Subjecthood and annotation: The cases of French and Wolof

Olivier Bondéelle, Sylvain Kahane

▶ **To cite this version:**

Olivier Bondéelle, Sylvain Kahane. Subjecthood and annotation: The cases of French and Wolof. M.C. de Marneffe, M. de Lhoneux, J. Nivre, & S. Schuster. COLING 2020 Fourth Workshop on Universal Dependencies (UDW 2020) Proceedings of the Workshop December 13, 2020 Barcelona, Spain (Online), International Conference on Computational Linguistics (ICCL) (2020), Association for Computational Linguistics, 2020, ACL Anthology, 978-1-952148-48-4. hal-03036242

HAL Id: hal-03036242
https://hal.science/hal-03036242v1

Submitted on 6 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Subjecthood and annotation: The cases of French and Wolof[1]

Olivier Bondéelle[1], Sylvain Kahane[2]

[1]Université de Picardie Jules Verne & CERCLL
[2]Université Paris Nanterre & Modyco, CNRS

Abstract

This article considers the annotation of subjects in UD treebanks. The identification of the subject poses a particular problem in Wolof, due to pronominal indices whose status as a pronoun or a pronominal affix is uncertain. In the UD treebank available for Wolof (Dione, 2019), these have been annotated depending on the construction either as true subjects, or as morphosyntactic features agreeing with the verb. The study of this corpus of 40 000 words allows us to show that the problem is indeed difficult to solve, especially since Wolof has a rich system of auxiliaries and several basic constructions with different properties. Before addressing the case of Wolof, we will present the simpler, but partly comparable, case of French, where subject clitics also tend to behave like affixes, and subjecthood can move from the preverbal to the detached position. We will also make a several annotation recommendations that would avoid overwriting information regarding subjecthood.

1. Introduction

In this article, we explore the identification of the subject in two languages with a rigid SVO order, French and Wolof. While these languages share no genetic relationship, they present similarities at the typological level and the identification of the subject position can become problematic in some constructions. In some languages, especially ergative languages, subject properties can be distributed onto different arguments (Keenan, 1976; Comrie, 1978). This is not the case for the languages we are considering, where the identification of the argument realized as a subject is very clear. What interests us here is the fact that the realisation of the same argument is distributed across several syntactic positions and subjecthood can move from one syntactic position to another. This should not be confused with the cases studied by Cole et al. (1980) for example, where subjecthood moves from one semantic argument to another.

To begin this discussion, we must first give a name to verb argument whose subjecthood we want to discuss. We will name it the *first actant*, following Tesnière (1959, 2015) and Mel'čuk (1988). The first actant is the semantic argument of the verbal form that can be realized as its subject: sometimes it is realized as its subject, but sometimes it is only realized as a pronominal affix in the verbal inflection (especially in pro-drop languages). In the languages we are studying, the first actant is easy to define, while the subject is more difficult to identify, because there are several positions were the first actant can appear. In English for instance, the first actant is characterized by the following set of traits: it can be realized in the preverbal position, it causes the verb to bear an agreement suffix, it controls the object position (*he washes himself*), and it can be realized by specific pronouns such as *she* or *we*.[2] It can appear in three positions as shown in (1), where the first actant of the verb *be* is realized as *my father, he*, and the pronominal index amalgamated in the *is* form of *be*.[3]

[1] This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

[2] Due to redistributions, such as the passive voice, the first actant is not always the same semantic argument of the verb (the first actant in *John was surprised by Mary* is *John*). Some verbs, such as *seem* in *it seems that Mary left*, do not really have a first actant, because none of their semantic arguments can be realized in the preverbal position.

[3] There is a fourth position where the first actant can appear (our thanks to a reviewer for highlighting this additional problem): a postverbal position in the so-called impersonal construction.

(i) **it** is also desirable **to retain them** [GUM_academic_exposure-5]

Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020), pages 34–45
Barcelona, Spain (Online), December 13, 2020

(1) "**My father, he's** an anthropologist," she said. [GUM_fiction_veronique-20][4]

Among these three positions, there is one position we will call the *subject*: the preverbal position, because it is obligatory and can be occupied by lexical NPs. The two other positions do not have such properties: one is part of the verb inflection and is not a syntactic position; the other is an optional, prosodically detached position, but always with a pronoun in the preverbal position reflecting it. In other word, the preverbal position is more canonical than the detached position.

We will now explore two languages where the identification of a canonical subject position is less clear: French in Section 2 and Wolof in Section 3. The fact that several positions can theoretically qualify for the subject label in such languages requires a certain degree of caution with regard to treebank annotations. We will also provide several proposals to resolve this issue in our conclusion in Section 4.

