Anotación morfosintáctica do Corpus Técnico do Galego

Xavier Gómez Guinovart
Universidade de Vigo
xgg@uvigo.es

Susana López Fernández
Universidade de Vigo
susanalopez@uvigo.es

Resumo

Neste traballo preséntanse a metodoloxía e os criterios empregados na anotación lingüística (etiquetaxe categorial e lematización) do Corpus Técnico do Galego, un corpus elaborado na Universidade de Vigo con textos monolingües especializados do galego contemporáneo nos eidos do dereito, da informática, da economía, das ciencias ambientais, da socioloxía e da medicina.

1. *Introdución*

O Corpus Técnico Anotado do Galego (CTAG) é a versión categorizada e lematizada do Corpus Técnico do Galego (CTG), unha colección de córpora do galego contemporáneo composta de textos monolingües especializados nos eidos do dereito, da informática, da economía, das ciencias ambientais, da socioloxía e da medicina, dispoñible en Internet desde 2006 para libre consulta (Gómez Clemente e Gómez Guinovart, 2006-2009). Cunha extensión actual de 12,5 millóns de palabras, o CTG reúne textos do ámbito xurídico-administrativo (2.516.846 palabras), textos de informática e telecomunicacións (2.027.816 palabras), textos de ecoloxía e ciencias ambientais (2.349.362 palabras), textos de economía (2.055.837 palabras), textos de socioloxía (2.442.765 palabras) e textos de medicina (1.154.071 palabras, aínda en fase de recompilación). A anotación do Corpus CTAG non é totalmente automática, senón que ten unha primeira fase na que se lle aplica un programa etiquetador e lematizador, e unha segunda fase na que se revisan manualmente os resultados deste procesamento automático. Os traballos de anotación lingüística do CTAG, en fase avanzada de elaboración, lévanse a cabo no marco de dous proxectos de investigación en curso[1], aínda que os seus resultados iniciais xa se poden consultar en Internet (Gómez Guinovart, 2006-2009). En concreto, xa se atopa dispoñible en Internet unha sección do CTAG de máis de 2 millóns de palabras, correspondente ao ámbito especializado da ecoloxía e das ciencias ambientais.

A etiquetaxe inicial do CTAG levouse a cabo empregando unha adaptación modificada do analizador morfolóxico do galego que forma parte do par español-galego do tradutor Apertium (Armentano Oller et al., 2006; Alegría Loinaz et al., 2006), con cambios no seu etiquetario, no tratamento das contraccións e no manexo das formas non normativas do galego. De maneira xeral, o conxunto de etiquetas deseñado para a anotación do CTAG constitúe unha adaptación ás características propias do galego dos principios elaborados polo grupo EAGLES (Leech e Wilson, 1996) para a creación dun estándar europeo de anotación morfosintáctica de léxicos e córpora. De maneira máis específica, o conxunto normalizado de etiquetas utilizado para o CTAG elaborouse tendo en conta as propostas realizadas por Civit para a lingua castelá (Civit, 2003) e adoptadas con algunhas modificacións no etiquetador morfolóxico do Freeling (Atserias et al., 2006). Nos seguintes apartados deste traballo, presentarase polo miúdo o etiquetario empregado na anotación do corpus, e as cuestións de deseño relacionadas coa codificación das formas anotadas e co tratamento das contraccións, formas enclíticas e formas non normativas presentes nos textos.

2. *Etiquetaxe do Corpus CTAG*

2.1. Codificación

A anotación do Corpus CTAG ten en conta todas as formas léxicas (galegas e non galegas, normativas e non normativas) que aparecen nos textos, e mais as cifras, abreviaturas, símbolos e signos de puntuación. Cada forma etiquetada consta de tres partes: a forma que aparece no texto, o

[1]Este traballo foi financiado polo Ministerio de Educación y Ciencia e o Fondo Europeo de Desenvolvemento Rexional (FEDER), dentro do proxecto *Deseño e implementacion dun servidor de recursos integrados para o desenvolvemento de tecnoloxías da lingua galega (RILG)* do Plan Nacional de I+D+I, 2006-2009 (ref. HUM2006-11125-C02-01/FILO); e pola Consellaría de Innovación e Industria da Xunta de Galicia, dentro do proxecto *Desenvolvemento e aplicación de recursos integrados da lingua galega* do Plan galego de investigación, desenvolvemento e innovación tecnolóxica (Incite), 2008-2011 (ref. INCITE08PXIB302185PR). Ambos son proxectos coordinados da Universidade de Vigo (Grupo TALG) coa Universidade de Santiago de Compostela (Instituto da Lingua Galega).

lema (ou representación abstracta da clase flexiva) e a etiqueta categorial, consonte o seguinte esquema: *^forma/lema_etiqueta*. Deste xeito, o adxectivo *transxénicos* vai ser anotado no corpus como *^transxénicos/transxénico_A0MP*.

2.2. Etiquetario

Para cada categoría inclúense dúas táboas. Na primeira táboa, recóllense as características lingüísticas ou atributos pertinentes para cada categoría (segunda columna), cos seus posibles valores (terceira columna), a abreviatura ou codificación dos valores na etiqueta (cuarta columna), e o lugar ou posición (primeira columna) que cada un dos valores vai ocupar na etiqueta resultante. Na segunda táboa, recóllese o inventario completo de etiquetas para cada categoría, cun exemplo de palabra e lema para cada caso. Esta descrición esquemática do etiquetario do CTAG, empregando táboas, está baseada no sistema utilizado en Civit (2003).

