Annotation guidelines of UD and SUD treebanks for spoken corpora: a proposal

Sylvain Kahane*, Bernard Caron, Emmett Strickland*, Kim Gerdes*****
*Modyco, Université Paris Nanterre & CNRS
**Llacan, CNRS & INALCO
***Lisn, Université Paris Saclay & CNRS

Abstract

This paper presents practical and theoretical guidelines for the development of treebanks for spoken languages in the UD and SUD annotation schemes. We discuss text-sound alignment, segmentation into "sentences", use of "punctuation", paradigmatic lists, disfluencies, and paratactic constructions. This proposal is based on the development of (Surface-Syntactic) Universal Dependencies treebanks for spoken French, Naija, and Beja.

1. Introduction

This paper presents our recommendations for the development of treebanks for spoken languages, based on our experience in the development of several treebanks for spoken French (Gerdes & Kahane 2009, Lacheret et al. 2014, 2019), Cantonese (Wong et al. 2017), and more recently a large treebank for spoken Naija, an English-lexifier pidgincreole from Nigeria which is spoken by more than 100 million people (Caron et al. 2019), as well as a small treebank for Beja, an Afro-Asiatic language spoken in Sudan (Kahane et al. 2021). All of these treebanks are dependency-based, and our more recent treebanks use the SUD (Surface-Syntactic Universal Dependencies) framework (Gerdes et al. 2018, 2019, 2021), which can be automatically converted into UD (Universal Dependencies) (de Marneffe et al. 2021). Previous attempts to provide guidelines for UD-based spoken treebanks include Dobrovoljc & Nivre (2016) and Øvrelid et al. (2018).

The need for special guidelines for spoken language treebanks arises from several particularities of spoken language corpora and spoken language grammar.

Spoken corpora are transcriptions of audio and video recordings. We will not discuss the issue of these transcriptions, as conventions can vary from language to language.[1] However, the relationship between the written text and the sound file which characterizes oral corpora introduces some specificities in the treebank itself, which are discussed in Section 2. The identification of speakers and the marking of overlaps between speakers are also considered.

The next challenge in the transcription of spoken languages is the fact that recorded speech does not contain any consistent sentence boundaries. A segmentation based on prosody is sometimes adopted, especially for interlinear glossed texts (IGTs), but prosodic units can be strongly divergent from syntactic units (Beliao et al. 2015, Kahane & Lacheret 2019). Section 3 is devoted to segmenting

[1] Each language we considered is bound by different constraints. French has a strong tradition of orthographic normalization (even if there remains some room for discussion, see Dister et al. 2019). Naija orthography has yet to be codified, even if the language now enjoys some official written media presence through outlets like BBC News Pidgin (https://www.bbc.com/pidgin). We therefore gave our transcriptors the freedom to choose how best to represent the language in writing. Beja has no written tradition whatsoever, which means that our Beja corpus is an interlinear glossed text with a phonetic transcription and a word segmentation done by a linguist (Vanhove 2014).

spoken texts into maximal syntactic units. Following the French tradition of studies of the syntax of spoken French (Blanche-Benveniste 1990, Berrendoner 1990), we consider two levels of analysis, referred to as *microsyntax* and *macrosyntax* in the French scholarly tradition: microsyntax, which is the syntax of government (Fr. *rection*), describes marked relations where one word imposes its form, category, or position on another; macrosyntax concerns looser relations, such as those involving detached/dislocated units, parentheses, inserts, discourse markers, etc. Macrosyntactic units generally correspond to units that are delimited by punctuation in written texts. The choice of the delimiters also requires some discussion.

One particularity of spoken productions is the great number of paradigmatic lists or "piles", which are successions of units piled up in the same syntactic position. This is the case with coordination, but also reformulation, or apposition. Section 4 proposes a homogeneous treatment of these constructions.

Section 5 is devoted to the analysis of disfluencies due to incomplete units.

The last characteristic of spoken languages we examine is the variety of paratactic constructions used in spoken production. In Section 6 we propose splitting the *parataxis* relation of UD into seven different subtypes.

