Status of morphosyntactic features Illustration with written and spoken French UD treebanks

Sylvain Kahane*△ **Bruno Guillaume**◊ **Léna Brun*** **Simeng Song***
*Modyco, Paris Nanterre Université & CNRS △Institut Universitaire de France
◊Université de Lorraine, CNRS, Inria, LORIA, France

Abstract

Morphosyntactic features used in UD treebanks have different status. If most of them correspond to values of inflectional morphemes, some describe lexical subclasses or are just conventional names of (polysemic) morphemes. Syncretism is also a challenge, because exact values are only deductible from contextual information. We propose an attempt at clarification and an implementation in the treebanks of written and spoken French.

1 Introduction

In Universal Dependencies (UD) annotation scheme for syntactic treebanks, syntax is encoded by relations between words, while morphosyntax is encoded by features on words (de Marneffe et al. 2021). For instance, in Fig.1, the noun *fille* 'lady' has three dependents: the determiner (det) *une* 'an', the adjectival modifier (amod) *jeune* 'young', and the past participle *habillée* (*en noir*) 'dressed (in black)', analyzed as an adjectival clause (acl). Each of the four words bears features indicating their POS (upos), their lemma, as well as morphosyntactic features, such as Gender, Number, Tense, etc.

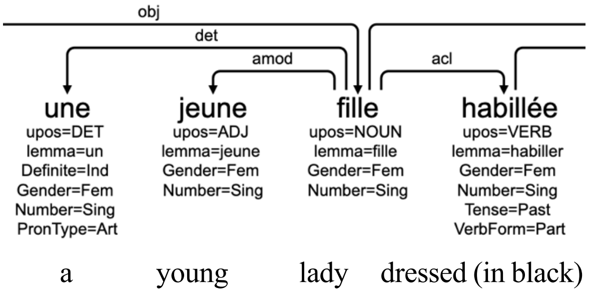

Fig. 1. Extract from UD_French-Rhpsodie@2.16

Morphosyntactic features can have different status. For instance, in French, adjectives agree with nouns in gender (and number): gender on French adjectives is an agreement morpheme, marking the relation with a noun, while gender on nouns is a pure lexical feature, an inherent feature of the lexeme, triggering the agreement of adjectives and determiners (Mel′čuk 1993:261, 2006; Corbett 2022; McCarthy et al. 2018). Moreover, if French adjectives always agree in gender, many of them have a common form for feminine and masculine and it is unclear whether they should bear a Gender feature. This phenomenon, which is already important in written French, become widespread in spoken French, where, for instance plural nouns are written with a *s* at the end, which is phonologically realized only in the rare case of the optional liaison with a following adjective beginning with a vowel (*des femmes* /de fam/ 'women'; *des femme***s** *illustres* /de fam **z** ilystr/ 'famous women').

Syncretism (Corbett 2011) is also a source of numerous discrepancies. When a form corresponds to several values of features, do we annotate the set of values associated to the form or the value that can be inferred from the context? For instance, if the English verbal form *thinks* clearly deserves the features Number=Sing, Person=3, what should be done with *think*? Currently, UD treebanks for English give the values inferred by the context, but these values do not have the same status as the value for *thinks*. As noted by Malaviya et al. (2018), the adjective *refrescante*, which is not inflected for gender in Spanish and Portuguese, has a Gender feature in Portuguese UD treebanks, but not in Spanish ones.

A last problem is posed by the traditional names of inflectional morphemes. UD morphosyntactic features are supposed to be universal features (de Marneffe et al. 2021), that is, comparative concepts, as opposed to language-specific categories (Haspelmath 2011). Romance and Germanic languages have two participles that are traditionally called the present and the past participles. Accordingly, particles in UD treebanks for English or French have a feature Tense=Pres or Tense=Past on every participle (see *habillée* 'dressed' in Fig. 1), while aspectual features, such as Aspect=Imp or Aspect=Perf, would have been much more motivated from the universal point of view. Only the feature Voice=Pass appears on past participles used in passive constructions.

In this paper we propose to distinguish four types of morphosyntactic features: plain features, such as Number, are used for the values of inflectional morphemes of the word, Number[ctxt] is used for the value inferred from the context, Gender[lex] is used for lexical features, and Tense[denom] is used for denominative features. [1]

In the following sections, we start by some examples (Section 2), then we discuss the formalism we use (Section 3). Section 4 presents possible problems of delimitation between lexical and inflectional features and Section 5, between presence and absence of a feature. Sections 6 and 7 are dedicated to the implementation of our annotation in treebanks of written and spoken French. Our conclusion (Section 8) comes back to the relevance of such an annotation for linguistic studies.

