Some preliminary observations on the preparation of a Saho treebank

Andrea Di Manno Università di Napoli «L'Orientale» – adimanno@unior.it
Paolo Milizia Università di Napoli «L'Orientale» – pmilizia@unior.it

ABSTRACT

In this paper we present some preliminary remarks on the preparation of a morphosyntactically annotated treebank for Saho within the Universal Dependencies (UD) framework, an annotation scheme used in a large collection of corpora, where Cushitic languages are still marginally represented. We discuss some issues concerning tokenization, morphology, and syntax, with particular reference to cases where deviations from the descriptions proposed in the literature on Saho are deemed necessary, in order to accommodate Saho data within the UD architecture, and cases where some peculiarities of the Saho language require an expansion of the inventories put forward in UD.

Keywords: *Saho language; treebank; syntax; morphosyntax; Universal Dependencies*

ISO 639-3 code: ssy

DOI: 10.23814/ethn.20.24.dim-mil

1. Introduction[1]

Saho is a Cushitic language spoken in central and south-eastern Eritrea and in the neighbouring area of Tigray, Ethiopia, and together with Afar constitutes the Saho-Afar group within Eastern Cushitic. The two languages form a dialectal continuum, and three major groups can be identified for Saho: Northern Saho, Central Saho, and Southern Saho. The latter is the closest to Afar, while for Central Saho one distinguishes varieties that share isoglosses with the northern varieties (CS1) and varieties that share features with the southern ones (CS2). Though in the last decade several contributions have improved our understanding of the language[2] and provided new material (including the Saho Corpus, cf. Jama Musse Jama 2022), a syntactically annotated treebank of the language is still missing.

This paper presents some preliminary remarks on the preparation of such a treebank within the Universal Dependencies (UD) framework (de Marneffe et al. 2021). UD adopts a dependency grammar perspective, meaning that it presumes direct links between linguistic units. Since it is based on a lexicalist view of syntax, dependency relations hold between words and there is no attempt to segment words into morphemes. Each word has a head on which it depends (except for the main predicate, which is considered the root) and can have one, zero or multiple dependents. In diagrams, dependency is represented through an arrow from the head to the (head of the) dependent and relation labels indicate the syntactic functions. Although providing a closed (universal) inventory of morphosyntactic categories for consistent cross-linguistic

[1] The authors thank Axmadsacad Maxammad Cumar, Giorgio Banti and Moreno Vergari for their support in the preparation of the present paper. Of course, the authors alone are responsible for all errors and omissions. The paper results from joint work by the authors. Paolo Milizia is responsible of Sections 2 and 3; Andrea Di Manno is responsible for the rest of the paper.

[2] See Banti – Vergari 2010 and Banti – Vergari 2023: 294-95 for previous contributions on the language. A comprehensive bibliography on the language and the people is in Vergari et al. 2023: 31-50.

annotation, UD allows language-specific extensions for the treatment of language-specific phenomena. Being an ongoing project, new treebanks are constantly added and, consequently, the guidelines keep being updated in order to deal with previously unattested phenomena. Although in the current version (2.13) UD consists of 259 treebanks in 148 languages, Cushitic languages are underrepresented, since only one treebank for Beja (Kahane et al. 2021) is available.

Annotations are encoded in plain text files in the CoNLL-U format. Each sentence consists of one or more word lines, and each word line contain the following ten fields: ID (the word index in the sentence), FORM (the word form), LEMMA (the lemma of the word form), UPOS (the universal part-of-speech tag), XPOS (the optional language-specific part-of-speech tag), FEATS (the list of morphological features), HEAD (the word index of the head of the current word), DEPS (the enhanced dependency graph, with additional relations and nodes to make implicit relations between words more explicit) and MISC, containing any other annotation. We use the MISC field to provide an English gloss of the Saho word; complying to the UD guidelines, we place an underscore whenever a field is empty[3].

2. Tokenization and lemmatization

An issue concerning tokenization is how to deal with words which correspond to a fused sequence of personal pronoun plus postposition: e.g. *yedde* (or *yodde*) = *yi* + *dde* 'on me'. A solution is to consider the fused word as a supertoken, e.g.:

1-2	yodde								
1	yo	yi	PRON	_	Case=Acc|Number=Sing|Person=1|PronType=Prs				
2	dde	dde	ADP	_	_	1	case	_	Gloss=on

A similar treatment can be applied to the so-called "free forms" of postpositions. According to Banti – Vergari (2023: 311), these are special forms of postpositions which are not combined with a pronoun and which, in their interpretation, entail an understood third-person pronoun. E.g., besides *-dde* 'on', a free form *adde* is found meaning 'on him/her/it'. According to a different analysis, such free postpositions can be viewed as portmanteau words comprising a bound pronoun *a-*/*aka-*/*aa-* plus the bound postposition[4]. Along these lines, an annotation of the following type may be proposed:

1-2	adde								
1	a	a	PRON	_	Case=Acc|Number=Sing|Person=3|PronType=Prs				
2	dde	dde	ADP	_	_	1	case	_	Gloss=on

A further issue concerns the nominalizing suffix *-m*. It is important, indeed, to distinguish the instances in which a final *-m* can be considered as an element of the morphosyntactic structure of a phrase from those in which a word ending in *-m* can be considered, from the synchronic point of view, as a non-analysable element even if that final *-m* is arguably to be identified with the nominalizing *-m* from the etymological point of view. Thus, for instance, adverbs such as *mangum* 'very much' and *dagum* 'a little, few' are better treated as simple ADV elements, though they are clearly historically connected with the stative verbs *mango* 'be much' and *dago* 'be few' (cf.

