A morph-based and a word-based treebank for Beja

Sylvain Kahane, Martine Vanhove, Rayan Ziane, Bruno Guillaume

A morph-based and a word-based treebank for Beja

Sylvain Kahane*, Martine Vanhove, Rayan Ziane***, Bruno Guillaume******
*Modyco, Université Paris Nanterre & CNRS
**Llacan, CNRS, INALCO, & EPHE
***Llacan, CNRS, INALCO, & Université d'Orléans
****Sémagramme, INRIA Nancy Grand Est

Abstract

The paper presents the first UD treebank for Beja, a Cushitic language spoken in Sudan. It has been built from the conversion and enhancement of an Interlinear Glossed Text (IGT). The paper's objectives are three-fold: we explain our choice to use a morph-based annotation and its consequences, we describe the processing chain from an IGT to a morph-based dependency treebank and a word-based treebank, and we present several interesting constructions in Beja.

1 Introduction

This paper presents a small treebank for Beja, a Cushitic language spoken in Sudan. Initially developed in SUD (Surface-Syntactic Universal Dependencies) (Gerdes et al. 2018, 2019, 2021), the treebank is also available in UD (de Marneffe et al. 2021). It has been built from an Interlinear Glossed Text (IGT) (Comrie et al. 2008) developed by Martine Vanhove (2014). The original corpus contains 5899 words and 12507 morphs representing slightly less than one hour of recordings divided into 18 files. Two files, containing 1101 morphs, 418 stems, and 56 sentences, have been completely annotated and constitute the UD2.8 Beja-NSC treebank.[1]

One of our goals was to avoid losing information contained in the original resource, which led us to adopt a morph-based tokenization rather than a word-based annotation.

The paper focuses on three aspects of developing this treebank. Section 2 presents the Beja corpus and the IGT annotation we started with, the UD annotation scheme, and the adjustments to it which were necessary for our morph-based annotation. An overview of the conversion to a word-based treebank is also provided. Section 3 explains the processing chain from an IGT to a UD treebank and the optimization of this chain. Section 4 introduces some challenges faced during the syntactic annotation of Beja.

2 A morph-based annotation for Beja

2.1 Beja and CorpAfroAs

Beja is the sole member of the North Cushitic branch of the Afroasiatic phylum. It is mostly spoken in eastern Sudan, as well as in southern Egypt and northern Eritrea. In Sudan, the country where the data were collected, the number of speakers is about 2,000,000, but the language has no official recognition and exists purely as an oral language. As explained by Martine Vanhove (2006), Beja is not poorly described compared to other Sudanese languages, and the most recent grammar, published in French, goes back to 2017 (Vanhove 2017). However, some elements required for a complete description of the language remain unavailable.

The data used for the development of this treebank comes from the CorpAfroAs project (Mettouchi & Chanard 2010). CorpAfroAs is a multilingual corpus which aimed at providing a structured database of natural records of Afroasiatic languages, transcribed, translated and annotated to allow for complex

[1] The first version of the treebank published on May 1st, 2021, for UD2.8, and released on November 1st, 2021, for UD2.9, is a morph-based treebank. New modifications have been done for the publication of this article that will be incorporated for UD2.10 on May 1st, 2022. We also plan to publish the word-based version of the treebank on the same occasion.

requests. CorpAfroAs is organized around two axes: prosodic analysis and morphosyntactic glossing. It is this morphosyntactic glossing that served as the raw material for our work.

2.2 Beja's Interlinear Glossed Text

All CorpAfroAs corpora use a common format for IGTs presented in Comrie (2015).

(1) *w=ʔi:d arraf-i / a-di=t a- ba i-ni //*
DEF=Aid congratulate-AOR.1SG 1SG-say\PFV=COORD 1SG-go 3SG.M-say\PFV
'I went to wish him a blessed Aid, he said.'

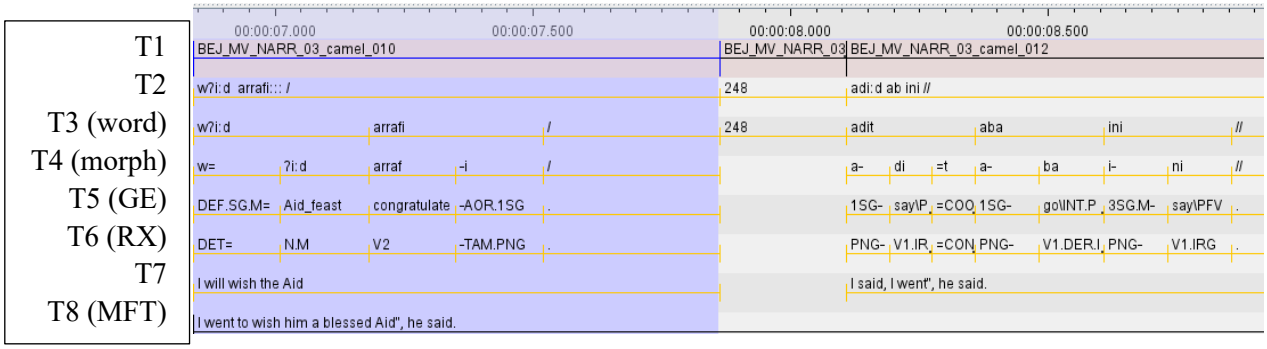


Figure 1. ELAN screenshot of the CorpAfroAs IGT for (1)

The top of Figure 1 contains the timeline. Tier 1 is the segmentation into prosodic units. Tier 2 is a broad phonetic transcription; / marks a minor prosodic break (rising final continuative intonation) and // a major prosodic break (falling final intonation). Tier 3 contains the prosodic words and Tier 4 a tokenization into the smaller units with a non-segmentable signifier (i.e. the morphs, as defined by Haspelmath 2020). The tokens considered are lexical stems, inflected forms of stems (when the inflectional "morphemes" are not affixes), inflectional affixes, and derivational affixes. For the sake of simplicity, we will call these tokens *morphs*, even if some of them are a combination of morphemes, such as *ni*, which is the perfective form of the stem of the verb *di* 'to say'. This inflected stem also combines with a prefix *i-*, which is a subject agreement morph, giving the verbal form *ini* at the end of example (1).