2. Text-sound alignment and speech turns

UD-based treebanks are encoded using the CoNLL-U format (https://universaldependencies.org/format.html). In this tabular format, each sentence is encoded separately. Each token occupies a row and information associated with that token is divided into 10 columns (see Fig. 1).[2] Metadata associated with a given sentence precede the table describing the dependency tree. These metadata must contain the text of the sentence (# text) and a sentence id (# sent_id). For spoken corpora, two main pieces of information should be added: a (permanent) link to the sound file (# sound_url) and, if the corpus is not a monologue, an id for each speaker (# speaker_id).

```
#sound_url = http://www.tal.univ-paris3.fr/trameur/iTrameur-naija/mp3/KAD_22_Chatting-At-The-Restaurant_DG.mp3
#speaker_id = Sp205
#text = na wa for dat woman o !//
#text_en = It's a shame for that woman!
#text_ortho = Na wa for dat woman o!
1   na  na  PART    _   PartType=Cop   0   root    _   AlignBegin=15240|AlignEnd=15382|ExtPos=INTJ|Gloss=be1|Idiom=Yes
2   wa  wa  INTJ    _   PartType=Cop   1   comp:pred  _   AlignBegin=15382|AlignEnd=15523|Gloss=wow|InIdiom=Yes
3   for for ADP _   _   1   comp:obl    _   AlignBegin=15523|AlignEnd=15665|Gloss=for
4   dat dat DET _   Number=Sing|PronType=Dem  5   det _   AlignBegin=15665|AlignEnd=15807|Gloss=SG.DEM
5   woman   woman   NOUN    _   _   3   comp:obj    _   AlignBegin=15807|AlignEnd=15948|Gloss=woman
6   o   o   PART    _   PartType=Disc  1   mod:emph    _   AlignBegin=15948|AlignEnd=16090|Gloss=EMPH
7   !// !// PUNCT   _   _   1   punct   _   AlignBegin=16090|AlignEnd=16090|Gloss=PUNCT
```

Figure 1. Encoding of text-sound alignment (from SUD_Naija-NSC)

With the sound url and an additional temporal alignment, it is possible to listen to the sound associated with each word. In (S)UD_Naija-NSC and in (S)UD_Beja-NSC, each word is time-aligned using the features AlignBegin and AlignEnd, with a value in milliseconds from the start of the sound file.[3] If the corpus features punctuation marks, the two features should typically be equal in value unless the symbol corresponds to a prosodic break. These word-level features also allow one to determine the timespan of a given utterance: the AlignBegin value of the first token correspond to the beginning of

[2] SUD_Naija-NSC uses a macrosyntactic markup (see Section 3) with special punctuation signs, such as < or //. This markup is part of the annotated text, given in #text. An orthographic variant is proposed in #text_ortho, with traditional punctuation signs and an uppercase at the beginning of the sentence. Moreover an English translation is given in # text_en, as well as a gloss for each token.

[3] In the case of SUD_Naija-NSC, only some time boundaries were stored during the transcription process; others have been roughly assessed by dividing each time segment by the number of words. It would have been also possible to keep the features on some words only.

the utterance, while the AlignEnd value of the final token corresponds to the end. For corpora with only a sentence-level alignment, we recommend adding a single AlignBegin feature to the first token, and an AlignEnd feature to the last.

For the time being, (S)UD_Naija-NSC, (S)UD_Beja-NSC, and (S)UD_French-ParisStories are the only (S)UD treebanks that provide a link towards a sound file. UD_English-GUM (Zeldes 2017), a third of which represents spoken data, contains # speaker and # addressee features, which are, to our knowledge, the only way to distinguish spoken from written utterances. UD_Polish-LFG (Patejuk & Przepiórkowski 2018) contains a feature # genre = spoken (prepared) to distinguish the spoken data. UD_Frisian_Dutch-Fame (Braggaar & Vander Goot 2021) and UD_Scottish_Gaelic-ARCOSG (Batchelor 2019) have a # speaker feature, while UD_Norwegian-NynorskLIA (Øvrelid et al. 2018) has metadata including compound features like # dialect: eidsberg speakerid: eidsberg_uio_03, which include both a location and a speaker id. UD_Latvian-LVTB (Pretkalniņa et al. 2018) has a feature # newpar id, which probably contains the speaker id of new participant. Other treebanks of spoken languages (UD_Turkish_German-SAGT, UD_Slovenian-SST, UD_Cantonese-HK, UD_Komi_Zyrian-IKPD, UD_Swedish_Sign_Language-SSLC, UD_Chukchi-HSE)[4] do not contain any specific metadata, while some treebanks (UD_English-Lines, UD_Greek-GDT, UD_Persian-Seraji) that are partially composed of spoken data do not contain any metadata allowing one to distinguish between speech and writing. For mixed corpora, we think that a feature # genre = spoken should be used to facilitate the identification of speech. Outside of (S)UD, the CHAT transcription system used in the CHILDES database of childhood speech marks each utterance with a three-letter speaker identifier, each of which is associated with a name and role (i.e., father, mother, child) in the file header. Additional information may be provided about a speaker's background (age, socioeconomic status) and the context of the recording, while special punctuation can be used to designate overlapping speech (MacWhinney 2000).