[3] See https://universaldependencies.org/format.html All the cited urls have been last accessed in March 2024.

[4] This pronoun might be etymologically connected with the anaphoric demonstrative *ay* (Banti – Vergari 2023: 309), *aa* in Southern Saho (Esayas Tajebe 2015: 161), *a-* both in Afar and Northern Saho according to Morin (1995: 88).

Banti – Vergari 2005: 124) and probably contain the nominalizing *-m*, etymologically. Analogously, some postpositional phrases with *-h* which have become fixed expressions can be considered as single ADV words. As a rule of thumb, they can be treated as single elements if Vergari – Vergari (2003) has a dedicated entry for them, as occurs, e.g., in the case of *rummah*, 'truly', but literally 'with truth'.

Sticking to Vergari – Vergari (2003), nouns and pronouns are lemmatized using the absolute form; verbs of classes 1, 2 and 4 are lemmatized using the first person singular of their perfect, since it is closer to the verbal stem and can be easily recognized; verbs of class 3 are lemmatized according to the form of the third person singular of the present.

3. Features and feature values

We prefer not to introduce new parts of speech (*xpos*) and therefore, when relevant, differences in the sub-parts of speech will be dealt in the field FEATS, which, in the UD scheme, contains pieces of information about a word's part-of-speech and morphosyntactic properties.

In order to account for the complexity of the number system of Saho nouns, we propose to expand the feature value inventory of UD by adding the options "Number = Gnrl" for the general number and "Number = Sgtv" for the singulative (a value which has also been proposed for the Beja treebank[5]). Indeed, Saho nouns fall into two subsystems: one exhibits the common opposition "singular : plural"; the other opposes a general number form, which is unspecified as to the "singular : plural" opposition, to a corresponding singulative which is used when the noun refers to one item, but is structurally marked as compared to the general number form (see Zaborski 1986: 21-53; Banti – Vergari 2023: 307). As a rule of thumb, the subsystem to which a noun belongs can be inferred from Vergari and Vergari's dictionary (2003): if the English rendering of a noun exhibits an optional "(s)" marker or some analogous expedient (e.g. "*zizzaale nf* honeybee(s)") then the noun belongs to the "general : singulative" subsystem, otherwise it can be treated as a common "singular : plural" noun. It should be mentioned, however, that the status of the "general : singulative" opposition as inflectional rather than derivational may be questioned[6]. Note that agreement targets such as verbs only exhibit the opposition "singular : plural".

The core cases of Saho are nominative, absolute and genitive, for which, following the UD guidelines[7], we use the values Nom, Acc and Gen respectively.

As far as verbal morphology is concerned, according to the use of UD, the specifications "Tense = Pres"[8] and "Tense = Past" will be used: these will also be applied to simultaneous and sequential converbs, respectively.

We use the polarity feature to distinguish negative verb forms from positive ones: while "Polarity = Neg" should be mandatorily present with negative verb forms such as the negative past or the negative imperative, the complementary specification "Polarity = Pos" may be omitted throughout.

---

[5] In fact, the Beja treebank adds a singulative feature, which is used to distinguish collective nouns that designate a single entity with a boolean value (Sgtv = Yes if it is a singulative).
See https://universaldependencies.org/bej/index.html

[6] For instance a "singular : plural" noun such as *fan* (M.SG) : *faanon* (M.PL) can also have a corresponding singulative, as *fanta* (F). Note also that a "general : singulative" noun such as *zizaale* can have more than one singulative counterparts: *zizzaaletto* (M) and *zizzaalettö* (F).

[7] See https://universaldependencies.org/u/feat/Case.html

[8] Note that in UD the use of the label "Pres" is also prescribed for non-past tenses.

Verbs of classes 1, 2 and 4 exhibit a gender distinction in the third singular: this will be accounted for by the values Masc or Fem of the feature Gender, that will be omitted for other persons and for verbs of class III, which have a single word form for the third singular both masculine and feminine.

The feature specification "VerbForm = Rel" may be reserved for cases in which the relative verb is formally different from the corresponding indicative mood. Otherwise, the field DEPREL will suffice to disambiguate between relative and non-relative uses of verbs.

Similarly to what has been proposed for the standard Romanian infinitives and negative imperatives[9], for which a short and a long variant exist, the distinction between the short and the long negative past (e.g. *malifo* vs. *maalifinna*, both meaning 'he/she did not close') can be accounted for by using the feature specification "Variant = Short" and "Variant = Long". The specification "Variant = Short" may also be used for signalling other verb variants, such as the common short form *le* of *leya* 'have' (Class 3 stative verb), e.g.:

```
4   le      leya    VERB    _
    Mood=Ind|Number=Sing|Person=3|Tense=Pres|Variant=Short|VerbClass=3    5
    acl:relcl   _   Gloss=have
```

A further use of "Variant = Short" is with the short variants of the enclitic postpositions (e.g. *-d* 'on, in' instead of *-dde*; cf. Banti – Vergari 2005, 123f.), while "Variant = Long" is used for the emphatic absolute forms of personal pronouns (*yoyya* as opposed to *yi*, cf. Banti – Vergari 2005: 115; Banti – Vergari 2023: 308).