Tier 5, labelled GE, is a gloss. Tier 6, labelled RX, contains morphosyntactic features, including POS (DET, N, V1…) and inflectional categories (TAM, PNG for Person-Number-Gender, …).

Tier 7 is a translation of each prosodic unit and Tier 8, called the "major free translation" (MFT), is a translation based on larger units, allowing for better translations. It is this last tier which was used as the basis for the sentence segmentation. Since the end of each MFT unit does not necessarily corresponds to the end of a sentence, understood as a coherent syntactico-semantic unit, we copied the MFT tier (Mft-cp) on which we signaled the end of each sentence by a # sign (Figure 2).[2]

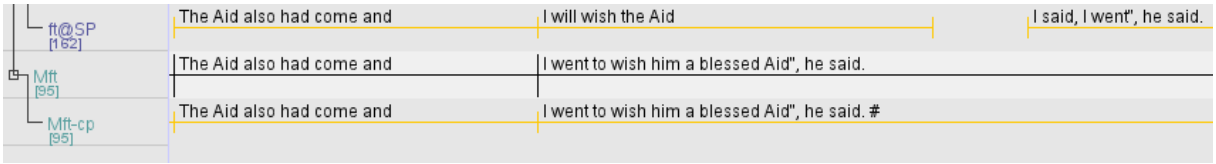


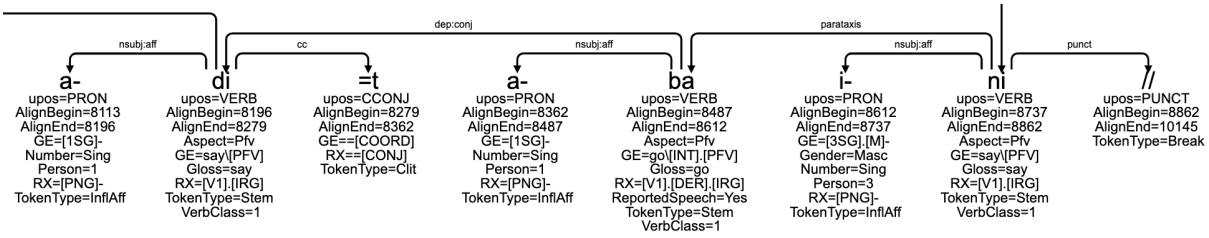
Figure 2. ELAN screenshot of the sentence segmentation

The IGT does not contain lemmas (which supposes a more advanced description and a lexicon), but contains glosses and translations. The corpus contains a time alignment and each IGT sample is coupled with a sound file accessible on the CorpAfroAs website.

[2] The segmentation of spoken corpora into major syntactic units (often called *sentences*, even if the notion can be problematic for spoken production) is a complex question that will not be addressed here. See Kahane et al. (2021) for some guidelines and Pietrandrea et al. (2014) for a more comprehensive study.

2.3 A morph-based annotation scheme for SUD and UD

We use the "morph" segmentation for tokenization. The content of the tiers GE and RX is kept in features GE and RX. The time alignment gives us the features AlignBegin and AlignEnd of each token, including the prosodic breaks (Figure 3) (see Kahane et al. 2021 for the conventions we use for spoken corpora).



'I said and I went, he said'

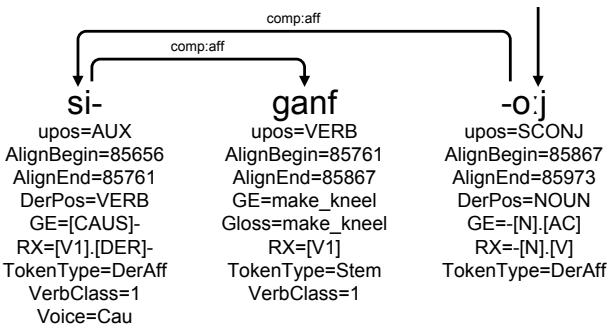
Figure 3. The UD annotation for the end of (1)

Some pieces of information of the GE tier are used to instantiate morphosyntactic features, such as Number, Person, Gender, Aspect, Definiteness… (see Section 3.2). A TokenType is added on each token, with five values: Stem for stems (including some inflected stems), Clit for clitics, InflAff for inflectional affixes, DerAff for derivational affixes, and Break for prosodic breaks.

In addition to the usual #text and #sent_id features (see online UD guidelines or de Marneffe et al. 2021), the metadata contain a #sound_url feature for the URL of the sound file corresponding to the transcription and #phonetic_text for the phonetic transcription from Tier 2 (Kahane et al. 2021). The #text value, which is the concatenation of the tokens, including simple and double hyphens, is distinct from the #phonetic_text value.

We decided to give each token a POS as if it was a word. In consequence, pronominal affixes are PRON, verb nominalizers are SCONJ, TAM or causatives are AUX, case markers are ADP, plurals are DET, and purely phonetic signs are PART.

