EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Université de Lille 3 SCIENCES HUMAINES ET SOCIALES

UR

ENS DE LYON

DFG

ANR

The common tagset *UD* for cross-linguistic queries in the Latin Subcorpus *PaLaFraLat 2.0*

Version 1.0

Rembert Eufe

April 2018

Content

1 Introduction

The annotation of the two subcorpora with the *lapos* and *frapos* tags, respectively, has been supplemented with tags of an additional tagset called *ud* in order to facilitate interlingual queries and comparisons. This tagset adopts the proposals of the annotation framework *Universal Dependencies*, dedicated to

> “cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective. The annotation scheme is based on an evolution of (universal) Stanford dependencies […], Google universal part-of-speech tags […], and the Interset interlingua for morphosyntactic tagsets […]. The general philosophy is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary.” (http://universaldependencies.org/introduction.html)

Although no syntactic annotation has been carried out in the PaLaFra project (so far), the annotation system of *Universal Dependencies* has proved very useful insofar as its cross-linguistically comparable syntactic annotation presupposes a certain harmonisation of the morphological annotation, which is exactly our concern.[1]

For each token, the *Universal Dependencies* annotation scheme requires a PoS tag filled with one of 17 possible parts of speech, representing a fixed and limited list. Additional morphological information can be indicated as *features*, permitting also to account for language specific categories.

2 The equivalences between *ud-pos* tags and *lapos* parts of speech

2.1 One-to-one correspondences

Six of the 17 *ud-pos* tags are equivalent to parts of speech used in *lapos* (irrespective of some very slight differences in spelling):

POS-Tag *lapos*	**category**	**POS-tag *ud-pos***	**category**
ADJ	adjective	**ADJ**	adjective
ADV	adverb	**ADV**	adverb
AP	adposition	**ADP**	adposition
ITJ	interjection	**INTJ**	interjection
NN	normal noun	**NOUN**	noun
PRO	pronoun	**PRON**	pronoun
V	verb	**VERB**	verb

[1] *Universal Treebanks* contains three different Latin *treebanks* (Latin Dependency Treebank 2.0 of the Perseus Project, the Latin-ITT of the Index Thomisticus and the Latin part of PROIEL (Pragmatic Resources in Old Indo-European Languages). For Modern French, it offers one treebank.

2.2 PoS-distinctions in *lapos* missing in *ud-pos*

In two cases, *lapos* marks distinctions on the PoS level which are not indicated as *ud-pos* tags. Nevertheless, these distinctions are registered as features in Universal Dependencies. They concern the differentiation between cardinal, ordinal and distributive numbers and between personal names opposed to other proper names:

PoS tag *lapos*	category	**PoS tag *ud-pos***	category	feature
NUM	cardinal number	**NUM**	numeral	NumType:Card
ORD	ordinal number	**NUM**	numeral	ud-numtype:Ord
DIST	distributive number	**NUM**	numeral	ud-numtype:Dis
NE	named entity	**PROPN**	proper noun	
NP	personal name	**PROPN**	proper noun	ud-nametype:Prs[2]

2.3 PoS-tag distinctions not available in *lapos*, but provided by *ud-pos*

The opposite case as the one just cited occurs in the form of a distinction in *ud-pos* which is not drawn by the *lapos* tagset, namely the difference between coordinating and subordinating conjunctions:

POS-Tag *lapos*	**category**	**POS-tag *ud-pos***	**category**
CON	conjunction	CCONJ	coordinating conjunction
CON	conjunction	SCONJ	subordinating conjunction

Please note that some Latin conjunctions still await sorting or correction to be corrected, hence they are provisionally marked as XCONJ.

[2] Cf. http://universaldependencies.org/u/feat/NameType.html.

2.4 Intersecting tag distinctions between *ud-pos* and *lapos*

For a series of three tags, the distinctions between the two tagsets overlap, resulting in a chain of non-biunique correspondences: While the *ud-pos* tag X corresponds to FM and some cases of XY in *lapos*, the latter is used not only for some X, but sometimes also for a part of SYM in *ud-pos*. This tag in turn is equivalent in some cases to XY, but in others to PUNCT in *lapos*. The latter again corresponds not only partially to SYM, but also to PUNCT in *ud-pos*:

POS-Tag *lapos*	**category**	**POS-tag *ud-pos***	**category**	**feature**
FM	foreign material	X	other	Foreign:Yes
XY	non word	X	other	
XY	non word	SYM	symbol	
PUNCT	… † § +	SYM	special characters	
PUNCT	. : , ; - ' " ? !	PUNCT	punctuation	

2.5 PoS tags of *ud-pos* not (yet) used for our annotation and consequently without equivalences in *lapos*

Three of the 17 Universal PoS tags do not occur in our annotation, simply because Latin lacks these categories or we had already developed other solutions in the meantime. These are AUX for auxiliaries (although this tag could prove useful for some compound tense forms), DET for determiners (mainly due to the lack of articles in Latin) and PART for particles.[3]

[3] Cf. http://universaldependencies.org/u/pos/PART.html for examples.

3 Features

In both tagsets, further specifications are supplied by a second level of features.