3.1. Features related to part-of-speech sub-classes

As concerns the part-of-speech classes, distinctions worth being accounted for are those between non-stative and stative verbs and between bound and free cardinal numerals. Indeed, stative verbs (class 3 verbs in Banti – Vergari 2023 subgrouping) can be considered as a distinct sub-part-of-speech since they have a reduced inflectional potential: they inflect only for the non-past indicative, the simultaneous converb and the infinitive and do not have separate forms for masculine and feminine third singular.
Although not directly separating stative verbs from non-stative ones, this difference can be accounted for by the VerbClass feature: the values range from 1 to 4, each indicating the corresponding verb class according to the subgrouping proposed in Banti – Vergari (2023).

```
5   mango   mango   VERB    _
    Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbClass=3    6    acl:relcl   _
    Gloss=be.much
```

As for numerals, we use the value Bound of the feature Variant to account for the distinction between free and bound elements (tagged as "Variant = Bound"). Since the distribution between the two types is syntactically determined, the appropriate tag can be assigned even if the numeral is a Roman or Arabic digit. E.g. for *5* in *5 iggida* 'five years':

```
_   5   koona   NUM    _    NumForm=Digit|NumType=Card|Variant=Bound
    _   nummod     _    Gloss=five
```

[9] See https://universaldependencies.org/ro/feat/Variant.html

It can be noted, incidentally, that there seems to be no need for a special class of ideophones, as far as UD morphosyntactic annotation is involved. Indeed, even if Banti – Vergari (2023: 300) posit such a word class for Saho, its members only appear in compound verbs and other derivatives and never show as autonomous syntactic nodes: an example is *sik* in *sik-erhxe* 'to be silent'. One may also add that the boundary between such ideophone class and the noun class is blurry since full-fledged nouns may also appear as first elements of compound verbs: e.g., the feminine noun *cafuu* 'forgiveness' appears in the compound verb *cafuu-erhxe* 'to forgive'. On the other hand, the ideophone *sik* also appears in the derived noun *sikko* 'silence'.

It is, moreover, advisable to stick to Banti's terminology (2010) and to classify as converbs not only the invariable converbs of Central and Southern Saho but also the subject-agreeing simultaneous converbs of Northern Saho. Indeed, treating them as participles, as in Banti – Vergari (2005: 105), would miss the fact that they are typically used as elements depending on the predicate of the matrix clause rather than on the element with which they agree: e.g. the structure *aliftak ku juble* opening.2SG 2SG.OBL saw.3SG.M 'he saw you while you were opening it' (cf. Banti – Vergari 2023) can be compared with the Italian infinitival clause of *ti ha visto aprirlo*, in which *ti… aprirlo* functions as a complement clause (cf. Mensching 2017: 382). This does not mean that the language is not able to express object clauses proper; as for the Irob dialect, an example is that given by Reinisch (1878: 18 = 104): *kāy yígdifa-m úbẹla* 3SG.M.OBL kill.PST.3SG.M-NMLZ see.PST.1SG 'I saw that he killed him', where the complement clause is marked by the presence of the nominalizing suffix *-m* on the verb (which is also its last word; see §5.2). As concerns the dependency relation between the converb and its head, it will be classified as advcl ("adverbial clause") independently of the issue of whether it forms a complement clause or a true adverbial clause, in accordance with the use prescribed in the UD guidelines, where, e.g., in the annotation of the sentence *They heard about you missing classes* the gerund *missing* is treated as an advcl dependent of the main verb[10].

Note that for converbs we use the "VerbForm = Conv" value and the agreeing or non-agreeing nature of a converb is inferable from the presence or absence of agreement features in the field FEATS.

4. Auxiliaries

According to UD guidelines, we posit a separate part-of-speech class for auxiliaries. The lexemes *ine* 'to be, exist', *kinni* 'to be', *leya* 'to have' and *waye* 'to lack, miss' belong here when used as copulas and when, together with the subjunctive, the infinitive or a converb, they form compound tenses.

Ine and *kinni* are the only two lexemes that can be used as copulas: e.g., in *roble abraahim barha kinni* 'Roble is Ibrahim's daughter' *kinni* is tagged as AUX and is a dependent, through the aux relation, of the nominal predicate *barha* 'daughter'.

The past tense of the class 1 verb *ine* is also used, together with the simultaneous converb, to form the past tense of class 3 verbs: e.g., in *kixinii yine* 'he loved' the converb is considered the root of the sentence, while *yine*, tagged with upos = AUX, depends on it, through the aux relation.

[10] See https://universaldependencies.org/u/dep/advcl.html

Class 3 verbs *kinni* and *leya* are used, together with subjunctive forms, for the two futures of Saho, and are accordingly considered auxiliaries: e.g., *yamaatoona kinon* / *linon* 'they will come'.