In our SUD annotation, SCONJ, ADP, and AUX affixes are treated as governors of their base, as if they were words (Gerdes et al. 2018). We use the corresponding SUD syntactic relation with the *aff* extension: In other words, subject pronominal affixes are *subj:aff* of their base, plurals are *det:aff*, while for SCONJ, ADP, or AUX, the base is *comp:aff* of the affix. Figure 4 gives the example of the deverbal noun *siganfo:j* 'settling' where the verbal stem *ganf* 'make kneel' combines with two derivational affixes, the causative prefix *si-* and the nominalizer *-o:j*,



'settling'

Figure 4. Derivational affixes (SUD-style)

Note that the fact that some affixes are treated as heads allows us to indicate in which order they combine. In the case of *siganfo:j*, *ganf* combine first with the causative *si-* and then with the nominalizer *-o:j*. Derivational affixes receive an additional feature *DerPos* indicating the POS of the derived form: AUX affixes have a DerPos=VERB and SCONJ affixes have a DerPos=NOUN. For inflected forms, the POS of the inflected form remains the POS of the stem. Note also that in the case of an inflected

form, the base can have its own dependents, while in the case of a derived form all dependents are on the derivational affix.[3]

In the UD version of the morph-based, all affixes are dependent of the stem. We use the UD syntactic relation corresponding to their functional role, with an additional *aff* extension: subject pronominal affixes are *nsubj:aff*, case markers are *case:aff*, nominalizers are *mark:aff*, AUX affixes are *aux:aff*, plurals are *det:aff*. Figure 5 gives the UD version of the two words of Figure 4. (See Figure 3 of the example of pronominal affixes in UD-style.) Note that the order in which the affixes combine with the stem is lost in the UD version. This is a problem for the conversion to the word-based version, because we cannot easily determine whether *siganfo:j* is a noun or a verb (see Gerdes et al. 2021 for a similar discussion about the fact that UD underspecifies the internal structures of nuclei). For this reason, the word-based UD version of the treebank is derived from the morph-based and word-based SUD versions and not from the morph-based UD version.

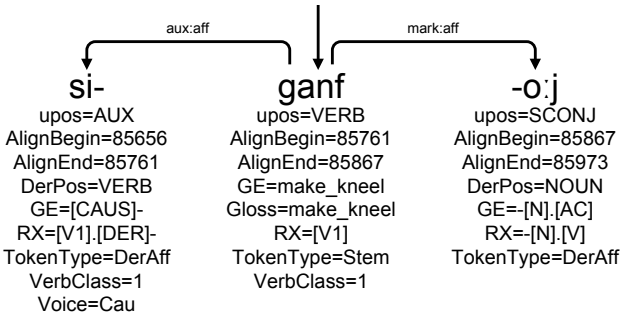


Figure 5. Derivational affixes and inflectional affixes (UD-style)

2.4 From a morph-based treebank to a word-based treebank and back

UD theoretically requires treebanks to be word-based. However, we think that the fact that our treebank is morph-based does not pose a problem because the morph-based annotation is explicit, due to the different features we introduced (*TokenType* on tokens and *aff* on relations) and because it is not difficult to merge every stem with its affixes to obtain a word-based treebank. Note that the question to have morph-based (generally called morpheme-based) treebank rather than word-based treebank has been discussed several times for different languages: see Tsarfaty & Goldberg (2008) for Modern Hebrew, Vincze et al. (2010) for Hungarian, Zhan et al. (2014) for Chinese, or Park (2017) for Korean.[4]

For the conversion into a word-based treebank, the lists of morphosyntactic features attached to the different parts of a word must be merged in different ways. Some features had to be concatenated, such as the feature *form*, containing the form, and the features *GE* and *RX* containing the morphosyntactic glosses. See Figure 6 for the word-based version of Figures 4 and 5.

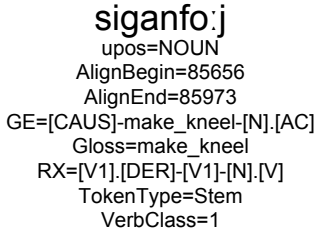


Figure 6. The word-based annotation of the derived word *siganfo:j*

[3] Compare in English:

(i) *He cleaning the table was impressive.*
(ii) *His cleaning of the table was impressive.*

In (i), *cleaning* is an inflected form and *clean* can have a subject and a direct object, while, in (ii), *cleaning* is a derived noun with a determiner and a noun complement.

[4] We would like to thank one of our reviewers, who pointed out that our segmentation was morph-based rather than morpheme-based.

The most challenging feature is *upos* (the UD feature for the "universal" POS) for derived forms. Thanks to the SUD annotation where the derivational affixes are head and to the *DerPos* feature, it becomes trivial to compute the *upos* of a derived form: it is the *DerPos* of the topmost derivational affix. Except for the features *form, GE,* and *RX*, which are concatenated, and *upos*, which is replaced by *DerPos*, the features of the derived form are the features of the topmost derivational affix.

For inflected forms, the *upos* of the word is the *upos* of the stem. The features *form, GE,* and *RX* are concatenated as for derived forms, but contrary to derived forms, other features are unified for inflected forms. Figure 7 shows the word-based version of the morph-based analysis of Figure 3, where we can see that the *Person* and *Number* features coming from the pronominal affixes are reported on the verb forms.

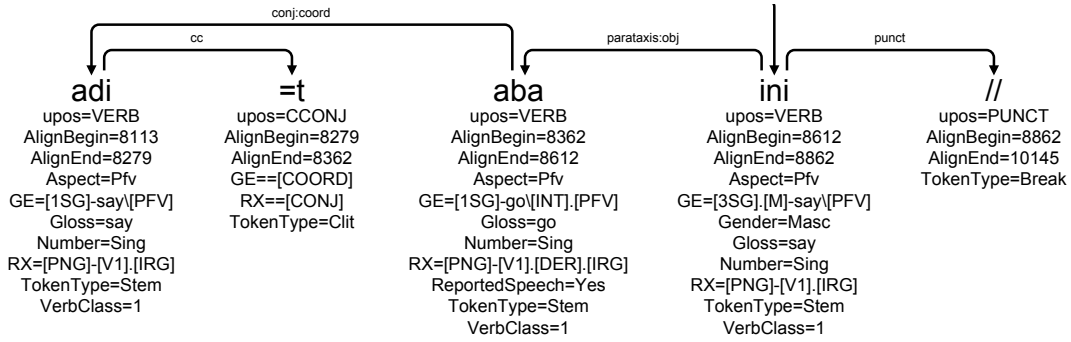
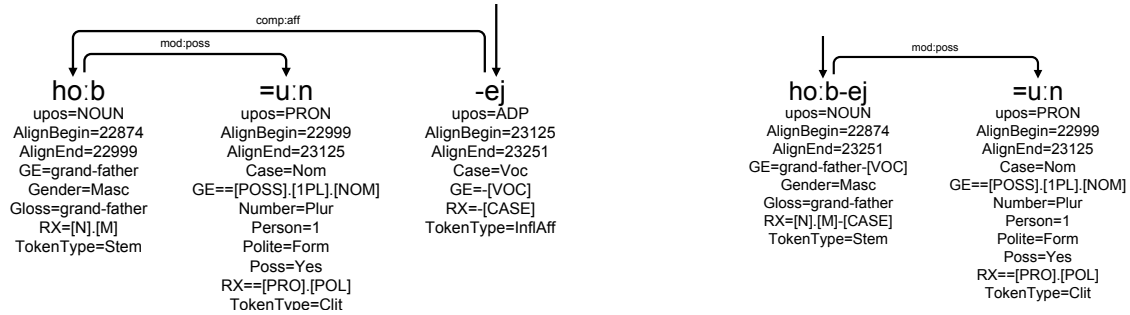


