Universal Dependencies Guidelines for the Galician-TreeGal Treebank

Marcos Garcia

LyS Group
Departamento de Galego-Portugués, Francés e Lingüística
Facultade de Filoloxía, Universidade da Coruña
marcos.garcia.gonzalez@udc.gal

v0.4 (November 15, 2016)

1 Introduction

The present document is a technical report that describes the annotation guidelines of the Galician-TreeGal treebank, a corpus for Galician language labeled according to the Universal Dependencies (UD) in its 1.4 version. As the treebank is not a finished project, this document is an ongoing work which can be eventually updated with new UD guidelines, with a larger version of the corpus, or with new examples from the treebank.

The UD Galician-TreeGal treebank is based on a subcorpus (called *xeral*) of the XIADA project (version 2.6) [2]. The *xeral* corpus consists of $190,742$ tokens ($7,489$ sentences) from general press news of Galician newspapers. The corpus was automatically converted to the CoNLL-U format, extracting from their original (fine-grained) POS-tags both the UD POS-tags and the morphological features. A preliminary version of this resource was presented in [1].

The current version of the treebank (0.3) consists of a manually reviewed part of $1,000$ sentences ($24,219$ tokens), splitted into *train*, *devel* and *test* sets:

- *train*: 800 sentences, $19,216$ tokens.
- *devel*: 100 sentences, $2,429$ tokens.
- *test*: 100 sentences, $2,574$ tokens.

2 Tokenization

The current version of the Galician-TreeGal treebank keeps the original tokenization of the XIADA corpus (version 2.6), which does not follow the UD guidelines in these two cases:

- Compound nouns:
  - Proper nouns: Enmanuel_Kant
  - Common nouns: tenente_de_alcalde
- Some multiword expressions:
  - Adverbial locutions: hoxe_por_hoxe
  - Conjunctive locutions: por_moito_que
  - Prepositional locutions: a_través_de
  - Pronominal locutions: cada_quen
  - Interjections: meu_Deus
  - Numerals: vinte_e_cinco

Compound proper nouns as well as adverbial, prepositional and conjunctive locutions are the most frequent cases, while the numbers of the other elements are almost marginal. This disagreement in tokenization with the UD guidelines is intended to be solved in further versions of the treebank.

3 Morphology

3.1 General principles

Tokens and lemmas of the Galician-TreeGal treebank were directly obtained from the original XIADA corpus, with the only modification of replacing a blank space by "_" in compound forms ("a tempo" > "a_tempo").

POS-tags and morphological features were automatically converted from the XIADA tagset, but a few of them are not perfectly mapped (e.g, acronyms are NOUNS instead of SYM), and they should be checked individually in further revisions of this resource.

3.2 Galician POS tags

The correspondence between the XIADA POS-tags[1] and the UD ones[2] is explained below (where * is used as a wildcard):

ADJ: adjective

The UD POS-tag ADJ in Galician-TreeGal derives from the A* XIADA POS-tags.

[1] http://corpus.cirp.es/xiada/etiquetario.html

[2] http://universaldependencies.org/u/pos/all.html

ADP: adposition

The Galician-TreeGal ADP was converted from the P and Lp0 (for prepositional locutions) XIADA POS-tags.

ADV: adverb

ADV comes from the W* and La0 (for adverbial locutions) XIADA POS-tags.

AUX: auxiliary verb

The current version of the corpus does not use the AUX POS-tag, so auxiliar verbs are POS-tagged as VERB (and labeled with the *aux* dependency relation).

CONJ: coordinating conjunction

The Galician-TreeGal CONJ derives from the Cc and Lcc (for conjunctive locutions) XIADA POS-tags.

DET: determiner

DET POS-tag is a conversion of the D* (articles), Ed* (demonstratives), Gd* (interrogatives and exclamatives), Ia* and Id* (indefinites), Md* (possessives), and Td* (relatives) XIADA POS-tags.

INTJ: interjection

The Galician-TreeGal INTJ comes from the Y XIADA POS-tag.