Finally, in negative relative clauses, while Northern Saho uses a special negative relative paradigm unmarked for time reference (these forms are marked as "VerbForm = Rel"), Southern Saho (like Afar) uses the infinitive with the auxiliary *waye*; the infinitive followed by a form of *waye* is also used for the negative jussive and to negate verbs in conditional sentences (Banti – Vergari 2023: 314).

As already stated, these lexemes are ambiguous between AUX and VERB: in the sentence *zizzaale fantat lacnale makaano kixina* 'Bees like places that have intermediate temperatures' the form *le* (in *lacnale*) is a VERB (see the CoNLL-U line in §3), depending, through the acl:relcl relation, on the feminine noun *makaano* 'place', and having the feminine noun *lacna* 'temperature' as its direct object dependent.

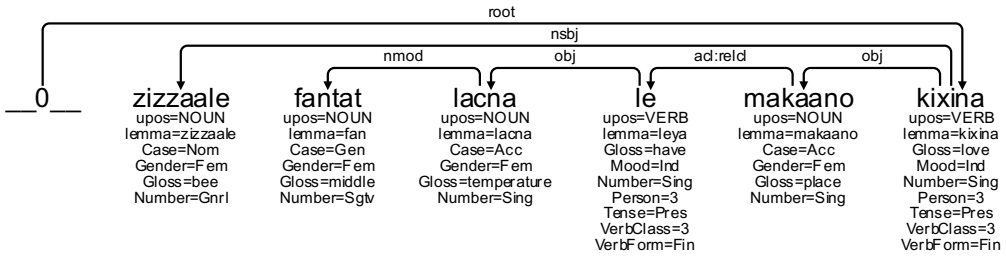

Figure 1. *zizzaale fantat lacnale makaano kixina* 'Bees like places that have intermediate temperatures' (from Banti - Axmadsacad Maxammad Cumar 2009: 106).

5. M-nominalisations

A relevant issue concerns the nominalizing suffix *-m* and the indefinite *tiya* m. (*tiyä* f., *mara* pl.), which are found both in what Parker and Hayward (1985: 287) call "M-nominalised clauses" and as the second element of independent pronoun forms. Although they are arguably historically interconnected, we propose to treat them separately.

5.1. Determiners and pronouns

We distinguish in Saho a class of determiners (upos = DET), which are only used as attributes and do not inflect for gender and number, from a class of pronouns (upos = PRON), which can be used independently. Each of these classes consists of possessive, demonstrative, indefinite and interrogative elements (distinguished through the different values of the feature "PronType" and through the feature "Poss", for possessives). Thus, for example, the possessives *ku* in *ku migac* 'your name' and *kuti* in *kuti yiggidile* 'yours is broken' are treated as two distinct lexemes (*ku* and *kutiya* respectively) with different POS tags:

1 ku ku DET _ Number = Sing|Person = 2|Poss = Yes|PronType = Prs 2 det:poss _ _

1 kuti kutiya PRON _ Case = Nom|Gender = Masc|Number = Sing|Number[psor] = Sing|Person = 2|Poss = Yes|PronType = Prs 2 nsubj _ _

As far as the feature layer is concerned, while possessive determiners are only tagged for the values of the possessor, possessive pronouns may have two different genders and

numbers: that of the possessed object (triggering agreement on the verb if it is a subject) and that of the possessor[11], indicated by *psor* in square brackets after the relevant feature attribute. In this way we can account for the differences between, e. g., *teetiya* (Case = Acc|Gender = Masc|Gender[psor] = Fem|Number = Sing|Number[psor] = Sing|Person = 3|PronType = Prs|Poss = Yes) and *kaatiyä* (Case = Acc|Gender = Fem|Gender[psor] = Masc|Number = Sing|Number[psor] = Sing|Person = 3|PronType = Prs|Poss = Yes)[12]. In the plural there is no gender distinction, but according to Banti and Vergari (2005: 116-17) forms ending in *-mara* are used for human referents, while forms ending in *-m* are used for non-human ones. In our annotation scheme this can be accounted for by adding the feature "Animacy", with values "Hum" and "Nhum" for humans and non-humans respectively[13], so that *yimara* (Animacy = Hum|Case = Acc|Number = Plur|Number[psor] = Sing|Person = 1|PronType = Prs|Poss = Yes) can be distinguished from *yim* (Animacy = Nhum|Case = Acc|Number = Plur|Number[psor] = Sing|Person = 1|PronType = Prs|Poss = Yes). Some pronouns also show the "singulative : general" opposition, e.g. *aketto* 'another one, m.', *akettö* f., *akim* pl. (Banti – Vergari 2005: 119)[14].

5.2. Subordination

The nominalizing suffix *-m* and the indefinite *tiya* (*tiyä* f., *mara* pl.) are also found in structures corresponding to free relative clauses ("headless relative clauses" in Banti – Vergari 2023: 317) in other languages, although in Saho these can be used in a variety of contexts. The analysis we propose is somehow different from the ones found in previous literature, since we treat the two elements in a different way.

The indefinite *tiya* is analysed as an indefinite pronoun representing the element modified by the relative clause in cases like the one in Figure 2, where the verbal form *orbishshe* is the dependent of the pronoun *tiya* (nominative *ti*), through the acl:relcl relation.