2 First examples

In the French example (1), the noun *exercice* 'exercise' is masculine; the indefinite article *un* agrees with it, while the adjective *utile* 'useful', which does not vary in gender, inherits the masculine from the context.

(1)			
	un	*exercice*	*utile*
	an	exercise	useful
	Gender=Masc		Gender[ctxt]=Masc
		Gender[lex]=Masc	

In Russian, nouns vary in case and can have six different singular forms. The noun *žurnal* 'magazine' has the same form in the nominative and accusative cases. We propose to distinguish the case value associated to the form (type-level) from the case value given by the context (token-level):

(2)		
	novyj	*žurnal*
	new	magazine
	Case=Nom	Case=Nom,Acc
		Case[ctxt]=Nom

The English verbal form *arrived* can be a preterit or past participle. In *she has arrived*, the word will receive the following features:

(3) *arrived*
VerbForm[ctxt]=Part, Tense[denom]=Past
Aspect[ctxt]=Perf

They indicate that, in this context, the form is a participle with a perfective aspectual value, which is denominated the *past* participle. Because the past participle has the same form as the preterit, we can indicate that VerbForm is a contextual feature using a feature VerbForm[ctxt], or we can consider that it is a case of homonymy and use the feature VerbForm without extension.

3 Formalization

We use the notation of layered features, which has been introduced for another purpose, when a word has two features of the same type (https://universaldependencies.org/u/overview/feat-layers.html). For instance, auxiliaries in Basque can agree with several arguments: the form *dute* marks the agreement in number and person with an absolutive and an ergative argument:

(4) *dute*
upos=AUX, lemma=edun,
VerbForm=Fin, Mood=Ind,
Number[abs]=Sing, Person[abs]=3
Number[erg]=Plur, Person[erg]=3

It is not the only possible formalization, but this one is already integrated in the query language of main query systems, such as Grew-Match (Guillaume 2021).[2] Note that we cannot exclude that a layered feature is syncretic and to have a feature such as Number[abs][ctxt].

We can also remark that the layer [psor] that has been introduced for personal determiners that agree both with their governor and with the possessor is in fact a lexical feature. The current annotation for the German possessive *seine* 'his.FEM' in (5a) can be replaced by (5b):

(5) a. *seine*: Number=Sing, Gender=Fem,
Number[psor]=Sing, Gender[psor]=Masc
b. *seine*: Number=Sing, Gender=Fem,
Number[lex]=Sing, Gender[lex]=Masc

[1] The Leipzig Glossing Rules also advocate a particular convention for lexical features: "Inherent, non-overt categories such as gender may be indicated in the gloss, but a special boundary symbol, the round parenthesis, is used."

[2] Because brackets are special symbols in Grew, the request uses a double underscore: pattern { X [Number__erg] }. See e.g. https://universal.grew.fr/?custom=684da8a0a4075.

4 Inflectional vs lexical features

Lemmas and the repartition between lexical and inflectional features both depend on what we consider as inflectional paradigms. In French, nouns denoting persons or animals can have a masculine and a feminine form: *un instituteur* 'a teacher. MASC', *une institutrice* 'a teacher.FEM'. Two choices are possible:

(6) a. lemma=institutrice, Gender[lex]=Fem
b. lemma=instituteur, Gender=Fem

Following Mel'čuk (2000), the first solution has been chosen for French treebanks.

Another case where it can be difficult to decide what the inflectional paradigms are is illustrated by pronouns. Unlike nouns, French personal pronouns have different forms in subject, object and oblique positions. The traditional analysis is to consider that they vary in case. We consider that we have different pronouns for 1st and 2nd person singular and plural. The lemma is the emphatic form, which is the only form that can stand alone (the form is also used after a preposition). See (7).

(7) a. *je* '1SG.NOM',
lemma=moi, Number[lex]=Sing,
Person[lex]=1, Case=Nom
b. *me* '1SG.ACC|DAT',
lemma=moi, Number[lex]=Sing,
Person[lex]=1, Case=Acc,Dat

We have considered that personal determiner *mon* 'my' is not part of the paradigm and has its own lemma, because it varies in gender and number. The personal pronouns for 3rd person pose an additional problem. The feminine and masculine forms are different for emphatic (*elle, lui, elles, eux*), nominative (*elle* 'she', *il* 'lui', *elles, ils* 'they'), singular accusative (*la* 'her' vs *le* 'him'), but gender is neutralized in plural accusative (*les* 'them') and dative (*lui* 'to her/him', *leur* 'to them'). Moreover, the singular and plural forms are morphologically related and we decided to have the same lemma (even if it was not the choice before v2.16).