2. The case of French

French has a basic structure similar to that of English, with a preverbal position identified as the subject, an SVO order, and a pronominal suffix on the verb in agreement with the subject (identified as s'). However, there are a few differences: pronominal objects (o) are placed before the verb and are cliticized. The pronominal subject is also cliticized on the verb: it has a weak form, which is distinct from the strong/tonic pronominal form in the detached position (D) (2b) and which cannot be separated from the verb (V) (2c,d). We therefore postulate the existence of different positions for the lexical subject (S) and the pronominal subject (s), since a non-cliticized element can be inserted between S and V but not between s and V, which gives us the topological scheme (2a).[5]

(2) a. D S s=o=V-s' O

b. **lui il** avait passé les quatre nuits ou trois nuits à à ramper dans les décombres [Rhap_D0003-18]

'him, he had spent the four nights or three nights crawling through the rubble'

c. **le programme** monsieur le premier ministre comporte un certain nombre de projets

'the program, Mister Prime Minister, includes a number of projects' [Rhap_D2006-1]

d. *** il** monsieur le premier ministre comporte un certain nombre de projets

'it, Mister Prime Minister, includes a number of projects'

The pronominal paradigms regarding D, s and s' are provided in Table 1. One may note that the s' agreement suffixes tend to disappear; at present, only the 2PL form is really marked. The *nous* V-*ons* form of the 1PL has been largely replaced by *on* V-Ø in spoken French. The future tense, which is the only tense where agreement is well marked in the s' position, is often replaced by a complex form with the auxiliary *aller* 'go', which is another way to move the agreement to a preverbal position.

This position is analyzed as the subject in UD. This issue falls outside of the scope of this paper, but we think that this annotation is quite problematic because this position does not display the same properties as the preverbal position, and should not be named in the same way according to traditional surface-syntactic criteria (see for instance criterion C in Mel'čuk 1988). In the Surface-Syntactic UD (SUD) annotation scheme, we have analyzed it as an object position (Gerdes et al. 2019).

[4] All our examples are extracted from UD treebanks with their sent_id identifier.

[5] By *topological scheme*, we refer to a linear template corresponding to a syntactic configuration. The topological model was first developed for the modeling of word order in Germanic languages during the 19th century, and was later implemented in dependency grammar (Duchier and Debusmann, 2001; Gerdes and Kahane, 2001).

Table 1. Pronominal indices in French

	D	s	s’
1SG	*moi*	*je*	Ø
2SG	*toi*	*tu*	Ø
3SG	*lui/elle/ça*	*il/elle/ce*	Ø
1PL	*nous* *nous*	*nous* *on*	*-ons* Ø
2PL	*vous*	*vous*	*-ez*
3PL	*eux/elles*	*ils/elles*	Ø

It can be noted that non-pronominal subjects are relatively rare in oral French. In the treebank UD_French-Spoken (Kahane et al., 2019), subjects are divided into 12% lexical subjects, 11% relative pronouns subjects and 77% pronominal subjects (we do not have a spoken English treebank for comparison).

French, especially in its spoken form, frequently uses dislocation, which concerns 10% of sentences in UD_French-Spoken. We do not know what proportion of these detached elements are first actants, as they have not been annotated for the moment. It has been argued by some authors (notably Culbertson and Legendre, 2008; Miller and Sag, 1997), that the first lexical actant tends to be realized in position D rather than in position S in spoken French. Data from the UD_French-Spoken corpus shows that S still dominates D in spoken French. Nevertheless, we can imagine a future form of French with a topological scheme D s=o=V O, where the subjecthood has moved to position D and s no longer commutes with S and thus becomes an agreement prefix.

French has several interrogative constructions. In the standard interrogative construction (3a), s and S do not commute either: both positions can be filled simultaneously (3b), s is mandatory, while S is optional (3c) and cannot accept personal pronouns (3d).[6] As only S can accept lexical realisations of the first actant, we consider S to be the subject and the interrogative construction is therefore a pro-drop construction where s has the status of an agreement suffix.

(3) a. interrogative: D S o=V-s'=s O
b. mais **l'acte d'écrire** est**-il** le prolongement de l'acte de penser ? [Rhap_D2009-9]
‘but is **the act of writing** an extension of the act of thinking?’
c. mais est**-il** le prolongement de l'acte de penser ?
d. * mais **il** est**-il** le prolongement de l'acte de penser ?

It is remarkable that French has both pro-drop constructions and non pro-drop constructions.

Currently, the two positions s and S are annotated **nsubj** in the French treebanks. In interrogative constructions, one can thus have two **nsubj** relations. On the other hand, the first actants in position D are annotated **dislocated** and are therefore not distinguished from the other NPs in this position. New proposals will be made in Section 4.