Figure 7. The word-based UD annotation for the end of (1)

We faced one unexpected problem with a clitic placed between a stem and an inflectional affix. We analyzed this case as an amalgam with only one word corresponding to two lexemes (Figure 8).



'our grand-father' 'our grand-father'

Figure 8. The morph-based and word-based SUD annotation of an incorporated clitic

We must also compute the *AlignBegin* and *AlignEnd* features of words, which are easily deductible from the corresponding features of the morphs. Note that in the word-based version of the treebank, some information is lost and it will not be possible to recover the segmentation into morphs, as well as the features associated to this morphs. This is why we decided to distribute the Beja UD@2.8 treebank in a morph-based version. On our side, we will maintain the morph-based SUD version, which is the most informative one (see in particular the discussion of Section 3 when the word contains two derivational affixes).

3 From IGT to UD

The construction of the UD treebank was carried out in three steps: the conversion of the IGT into a CoNLL-U (Section 3.1); the automatic pre-annotation by enrichment of the CoNNL-U (Section 3.2); the manual SUD annotation, the conversion to UD and the validation of the treebank (Section 3.3).

3.1 From IGT to CoNLL-U

The first obstacle to the conversion to CoNLL-U is the fact that the corpus is segmented into prosodic units that do not necessarily correspond to syntactic units. We decided to base our major segmentation on the "major free translation" segmentation, which corresponds more or less to illocutionary units as illustrated in Figure 2 in section 2.2. The tokenisation is based on the segmentation into "morphs".

Once the choice of the tiers for the tokenization and the segmentation into sentences is made, the conversion of the IGT to a CoNLL-U is straightforward and loses no information from the IGT format. For each token, the time alignment is stored in the features *AlignBegin* and *AlignEnd*, the content of the GE and RX tiers is stored in the features *GE* and *RX*.

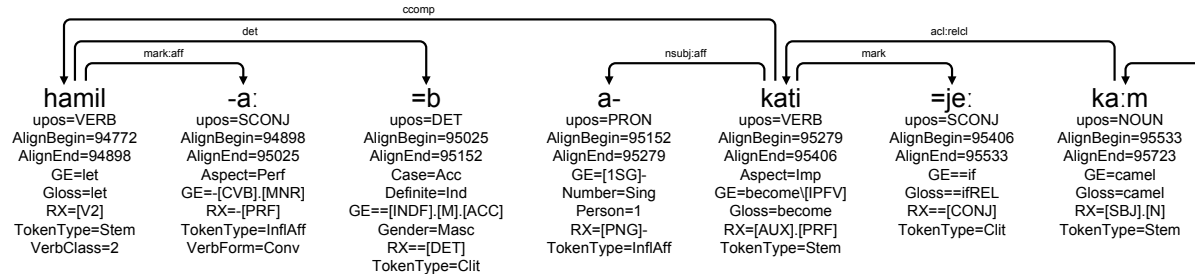
3.2 Automatic pre-annotation

The first CoNLL-U we obtain is almost similar to the IGT. The second step consists in enriching this CoNLL-U by transferring the content of the GE and RX tiers into UD features in order to fit the UD annotation scheme. The annotation specific to the morph-based level was introduced entirely automatically.

As the GE and RX formats of CorpAfroAs IGTs are enriched versions of the Leipzig Glossing Rules (Comrie 2015), they allow us to infer the UD POS and all the UD morphosyntactic features that must be associated with the tokens. We built a lexicon that proposes a translation into a UD feature for each label used in the GE and RX tiers. It was also possible to infer the syntactic relation for many tokens. The Grew tool (Guillaume, 2021), through its graph rewriting function, makes it possible to write a grammar of rules matching elements within dependency trees.

The feature *TokenType*, which distinguishes stems, affixes, clitics, and prosodic breaks, is based on the form of the token: As usual in IGTs, affixes have a hyphen (*a-* or *-a*) and clitics a double hyphen (*ba=* or *=i*), while prosodic breaks are assigned to special symbols (/ and //). For affixes and clitics, the governor was the closest stem and the positions of the hyphens indicate if the stem occurs after or before them. The distinction between inflectional and derivational affixes can be computed from the syntactic *RX* feature (most derivational affixes have a DER value in RX).