(8) *elle*, lemma=lui,
Person[lex]=3, Number=Sing,
Gender=Fem, Case[ctxt]=Nom

The genitive form *en* is currently analyzed as a separate lemma, because personal pronouns of 1st and 2nd person do not have a genitive and *en* is not related morphologically to other 3rd person pronouns.

A contrast between lexical and inflectional features is illustrated by simple vs complex verbal forms. Compare Haitian Creole and French.

(9) a. *lavi m pra chanje*
life my will change
'my life will change'
b. *ma vie changera*
my life will_change

French has a morphological future, while Haitian Creole has a separate auxiliary for future, like English. In the Haitian Creole treebank (Kahane et al. 2024), a feature Tense=Fut has been attached to the marker, but it is clearly a lexical feature:

(10) a. *pra*: upos=AUX, Tense[lex]=Fut
b. *changera*: upos=VERB, Tense=Fut

In English UD treebanks, modals have a VerbForm=Fin feature, but this feature is a lexical feature because modals do not inflect but they can only be used in finite clause:

(11) *must*: upos=AUX, VerbForm[lex]=Fin

In the same way, definiteness on article is a lexical feature: *the*, Definite[lex]=Def. (The article also has a feature PronType=Art, but such a feature is lexical by nature, and we don't need to add [lex] in such a case.)

French has a past tense, called *passé composé*, which is built like English present perfect, but is semantically more similar to the preterit. In this case also, the auxiliary can be considered as a lexical marker of the past, the inflection of the lexical verb being imposed by the auxiliary and being part of the semantics of the auxiliary.

(12) *elle est* *venue*
VerbForm=Fin VerbForm=Part
Tense=Pres Tense[denom]=Past
Tense[lex]=Past
'she came'

Such an analysis is not very different from the analysis we can do for Haitian Creole, where the auxiliary *te* is Tense[lex]=Past and the lexical verb is invariable. But in the case of French, this analysis allows us to indicate that the complex verbal form is past, even if the tense of the auxiliary is present.

5 When to annotate a feature

One question is when to annotate a feature. For instance, the French definite determiner has three form: *le* 'the.SG.MASC', *la* 'the.SG.FEM' and *les* 'the.PL'. For the syncretic plural, do we want to have features Gender or Gender[ctxt]? A feature Gender=Fem,Masc could seem useless, but it can indicate that *les* is a form of a lemma that can vary in gender and contrast it with an adjective such as *utile* 'useful' that is not inflected in gender. And when *les* is combined with a noun, do we want to add a feature Gender[ctxt]?

Because a majority of French adjectives vary in gender, we have decided to add a feature Gender[ctxt] for adjectives that are not inflected in gender. But it is clear that such features are not very useful and could be omitted. Nevertheless, UD treebanks are full of such features. For a case of syncretism such as Russian *žurnal* 'magazine' it is more interesting to indicate that the form is Case=Nom,Acc, because this form contrasts with other forms for dative or locative. And because UD annotation is token-based, it also makes sense to indicate in particular contexts whether it is Case[ctxt]=Nom or Case[ctxt]=Acc.

For English verbal form such as *think*, it is complicated, because the form can correspond to infinitive or present tense and in present tense it can correspond to any number or any person, with the exception of the combination Number=Sing, Person=3. This cannot easily be indicated in the features. Moreover all English infinitive forms will always have VerbForm[ctxt]=Inf, because it is not possible to know that they are infinitive without the context. This is a general property of English morphology and it seems not necessary to indicate it for each occurrence of a verb. In English, many forms are polycategorial, such as *love*, which can be a verb or a noun. The conllu encoding does not allow us to have upos[ctxt], but it is not sure that we want to indicate such syncretisms.

In French, past participles of transitive verbs vary in number and gender and agree with their subject when they are passive forms and with their object when it is placed before the verb. But past participles of intransitive verbs are invariable. It is not always easy to decide whether a verb is transitive or intransitive and for the sake of simplicity, all past participles have features Number and Gender.