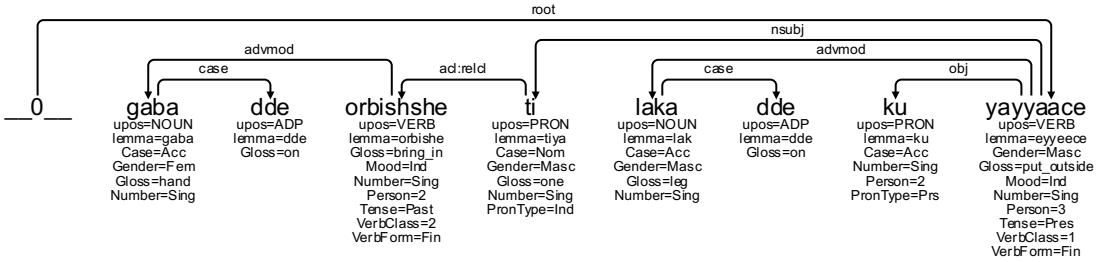

Figure 2. *Gabadde orbishshe ti lakadde ku yayyaashe* 'Who you brought in with your hand kicks you out with his foot' (from Banti – Vergari 2023: 312).

[11] The same distinction could apply to the Person feature, since the Person value tagged is always the possessor's one and hides the fact that nominative possessive pronouns always trigger a third person agreement on the verb. Nonetheless, we do not use it, since its usage is discouraged in UD: the issue is discussed in https://universaldependencies.org/u/overview/feat-layers.html

[12] Note that the "Gender[psor]" feature is relevant only when the possessor is a third person singular.

[13] In Irob (Southern Saho of Ethiopia), there seem to be a parallel distinction also in the singular, where the forms ending in *-tiya/-tiyä* are used for human referents, while a form in *-iyya* is used for non-human ones (Esayas Tajebe 2015: 150).

[14] The corresponding DET is *aki* 'other'. The singulative forms of the pronoun *aketto*, *akettö* seem lexicalized, since the element *-tto*, when it functions as a singulative suffix, does not trigger a change in the vowel, see e.g. *alaaki* 'specie(s) of shrub(s)' ~ *alaakitto* 'seed, fruit'. For similar phenomena in Afar see Parker and Hayward (1985: 237 n.; 238 n.2).

In this way the annotation does not differ from a relative clause modifying a noun (cf. Figure 1 in §4); in fact, the only difference seems to be that relative clauses with *tiya* can be used in pseudoclefts, as in Figure 3, where the copula, according to UD guidelines, is a dependent of the nominal predicate (in this case *miyatto*) and the pronoun in the nominative (*ti*) is considered the subject of the sentence.

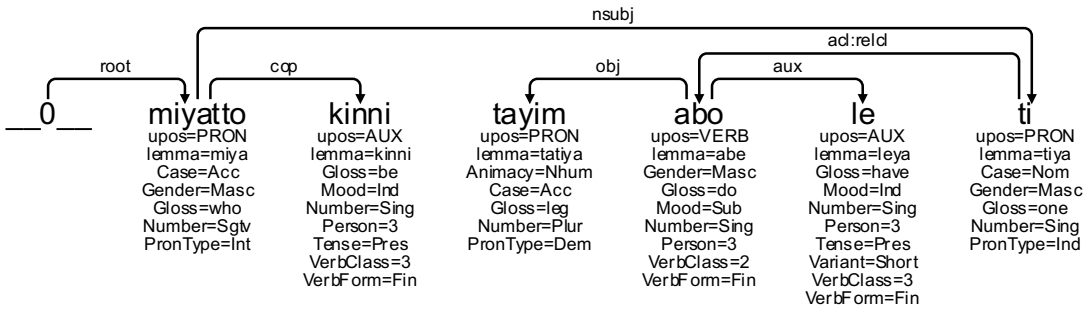

Figure 3. *Miyatto kinni tayim abo le ti* 'Who is it who will do this?' (from Banti – Vergari 2023: 314).

If a relative clause follows its head, *-ya* is suffixed to the verb: this is treated as a relativizing subordinating conjunction and considered a dependent of the verb through the mark relation, as in Figure 4.

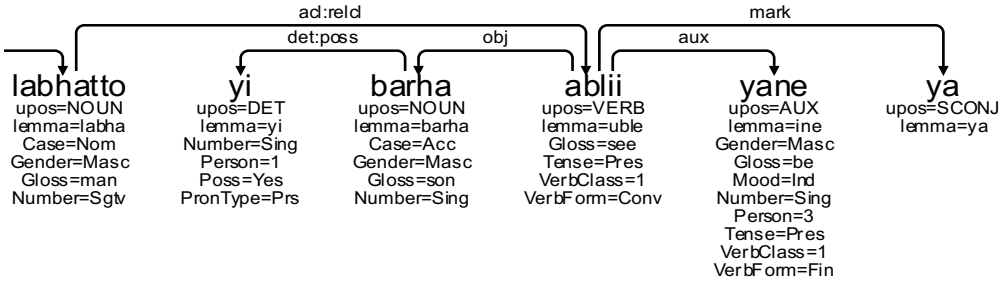

Figure 4. *Labhatto yi barha ablii yaneya* 'The man who is seeing my son' (from Banti – Vergari 2023: 317).

