cidrás

José Antonio Cutrín Garabal

Xosé Antonio Fernández Salgado

Ernesto González Seoane

Maarten Janssen

Natalia Lodeiros Souto

María López Sández

Vítor Míguez Rego

Cristian Pernas Rubal

Nereida Prada Piñeiro

Tamara Rial Montes

Reyes Rodríguez Rodríguez

Xulio Sousa

CONTACTO

maria.alvarez.delagranja@usc.gal