NOUN: noun

The NOUN tag was converted from the Sc* (common nouns), Zaf* and Zam* (abbreviations), and Zga*, Zgf, and Zgm* (acronyms) XIADA POS-tags.

NUM: numeral

The Galician-TreeGal NUM tag derives from the N* XIADA POS-tags.

PART: particle

The current version of the treebank does not make use of the PART POS-tag.

PRON: pronoun

The Galician-TreeGal PRON was converted from the En* (demonstrative), Gn* (interrogatives and exclamatives), In* (indefinites), Mn* (possessives), R* (pronouns), Tn* (relatives), and Zaa* (pronominal abbreviations) XIADA POS-tags.

PROPN: proper noun

The POS-tag PROPN comes from the Sp* (proper nouns), and Zg0*, Zgf* and Zgm* (abbreviations) XIADA POS-tags.

PUNCT: punctuation

The PUNCT tag derives from the Q* XIADA POS-tags.

SCONJ: subordinating conjunction

The Galician-TreeGal SCONJ was converted from the Cs and Lcs (for subordinating conjunctive locutions) XIADA POS-tags.

SYM: adjective

The UD POS-tag SYM was converted from the Z0*, Zo* and Zs* XIADA POS-tags.

VERB: verb

The Galician-TreeGal VERB comes from the V* XIADA POS-tags.

X: other

The Galician-TreeGal X POS-tag was converted from the Za00 XIADA POS-tag.

3.3 Galician features

This section includes the list of the UD features and their possible values, together with some examples in the treebank.

Animacy: animacy

The Animacy feature is not used in the Galician-TreeGal corpus.

Aspect: aspect

The Aspect feature is not used in the Galician-TreeGal corpus.

Case: case

The Case feature is used in the Galician-TreeGal corpus to characterize personal pronouns. They can have three different values:

- **Nom**: for nominative pronouns: “*Eles* saben música”.

- **Dat**: for dative pronouns. “A historia demóstra*nos* que…”.
- **Acc**: for accusative pronouns. “Quen *o* alterou…”.

Definite: definiteness or state

The Definite feature is used in the Galician-TreeGal treebank for distinguishing between two types of determiners:

- **Def**: definite determiners: “*A* casa é vermella”.
- **Ind**: indefinite determiners: “aceptan *un* acordo constitucional”.

Degree: degree of comparison

The Degree feature is used to characterize two types of adjectives:

- **Cmp**: comparative adjectives: “Poden ir en *maior* medida”.
- **Sup**: superlative adjectives: “É unha cuestión *importantísima* para a cidade”.

Gender: gender

Four different Gender features are used in the Galician-TreeGal corpus, for characterizing the following categories: ADJ, DET, NOUN, NUM, PRON, PROPN and VERB:

- **Masc**: masculine: “*Estes* son algúns.”.
- **Fem**: feminine: “en *estas imaxes*”.
- **Neut**: neutral: “*isto* non é certo”.
- **Com**: common: “somos *diferentes* e *universais*”.

Mood: mood

The feature Mood is used for classifying four modality types of verbs:

- **Ind**: indicative: “só *dedican* máis tempo nos EUA”
- **Imp**: imperative: “*imaxina* que non estivese traducido”
- **Cnd**: conditional (which is also Ind: Mood=Cnd,Ind): “non *deberiamos* falar de pegadas dactilares”
- **Sub**: subjunctive: “aínda que privadamente algúns *falen* de chantaxe”

Negative: whether the word can be or is negated

The Negative feature is not used in this version of the Galician-TreeGal corpus.

NumType: numeral type

The NumType feature distinguishes between two types of numbers:

- **Card**: cardinal numbers: “configurada en *17* bases”
- **Ord**: ordinal numbers: “reclama o *primeiro* posto”

Number: number

The Number feature is used in the Galician-TreeGal treebank for distinguishing between singular and plural nouns (NOUN and PROPN), adjectives, determiners, pronouns and verb forms:

- **Sing**: singular: “vivir en *liberdade*”
- **Plur**: plural: “equipo de *expertos*”

Person: person

The Person feature is used in our corpus in verb forms (1st for the speaker(s), 2nd for the addressee(s), 3*rd* for other person(s) different than the speaker and the addressee)[3], (possessive) pronouns and determiners.