2.2.1. Nomes

NOMES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Nome	N
2	Tipo	Común	C
		Propio	P
3	Xénero	Masculino	M
		Feminino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N
5	Grao	Apreciativo	A

Táboa 1: Etiquetas para nomes

Forma	**Lema**	**Etiqueta**
neno	neno	NCMS0
nenos	neno	NCMP0
nena	neno	NCFS0
nenas	neno	NCFP0
xornalista	xornalista	NCCS0
xornalistas	xornalista	NCCP0
microondas	microondas	NCMN0
Breogán	Breogán	NP000
neniño	neno	NCMSA
neniños	neno	NCMPA
neniña	neno	NCFSA
neniñas	neno	NCFPA

Táboa 2: Exemplos de nomes

O lema dos nomes vai ser sempre a forma masculina singular (*neno*) ou a forma singular común se o nome é de xénero común (*xornalista*). Nos nomes invariables, isto é, naqueles que presentan a mesma forma tanto no singular coma no plural (*microondas*), o lema e a forma van ser sempre coincidentes.

O atributo *grao* con valor **A** especifícase nos nomes con sufixación apreciativa (aumentativos, diminutivos, pexorativos, etc.) (*neniño, nenón*). No resto de nomes, o valor do atributo *grao* é de non especificado ou **0**.

Finalmente, os nomes propios levan no CTAG a etiqueta NP000, cos valores de xénero, número e grao sen especificar.

2.2.2. Adxectivos

ADXECTIVOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adxectivo	A
2	Grao	Apreciativo	A
3	Xénero	Masculino	M
		Feminino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N

Táboa 3: Etiquetas para adxectivos

Forma	**Lema**	**Etiqueta**
febles	feble	A0CP
feble	feble	A0CS
ecolóxicas	ecolóxico	A0FP
ecolóxica	ecolóxico	A0FS
ecolóxicos	ecolóxico	A0MP
ecolóxico	ecolóxico	A0MS
choromicas	choromicas	A0CN
grandiñas	grande	AAFP
grandiña	grande	AAFS
grandiños	grande	AAMP
grandiño	grande	AAMS

Táboa 4: Exemplos de adxectivos

O lema dos adxectivos vai ser sempre a forma masculina singular (*ecolóxico*) ou a forma singular común se o adxectivo é de xénero común (*fértil*). Nos adxectivos invariables (*choromicas*), o lema e a forma van ser sempre coincidentes.

O atributo *grao* especificarase para os adxectivos con grao comparativo (*meirande*) ou superlativo (*altísimo*), ou con sufixación apreciativa (diminutivos, aumentativos, pexorativos, etc.) (*pequeniño, fermosón*). Estes dous tipos de adxectivos vanse distinguir porque o valor do segundo atributo da etiqueta vai ser **A**, mentres que no resto de adxectivos vai ser sempre **0**.

2.2.3. Verbos

O lema dos verbos é sempre o infinitivo. O atributo de xénero só afecta aos participios. Nas formas de infinitivo e xerundio non conxugados non se especifican os atributos de tempo, persoa, número e xénero, polo que o seu valor vai ser

sempre de **0**. Só os participios e os xerundios poden levar o atributo de apreciativo, nos resto dos casos o valor na etiqueta é **0**.

VERBOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Verbo	V
2	Modo	Indicativo	I
		Subxuntivo	S
		Imperativo	M
		Infinitivo	N
		Xerundio	X
		Participio	P
3	Tempo	Presente	P
		Copretérito	I
		Futuro	F
		Pretérito	S
		Pospretérito	C
		Antepretérito	A
4	Persoa	Primeira	1
		Segunda	2
		Terceira	3
5	Número	Singular	S
		Plural	P
6	Xénero	Masculino	M
		Feminino	F
7	Grao	Apreciativo	A