We will see that the situation is more complex in the case of Wolof.

3. The case of Wolof

Our study of Wolof is essentially based on the analysis of the treebank UD_Wolof-WTB, annotated by Dione (2019). In Wolof, the s position of pronominal subjects must also be distinguished from the S position of lexical subjects. For example, in relative clauses, a very frequent construction in Wolof due to the absence of an adjective class (1739 relatives for 2107 sentences, i.e. 82 relatives for 100

[6] In spoken French, s is optional, but not in standard written French. The prosody, as well as the position of the interrogative pronoun, makes possible the distinction between the S and D positions (i a,b).

(i) a. S : A qui **Pierre** parle-t-il ? ‘Who does Pierre speak to?’
b. D : **Pierre**, à qui parle-t-il ? ‘Pierre, who does he speak to?’

sentences), the order is highly constrained: s is placed before clitic complements o, and S between o and V (4a).[7] Let us develop upon this description of relative clauses in Wolof.

The relative pronouns and determiners of Wolof are constructed with the combination of a nominal class marker (corresponding to the determined or antecedent noun) and one of the three morphemes *a, i* or *u* that structure the entire grammar of Wolof (they are also present in the verbal domain), giving the words CL-*i*, CL-*a* and CL-*u*. There are 10 nominal noun classes: 8 for the singular *(b, k, w, m, g, l, s, j*) and 2 for the plural (*ñ, y*). The classes *b* and *y* are becoming the default classes for the singular and plural. The morphemes *i* and *a* mark respectively a proximal or distal (4b), while *u* marks an indefinite and tends to become the default marker for the relative pronoun (4c).

Headless relative clauses are very frequent (about 1000 in the corpus). The pronoun can have an anaphoric value and agree with a distant antecedent or be a generic pronoun introducing a new referent. In this case, one of the five noun class markers which designate a human singular (*k*) or plural (*ñ*), an inanimate (*l*) (4d), a temporal (*b*) and a conditional (*s*) (4d) is used. In addition, there are two former nominal classes that indicate location (*f*) and manner (*n*). We gloss the generic marker of the conditional by CND, and the relative and integrative pronouns by REL.

(4) a. relative: R=s=o S V O

b. jigéen **ji**
woman CL.DEF
'the woman (close from me)'

c. jigéen **ju** ko am-e [wo_wtb-ud-train-1530]
woman CL.REL O3sg have-TR
'a woman who takes care of it'

d. **li** nga moom-ul [wo_wtb-ud-train-1106]
INA.REL S2SG possess-NEG
'what you don't possess'

e. **so**o ko yeexe gis [wo_wtb-ud-train-933]
CND.S2SG O3SG delay see
'if you're slow to see it'

f. buum **bi** nu kolonisatëër bi nas=oon [wo_wtb-ud-train-1094]
rope CL.REL O1PL colonizer CL.DEF thread=PAST
'the rope that the settlers put around our necks'

Verbs in a relative clause are always preceded by a subject realized in one of the three possible positions: the relative pronoun (4c), s (4d,e) or S (4f). Relative clauses do not have a D position. The pronominal subject s cliticizes on the relative pronoun and can be amalgamated with it (see (4e), where *soo=su*.S2SG). Positions s and S are distinguished by the position of the pronominal object, which occurs after s (4e) and before S (4f). Wolof has many auxiliaries, but only two of them appear in relative clauses: *di*, the preverbal marker of imperfective, and *woon*, the postverbal marker of past tense.[8]

In contrast to relative clauses, Wolof has several constructions in the main clause, each one controlled by a particular marker, which can be an auxiliary or a verbal suffix (Robert, 1991; Torrence, 2005; Torrence, 2013; Bondéelle, 2015; Martinovic, 2015; Robert, forthcoming). In all these constructions, the first actant can be realized as a pronominal index in position s or in a detached position D, and for some constructions a third position S is available. We will now study the main constructions and discuss the subjecthood for each of them.

---

[7] All our assertions have been verified by requests on UD_Wolof-WTB with grew-match (Guillaume et al., 2012; Bonfante et al., 2018). For instance, we can verify that there are no subject after the verb with a request such as:

pattern { H -[acl:relcl]-> V ; V-[nsubj]-> S ; V << S }.

[8] The past tense *woon* is only analyzed AUX when it is spelled as a separate word; there are also many verbs with the feature Tense=Past where *woon* is amalgamated with the verb, as in (4f).