6 Annotation of written French

One of our main motivations to distinguish contextual vs overt values of features was the fact that many adjectives in French do not inflect in gender. It was easy to make this distinction because there are resources indicating whether each adjective inflects in gender or number, such as the Lefff (Lexique des formes fléchies du français 'Lexicon of inflected forms of french') (Sagot 2010). A Grew script (Guillaume 2021) based on Lefff has been applied on French-GSD (Guillaume et al. 2019). On the 23817 adjectives of the corpus, 16949 occurrences (71%) (for 2472 lemmas) were covertly marked for gender and number, but 6796 (28%) (for 1124 lemmas) were only marked for number and receive a feature Gender[ctxt], 975 (4%) (for 157 lemmas) were not marked for number at the masculine and receive a feature Number[ctxt], and 72 occurrences (0.3%) (for 22 lemmas) are from invariable adjectives:

- Most common adjectives without gender inflection: *autre* 'other', *même* 'same', *jeune* 'young', *propre* 'proper, clean', *politique* 'political' …
- Most common adjectives with unique form at the masculine: *français* 'French', *nombreux* 'numerous', *anglais* 'English', *vieux* 'old' …
- Most common invariable adjectives: *super* 'super', *standard* 'normal', *arrière* 'back', *cool* 'cool' …

A multilingual lexicon such as UniMorph (Sylak-Glassman et al. 2015) could allow us to do the same thing for other languages.

Some determiners are lexically singular (*chaque* 'each', Number[lex]=Sing) or plural (*pusieurs* 'several', Number[lex]=Plur). Articles vary in number and gender but have a syncretic form for plural. Beyond a vowel, the definite article and possessive determiners have a different form. As shown in (13), the masculine form of the possessive is used before a vowel whatever the gender of the noun.

(13)	*mon*	*étoile*
	Gloss=my	Gloss=star
	Gender=Masc	Gender[lex]=Fem
	Gender[ctxt]=Fem	

Numerals are interesting. They are lexically plural when they are used as determiners/cardinals, but they are singular when they are used as proper nouns:

(14)	*2025*	*est*	*une*	*année*	*très*	*chaude*
	2025	is	a	year	very	hot
	upos=NUM ExtPos=PROPN Number[lex]=Sing					

The French treebanks have some denominative features, such as Tense[denom] for participles (see Section 2). The Tense=Imp feature for *imparfait* tense is another example of denominative feature. We propose to replace it by Tense[denom]=Imp, Tense=Past, Aspect=Imp (imperfective past).

7 Annotation of spoken French

If the corpus is a spoken corpus, we must annotate the morphosyntactic properties of the spoken form and not of its written transcription. We think that it is important to state this, because it is not what was done in spoken French UD corpora before we started this study.

The question is delicate in French, because orthography marks a lot of things that are not pronounced. For instance, plural on the majority of nouns is marked by a grapheme *s*, which would only be audible if a liaison with a following adjective beginning with a vowel is realized and the liaison is almost never attested in spontaneous speech. In consequence, we consider that nouns in spoken French have no number, except for a small set of nouns finishing in *-al* or *-ail*, which have a plural in *-aux* /o/: *un cheval* /œ̃ ʃəval/ 'a horse', *des chevaux* /de ʃəvo/ 'horses'. Among the 5195 nouns in French-Rhapsodie (Lacheret et al. 2019), we have only 49 occurrences (for 9 lemmas) of such nouns.

Adjectives have also a plural marked by a grapheme *s*, but the number is not marked on adjectives in most cases and is only contextual (15a. For prenominal adjectives the liaison with a following noun starting with a vowel is obligatory and the plural will be marked in this case (15b).

(15)	a.	*des oiseaux*	*très*	*petits*
		de zwazo	trɛ	pəti
		Number=Plur		Number[ctxt]=Plur
		'very small birds'		

	b.	*des*	*petits*	*oiseaux*
		de	pəti	**z**wazo
		Number=Plur	Number=Plur	
		'small birds'		

If the gender remains marked for a majority of adjectives (*vert* /vɛʁ/ 'green.MASC', *verte* /vɛʁt/ 'green.FEM'), it is no longer marked for adjectives finishing by a vowel (*joli* /ʒoli/ 'nice', written *jolie* /ʒoli/ in the feminine form), which also concerns past participles. In some dialect such as Belgian French, the final vowel of feminine forms such as *jolie* is lengthen, but it is not the case in the spoken corpora currently in UD.