3.1 Pronoun type

***pronountype* (lapos)**	***feature* (ud-pos)**
CORRELATIVE	PronType=Rel
DEMONSTRATIVE	PronType=Dem
INDEFINITE	PronType=Ind
INTENSIVE	PronType=Ind
INTERROGATIVE	PronType=Int
PERSONAL	PronType=Prs
POSSESSIVE	Poss=Yes
REFLEXIVE	Reflex=Yes
RELATIVE	PronType=Rel

The table shows that for all the *lapos* pronountype values equivalents exist in *ud-pos* – with the exception of the correlative pronouns, marked as relative pronouns in *ud-pos*. Since it is possible to enlarge the set of features (in contrast to the fixed list of 17 PoS tags) for the annotation of special languages, it would be possible to introduce PronType=Cor in *ud-pos* in the future.

3.2 Verb type

No equivalents are available in *ud-pos* for the *lapos* feature *verbtype* with its values DEPONENT, IMPERSONAL, INTRANSITIVE, SEMIDEPONENT, TRANSITIVE, VERBA_ANOMALA and VERBA_DEFECTIVA.

3.3 Mood

For all the values of *mood* in *lapos*, equivalents are provided by *ud-pos*:

***mood* (lapos)**	***feature* (ud-pos)**
GERUND	VerbForm=Ger
GERUNDIVE	VerbForm=Gdv
IMPERATIVE	Mood=Imp
INDICATIVE	Mood=Ind
INFINITIVE	VerbForm=Inf
PARTICIPLE	VerbForm=Part
SUBJUNCTIVE	Mood=Sub
SUPINE	VerbForm=sup

ud-pos offers a value *Fin* of the feature *VerbForm* for finite verb forms like those with Mood=Imp, Mood=Ind and Mood=Sub.[4] However, it was not necessary to use it for our annotation.

3.4 Case

Almost all the case form values **(*casus*)** of *lapos* appear also in *ud-pos* – only OBLIQUE is conflated with ACCUSATIVE into *Acc* in *ud-pos*[5] and INDECLINABLE has no equivalent:

***casus* (lapos)**	***feature* (ud-pos)**
ABLATIVE	Case:Abl
ACCUSATIVE	Case:Acc
DATIVE	Case:Dat
GENITIVE	Case:Gen
INDECLINABLE	Ø
LOCATIVE	Case:Loc
NOMINATIVE	Case:Nom
OBLIQUE	Case:Acc

3.5 Gender

The three values for *genus* in *lapos* correspond to the same values in *ud-pos*:

***case* (lapos)**	***feature* (ud-pos)**
MASCULINE	Gender:Masc
FEMININE	Gender:Fem
NEUTER	Gender:Neut

3.6 Number

The two values for *numerus* in *lapos* correspond to the same values in *ud-pos*:

***case* (lapos)**	***feature* (ud-pos)**
SINGULAR	Number:Sing
PLURAL	Number:Plur

[4] Cf. http://universaldependencies.org/u/feat/VerbForm.html#Fin.

[5] Cf. http://universaldependencies.org/u/feat/Case.html#Acc.

3.7 Person

The three values for *person* in *lapos* correspond to the same values in *ud-pos*:

case **(lapos)**	***feature*** **(ud-pos)**
FIRST_PERSON	Person:1
SECOND_PERSON	Person:2
THIRD_PERSON	Person:3

3.8 Declension type

No equivalents are available in *ud-pos* for the *lapos* feature *declensiontype* with its values FIRST_DECLENSION, SECOND_DECLENSION, THIRD_DECLENSION, FOURTH_DECLENSION, FIFTH_DECLENSION, INDECLINABLE and GREEK_DECLENSION.

3.9 Comparison degree

The three values for *comparisondegree* in *lapos* correspond to the same values in *ud-pos*:

case **(lapos)**	***feature*** **(ud-pos)**
COMPARATIVE	Degree:Cmp
POSITIVE	Degree:Pos
SUPERLATIVE	Degree:Sup

3.10 Tense

Some of the *tense* distinctions in *lapos* are indicated by combinations of *Tense* and *Aspect* in *ud-pos*:

case **(lapos)**	***feature*** **(ud-pos)**
FUTURE	Tense:Fut, Aspect:Imp
FUTURE PERFECT	Tense:Fut, Aspect:Perf
IMPERFECT	Tense:Past, Aspect:Imp
PERFECT	Tense:Past, Aspect:Perf
PLUPERFECT	Tense:Pqp
PRESENT	Tense:Pres

3.11 Voice

The two values for *voice* in *lapos* correspond to the same values in *ud-pos*:

case **(lapos)**	***feature*** **(ud-pos)**
ACTIVE	Voice:Act
PASSIVE	Voice:Pass

3.12 Conjugation type

No equivalents are available in *ud-pos* for the *lapos* feature *conjugationtype* with its values FIRST_CONJUGATION, SECOND_CONJUGATION, THIRD_CONJUGATION and FOURTH_CONJUGATION.

3.13 Use

No equivalents are available in *ud-pos* for the *lapos* feature *use* with its values ADJECTIVAL and SUBSTANTIVAL.