There is a minimal SVO construction without an auxiliary used in only 10.4% of sentences (5a). We will mainly focus on constructions with auxiliaries. The auxiliaries of Wolof, other than *di* and *woon*, focalize one of the elements of the verbal construction: *a* focalizes the subject, *la* one of the complements, *na* the verb, and *da* the VP. Negation is marked by the suffix *-u* which attaches to the verb and focalizes it. We leave aside the auxiliary *ngi*, which behaves like *a*, as well as different compound forms which behave more or less like *da*.

3.1 The forms of s

As in French, pronouns realized in position D have a strong form that is different from the form of pronouns in position s. In addition, the s pronouns cliticize on the auxiliary, which produces amalgams and some zero forms in the third person (see for example the zero forms *la* and *na* in columns 5 and 6, and the amalgam *moo* that results from the fusion of the pronoun *mu* and the auxiliary *a* in column 4 of Table 2). The 1PL and 3PL forms are regular for all auxiliaries (the forms *noo* = *nu.a* and *ñoo* = *ñu.a* obey a regular morphophonology rule of Wolof). The 1SG forms are quite regular, even if the consonant /m/ disappears with *la* and *na*. On the other hand, the 2SG and 2PL forms are highly irregular, due notably to the disappearance of the particles *la* and *na* and the use of strong pronouns as a basis for *a*. Finally, for 3SG, the index is only expressed for *a*. This is undeniably a sign of a tendency for the s position to become an agreement suffix. The choice of Dione in UD_Wolof-WTB was to analyze s as a subject with *a* (5b-b") and *da* and as a morphosyntactic feature on the auxiliary with *la* (5c-c"), *na* and *-u* (which are the three cases where 3SG has a zero form).

Table 2. Pronominal indices in Wolof

	D	s V	s=*a*	*la*=s	*na*=s	*da*=s	V-*u*=s
1SG	*man*	*ma* V	*maa*	*laa*	*naa*	*dama*	V-*uma*
2SG	*yow*	*nga* V	*yaa*	*nga*	*nga*	*danga*	V-*uloo*
3SG	*moom*	*mu* V	*moo*	*la*	*na*	*da(fa)*	V-*u(l)*
1PL	*nun*	*nu* V	*noo*	*lanu*	*nanu*	*danu*	V-*unu*
2PL	*yeen*	*ngeen* V	*yeena*	*ngeen*	*ngeen*	*dangeen*	V-*uleen*
3PL	*ñoom*	*ñu* V	*ñoo*	*lañu*	*nañu*	*dañu*	V-*uñu*

We will now look at the different topological schemes of the auxiliaries and the question of the lexical realization of the first actant. We will see that the S position behaves differently depending on the constructions.

3.2 Verbal constructions in the main clause

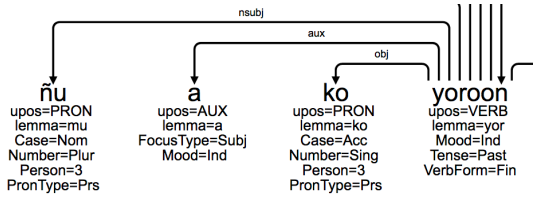
Here are the topological schemes of the different constructions in the main clause:

(5) a. minimal construction: D S/s V o O

b. auxiliary *a* construction (subject focalization): D S/s=*a* o V O

b'. **ñoo** ko yor=oon. [wo_wtb-ud-train-3] (*ñoo* = *ñu.a*)
S3PL.AUX O3SG possess=PAST
'They were the ones who detained him'

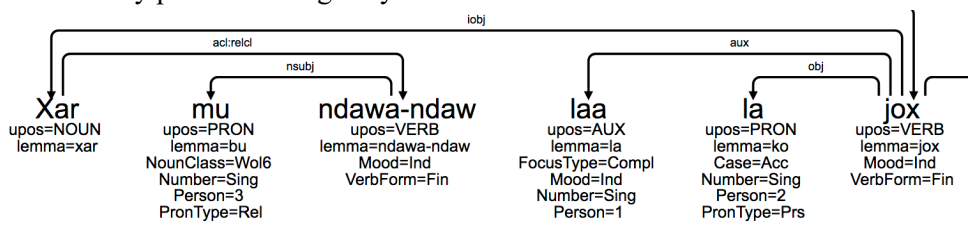
b".



c. auxiliary *la* c: D O! *la*=s o S V O (object focalization of a unique element in O!)

c'. Xar mu ndawa-ndaw **la**a la jox. [wo_wtb-ud-train-840]
portion CL.REL be_tiny AUX:S1SG O2SG give
'This is a tiny portion that I gave you'

c".



d. auxiliary *na* construction (verb focalization): D V *na*=s o O

e. auxiliary *da* construction (VP focalization): D *da*=s o V O

f. suffix -*u* construction (negation): D V-*u*=s o O

The three assertive constructions with *na*, *da* and -*u* (respectively focalization of V, VP, and negation) block the realization of a lexical subject in position S (there is no possible confusion with position D which, unlike S, can accommodate strong pronouns). In other words, for these constructions, the paradigm of weak subject pronouns, those occupying the position s, no longer commutes with a lexical subject. This leads most authors to consider that weak subject pronouns have become pronominal indices belonging to the verb inflection and that an element in position D co-referring with the index s is therefore the true subject (Sauvageot, 1965; Church, 1981; Diouf, 1985; Ndiaye-Corréard, 1989; Robert, 1991; Fal, 1999; Ndiaye-Corréard, 2003; Guérin, 2016).