Táboa 5: Etiquetas para verbos

Tempo	**Forma**	**Lema**	**Etiqueta**
Pres. Ind.	canto	cantar	VIP1S00
	cantas	cantar	VIP2S00
	canta	cantar	VIP3S00
	cantamos	cantar	VIP1P00
	cantades	cantar	VIP2P00
	cantan	cantar	VIP3P00
Copretérito	cantaba	cantar	VII1S00
	cantabas	cantar	VII2S00
	cantaba	cantar	VII3S00
	cantabamos	cantar	VII1P00
	cantabades	cantar	VII2P00
	cantaban	cantar	VII3P00
Pret. Ind.	cantei	cantar	VIS1S00
	cantaches	cantar	VIS2S00
	cantou	cantar	VIS3S00
	cantamos	cantar	VIS1P00
	cantastes	cantar	VIS2P00
	cantaron	cantar	VIS3P00
Fut. Ind.	cantarei	cantar	VIF1S00
	cantarás	cantar	VIF2S00
	cantará	cantar	VIF3S00
	cantaremos	cantar	VIF1P00
	cantaredes	cantar	VIF2P00
	cantarán	cantar	VIF3P00
Pospretérito	cantaría	cantar	VIC1S00
	cantarías	cantar	VIC2S00
	cantaría	cantar	VIC3S00
	cantariamos	cantar	VIC1P00
	cantariades	cantar	VIC2P00
	cantarían	cantar	VIC3P00
Antepretérito	cantara	cantar	VIA1S00
	cantaras	cantar	VIA2S00
	cantara	cantar	VIA3S00
	cantaramos	cantar	VIA1P00
	cantarades	cantar	VIA2P00
	cantaran	cantar	VIA3P00
Pres. Subx.	cante	cantar	VSP1S00
	cantes	cantar	VSP2S00
	cante	cantar	VSP3S00
	cantemos	cantar	VSP1P00
	cantedes	cantar	VSP2P00
	canten	cantar	VSP3P00
Pret. Subx.	cantase	cantar	VSI1S00
	cantases	cantar	VSI2S00
	cantase	cantar	VSI3S00
	cantásemos	cantar	VSI1P00
	cantásedes	cantar	VSI2P00
	cantasen	cantar	VSI3P00
Fut. Subx.	cantar	cantar	VSF1S00
	cantares	cantar	VSF2S00
	cantar	cantar	VSF3S00
	cantarmos	cantar	VSF1P00
	cantardes	cantar	VSF2P00
	cantaren	cantar	VSF3P00
Imperativo	canta	cantar	VM02S00
	cante	cantar	VM03S00
	cantemos	cantar	VM01P00
	cantade	cantar	VM02P00
	canten	cantar	VM03P00
Infinitivo	cantar	cantar	VN00000
Xerundio	cantando	cantar	VX00000
	cantandiño	cantar	VX0000A
Participio	cantada	cantar	VP00SF0
	cantado	cantar	VP00SM0
	cantadas	cantar	VP00PF0
	cantados	cantar	VP00PM0
	cantadiña	cantar	VP00SFA
	cantadiño	cantar	VP00SMA
	cantadiñas	cantar	VP00PFA
	cantadiños	cantar	VP00PMA
Inf. conxugado	cantar	cantar	VN00000
	cantares	cantar	VN02S00
	cantar	cantar	VN00000
	cantarmos	cantar	VN01P00
	cantardes	cantar	VN02P00
	cantaren	cantar	VN03P00
Xer. conxugado	cantándomos	cantar	VX01P00
	cantándodes	cantar	VX02P00

Táboa 6: Exemplos de verbos

2.2.4. Adverbios

A indicación de **R** no CTAG serve para etiquetar tanto os adverbios coma as locucións adverbiais. Por outra banda, os adverbios rematados en *-mente*, derivados de adxectivos, manteñen como lema a súa forma derivada.

ADVERBIOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adverbio	R
2	Grao	Apreciativo	A

Táboa 7: Etiquetas para adverbios

Forma	Lema	Etiqueta
xa	xa	R0
hoxe	hoxe	R0
sempre	sempre	R0
tecnoloxicamente	tecnoloxicamente	R0
ambientalmente	ambientalmente	R0
de_acordo	de_acordo	R0
ao_chou	ao_chou	R0
non	non	R0
loguiño	logo	RA
a_modiño	a_modiño	RA

Táboa 8: Exemplos de adverbios

2.2.5. Numerais

NUMERAIS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Numeral	M
2	Tipo	Cardinal	C
		Ordinal	O
		Partitivo	P
3	Grao	Apreciativo	A
4	Xénero	Masculino	M
		Feminino	F
		Común	C
5	Número	Singular	S
		Plural	P
		Invariable	N

Táboa 9: Etiquetas para numerais

A diferenza da proposta de Civit (2003), na que os numerais se inclúen entre os determinantes, no etiquetario do CTAG aparecen como unha categoría de seu, consonte coa tradición gramatical galega e coas recomendacións de EAGLES (Leech e Wilson, 1996). Como no caso dos adxectivos e do resto das categorías posuidoras de flexión, para os numerais con xénero e número morfoloxicamente marcado, o lema indicado no CTAG vai ser a forma masculina singular.

Forma	Lema	Etiqueta
un	un	MC0MN
unha	un	MC0FN
tres	tres	MC0CN
primeiras	primeiro	MO0FP
primeira	primeiro	MO0FS
primeiros	primeiro	MO0MP
primeiro	primeiro	MO0MS
primeiriñas	primeiro	MOAFP
primeiriña	primeiro	MOAFS
primeiriños	primeiro	MOAMP
primeiriño	primeiro	MOAMS
medio	medio	MP0MS
media	medio	MP0FS

Táboa 10: Exemplos de numerais

2.2.6. Determinantes

No etiquetario do CTAG, só se inclúen na categoría dos determinantes as formas do artigo definido. A categoría de artigo indeterminado (*un*) trátase dentro da dos pronomes indefinidos. Tampouco se inclúen entre os determinantes os demostrativos, posesivos, indefinidos, relativos, exclamativos ou interrogativos, sendo todos eles tratados como categorías independentes.