- **1**: first person: “a *nosa* editorial”
- **2**: second person: “se *ti* vas a un hotel”
- **3**: third person: “dous articulistas *afirmaban*”

Poss: possessive

The Poss feature is used for classifying possessive determiners and pronouns.

- **Yes**: possessive: “os *seus* libros”

[3] The 3*rd* person is also used as a courtesy form to treat the addressee.

PronType: pronominal type

The PronType feature is used in Galician-TreeGal corpus to distinguish between several types of pronouns, nouns, determiners, and adverbs.

- **Prs**: personal and possessive pronouns and determiners (the Poss feature disambiguates betweem personal and possessives): “a valoración da lingua *nosa*”
- **Art**: article: “*a* política exterior”
- **Int**: interrogative: “eles *que* fan?”
- **Rel**: relative: “incluír os terros *onde* está a empresa. . . ”
- **Dem**: demonstrative: “que *isto* o diga alguén”
- **Ind**: indefinite: “*outros* participantes no foro”

Apart from that, we use a language-specific PronType value for clitics (**Clit**): “ás cinco remáta*se* de traballar”.

Reflex: reflexive

The Reflex feature is not used in the Galician-TreeGal corpus.

Tense: tense

Tense is used for specifying when the verb action took (or takes, or will take) place:

- **Past**: past tense: “a dirección *adiou* a elección”
- **Pres**: present tense: “*teñen* a condición de cidadáns”
- **Fut**: future tense: “non *poderá* privarse dela”
- **Imp**: imperfect: “non lle *quedaba* outro remedio”
- **Pqp**: pluperfect: “nunca *vira* a morte tan de perto”

Note that the conditional is classified as Mood instead of as Tense (as it is usually described in grammars of Galician).

VerbForm: form of verb or deverbative

In our corpus, VerbForm is used only in verbs, and it can have the following values:

- **Fin**: finite verb: "*rematou* co sistema socialista"
- **Inf**: infinitive: "vontade de *integrármo*nos na estrutura"
- **Part**: participle: "*decididos* a progresar no camiño"
- **Ger**: gerund: "*afirmando* que vai gañar por maioría"

Galician infinitives can be inflected in number and person, so VerbForm may be combined with Number and Person features.

Voice: voice

The Voice feature is not used in the Galician-TreeGal corpus.

3.3.1 Galician-specific features

AdpType

This is a language-specific feature kept from the original information of the XI-ADA corpus, also used in other languages. In the current version of Galician-TreeGal, it only has one possible value (Prep), so it does not provide new information to the tokens labeled with the ADP POS-tag.

Number[psor]

This language-specific feature (also used in other languages, but not present in the universal tagset) is utilized in possessive determiners and pronouns for specifying the number of possessors, and it has the same values as Number (which in these cases is used for specifying the number of the possessed):

- **Sing**: singular: "o *seu* marco relacional"
- **Plur**: plural: "a *nosa* identidade política"

4 Syntax

4.1 General principles

The Galician-TreeGal treebank was syntactically annotated following the UD recommendations. The guidelines for labeling the corpus follow three main foundations:

1. Use the UD recommendations whenever possible.

2. Use the shortest possible number of language-specific relations.

3. For labeling structures with more than one possibility of analysis, make the corpus coherent with the European Portuguese and Spanish ones (in this particular order).

Taking the above into account, the main properties of the dependency annotation for Galician are the following:

- Pseudo-copulative verbs: Verbs belonging to this class are tagged as *cop* (copulative) when they function as copulas (e.g., “Miguel_Barros permanecerá relegado”).

- Modal, temporal and aspectual verbs: These verbs are considered *aux* (auxiliary) of the verb they depend on (e.g., “debe conducirnos”or “deixa de ser”). Similarly, auxiliary verbs in verbal periphrases are also tagged as *aux* (e.g., “vai gañar”).