Let us see how position D is like a subject. Position S is possible with *a* and *la*. Note that with *a*, the pre-auxiliary position S/s (see 5b) is filled by weak pronouns 81.5% of the time (387 out of 475), compared to 14.5% for NPs and 4% for other pronouns. The proportion of weak pronouns is very high, especially if we consider that *a* focalizes the subject and that it is a written corpus. In comparison, UD_French-GSD and UD_French-FTB have 41% and 33.5% subject pronouns (including relative pronouns) and UD_English-GUM and UD_English-EWT have 53.5% and 57.5% subject pronouns.

The case of *la* (focalization of the object) is particularly interesting, since it opens a position S distinct from s and allows for the realization of a lexical first actant in both the S and D positions. Dione's corpus contains 456 occurrences of the particle *la*, including 115 occurrences of the form *lañu* (*la*=s3PL). None of them contain an NP in position S. On the other hand, we have 78 lexical subjects in position S including 11 with a plural determiner.[9] All are with the form *la*. We conclude from this that it is not possible to have both a pronoun in position s and a lexical subject in position S at the same time. As S is considered a subject position and s is in complementary distribution with S, we must consider that s is a subject in this construction. Example (6) illustrated the case of a plural subject in S: the initial pronoun *moom* 3SG is a strong pronoun in the focalized object position (O! in (5c)) and the position S is occupied by the syntagm *yaakaar yi* 'hopes' whose determiner *yi* marks the plural. The form *la* in this case does not combine with a s3SG zero form and is glossed only by AUX.

[9] Nouns without determiners are not marked in numbers. The query to retrieve plural lexical subjects in position S is:

pattern { L [upos=AUX, lemma=la] ; V -[aux]-> L ; V -[nsubj]-> S ;
S [upos = NOUN|PROPN] ; S -[det]-> D ; D [Number=Plur] ; L << S }.

(6) […] moom **la** **yaakaar yi** tas-e […] [wo_wtb-ud-train-488]
3SG AUX hope PL.DEF be_spread-TR
'[…] it's with him (that) hopes have been dashed […]'

Conversely, there are 16 assertive constructions with the auxiliary *la* where a nominal group in position D is analyzed by Dione as a subject and is therefore the realization of the first actant of the verb: only one has a plural determiner and the form of the auxiliary is *lañu* (*la*=S3PL).

We conclude that the first actants in positions S and D do not behave in the same way with *la*: position S blocks the realization of s and position D requires the realization of s. For constructions with *la*, we would therefore tend to consider that s is the realization of the subject (contrary to the analysis of Dione 2019).

With the negation marker *-u*, the situation is more confusing. Unlike the auxiliaries *a, la, na* and *da* (respectively focalization of subject, object, V, and VP), negation is possible in relative clauses. When the verb is in a relative clause, there is no D position, whereas when the verb is in a main clause, the S position is blocked and the D position is accessible to the first actant. There are 13 occurrences of the first actant of a negative verb being analyzed as a subject by Dione and which have a plural determiner. There are 7 instances where position s is instantiated by *ñu* 's3PL' and 6 where position s is empty. If the situation were the same as for *la*, the first actant would be expected to be in the D position in the first case and in the S in the second. 9 cases behave as expected; for 1 case, after the temporal conjunction *ba*, nothing can be said because both constructions are possible. Some cases are clearly deviant: in (7a), the verb *amag-ul* 'still_have-NEG' is the main verb and s is empty (the *-ul* form of the negative suffixed to V is not marked in number while the syntagm *ay arondismaa* 'boroughs' in position S bears the mark of the plural determiner *-y*); in (7b,c), we have two examples of the same interrogative construction where the verb depends on the verb *tax* 'cause' without a complementizer. In one of these examples, position s is empty (the *-ul* form of the negative in (7b)) while in the other (in (7c)) it is instantiated (the form *-uñoo* = *-u=ñu=a* = NEG=S3PL=PART, where *a* is a subordinating particle).