We can also note that in spoken French, the singular present and imparfait forms of almost all verbs are similar. In consequence, for these verbs, Person is only contextual. Moreover, for most verbs the 3rd person plural is also similar, which means that Number is also contextual. In other words, only the 1st and 2nd person plural are marked. Moreover, the 1st person plural is rarely used: only 2 occurrences of *nous* 'we' subject for 755 occurrences of the indefinite pronoun *on* 'one' in French-ParisStories (Kahane et al..

In conclusion, without taking into account the specificity of spoken data and differentiating the contextual values, the treebank would have been completely misleading concerning number and gender marking.

8 Conclusion

The distinction between inflectional, lexical, and denominative features allows us to clarify the status of morphosyntactic features in UD treebanks. If we study Tense in English (without any knowledge of the language), we would have strange results due to the Tense feature on participles and the absence of lexical feature on auxiliaries would give us that idea that the language has no future.

It is also very useful for linguists exploiting the treebanks to know whether a feature is overt or it has been inferred from the context. Without such an annotation it is not possible to evaluate the range of a given feature. For instance, in French, the subject position is marked by the preverbal position, the agreement of the verb in person and number, and case on personal pronouns, but without a precise annotation it would not be possible to know which features are really effective. Same thing for the range of the noun-adjective agreement in French.

Our proposition of a more precise annotation of morphosyntactic features is a first attempt in UD treebanks and it will certainly evolve in the future. But we hope that such annotation will spread in treebanks of other languages, allowing a more accurate comparison between languages.

Acknowledgements

We would like to thank our three reviewers, as well as Nathan Scheider, Amir Zeldes, and Flavio Massimiliano Cecchini for fruitful discussions on UD issues. This work has been supported by Sylvain Kahane's Institut Universitaire de France grant, which funded Léna Brun and Simeng Song internships.

References

Corbett, Greville G. 2011. The penumbra of morphosyntactic feature systems. *Morphology* 21.2: 445-480.

Corbett, Greville G. 2012. Canonical morphosyntactic features. In Dunstan Brown, Marina Chumakina, and Greville G. Corbett (eds), *Canonical Morphology and Syntax*, 48-65.

de Marneffe, Marie-Catherine, Christopher Manning, Joakim Nivre, Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics* 47(2): 255–308.

Guillaume, Bruno. 2021. Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of EACL: System Demonstrations*.

Guillaume, Bruno, Marie-Catherine de Marneffe, and Guy Perrier. 2019. Conversion et améliorations de corpus du français annotés en Universal Dependencies. *Traitement Automatique des Langues* 60 (2), pp.71-95.

Malaviya, Chaitanya, Matthew R. Gormley, Graham Neubig. 2018. Neural factor graph models for cross-lingual morphological tagging. In *Proceedings of the 56th ACL*, 2652-2662.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86.3: 663-687.

Kahane, Sylvain, Claudel Pierre-Louis, Sandra Jagodzińska, Agata Savary (2024). The first Haitian Creole treebank. In *2nd UniDive Workshop*, Naples.

Kahane S., Caron B., Gerdes K., Strickland E. (2021) Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. In *Proceedings of 19th international conference on Treebanks and Linguistic Theories (TLT)*, SyntaxFest, ACL.

Lacheret-Dujour, Anne, Sylvain Kahane, Paola Pietrandrea (eds) (2019), *Rhapsodie – A Prosodic and Syntactic Treebank for Spoken French*, John Benjamins, Amsterdam.

de Marneffe, Marie-Catherine, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics* 47(2): 255-308.

McCarthy, Arya D., Miikka Silfverberg, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2018. Marrying Universal Dependencies and Universal Morphology. In *Proceedings of the Second Workshop on Universal Dependencies (UDW)*, 91-101.

Mel′čuk, Igor A. 1993. *Cours de morphologie générale: Introduction et première partie: le mot.* Presses de l'Université de Montréal.

Mel'čuk, Igor A. 2000. Un FOU/une FOLLE: un lexème ou deux?. *Lexique, syntaxe et sémantique. Mélanges offertes à Gaston Gross à l'occasion de son soixantième anniversaire* [BULAG, numéro hors série], 95-106.

Mel′čuk, Igor A. 2006. *Aspects of the Theory of Morphology*. Vol. 146. Walter de Gruyter.

Sagot, Benoît. 2010. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC)*.

Sylak-Glassman, John, Christo Kirov, David Yarowsky, and Roger Que. 2015. A Language-Independent Feature Schema for Inflectional Morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL)*, 674-680.