- Objects: dative objects are labeled as *iobj* even if they are the only explicit object in the sentence (“a tarefa$_{\text{nsubj}}$ corresponde$_{\text{root}}$ lle$_{\text{expl}}$ a o goberno$_{\text{iobj}}$”), despite the fact that UD recommends to label them as *dobj* (these cases can be automatically converted to *dobj* in case it could be needed).

- Following the UD recommendations, *Reflexive*, *reciprocal* and *expletive* pronouns are labeled as *expl* (expletive). This includes the non-argumental clitic pronouns often described as *dativos de solidariedade e de interese*.

4.2 Specific constructions

This section describes some syntactic structures which are frequent in Galician. As this is an ongoing work, we intend to enlarge it with new information in further versions of this report.

Predicates

The main predicates in Galician are verbs, so they are labeled as *root*, *ccomp*, *xcomp*, *advcl* and *acl*. They can also be labeled as *conj* (in coordination structures), or as *parataxis*.

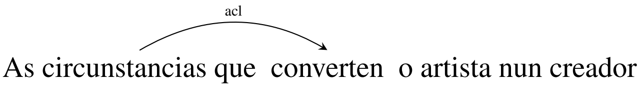
xcomp is also used in predicative complements, so it can label other elements such as adjectives (“que non nos pillen *adormecidos*”) or nouns (“que acabou sendo *secretario*”).

Copulative verbs (*ser*, *estar*, *parecer*, as well as other pseudo-copulative verbs) are treated as copulas, so they depend on the lexical predicates, which are analyzed as *root*: “non só é *lexítimo*”.

4.3 Galician relations

acl: clausal modifier of noun

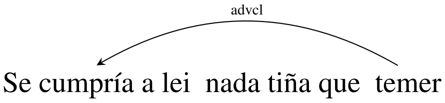
The *acl* relation is used for labeling clauses that modify a noun. Current version of Galician-TreeGal also labels relative clauses as *acl*:



As circunstancias que converten o artista nun creador (1)

advcl: adverbial clause modifier

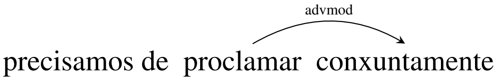
The *advcl* relation is used for clauses that modify a predicate (temporal clauses, causal clauses, purpose clauses, etc.):



Se cumpría a lei nada tiña que temer (2)

advmod: adverbial modifier

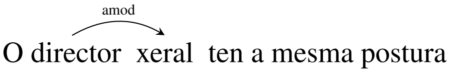
advmod is used for adverbs (non-clausal) which can modify verbs, adjective or nouns:



precisamos de proclamar conxuntamente (3)

amod: adjectival modifier

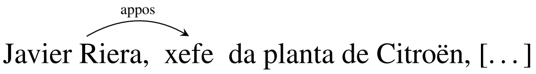
The *amod* relation is used for labeling adjectives that modify a noun:



O director xeral ten a mesma postura (4)

appos: appositional modifier

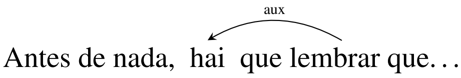
The appositional modifier is a noun that defines or modifies the noun at its left. It is used in parenthetical structures, abbreviations or appositions between commas:



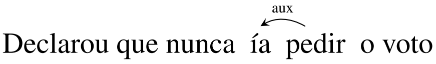
Javier Riera, xefe da planta de Citroën, [. . .] (5)

aux: auxiliary

aux is used for auxiliary verbs (i.e., non-main verb of a clause), which are dependents on the main verb. In the Galician-Lys treebank, *aux* is used for modal, aspectual, and temporal verbs as well as for other structures such as verbal periphrasis:



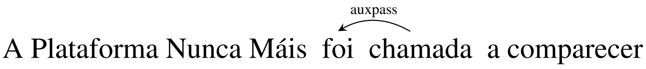
Antes de nada, hai que lembrar que. . . (6)