The syntactic label set of CorpoAfroAs, corresponding to the RX tier, is richer than the UD *upos* set of POS and the POS conversion was trivial for most of the labels. For instance, the labels V1, V2, LV, and IRG are all converted to the VERB *upos* tag. In order not to lose information, V1 and V2 receive a *VerbClass* feature with values 1 and 2 according to the original label. LV is provided with a VerbType=Light feature. In a similar way, the label DEM is converted into a DET *upos* with the PronType=Dem feature. The label REL for relativizers gives us an SCONJ *upos*, as well as the SUD relation *mod@relcl* (translated into UD *acl:relcl*). Moreover, due to the head-final behaviour of Beja, the relativizer can be linked to the verb preceding it (Figure 9).[5]



'a camel that I had let loose'

Figure 9. Relative clause (UD-style)

[5] This figure, as well as all the following figures, is extracted from the morph-based UD version of the treebank, revised for the paper, and which will be distributed on May 1st, 2022, for the UD2.10 release. All the SUD versions of the treebanks are available on the SUD website https://surfacesyntacticud.github.io/ and all versions can be requested on the Grew-match website (the latest versions of UD treebanks converted from SUD treebanks are at the end of the list of UD treebanks).

3.3 Manual SUD annotation and conversion from SUD to UD

The manual annotation was carried out in the SUD annotation scheme by a linguist specializing in Beja (Martine Vanhove), with the help of a specialist of treebank annotation (Sylvain Kahane) and a master student in NLP (Rayan Ziane), as well as some feedback from two native speakers (Ahmed Mohamed-Tahir Hamid and Mohamed-Tahir Hamid Ahmed). It was not possible to have a double annotation for this language. Some problems of analysis we faced during the annotation process are discussed in the next section. For the conversion from SUD to UD, we had to customize the conversion of the relations introduced for the affixes. The fact that UD forces the coordination relation *conj* to be head-initial was also a problem and SUD head-final *conj* relations were converted into an ad hoc *dep:conj* UD relation (see Section 4.2). The different conversions were mastered by Rayan Ziane and Bruno Guillaume.

4 Some constructions of Beja

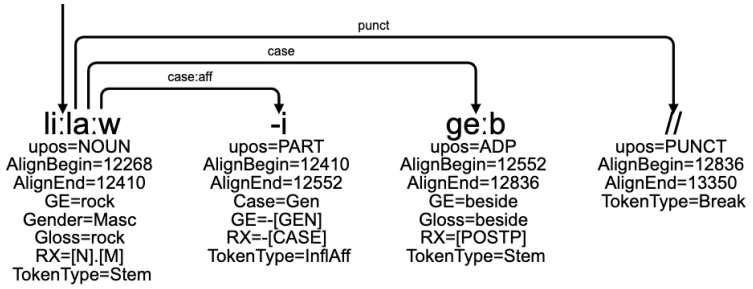
Below, we discuss four features of Beja syntax: affixes and word order (4.1), coordination (4.2), relative clauses (4.3) and serial verb constructions (4.4).

4.1 Affix and word order

The Beja treebank contains 684 words if we count both stems and clitics; 39% of words are clitics. The treebank contains 244 affixes for 418 stems, or a proportion of 58%. 59% of them are suffixes and 41% prefixes. 88% of the affixes are on verbs, 7% on nouns, and 5% on auxiliaries. All affixes on nouns are suffixes. Not all inflectional morphemes are affixes: 44% of the stems are in fact inflected forms, containing an inseparable inflectional morpheme, which increases the proportion of inflectional morphemes to 102% (102 inflectional morphemes for 100 stems).

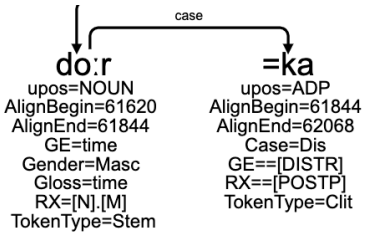
Beja is a head-final language: only 11% of the dependencies between two stems have the governor before the dependent in the SUD version of the treebank. Among the 31 dependencies concerned, 11 are for modifiers, 6 for discourse markers, 4 for dislocated objects, 2 for objects in canonical position, and 2 for determiners. Clitics occur on both sides: 47% are proclitics and 53% are enclitics. Clitics are mainly on verbs (56%) and nouns (38%). Clitics on nouns are determiners (70%), possessives (15%), postpositions (11%), and coordinating conjunctions (3%). Clitics on verbs are subordinating conjunctions (35%), object pronouns (25%), coordinating conjunctions (14%), an optative particle (2%), and, on nominalized forms, determiners (15%) and copulas (8%).

Beja adpositions are postpositions, either as independent words (Figure 10) or as enclitics (Figure 11):



'next to a rock'

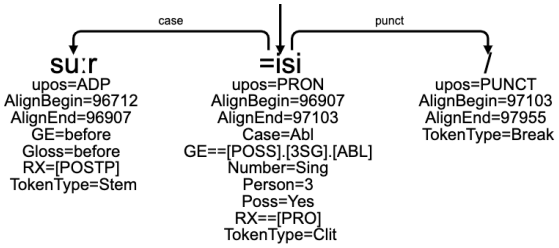
Figure 10. A Beja independent postposition (UD-style)



'in turn'

Figure 11. A Beja enclitic postposition (UD-style)

When the postposition complement is a pronoun, it is an enclitic and the postposition precedes its complement (Figure 12):

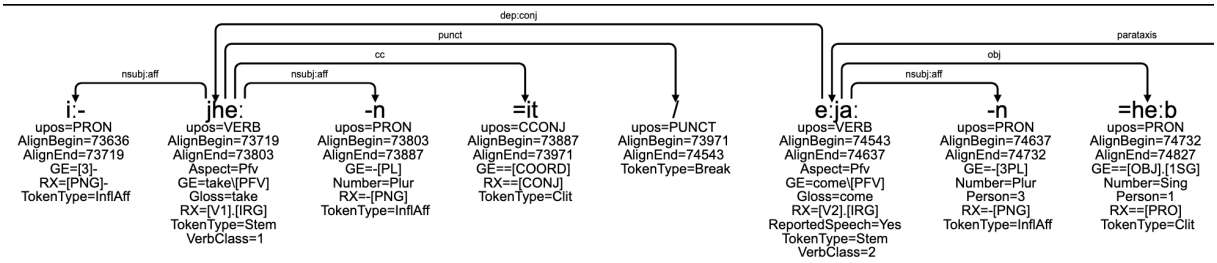


'in front of her'