In a similar way, the nominalizing suffix *-m* is treated as a subordinating conjunction (SCONJ), as in Figure 5: note that a verbal form modified by *-m* can have a determiner as a dependent, as is the case with infinitives in Italian (*il dire*).

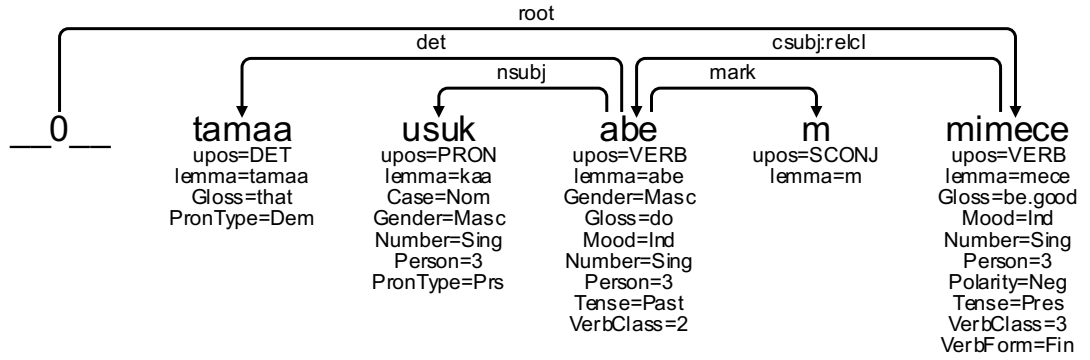

Figure 5. *Tamaa usuk abem mimece* 'What he did isn't good' (from Banti – Vergari 2023: 317).

Most importantly, these relative clauses can be used as sentential complements (see Figure 6 for a subjective clause, and Figure 7 for an objective clause): we use the

dependency relations csubj:relcl, ccomp:relcl and xcomp:relcl in order to distinguish them from the complement clauses with the subjunctive (Figure 8). In UD the relation xcomp is reserved to clausal complements whose subject is controlled (that is, must be the same as the higher subject or object, with no other possible interpretation, as is the case in Figure 7 with *aba* and *farha*), while the ccomp relation is reserved to all the other cases (e.g., *mece* in Figure 7).

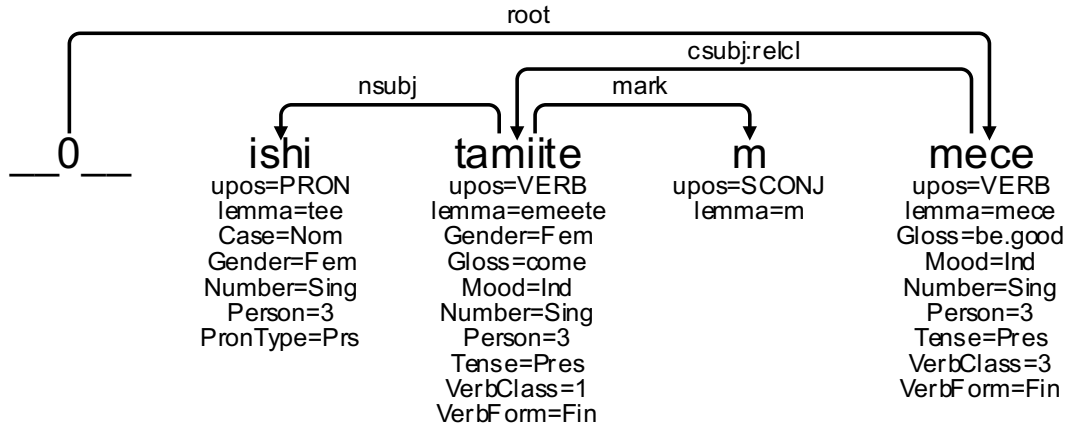

Figure 6. *Ishi tamiitem mece* 'It is good that she comes' (from Banti – Vergari 2023: 318).

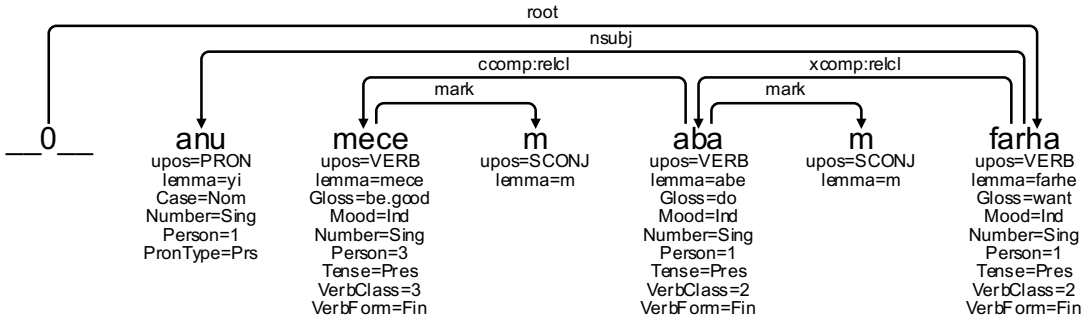

Figure 7. *Anu mecem abam farha* 'I want to do what is good' (from Banti – Vergari 2023: 318).