Declarou que nunca ía pedir o voto (7)

auxpass: passive auxiliary

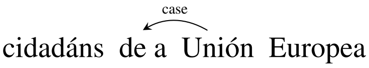
In passive clauses, the auxiliar verb is labeled as *auxpass*:



A Plataforma Nunca Máis foi chamada a comparecer (8)

case: case marking

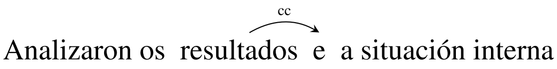
The *case* relation is used for preposition that introduce nominals, where prepositions are analyzed as dependents. Prepositions that introduce clauses are labeled as *mark*:



cidadáns de a Unión Europea (9)

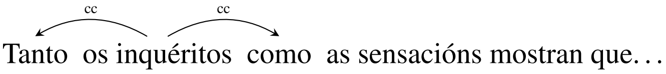
cc: coordination

cc is used for labeling the conjunction of coordination structures, which is a dependent of the first element of the coordination:



Analizaron os resultados e a situación interna (10)

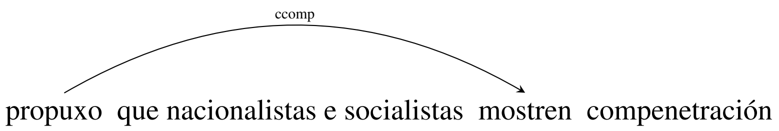
It is also used for adverbs and conjunctions which behave as coordination elements:



Tanto os inquéritos como as sensacións mostran que. . . (11)

ccomp: clausal complement

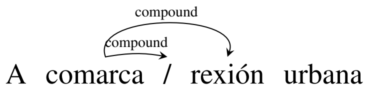
ccomp is a complement clause of a verb or an adjective which functions as an object of its nucleus:



propuxo que nacionalistas e socialistas mostren compenetración (12)

compound: compound

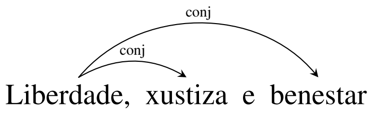
The *compound* relation is used for linking the individual elements of compound nouns. It is used only in few cases in our treebank, since some compound structures are already unified due to tokenization:



A comarca / rexión urbana (13)

conj: conjunct

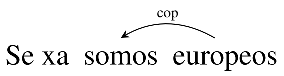
conj is used for linking the elements of a coordinated structure, which are dependents of the first conjunt:



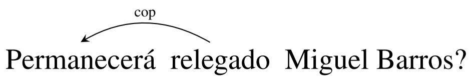
Liberdade, xustiza e benestar (14)

cop: copula

We use *cop* for annotating the relation between a copular verb (the dependent) and its complement (the head). Note that we also use *cop* for labeling pseudo-copulas:

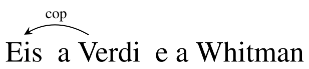


Se xa somos europeos (15)



Permanecerá relegado Miguel Barros? (16)

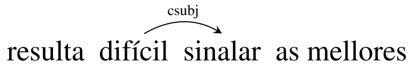
It is also used for sentences with the adverbs *eis* and *u*:



Eis a Verdi e a Whitman (17)

csubj: clausal subject

This label is used for annotating clausal subjects of another clause:

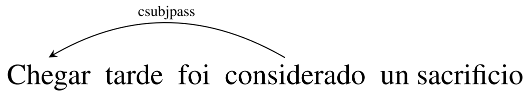


resulta difícil sinalar as mellores

(18)

csubjpass: clausal passive subject

The *csubjpass* relation is used for clausal subjects of passive clauses:

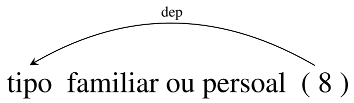


Chegar tarde foi considerado un sacrificio

(19)

dep: dependent

dep is used for linking two elements that do not have a precise syntactic relation, in order to ensure a full parse of the sentence:

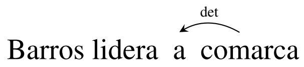


tipo familiar ou persoal (8)

(20)

det: determiner

The *det* relation links the head of a noun phrase to its determiner:

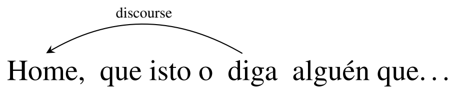


Barros lidera a comarca

(21)

discourse: discourse element

discourse is used for linking interjections and other elements which are not clearly linked to the syntactic structure of the sentence:

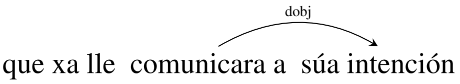


Home, que isto o diga alguén que. . .