Figure 12. Postposition with an enclitic pronoun (UD-style)

4.2 Coordination in a head-final language

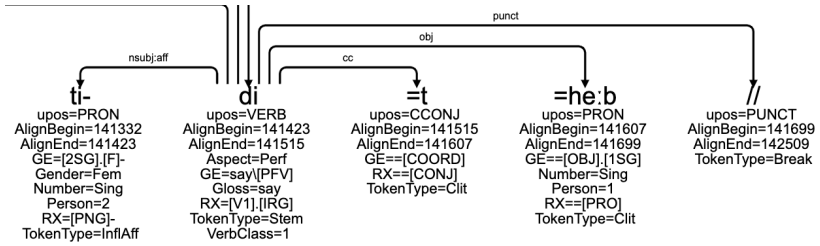
In Beja, verbal and nominal coordination are expressed with different enclitic morphemes. The texts contain 20 tokens of the verbal coordinating conjunction *=t* (and its allomorphs *=it* and *=ajt*) (Figure 13). For half of the tokens, the conjunctions occur at the end of a prosodic unit, be it a major or a minor prosodic break, or a sentence. For this reason, we attach the coordinating conjunctions to the first conjunct and we consider that the second conjunct is the head of the coordination. See Kanayama et al. (2018) for a similar analysis in two other head-final languages, Japanese and Korean. As the *conj* relation is forbidden from right to left in UD, we introduced a *dep:conj* relation.



'they took them and they came to me'

Figure 13. Verbal coordination (UD-style)

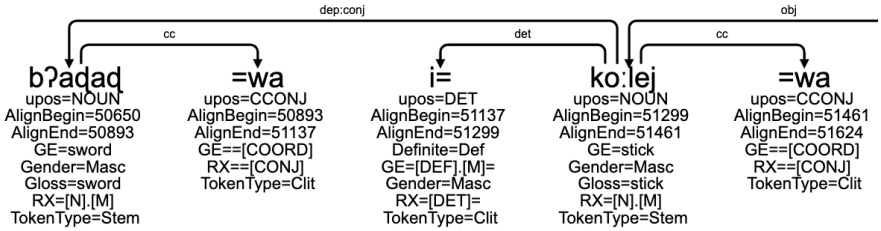
The verbal coordinating conjunction is tightly linked to the right of the verb. It even occurs before enclitic object pronouns (6 tokens), as in Figure 14.



‘you told me and …’

Figure 14. Position of the verbal coordinating conjunction (UD-style)

The nominal coordinating conjunction is the enclitic *wa*. It is expressed on each conjunct as shown in Figure 15.



‘a sword and the stick’

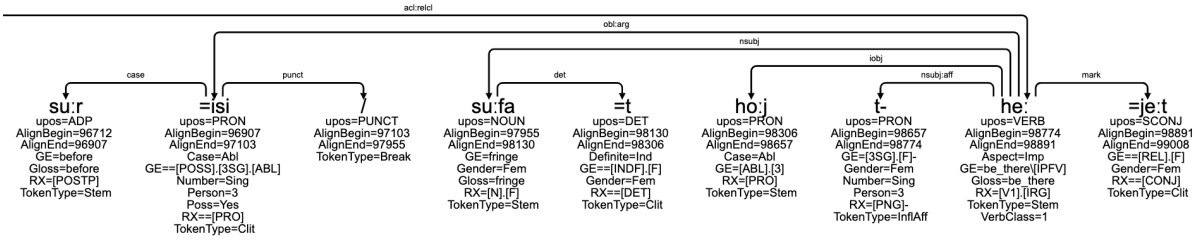
Figure 15. Nominal coordination (UD-style)

4.3 Relative clauses

In the texts, only object relative clauses occur, for which the number of tokens amounts to 18. They are marked by several clitics, either proclitics (*ji=, j=, wi=, w=*), or enclitics (*=e:b, =e:t, =t, =b, =e:*). There are also instances of a zero morph. 7 tokens were found with a preposed antecedent and 11 tokens with a postposed antecedent.

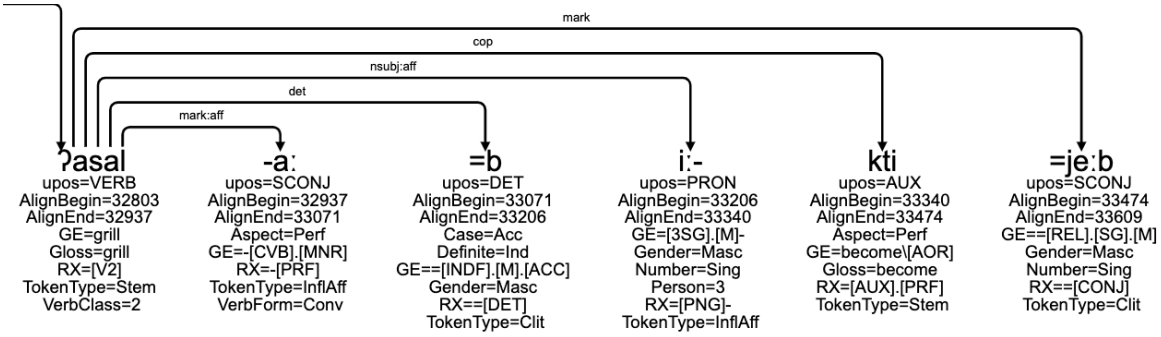
The anteposition of an antecedent is an unusual word order in verb-final languages. The SUD annotation revealed that this construction occurs in two contexts linked to information structuring:

1. when the object of the transitive verb of the matrix clause is topicalized (in Figure 16 two relative clauses precede the verb of the matrix clause)
2. when the relative comes as an afterthought (Figure 17).