DETERMINANTES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Artigo	G
2	Xénero	Masculino	M
		Feminino	F
		Común	C
3	Número	Singular	S
		Plural	P

Táboa 11: Etiquetas para determinantes

Forma	Lema	Etiqueta
o	o	GMS
os	o	GMP
a	o	GFS
as	o	GFP
@s	o	GCP

Táboa 12: Exemplos de determinantes

2.2.7. Pronomes

Malia que Civit (2003) inclúe nesta categoría os pronomes demostrativos, posesivos, indefinidos, relativos, interrogativos, exclamativos e numerais, na etiquetaxe do CTAG todas estas categorías considéranse categorías independentes, resérvandose a categoría pronominal do etiquetario para os denominados tradicionalmente pronomes persoais. Na anotación do CTAG, o atributo de cortesía, marcado con valor **P**, especifícase soamente para as formas *vostede* e *vostedes*.

PRONOMES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Pronome	P
2	Persoa	Primeira	1
		Segunda	2
		Terceira	3
3	Xénero	Masculino	M
		Feminino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N
5	Caso	Nominativo	N
		Nom/Recto	B
		Acusativo	A
		Dativo	D
		Oblicuo	O
		Acus/Dat/Reflex	C
		Reflexivo	R
6	Cortesía	Cortés	P

Táboa 13: Etiquetas para pronomes

Forma	**Lema**	**Etiqueta**
eu	eu	P1CSN0
min	min	P1CSO0
nós	nós	P1CPB0
nosoutros	nosoutros	P1MPB0
nos	nos	P1CPC0
te	te	P2CSA0
che	che	P2CSD0
ti	ti	P2CSB0
vostede	vostede	P3CSBP
vostedes	vostede	P3CPBP
vós	vós	P2CPB0
vos	vos	P2CPC0
el	el	P3MSB0
ela	el	P3FSB0
elas	el	P3FPB0
eles	el	P3MPB0
a	o	P3FSA0
as	o	P3FPA0
o	o	P3MSA0
os	o	P3MPA0
lle	lle	P3CSD0
lles	lle	P3CPD0
se	se	P3CNR0
si	si	P3CNO0

Táboa 14: Exemplos de pronomes

2.2.8. Posesivos

No CTAG, o atributo de *posuidor* utilízase cos pronomes posesivos para marcar o número do posuidor: singular para *meu* e *teu*, plural para *noso* e *voso*. Os pronomes en que o posuidor é unha terceira persoa (*seu*) reciben como valor **0** para este atributo, dada a dificultade de distinguir o seu número gramatical singular ou plural, isto é, se fai referencia a *el/ela* ou a *eles/elas*.

POSESIVOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Posesivo	X
2	Persoa	Primeira	1
		Segunda	2
		Terceira	3
3	Xénero	Masculino	M
		Feminino	F
4	Número	Singular	S
		Plural	P
5	Posuidor	Singular	S
		Plural	P

Táboa 15: Etiquetas para posesivos

Forma	**Lema**	**Etiqueta**
miña	meu	X1FSS
miñas	meu	X1FPS
meus	meu	X1MPS
meu	meu	X1MSS
nosa	noso	X1FSP
nosas	noso	X1FPP
noso	noso	X1MSP
nosos	noso	X1MPP
súa	seu	X3FS0
súas	seu	X3FP0
seu	seu	X3MS0
seus	seu	X3MP0
túa	teu	X2FSS
túas	teu	X2FPS
teu	teu	X2FSS
teus	teu	X2MPS
vosa	voso	X2FSP
vosas	voso	X2FPP
voso	voso	X2MSP
vosos	voso	X2MPP

Táboa 16: Exemplos de posesivos

2.2.9. Demostrativos

Forma	**Lema**	**Etiqueta**
aquelas	aquel	DFP
aquela	aquel	DFS
aqueles	aquel	DMP
aquel	aquel	DMS
aquilo	aquel	DNS
esas	ese	DFP
esa	ese	DFS
eses	ese	DMP
ese	ese	DMS
iso	ese	DNS
estas	este	DFP
esta	este	DFS
estes	este	DMP
este	este	DMS
isto	este	DNS

Táboa 17: Exemplos de demostrativos

DEMOSTRATIVOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Demostrativo	D
2	Xénero	Masculino	M
		Feminino	F
		Neutro	N
3	Número	Singular	S
		Plural	P

Táboa 18: Etiquetas para demostrativos

2.2.10. Interrogativos

INTERROGATIVOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Interrogativo	T
2	Xénero	Masculino	M
		Feminino	F
		Común	C
3	Número	Singular	S
		Plural	P
		Invariable	N
4	Grao	Apreciativo	A

Táboa 19: Etiquetas para interrogativos

Forma	**Lema**	**Etiqueta**
cal	cal	TCS0
cales	cal	TCP0
que	que	TCN0
canto	canto	TMS0
cantos	canto	TMP0
canta	canto	TFS0
cantas	canto	TFP0
cantiño	canto	TMSA

Táboa 20: Exemplos de interrogativos

2.2.11. Relativos

RELATIVOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Relativo	Q
2	Xénero	Masculino	M
		Feminino	F
		Común	C
3	Número	Singular	S
		Plural	P
		Invariable	N
4	Grao	Apreciativo	A

Táboa 21: Etiquetas para relativos

Forma	**Lema**	**Etiqueta**
cal	cal	QCS0
cales	cal	QCP0
canta	canto	QFS0
cantas	canto	QFP0
canto	canto	QMS0
cantos	canto	QMS0
cantiño	canto	QMSA
que	que	QCN0

Táboa 22: Exemplos de relativos

2.2.12. Indefinidos

Con esta categoría etiquétanse tamén no CTAG os artigos indeterminados (*un*), alén dos catalogados tradicionalmente como pronomes indefinidos.

INDEFINIDOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Indefinido	I
2	Xénero	Masculino	M
		Feminino	F
		Neutro	N
3	Número	Singular	S
		Plural	P
4	Grao	Apreciativo	A

Táboa 23: Etiquetas para indefinidos

Forma	**Lema**	**Etiqueta**
algo	algo	IMS0
alguén	alguén	IMS0
algunha	algún	IFS0
algunhas	algún	IFP0
algún	algún	IMS0
algúns	algún	IMP0
calquera	calquera	INS0
mesma	mesmo	IFS0
mesmas	mesmo	IFP0
mesmo	mesmo	IMS0
mesmos	mesmo	IMP0
mesmiño	mesmo	IMSA
mesmiña	mesmo	IFSA
mesmiñas	mesmo	IFPA
nada	nada	IMS0
nadiña	nada	IMSA
ninguén	ninguén	INS0
ningunha	ningún	IFS0
ningunhas	ningún	IFP0
ningún	ningún	IMS0
ningúns	ningún	IMP0
pouca	pouco	IFS0
poucas	pouco	IFP0
pouco	pouco	IMS0
poucos	pouco	IMP0
pouquiño	pouco	IMSA
unha	un	IFS0
unhas	un	IFP0
un	un	IMS0
uns	un	IMP0
varias	varios	IFP0
varios	varios	IMP0

Táboa 24: Exemplos de indefinidos

2.2.13. Preposicións

PREPOSICIÓNS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Preposición	S

Táboa 25: Etiquetas para preposicións

Forma	Lema	Etiqueta
a	a	S
de	de	S
ante	ante	S
baixo	baixo	S
con	con	S
cara_a	cara_a	S

Táboa 26: Exemplos de preposicións

2.2.14. Conxuncións

CONXUNCIÓNS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Conxunción	C
2	Tipo	Coordinativa	C
		Subordinativa	S

Táboa 27: Etiquetas para conxuncións

Forma	Lema	Etiqueta
e	e	CC
e_mais	e_mais	CC
nin	nin	CC
ou	ou	CC
pero	pero	CC
senón	senón	CC
aínda_que	aínda_que	CS
porque	porque	CS
pois	pois	CS
que	que	CS
se	se	CS
xa_que_logo	xa_que_logo	CS

Táboa 28: Exemplos de conxuncións

2.2.15. Interxeccións

INTERXECCIÓNS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Interxección	O

Táboa 29: Etiquetas para interxeccións

Forma	Lema	Etiqueta
ou	ou	O
xe	xe	O
bo	bo	O
vaites	vaites	O

Táboa 30: Exemplos de interxeccións

2.2.16. Puntuación

A respecto da puntuación, o etiquetario do CTAG segue o utilizado no FreeLing (Atserias et al., 2006), baseado en Civit (2003).

PUNTUACIÓN			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Puntuación	F

Táboa 31: Etiquetas para puntuación

Forma	Lema	Etiqueta
¡	¡	Faa
!	!	Fat
,	,	Fc
[	[	Fca
]	]	Fct
:	:	Fd
"	"	Fe
-	-	Fg
¿	¿	Fia
?	?	Fit
{	{	Fla
}	}	Flt
.	.	Fp
(	(	Fpa
)	)	Fpt
«	«	Fra
»	»	Frc
...	...	Fs
%	%	Ft
;	;	Fx
_	_	Fz
+	+	Fz
=	=	Fz

Táboa 32: Exemplos de puntuación

2.2.17. Cifras

As cifras etiquétanse no CTAG co código **Z**. Con esta categoría, abránguense anos, enderezos, números de teléfono, etc.

CIFRAS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Cifra	Z

Táboa 33: Etiquetas para cifras

Forma	Lema	Etiqueta
10'2	10'2	Z
1.998	1.998	Z

Táboa 34: Exemplos de cifras

2.2.18. Abreviaturas

No CTAG emprégase a etiqueta **Y** para as abreviaturas de resolución incerta e tamén para os enderezos electrónicos e indicacións de unidades de temperatura (^oC) e outras. Porén, etiquétanse como nomes propios formas como M^a ou siglas que corresponden a entidades propias e individualizadas, como *CEE* ou *EEUU*; como nomes comúns formas abreviadas como n^o ou *ex.* ou siglas do tipo *SA*, *SP* ou *PEMES* (sic); e como numerais ordinais as abreviaturas como 1^o ou 3^a.

ABREVIATURAS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Abreviatura	Y

Táboa 35: Etiquetas para abreviaturas

Forma	Lema	Etiqueta
ºC	graos Celsius	Y
sli.uvigo.es	sli.uvigo.es	Y
Mª	Mª	NP000
1º	1º	NO0MS
S.A.	S.A.	NCFS0
PEMES	PEMES	NCFP0

Táboa 36: Exemplos de abreviaturas

2.2.19. Símbolos

Inclúense na categoría dos símbolos todas as formas abreviadas que representan símbolos químicos da táboa periódica e formas compostas por eles. O lema vai coincidir coa forma plena estándar que corresponde a cada símbolo.

SÍMBOLOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Símbolo	L

Táboa 37: Etiquetas para símbolos

Forma	Lema	Etiqueta
Fe	ferro	L
Ni	níquel	L
O	osíxeno	L
ClH	ácido clorhídrico	L

Táboa 38: Exemplos de símbolos

2.2.20. Estranxeirismos

Todos os estranxeirismos pertencentes a calquera lingua distinta do galego etiquétanse no CTAG como **E**, sen especificar o idioma de orixe.

ESTRANXEIRISMOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Estranxeirismo	E

Táboa 39: Etiquetas para estranxeirismos

Forma	Lema	Etiqueta
monsieur	monsieur	E
and	and	E

Táboa 40: Exemplos de estranxeirismos

2.2.21. Palabras non clasificadas

As formas que resultan descoñecidas ou de difícil clasificación codifícanse no CTAG coa etiqueta **U**.

NON CLASIFICADAS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Non clasificada	U

Táboa 41: Etiquetas para palabras non clasificadas

Forma	Lema	Etiqueta
R50	R50	U
LOUFungi	LOUFungi	U

Táboa 42: Exemplos de palabras non clasificadas

2.3. Contraccións e enclises

O galego ofrece moitas posibilidades de contraccións, por iso cómpre precisión á hora de describilas, tendo en conta as especificacións de cada un dos seus compoñentes.

O sistema de anotación do CTAG equipara formalmente a codificación dos diversos casos onde se produce a unión de dúas ou máis formas como ocorre, por exemplo, na segunda forma do artigo, na enclise dos pronomes átonos, nas contraccións propias das preposicións, ou na contracción con artigo da conxunción comparativa *ca*.

Dun modo xeral, o método de codificación das contraccións e enclises no CTAG é o seguinte: se *F* é unha forma contracta ou enclítica formada pola unión das palabras *P1+P2+...+Pn*, sendo *L1,L2...Ln* os lemas das palabras compoñentes e *C1,C2...Cn* as súas etiquetas categoriais, a forma codificada xenérica da forma contracta sería *F/L1_ C1 ∼/L2_ C2 ... ∼/Ln_ Cn*, como se ilustra a seguir na etiquetaxe das formas contractas *facelas*, *nesoutra* e *entrámbolos*:

- facelas/facer_VN0000 ∼/o_P3FPA0
- nesoutra/en_S ∼/ese_DFS ∼/outro_IFS0
- entrámbolos/entre_S ∼/ambos_IMP0 ∼/o_GMP

Xa que logo, as formas contractas e enclíticas están analizadas no CTAG como secuencias de palabras aptas para a posterior análise sintáctica. O til (∼) indica que a forma fónica da palabra está subsumida na contracción anterior.

A mesma codificación aplícase coherentemente ás enclises da segunda forma do artigo determinado, como se amosa nos seguintes exemplos:

- face-lo/facer_VN0000 ∼/o_GMS
- perdíche-los/perder_VIS2S00 ∼/o_GMP
- collémo-la/coller_VIP1P00 ∼/o_GFS
- protexe-las/protexer_VN0000 ∼/o_GFP
- tódolos/todo_IMP0 ∼/o_GMP
- mailas/mais_CC ∼/o_GFP
- nó-los/nós_P1CPB0 ∼/o_GMP

A mesma codificación aplícase aos pronomes enclíticos, mesmo cando estes van seguidos dunha segunda forma do artigo determinado:

- lévannos/levar_VIP3P0 ∼/nos_P10PC0
- permítenlles/permitir_VIP3P0 ∼/lle_P30PD0

- débellelo/deber_VIP3S0 ∼/lle_P30PD0 ∼/o_P3MSA0
- dóuvo-la/dar_VIP1S00 ∼/vos_P20PC0 ∼/o_GFS
- quitóulle-las/quitar_VIS3S00 ∼/lle_P30PD0 ∼/o_GFP

O mesmo método de anotación utilízase tamén para as unións dos pronomes en dativo co acusativo de terceira persoa:

- cho/che_P2CSD0 ∼/o_P3MSA0
- nola/nos_P10PC0 ∼/o_P3FSA0
- lla/lle_P30SD0 ∼/o_P3FSA0
- llela/lle_P30PD0 ∼/o_P3FSA0

Tamén nos diversos casos de contracción de preposicións con artigos determinados e indeterminados, demostrativos, pronomes persoais, indefinidos, etc.:

- das/de_S ∼/o_GFP
- polos/por_S ∼/o_GMP
- coa/con_S ∼/o_GFS
- no/en_S ∼/o_GMS
- cara á/cara a_S ∼/o_GFS

E tamén nos casos de contracción da conxunción comparativa *ca* coas diferentes formas do artigo determinado:

- cá/ca_CS ∼/o_GFS
- cás/ca_CS ∼/o_GFP
- có/ca_CS ∼/o_GMS
- cós/ca_CS ∼/o_GMP

Isto é, o CTAG utiliza un método uniforme analítico para a etiquetaxe de toda a ampla variedade de formas enclíticas e contractas do galego.