(22)

dobj: direct object

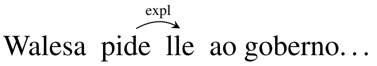
In the Galician-TreeGal treebank, nominal accusative objects of verbs are labeled as *dobj*:



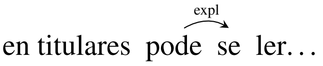
que xa lle comunicara a súa intención (23)

expl: expletive

In our corpus, we use the *expl* relation for labeling clitics that refer to an object already present in the sentence, such as reciprocal pronouns, as well as to reflexive elements:



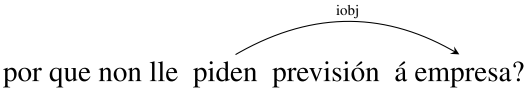
Walesa pide lle ao goberno… (24)



en titulares pode se ler… (25)

iobj: indirect object

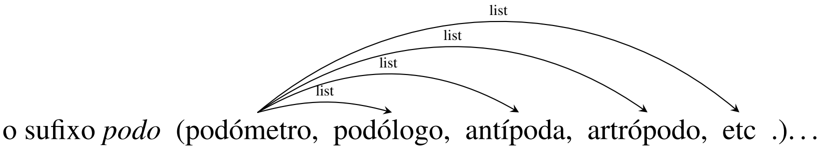
Indirect objects in Galician-TreeGal are the dative objects of the verb (except dative pronouns labeled as *expl*):



por que non lle piden previsión á empresa? (26)

list: list

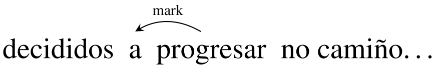
list is used for linking elements of a list that are not in a coordinated structure, the first element being the head and all the others their dependents:



o sufixo *podo* (podómetro, podólogo, antípoda, artrópodo, etc .)… (27)

mark: marker

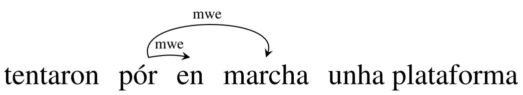
Markers are the words (usually adpositions or conjunctions) that introduce a subordinate clause, and are dependents on the head of the subordinate clause:



decididos a progresar no camiño… (28)

mwe: multi-word expression

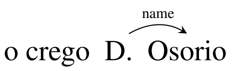
The *mwe* label is used for linking the internal tokens of multi-word expressions, which are fixed expressions that behave as single words. Note, however, that current version of Galician-TreeGal has some multi-word expressions already tokenized



(29)

name: name

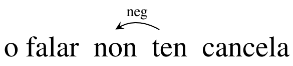
The *name* dependency is used for linking the different tokens of proper nouns with more than on element. As in *mwe*, most of these nouns are already tokenized in the current version of the corpus, so there are few occurrences of *name*:



(30)

neg: negation modifier

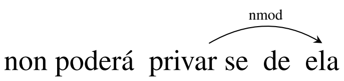
The negation modifier relation, *neg*, links a negation word (usually and adverb) with the word it modifies (the nucleus):



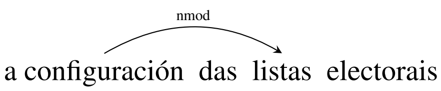
(31)

nmod: nominal modifier

The *nmod* relation is used for nominals that modify nouns or clausal predicates. It is used in Galician-TreeGal corpus to both label arguments and modifiers of clauses, apart from noun modifiers:



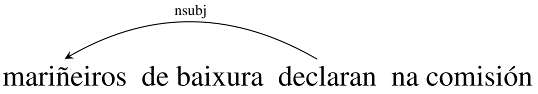
(32)



(33)

nsubj: nominal subject

Nominal subjects (in active mode) are labeled with the *nsubj* relation.