‘(a girl) who has a fringe in front of her’

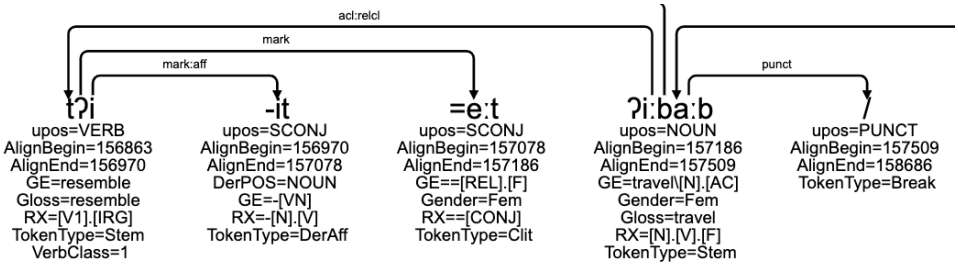
Figure 16. Topicalized preposed antecedent (UD-style)



‘(The man was carrying on his shoulder a lamb.) That he had grilled.’

Figure 17. Preposed antecedent in an afterthought (UD-style)

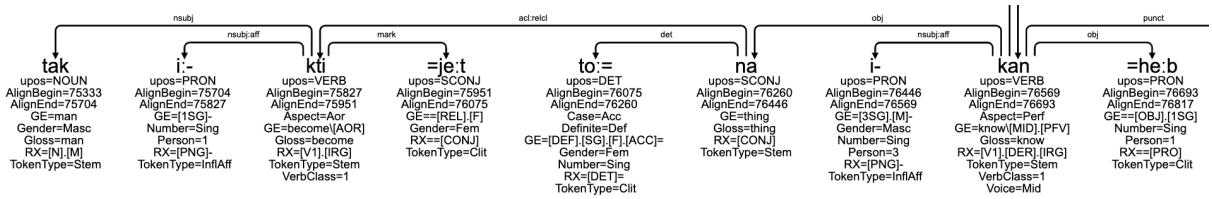
Otherwise the canonical constituent order is used: relative clause – object antecedent – main verb (Figure 18).



‘a story like that (happened to me)’, lit. a travel that resembles

Figure 18. Antecedent in canonical word order (UD-style)

Complement clauses may also be formed on the basis of a relative clause. In such cases, the antecedent, which is the dummy noun *na* ‘thing’, is always placed after the relative clause, i.e. the canonical constituent order (Figure 19).

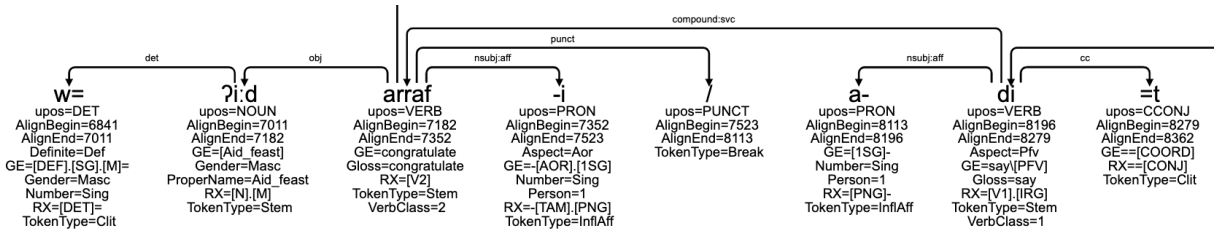


‘he realized that I was a man’

Figure 19. Relative-based complement clause (UD-style)

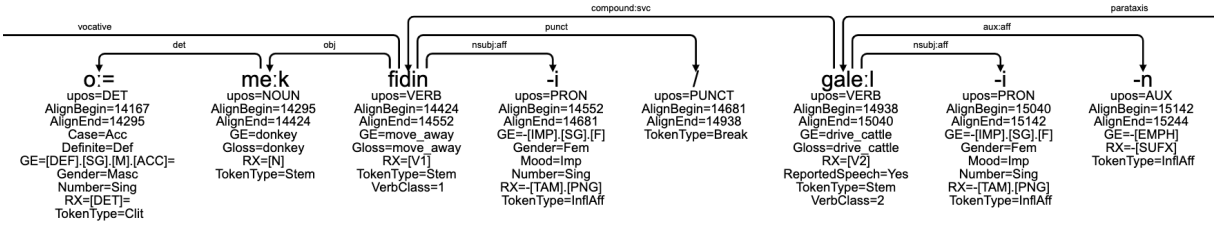
4.4 Non canonical Serial Verb Constructions

We chose to label as SVCs any series of two finite verbs of the same semantic domain which are not coordinated and do not have a predicate–argument relation. This characterization only partly complies with e.g. Haspelmath’s (2016: 296) narrow definition of serial verb constructions as a comparative concept: “A serial verb construction is a monoclausal construction consisting of multiple independent verbs with no element linking them and with no predicate–argument relation between the verbs.” Such a definition does not impose any semantic restriction on the semantic domains of SVCs (apart from expressing a dynamic event, Cleary-Kemp 2015: §4.2.1.3; Haspelmath 2016: 302), as is the case in Beja. Moreover, SVCs seem to be unproductive, limited to very few semantic domains, and restricted to series of two verbs. There are 3 occurrences of SVCs in our data, 1 with a verb of saying (Figure 20), and 2 with motion verbs (Figures 21) conjugated at various TAM.



‘(I went) to wish him a blessed Aid’

Figure 20. SVC with verbs of saying (aorist + perfective) (UD-style)



‘chase the donkey away!’

Figure 21. SVC with motion verbs (both imperfective) (UD-style)

5 Conclusion

Beja belongs to a sub-family of Afroasiatic languages, the Cushitic languages, for which there were no treebanks. It is a language with a rich morphology and which, unlike its cousins, the Semitic languages, is a head-final language with non-canonical serial verb constructions.