(7) a. Booba jamono **ay** **arondismaa** amag-**ul**. [wo_wtb-ud-train-627]
CL.ANAPH period CL.IND boroughs still_have-NEG
'At that time, there weren't any boroughs yet.'

b. Lu tax **tooñaange yii** **yépp** jur-**ul** coow? [wo_wtb-ud-train-2007]
CL.INT cause_that teasing CL.DEF CL.everything produce-NEG noise
'Why is it that all the vexations don't make any noise?'

c. Lu tax, **daamar yooyu** mën-**u=ñoo** daw ci Senegaal? [ud-train-2048]
CL.INT cause_that vehicle CL.ANAPH can-NEG=S3PL:PART run LOC Senegal
'Why can't these vehicles circulate in Senegal?'

These deviant cases show a certain wavering in the instantiation of s in relation to *-u* and nevertheless accredit the fact that the functioning of s in relation to the positions S and D tends to harmonize and D to be treated as a subject position equivalent to S.

The correlation between the instantiation of s (presence of *ñu*) and the presence of a comma can be seen in example (7c). One can imagine that this is also correlated with different prosodies. It is probably necessary to distinguish, among the phrases in position D, between those which are actually prosodically detached and those which are prosodically integrated into the verbal nucleus. If the first actants in position D are subjects, it is expected that they are prosodically integrated into the verbal nucleus. It appears from the literature that both situations are possible (Rialland and Robert, 2004). We only have a written corpus and we cannot study prosody, but we can study the presence or absence of a comma after the D position, which is usually the marker of a prosodic boundary. Dione's corpus contains 80 negative verbs which are roots and are preceded by a phrase analyzed as a subject; among them, 7 are followed by a comma. With the root verbs accompanied by the auxiliaries *la* or *na*, we have only 35 subject phrases followed by a comma. Note that there is one case of a comma after a lexical subject in position S with the auxiliary *a* (for 100 without a comma).

We can therefore observe that the first actants in position D of the assertive constructions involving *na*, *la* or negation are rarely followed by a comma (only 10%). For comparison, there are 168 phrases annotated as dislocated (dislocated relation of UD) with the assertive constructions involving *na* and *la* and 73 (i.e. 43%) of them are followed by a comma. Out of 329 adverbial clauses in position D, there are 262 (80%) followed by a comma.

Consequently, it must be considered that there are two types of subjects in Wolof: subjects in position S, which are obligatory and commute with pronouns and do not trigger agreements, and subjects in position D, which are optional and trigger agreements. We are therefore looking at a hybrid system with two subject functions with very different properties.[10] The problem is therefore the analysis of position s which becomes heterogeneous: in constructions where the position S is accessible and in mutual exclusion with s, s must be analyzed as a subject, whereas when the position S is no longer accessible and D is analyzed as a subject, s must be analyzed as an agreement morpheme.

4. Conclusion

We looked at the issue of subjecthood in three rigidly ordered nominative-accusative SVO languages, English, French and Wolof. In English, the pronominal first actants occupy the same linear place as lexical first actants and are in complementary distribution. In such a case, it is clear that their syntactic position is one and the same. The situation may be more confusing in other languages, such as French or Wolof. The potential problems are as follows:

- the first lexical actant can occur in two different positions, which we have named S and D, position D being a detached position that can be occupied by other NPs;
- position S tends to be used less and less in favor of D, or even to disappear;
- s and S occupy different linear positions;
- s and S can co-occur and are no longer in complementary distribution;
- the forms in position s differ from the pronominal forms in position D (weak vs. strong forms);
- s tends to merge with the verb (or a verbal auxiliary) by becoming inseparable from the verb, resulting in varying forms according to the verb as well as zero forms.

Each of these elements accompanies a shift of the subjecthood from the S position to the D position. It is interesting to note that this shift does not occur homogeneously, but can be faster in some

---

[10] This has already been considered for the SVO and VSO orders of Classical Arabic. With SVO, the verb agrees in gender, person and number with the subject (strong agreement) (El Kassas and Kahane, 2004; Attia, 2008):

(i) al'awlad akal**uu** al-mawz
DEF-boy.PL eat.PASS.MASC.3PL DEF-banana.PL
'The boys ate the bananas.'

(ii) al-banaat 'akal**naa** al-mawz
DEF-girl.PL eat.PASS.FEM.3PL DEF-banana.PL
'The girls ate the bananas.'

But if the order is VSO, the verb agrees only in gender and in person with the subject (weak agreement):

(iii) 'akal**at** al-banaat al-mawz
eat.PASS.FEM.3 DEF-girl.PL DEF-banana.PL
'The girls ate the bananas.'