2.4. Problemas normativos

Non é infrecuente atopar nos textos do corpus CTAG exemplos de palabras que non se adaptan á normativa ortográfica oficial para o galego, vixente desde o ano 2003 (Real Academia Galega e Instituto da Lingua Galega, 2003). Nalgúns casos, as formas non normativas identificadas son froito do descoñecemento da norma ou do *lapsus calami*; mais noutros casos trátase de formas documentadas en textos escritos en datas anteriores á reforma normativa de 2003, correctas na normativa vixente no momento en que foron escritas; ou mesmo de formas pertencentes a normativas distintas da oficial. En todos estes casos, a anotación do CTAG inclúe, ao carón da forma non normativa documentada, a forma normativa da palabra precedida do símbolo '#'. Doutra banda, cando esta corrección implica un cambio categorial na forma non normativa documentada, a anotación inclúe tamén a etiqueta morfolóxica da forma incorrecta precedida do símbolo '|'. Véxase a aplicación destas convencións na etiquetaxe do corpus CTAG nos seguintes exemplos:

- presencia/presencia#presenza_NCFS0
- productos/producto#produto_NCMP0
- efeitos/efeito#efecto_NCMP0
- meio/meio#medio_NCMS0
- desbroce/desbroce|NCMS0#roza_NCFS0
- aporte/aporte|NCMS0#achega_NCFS0
- fango/fango|NCMS0#lama_NCFS0
- promedio/promedio|NCMS0#media_NCFS0
- llo/llo|lle_P30SD0#lle_P30PD0 ∼/o_P3MSA0 (*llo* por *llelo*)

Outra causa frecuente de conflito coa normativa provén dos recursos gráficos utilizados en relación co uso non sexista da linguaxe. A efectos da etiquetación do corpus, as arrobas e as formas alternativas con barra inclinada tipo que se documentan en exemplos como *@s europe@s*, *o/a consumidor/a* ou *os/as destinatarios/as* son tratatadas como grafías que indican un xénero común do lema (inexistente na súa morfoloxía), xénero que se recolle para cada caso na etiqueta correspondente, como se pode observar nos seguintes exemplos:

- @s/o_GCP
- europe@s/europeo_NCCP0
- o\a/o_GCS consumidor\a/consumidor_NCCS0
- destinatarios\as/destinatario_AOCP

3. *Fragmentos ilustrativos*

Seguen algúns fragmentos ilustrativos dos principios metodolóxicos expostos neste artigo tirados do Corpus CTAG.

```
<frase>^A/O_GFS ^expansión/expansión_NCFS0 ^dos/de_S ^∼/o_GMP ^cultivos/cultivo_NCMP0 ^transxénicos/transxénico_AOMP ^ameaza/ameazar_VIP3S00 ^a/o_GFS ^diversidade/diversidade_NCFS0 ^xenética/xenético_AOFS ^pola/por_S ^∼/o_GFS ^simplificación/simplificación_NCFS0 ^dos/de_S ^∼/o_GMP ^sistemas/sistema_NCMP0 ^de/de_S ^cultivos/cultivo_NCMP0 ^e/e_CC ^a/o_GFS ^promoción/promoción_NCFS0 ^da/de_S ^∼/o_GFS ^erosión/erosión_NCFS0 ^xenética/xenético_AOFS ^./._Fp </frase>
```

Méndez, Lucía, "Queres comer alimentos transxénicos?". *Terra: Boletín da Federación Ecoloxista Galega*, 4, 1999.

```
<frase>^Por exemplo/Por exemplo_R0 ^non/non_R0 ^podemos/poder_VIP1P00 ^dicir/dicir_VN00000 ^que/que_CS ^Galicia/Galicia_NP000 ^sexa/ser_VSP3S00 ^moi/moi_R0 ^diversa/diverso_AOFS ^en/en_S ^aves/ave_NCFP0 ^,/,_Fc ^lévanse/levar_VIC3P00 ^∼/se_P3CNR0
```

```
^rexistradas/rexistrar_VP00PF0 ^unhas/un_IFP0
^250/250_Z ^habituais/habitual_A0CP ^ó/a_S
^∼/o_GMS ^longo/longo_A0MS ^dun/de_S ^∼/un_IMS0
^ano/ano_NCMS0 ^como moito/como moito_R0 ^,/,_Fc
^mentres que/mentres que_CS ^en/en_S ^toda/todo_IFS0
^Europa/Europa_NP000 ^hai/hai_VIP3S00 ^un-
has/un_IFP0 ^500/500_Z ^especies/especie_NCFP0
^e/e_CC ^en/en_S ^países/país_NCMP0 ^como/como_CS
^Perú/Perú_NP000 ^a/o_GFS ^cifra/cifra_NCFS0
^ascende/ascender_VIP3S00 ^a/a_S ^máis/máis_R0
^de/de_S ^1600/1600_Z ^para/para_S ^un/un_IMS0
^total/total_NCMS0 ^de/de_S ^9000/9000_Z
^aves/ave_NCFP0 ^de/de_S ^diferentes/diferente_A0CP
^especies/especie_NCFP0 ^existentes/existente_A0CP
^no/en_S ^∼/o_GMS ^planeta/planeta_NCMS0 ^./._Fp
</frase>
```