The Beja treebank that we present is a very small treebank, but richly annotated, with a segmentation into morphs, glosses, and an alignment to the sound file.

While developing a morph-based treebank for Beja we have been led to bring some enrichments to the SUD and UD annotation schemes. We introduced the *TokenType* feature which takes five possible values (*Stem, InflAff, DerAff, Clitic, Break*) and an *aff* extension for syntactic relations to indicate more explicitly the internal relations of words. We also introduced the feature *DerPos* on derivational morphs for indicating the POS of the derived form.

We have also seen that the SUD version of the morph-based treebank makes it easier to compute the word-based version of the treebank, since it explicitly indicates the internal structure of the word and the order in which the affixes combine with the stem.

As it is possible to convert the morph-based annotation into a word-based annotation, we think it is better to distribute the morph-based annotation, which contains more information and is closer to the format that field linguists use. This format allows us to extract qualitative and quantitative information about the inflectional morphology of the language, which is extremely useful for typological studies (Greenberg 1960).

The Universal Dependency project was initially developed to unify treebank annotation schemes in order to have a common format for the development of NLP tools. The UD annotation scheme is heavily based on the output format developed for the Stanford parser for English (de Marneffe et al. 2006). The 33 languages of the UD1.2 (Nivre et al. 2016) were all languages with long-standing writing traditions, and all corpora were written corpora following well-established orthographic conventions, most of them with a segmentation into words.

UD is now integrating a wide range of new languages coming from different families. Many field linguists having data that are already analyzed in IGT are ready to enrich their corpus with a syntactic annotation. It is necessary that UD offer the possibility of a morpheme-based view of annotation, which allows them to keep the IGT structure. This paper is a first step in this direction by setting up a processing chain to convert an IGT into a morph-based treebank, then a word-based treebank.

References

Cleary-Kemp, J. 2015. *Serial Verb Constructions Revisited: A Case Study from Koro.* PhD dissertation, University of California at Berkeley.

Comrie, B. 2015. From the Leipzig Glossing Rules to the GE and RX lines. In A. Mettouchi, M. Vanhove & D. Caubet (eds.), *Corpus-based Studies of Lesser-described Languages*. John Benjamins, 207-219.

Comrie, B., Haspelmath, M., & Bickel, B. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology & the Department of Linguistics of the University of Leipzig. Retrieved January, 28, 2010. https://www.eva.mpg.de/lingua/resources/glossing-rules.php.

CorpAfroAs: The CorpAfroAs Corpus of Spoken AfroAsiatic Languages. http://dx.doi.org/10.1075/scl.68.website.

de Marneffe, M.-C., MacCartney, B., & Manning, C. D. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th Language Resources and Evaluation Conference (LREC)*.

de Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. 2021. Universal dependencies. *Computational Linguistics*, 47(2), 255-308.

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. 2018. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Universal Dependencies Workshop (UDW)*.

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. 2019. Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic relations and deep syntactic features. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT)*.

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. 2021. Starting a new treebank? Go SUD!. In *Proceeding of the 6th conference on Dependency Linguistics (Depling)*.

Gerdes, K., Kahane, S., & Chen, X. (2021). Typometrics: From Implicational to Quantitative Universals in Word Order Typology. *Glossa: a journal of general linguistics*, 6(1).

Greenberg, J. H. 1960. A quantitative approach to the morphological typology of language. *International journal of American linguistics*, 26(3), 178-194.

Guillaume, B. 2021. Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL)*, 168–175.

Haspelmath, M. 2016. The serial verb construction: Comparative concept and cross-linguistic generalizations. *Language and Linguistics* 17(3), 291–319.

Haspelmath, M. 2020. The morph as a minimal linguistic form. *Morphology* 30: 117–134. https://doi.org/10.1007/s11525-020-09355-5.

Kahane, S., Caron, B., Gerdes, K., & Strickland, E. 2021. Annotation guidelines of UD and SUD treebanks for spoken corpora. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT)*.

Kanayama, H., Han, N. R., Asahara, M., Hwang, J. D., Miyao, Y., Choi, J. D., & Matsumoto, Y. 2018. Coordinate structures in universal dependencies for head-final languages. In *Proceedings of the Second Universal Dependencies Workshop (UDW)*, 75-84.

Mettouchi, A., & Chanard, C. 2010. From Fieldwork to Annotated Corpora: The CorpAfroAs Project. *Faits de Langue-Les Cahiers n°2*, 255-265.

Park, J. 2017. Segmentation granularity in dependency representations for Korean. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, 187-196.

Pietrandrea, P., Kahane, S., Lacheret-Dujour, A., & Sabio, F. (2014). The notion of sentence and other discourse units in corpus annotation. In T. Raso, H. Mello, M. Pettorino (eds.), *Spoken Corpora and Linguistic Studies,* John Benjamins, Amsterdam, 331-364.

Tsarfaty, R., & Goldberg, Y. 2008. Word-Based or Morpheme-Based? Annotation Strategies for Modern Hebrew Clitics. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC)*.

Vanhove, M. 2006. The Beja language today in Sudan: The state of the art in linguistics. *Proceedings of the 7th International Sudan Studies Conference*. Bergen: University of Bergen, CD Rom.

Vanhove, M. 2014. The Beja Corpus. In Mettouchi, A. and C. Chanard (eds.). *The CorpAfroAs Corpus of Spoken AfroAsiatic Languages*. http://dx.doi.org/ 10.1075/scl.68.website.

Vanhove, M. 2017. *Le Beja*. Leuven, Paris: Peeters (coll. Les Langues du Monde 9).

Vincze, V., Szauter, D., Almási, A., Móra, G., Alexin, Z., & Csirik, J. 2010. Hungarian dependency treebank. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC)*.

Zhang, M., Zhang, Y., Che, W., & Liu, T. 2014. Character-level Chinese dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1326-1336.