There are thus two types of subjects with different agreement properties according to the word order. In the so-called dialectical Arabics, the situation has become simpler. For example, in Egyptian Arabic, both orders (SVO and VSO) are possible (though the SVO order is dominant) and the verb is inflected in the same way in both cases, in gender and number only:

(iv) el-banaat 'akal**et** el-moz
DEF-girl.PL eat.PAST.FEM.PL DEF-banana.PL
'The girls ate the bananas.'

(v) 'akal**et** el-banaat el-moz
eat.PAST.FEM.PL DEF-girl.PL DEF-banana.PL
'The girls ate the bananas.'

(We thank Mohamed Galal for the data.)

constructions and lead to a hybrid system where the first actant can be realized in the S or D positions according to the construction, and where s can function more or less independently from S. Thus in French, in the interrogative construction, s is no longer in complementary distribution with S and becomes an affix, whereas in the basic declarative construction, the realization of the first actant tends to move from S to D. In Wolof, s tends to merge with the various auxiliaries and S has disappeared from some constructions in favour of D. However, in other constructions, such as the relative clause, D is not accessible and s and S are in complementary distribution.

In terms of annotation, our recommendations are as follows.

1) As soon as there is a suspicion of a shift in subjecthood from S to D, it is advisable to use a **dislocated:subj** relation to be able to identify the realizations of the first actant in position D.[11]

2) When there is a suspicion that a pronominal index may not be an affix, it is best to treat it as a pronoun, that is a separate word. One can use the **nsubj** function (and thus sometimes have two subjects), but it may be desirable to distinguish the function of elements in position s (e.g. by an **nsubj:weak** relation), as some pronouns may occupy the position S.[12]

[11] Such an annotation may also relate to the object. For example, in Mandarin and Cantonese, the second actant may be detached on the left, without it being clear whether it is a topicalized or dislocated object. This was annotated **dislocated** in UD_Cantonese-HK, making it difficult to study (Wong et al., 2017). A **dislocated:obj** relation would have allowed for a better exploitation of the corpus and comparison with the Mandarin corpus.

[12] As we said Dione has opted for a heterogeneous annotation of s. Despite this, it was possible to identify all occurrences, sometimes at the cost of rather complex queries. The main problem for our study of subjecthood has been the use of **dislocated** regardless of the role of the detached NP.

Abbreviations for glosses

ANAPH: anaphoric
AUX: auxiliary
CL : nominal class
DEF : definite
IMP : imperfective
INT: interrogative
NEG : negative
O : object
PAST : past
PL : plural
REL : relative
S : subject
SG : singular

Abbreviations for topological positions

D : detached item field on the left
o : clitic complements field
O : non-clitic complements field
O! : field accommodating exactly one complement
s : weak pronominal subject field
S : subject field
V : verb field

Acknowledgements

We thank Bernard Caron, Jasmina Milićević and Emmett Strickland, as well as the two reviewers for their comments and suggestions that helped us improve the initial text.

References

Mohamed Attia. 2008. Alternate Agreement in Arabic. *Proceedings of Parallel Grammar Meeting (ParGram)*, Istanbul, Turkey.

Olivier Bondéelle. 2015. *Polysémie et structuration du lexique : le cas du wolof*. Utrecht : LOT.

Guillaume Bonfante, Bruno Guillaume, and Guy Perrier. 2018. *Application of Graph Rewriting to Natural Language Processing*, Wiley-ISTE.

Eric Church. 1981. *Le système verbal du wolof*. Dakar : Université de Dakar.

Peter Cole, Wayne Harbert, Gabriella Hermon, and S. N. Sridhar. 1980. The acquisition of subjecthood. *Language* 56(4) 719-743.

Bernard Comrie. 1978. Ergativity. In *Syntactic typology*, W. P. Lehma (ed.), 329-393. Austin: University of Texas.

Jenny Culbertson and Géraldine Legendre. 2008. Qu'en est-il des clitiques sujet en français oral contemporain ? In Durand J. Habert B., Laks B. (eds.) *Congrès Mondial de Linguistique Française - CMLF'08*. Paris : Institut de Linguistique Française.

Cheikh Bamba Dione. 2019. Developing Universal Dependencies for Wolof, *Proceedings of the Third Workshop on Universal Dependencies (UDW)*, SyntaxFest, Association for Computational Linguistics, 12-23.

Jean-Léopold Diouf. 1985. *Introduction à une étude du système verbal wolof*. Dakar : CLAD.

Denys Duchier and Ralph Debusmann. 2001. Topological dependency trees: A constraint-based account of linear precedence. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (*ACL*), 180-187.