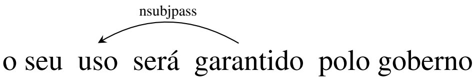


(34)

Note that nominal subjects in passive clauses are labeled as *nsubjpass*, while *csubj* and *csubjpass* are used for clause subjects.

nsubjpass: passive nominal subject

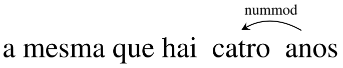
Nominal subjects of passive clauses are labeled as *nsubjpass*:



(35)

nummod: numeric modifier

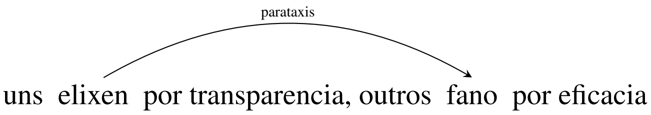
The *nummod* relation is used for linking a numerical expression (as dependent) that modifies a noun with a quantity:



(36)

parataxis: parataxis

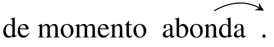
The *parataxis* relation is used in Galician-TreeGal corpus for linking two clauses at the same level, which may occur between *:*, or placed side by side without explicit coordination:



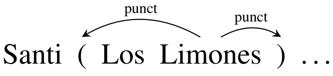
(37)

punct: punctuation

punct is used for linking punctuation elements, usually to the head of the clause, or of the phrase:



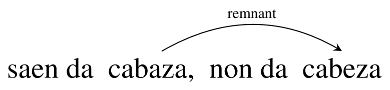
(38)



(39)

remnant: remnant in ellipsis

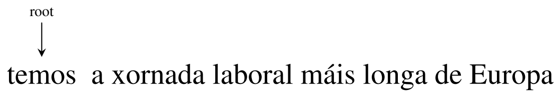
The *remnant* relation is used for treating some cases of ellipsis (namely those where a verb gets elided), without creating an empty node in the representation:



saen da cabaza, non da cabeza (40)

root: root

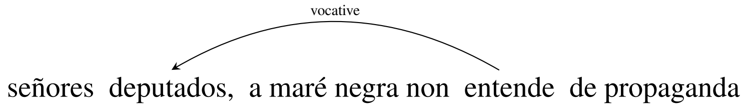
root is the root of the sentence. It is the main verb of the sentence, except if it is a copular verb, whose head will be the root. For non-clausal sentence, *root* is the head element:



temos a xornada laboral máis longa de Europa (41)

vocative: vocative

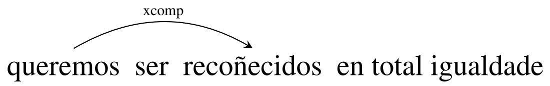
The *vocative* relation is used to label a vocative element in structures like dialogues or direct references to the addressee:



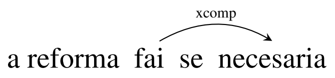
señores deputados, a maré negra non entende de propaganda (42)

xcomp: open clausal complement

The *xcomp* relation is used for annotating clausal complements or predicatives which do not have an own subject. Thus, its subject or object (if any) are often the *nsubj* (or the *dobj*) of the next higher clause:



queremos ser recoñecidos en total igualdade (43)



a reforma fai se necesaria (44)

References

[1] Garcia, Marcos and Carlos Gómez-Rodríguez and Miguel A. Alonso. Creación de un treebank de dependencias universales mediante recursos existentes para lenguas próximas: el caso del gallego. *Procesamiento del Lenguaje Natural*, 57:33–40, 2016.

[2] Guillermo Rojo, Marisol López Martínez, Eva Domínguez Noya, and Fco. Mario Barcala. Corpus de adestramento do Etiquetador/Lematizador do Galego Actual (XIADA), versión 2.6, 2015. `http://corpus.cirp.es/xiada/corpus_xiada_2_6.tar.gz`.