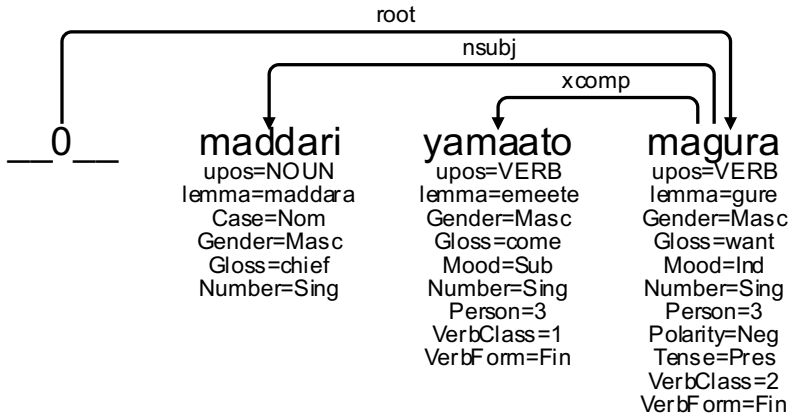

Figure 8. *Maddari yamaato magura* 'The chief doesn't want to come' (from Esayas Tajebe 2015: 327).

Followed by a postposition, free relative clauses with *-m* can function as subordinate adverbial clauses: in these cases, we use the advcl:relcl relation, and both the *-m* and the postposition are considered dependent on the verb, with the mark relation, as in Figure 9.

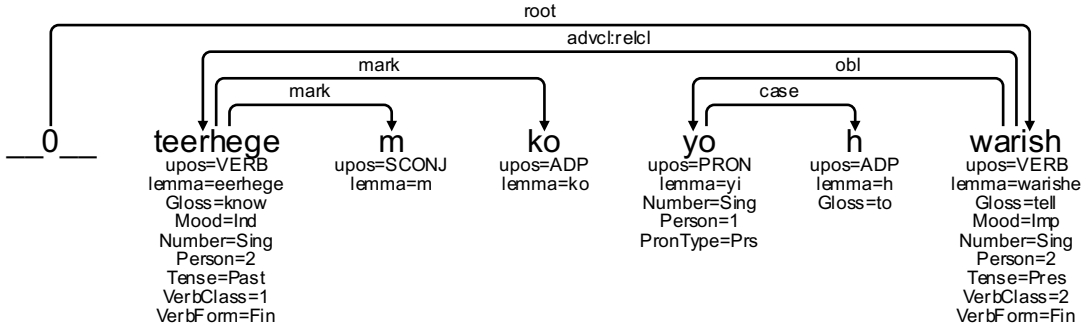

Figure 9. *Teerhegemko yoh warish* 'If you know it, tell me!' (from Banti – Vergari 2023: 318).

Sometimes the nominalizing *-m* is omitted (the phenomenon is widespread in Afar, according to Lamberti 1990: 155): in these cases, the verb of the subordinate clause is considered a dependent of the main verb through the advcl relation and the postposition receives the mark relation (Figure 10).

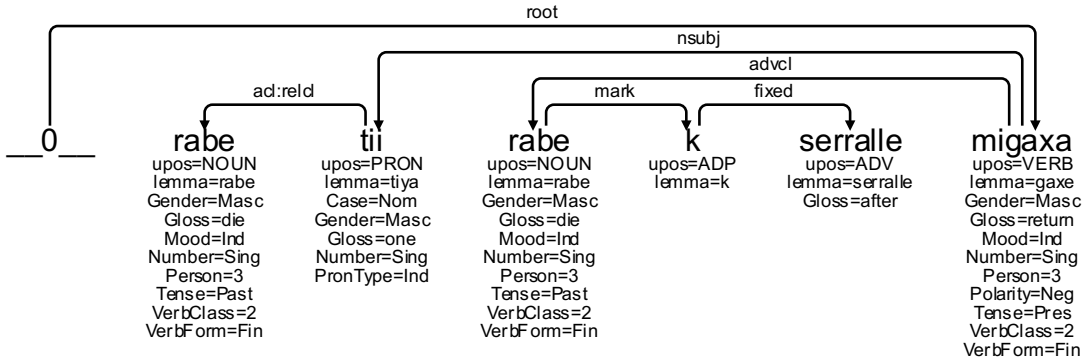

Figure 10. *Rabe tii rabek saralle migaxa* 'Whoever dies will not come back after death' (from Banti – Vergari 2023: 318).