Dina El Kassas and Sylvain Kahane. 2004. Modélisation de l'ordre des mots en arabe standard. *Actes de l'Atelier sur le traitement automatique de la langue arabe, JEP-TALN.*

Arame Fal. 1999. *Précis de grammaire fonctionnelle de la langue wolof*. Dakar.

Kim Gerdes and Sylvain Kahane. 2001. Word order in German: A formal dependency grammar using a topological hierarchy, *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (*ACL*).

Kim Gerdes and Sylvain Kahane. 2006. Phrasing It Differently, in Leo Wanner (ed.), *Selected lexical and grammatical issues in the Meaning-Text Theory*, Amsterdam / New-York: John Benjamins, 297-335.

Kim Gerdes and Sylvain Kahane. 2016. Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies, *Proceedings of Linguistic Annotation Workshop (LAW)*, ACL, Berlin.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2019. Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic functions and deep-syntactic features, *Proceedings of the 17th international conference on Treebanks and Linguistic Theories (TLT)*, SyntaxFest, Paris.

Maximilien Guérin. 2016. *Les constructions verbales en wolof : Vers une typologie de la prédication, de l'auxiliation et des périphrases*. Thèse de doctorat. Paris, Université Sorbonne Nouvelle Paris 3.

Bruno Guillaume, Guillaume Bonfante, Paul Masson, Mathieu Morey, and Guy Perrier. 2012. Grew : un outil de réécriture de graphes pour le TAL. *Actes de la 12e Conférence annuelle sur le Traitement Automatique des Langues (TALN)*, Grenoble, France.

Sylvain Kahane, Kim Gerdes, and Rachel Bawden. 2019. The microsyntactic annotation, In Lacheret-Dujour A., Kahane S., Pietrandrea P. (eds), *Rhapsodie – A Prosodic and Syntactic Treebank for Spoken French*, John Benjamins, Amsterdam, 49-68.

Edward Keenan. 1976. Towards a universal definition of 'subject'. In *Subject and Topic*, C. N. Li (ed.), 303-334. New York: Academic Press.

Martina Martinovic. 2015. *Feature geometry and head-splitting: Evidence from the morphosyntax of the Wolof clausal periphery*, Doctoral dissertation, University of Chicago.

Igor A. Mel'cuk. 1988. *Dependency syntax: theory and practice*. SUNY press.

Igor A. Mel'čuk. 2013. Syntactic subject, once again. *Proceedings of the 6th International Conference on Meaning-Text Theory*, Prague.

Philip H. Miller and Ivan A. Sag. 1997. French clitic movement without clitics or movement. *Natural Language & Linguistic Theory*, *15*(3), 573-639.

Geneviève N'Diaye-Corréard. 1989. Focalisation et système verbal en wolof. *Annales de la Faculté des Lettres et Sciences Humaines,* 19, Dakar, 177-190.

Geneviève N'Diaye-Corréard. 2003. Structure des propositions et système verbal en wolof. *SudLangues,* 3. 163-188.

Annie Rialland and Stéphane Robert. 2004. La focalisation en wolof : morphosyntaxe et intonation. In Anne Lacheret-Dujour, Jacques François (éd.) *Focalisation et moyens d'expression de la focalisation à travers les langues*, Mémoires de la Société de Linguistique de Paris, Peeters, 138-160.

Stéphane Robert. 1991. *Approche énonciative du système verbal : Le cas du wolof.* Paris : CNRS Éditions.

Stéphane Robert. To appear. Wolof: A grammatical sketch. In F. Lüpke (ed.), *The Oxford guide to the Atlantic languages of West Africa*. Oxford University Press.

Serge Sauvageot. 1965. *Description synchronique d'un dialecte wolof : Le parler du Dyolof.* Dakar : IFAN.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Paris : Klincksieck.

Lucien Tesnière. 2015. *Elements of structural syntax*, transl. by T. Osborne and S. Kahane, Amsterdam: John Benjamins.

William H. Torrence. 2005. *On the Distribution of Complementizers in Wolof*. Doctoral dissertation, University of California, Los Angeles.

William H. Torrence. 2013. *The Clause Structure of Wolof: Insights into the Left Periphery*. Amsterdam: John Benjamins.

Tak-Sum Wong, Kim Gerdes, Herman Leung, and John Lee. 2017. "Quantitative Comparative Syntax on the Cantonese-Mandarin Parallel Dependency Treebank" *Proceedings of the Fourth International Conference on Dependency Linguistics*, pp. 266−275, Pisa, Italy, September 2017.

Marina Yaguello. (ed.) 1994. *Subjecthood and subjectivity: the status of the subject in linguistic theory* [proceedings of the Colloquium" The status of the subject in linguistic theory"], London, 19-20 March 1993. Editions Ophrys.