Vázquez Pumariño, Xabier, *Que é a biodiversidade.* Documento electrónico dispoñible na web da Asociación para a Defensa Ecolóxica de Galiza (ADEGA).

```
<frase>^Galicia/Galicia_NP000 ^é/ser_VIP3S00
^a/o_GFS ^primeira/primeiro_M00FS ^Comu-
nidade/Comunidade_NCFS0 ^Autónoma/Autónomo_A0FS
^pesqueira/pesqueira_A0FS ^do/de_S ^∼/o_GMS
^Estado/estado_NCMS0 ^español/español_A0MS
^,/,_Fc ^o/o_GMS ^sector/sector_NCMS0
^pesqueiro/pesqueiro_A0MS ^represen-
ta/representar_VIP3S00 ^o/o_GMS ^8/8_Z ^%/%_Ft
^do/de_S ^∼/o_GMS ^PIB/PIB_NCMS0 ^e/e_CC
^o/o_GMS ^5/5_Z ^%/%_Ft ^da/de_S ^∼/o_GFS
^poboación/poboación_NCFS0 ^activa/activo_A0FS
^,/,_Fc ^estas/este_DFP ^cifras/cifra_NCFP0 ^a
pesar de/a pesar de_CS ^estar/estar_VN00000
^en consonancia/en consonancia_R0 ^coa/con_S
^∼/o_GFS ^importancia/importancia_NCFS0
^do/de_S ^∼/o_GMS ^litoral/litoral_A0CS
^a/a_S ^nivel/nivel_NCMS0 ^mundial/mundial_A0CS
^,/,_Fc ^o/o_GMS ^40/40_Z ^%/%_Ft ^da/de_S
^∼/o_GFS0 ^poboación/poboación_NCFS0 ^do/de_S
^∼/o_GMS ^mundo/mundo_NCMS0 ^vive/vivir_VIP3S00
^nas/en_S ^∼/o_GFP ^zonas/zona_NCFP0
^costeiras/costeiro_A0FP ^,/,_Fc ^pre-
senta/presentar_VIP3S00 ^unhas/un_IFP0
^cifras/cifra_NCFP0 ^moi/moi_R0 ^por/por_S ^en-
riba/enriba_R0 ^de/de_S ^calquera/calquera_INS0
^dos/de_S ^∼/o_GMP ^outros/outro_IMP0 ^paí-
ses/país_NCMP0 ^comunitarios/comunitario_A0MP
^./._Fp </frase>
```

López Fernández, Alfredo, *Estatus dos pequenos cetáceos da plataforma de Galicia.* Tese de doutoramento, Universidade de Santiago de Compostela, 2003.

4. Conclusións

Neste artigo presentamos as bases para a anotación lingüística (etiquetaxe categorial e lematización) do Corpus CTAG (Corpus Técnico Anotado do Galego) da Universidade de Vigo. Aínda que se trata dun proxecto en curso, algúns dos seus resultados xa se poden consultar libremente en Internet (Gómez Guinovart, 2006-2009) mediante unha interface web de consulta accesible en <http://sli.uvigo.es/CTAG/> que dá acceso a unha sección do corpus de máis de 2 millóns de palabras, constituída por textos pertencentes aos eidos da ecoloxía e das ciencias ambientais. Ao remate do proxecto, está prevista a dispoñibilización en Internet do resultado da anotación morfosintáctica da totalidade do Corpus Técnico do Galego.

Referencias

Alegría Loinaz, Iñaki, Iñaki Arantzabal, Mikel L. Forcada, Xavier Gómez Guinovart, Lluis Padró, José Ramom Pichel Campos, e Josu Waliño. 2006. Opentrad: Traducción automática de código abierto para las lenguas del estado español. *Procesamiento del Lenguaje Natural*, 37:357–358.

Armentano Oller, Carme, Rafael C. Carrasco, Antonio M. Corbí Bellot, Mikel L. Forcada, Mireia Ginestí Rosell, Sergio Ortiz Rojas, Juan Antonio Pérez Ortiz, Gema Ramírez Sánchez, Felipe Sánchez Martínez, e Miriam A. Scalco. 2006. Open-source portuguese-spanish machine translation. En *Lecture Notes in Computer Science 3960 (Computational Processing of the Portuguese Language, Proceedings of the 7th International Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR 2006)*, páxinas 50–59, Itatiaia, Rio de Janeiro.

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

Civit, Montserrat. 2003. *Criterios de etiquetación y desambiguación morfosintáctica de corpus en español.* SEPLN (Colección Monografías, 3), Alacante.

Gómez Clemente, Xosé María e Xavier Gómez Guinovart, editores. 2006-2009. *Corpus Técnico do Galego.* Universidade de Vigo, Vigo. <http://sli.uvigo.es/CTG/>.

Gómez Guinovart, Xavier, editor. 2006-2009. *Corpus Técnico Anotado do Galego.* Universidade de Vigo, Vigo. <http://sli.uvigo.es/CTAG/>.

Leech, Geoffrey e Andrew Wilson. 1996. Recommendations for the morphosyntactic annotation of corpora. Eagles guidelines. <http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html>.

Real Academia Galega e Instituto da Lingua Galega. 2003. *Normas ortográficas e morfolóxicas do idioma galego.* RAG/ILG, Santiago de Compostela.