Note that in (S)UD corpora only one speaker and one # speaker_id is allowed for each utterance. In instances of co-constructions between two speakers, we use the special features AttachTo and Rel. By co-constructions, we mean two trees T1 and T2 from two different speakers that form a cohesive syntactic construction, as in (1) from SUD_French-Rhapsodie.[5]

(1) *$L3 ^et c'est récupéré (bien sûr) par l'équipe $- argentine //+*
$L2 qui -$ sont de nouveau en possession (les Argentins) du "euh" ballon //
'$L3 ^and it is recovered (of course) by the $- Argentinian team //+.
$L2 who -$ are again in possession (the Argentines) of the "uh" ball //'

One token from T1 is the governor of the root of T2 in this new construction. In (1), L2's utterance is a relative clause that modifies the noun *équipe* 'team' from L3's utterance. This is indicated on the root of the second tree by the feature AttachTo=11@Rhap_D2003-92bis indicating that this utterance could be attached to the 11th token of tree Rhap_D2003-92bis, and by the feature Rel=mod@relcl indicating the relation between the two tokens (see Fig. 2 and 3).[6]

[4] UD-Chukchi-HSE (Tyers & Mishchenkova 2020) has a feature # text[phon], with a phonetic transcription, but it is not a feature specific to spoken corpora.

[5] SUD_French-Rhapsodie is the former SUD_French-Spoken, which was initially the Rhapsodie treebank (Lacheret et al. 2019). The treebank has been renamed due to the introduction of a new treebank of spoken French, SUD_French-ParisStories. Rhapsodie uses a markup, where sentence boundaries are indicated by // and co-constructions by //+. Overlaps are indicated by $- …-$. But contrary to SUD_Naija-NSC, # text is based on standard punctuations and the macrosyntacic markup is stored in # macrosyntax.

[6] UD requires the root of a tree to have the relation root.

```
# macrosyntax = ^et c'est récupéré ( bien sûr ) par l'équipe $- argentine //+
# text_ortho = et c'est récupéré, bien sûr, par l'équipe, euh, argentine.
# speaker = L3
# sent_id = Rhap_D2003-92bis
1  et  et  CCONJ  _  _  3  cc  _  _
2  c'  ce  PRON  _  Gender=Masc|Number=Sing|Person=3|PronType=Dem  3  subj@pass  _  _
3  est être  AUX _  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin  0  root  _  _
4  récupéré  récupérer  VERB  _  Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part  3  comp:aux@pass  _  _
5  ,  ,  PUNCT  _  _  6  punct  _  _
6  bien  bien  ADV _  _  7  mod _  ExtPos=ADV|InIdiom=Yes
7  sûr sûr ADJ _  Gender=Masc|Number=Sing  3  mod _  PhraseType=Idiom
8  ,  ,  PUNCT  _  _  9  punct  _  _
9  par par ADP _  _  4  comp:obl  _  _
10 l'  le  DET _  Definite=Def|Number=Sing|PronType=Art  11  det _  _
11 équipe  équipe  NOUN  _  Gender=Fem|Number=Sing  9  comp:obj  _  _
12 ,  ,  PUNCT  _  _  13  punct  _  _
13 euh euh INTJ  _  _  11  discourse  _  _
14 ,  ,  PUNCT  _  _  15  punct  _  _
15 argentine  argentin  ADJ _  Gender=Fem|Number=Sing  11  mod _  Overlap=Rhap_D2003-92ter
16 .  .  PUNCT  _  _  3  punct  _  _

# macrosyntax = qui -$ sont de nouveau en possession ( les Argentins ) du "euh" ballon //
# text_ortho = qui sont de nouveau en possession du, euh, ballon.
# speaker = L2
# sent_id = Rhap_D2003-92ter
1  qui qui PRON  _  _  2  subj  _  Overlap=Rhap_D2003-92bis
2  sont  être  VERB  _  Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin  0  root  _
AttachTo=11@Rhap_D2003-92bis|Rel=mod@relcl
3  de  de  ADP _  _  5  mod _  ExtPos=ADV|PhraseType=Idiom
4  nouveau nouveau ADJ _  Gender=Masc|Number=Sing  3  unk _  InIdiom=Yes
5  en  en  ADP _  _  2  comp:obl  _  _
6  possession possession NOUN  _  Gender=Fem|Number=Sing  5  comp:obj  _  _
7  les le  DET _  Definite=Def|Number=Plur|PronType=Art  8  det _  _
8  Argentins  Argentin  PROPN  _  _  2  dislocated _  _
9-10  du  _  _  _  _  _  _  _  _
11 de  de  ADP _  _  6  udep  _  _
12 le  le  DET _  Definite=Def|Gender=Masc|Number=Sing|PronType=Art  16  det _  _
13 ,  ,  PUNCT  _  _  12  punct  _  _
14 euh euh INTJ  _  _  16  discourse  _  _
15 ,  ,  PUNCT  _  _  14  punct  _  _
16 ballon  ballon  NOUN  _  Gender=Masc|Number=Sing  11  comp:obj  _  _
17 .  .  PUNCT  _  _  2  punct  _  _
```