Although temporal clauses with *gedda*/*ged* 'time' in Northern Saho or *gul* in Southern Saho can be treated as relative clauses modifying the noun *gedda/gul* (Lamberti 1990: 151-58; Banti – Vergari 2023: 318), we consider them adverbial clauses introduced by a subordinating conjunction *gedda/gul* 'when' (lemmatised as such in Vergari – Vergari 2003):

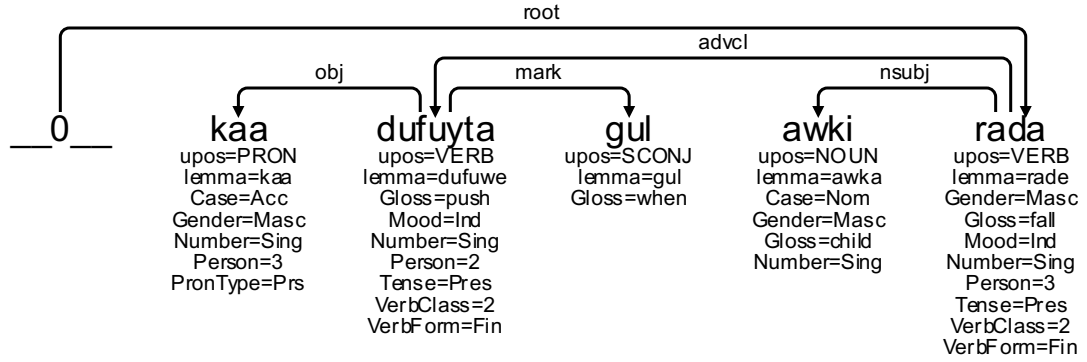

Figure 11. *Kaa dufuyta gul awki rada* 'When you push him, the child falls' (from Esayas Tajebe 2015: 288).

As already discussed in §3.1 temporal clauses are expressed also through converbs.

6. Final remarks

In this paper, we have addressed some preliminary issues on the preparation of a Saho treebank within the framework of Universal Dependencies. We have discussed some problems concerning tokenization, morphology, and syntax. The challenge, in this respect, was twofold. On the one hand, to account for morphological and part-of-speech peculiarities of the Saho language, we had to expand the feature and feature value inventories of UD. In this process, we tried to stick to features already used in other treebanks, especially in the Beja one, since Beja is the only Cushitic language at present available in UD. On the other hand, to stay within the architecture underlying the UD framework, significant deviations from the descriptions proposed in the literature were necessary in some cases.

This effort is useful and necessary both because it will enrich and increase the flexibility of UD, providing new material from a language belonging to an underrepresented family, and because it will make Saho data (for example, the material in the Saho Corpus; cf. Jama Musse Jama 2022) more accessible to the wider scientific community. Finally, although the specimen presented here primarily uses data from the Northern Saho variety described in Banti and Vergari (2023), it is hoped that a treebank of Saho-Afar could contain material from the different dialect groups of these languages: in the CoNLL-U format, the multilingual nature of such a treebank can be easily dealt with through sentence-level comments (lines starting with a hash), where metadata providing information about the dialectal provenance of a text can be stored.

References

BANTI, Giorgio (2010) "Remarks on the typology of converbs and their functional equivalents in East Cushitic". In VÖLLMIN, Sasha, Azeb AMHA, Christian RAPOLD, and Silvia ZAUG-CORETTI (eds.), *Converbs, Medial Verbs, Clause Chaining and Related Issues*. Pp. 31-80. Köln: Köppe.

BANTI, Giorgio, and AXMADSACAD MAXAMMAD CUMAR (2009) "A few Saho texts about bees and honey", *Ethnorêma* 5:89-108.

BANTI, Giorgio, and Moreno VERGARI (2005) "A Sketch of Saho Grammar", *Journal of Eritrean Studies* 1-2: 100-131.

BANTI, Giorgio, and Moreno VERGARI (2023) "Saaho". In MEYER, Ronny, Bedilu WAKJIRA, and Zelealem LEYEWM (eds.), *The Oxford Handbook of Ethiopian Languages*. Pp. 294-319. Oxford: Oxford University Press.

DE MARNEFFE, Marie-Catherine, Christopher MANNING, Joakim NIVRE, and Daniel ZEMAN (2021) "Universal Dependencies", *Computational Linguistics*, 47(2): 255-308.

ESAYAS TAJEBE (2015) *Descriptive Grammar of Saaho.* PhD dissertation.

JAMA MUSSE JAMA (2022) "Saho Corpus: Semi-automation of Verb Conjugation in Saho: Verbs Class I", *Ethnorêma* 18: 69-87.

KAHANE, Sylvain, Martine VANHOVE, Rayan ZIANE, and Bruno GUILLAUME (2021) "A morph-based and a word-based treebank for Beja". In DAKOTA, Daniel, Kilian EVANG, and Sandra KÜBLER (eds.), *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021).* Pp. 48–60. Sofia: Association for Computational Linguistics.

LAMBERTI, Marcello (1990) "Some word order principles of Saho-Afar", *Rassegna di Studi Etiopici* 34: 127-168.

MORIN, Didier (1995) *«Des paroles douces comme la soie». Introduction aux contes dans l'aire couchitique (bedja, afar, saho, somali).* Paris: Peeters.

REINISCH, Leo (1878) "Die Sahosprache", *Zeitschrift der Deutschen Morgenländischen Gesellschaft* 32: 415-464.

VERGARI, Moreno, and Roberta VERGARI (2003), *A basic Saho-English-Italian dictionary*. Asmara: Sabur Printing Services.

ZABORSKI, Andrzej (1986) *The Morphology of Nominal Plural in the Cushitic Languages*. Wien.