Figure 2. Encoding of a co-construction (from SUD_French-Rhapsodie)

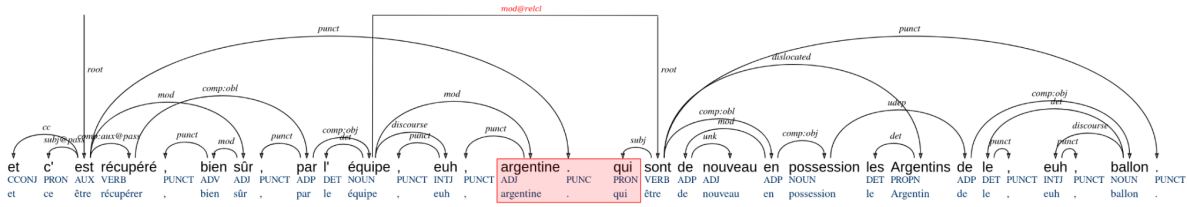


Figure 3. Visualization of the encoded information. In red: the AttachTo link and the Overlap.

In cases of overlap between two speech turns, words that overlap have a feature Overlap whose value is the id of the overlapping sentence. In our previous example, *argentine*, of sentence Rhap_D2003-92bis, has a feature Overlap=Rhap_D2003-92ter, indicating that it overlaps with sentence Rhap_D2003-92ter, more specifically with the word *qui* 'who', which itself carries the feature Overlap=Rhap_D2003-92bis. Overlaps can also be deduced from time alignment, but some corpora indicate overlaps without also containing temporal information, and for most query engines it is much simpler to request overlaps if a feature is present.

3. Sentence segmentation and punctuation

With the exception of rehearsed speeches and other kinds of highly prepared oratory, oral language cannot typically be segmented into units equivalent to the written sentence. Instead, we segment our corpora into illocutionary units (IUs) (Cresti 1995, Pietrandrea et al. 2014). An IU is defined as a speech

segment that corresponds to a single speech act, i.e., a single question, declaration, or command. Consider the following examples from French and Naija.

(2) *ceux qui sont en location < la moyenne < c'est environ trois ans //*
those who are on lease < the average < it is about three years //
'For those that rent, the average lease is about three years.'

(3) *e get one lady < hm she just enter inside shop o //*
there was one lady < hm she just came into the shop //

These utterances represent cohesive linguistic units serving to express a single primary idea. One can split these into internal units sharing no marked relations of dependency with one another, separated in these examples by the character <. The cohesiveness of these units can nevertheless be demonstrated by their varying degrees of autonomy. For example, the utterances *c'est environ trois ans* 'it is about three years' and *hm she just enter inside shop o* 'hm she just came into the shop' form perfectly acceptable utterances in isolation that carry the same basic message as the original utterance. However, *la moyenne* 'the average' or *e get one lady* 'there was one lady' would either be unacceptable or serve a very different illocutionary purpose in isolation. In oral corpora, we recommend preserving this cohesiveness by using the IU as the primary unit of segmentation. Within each IU, internal components can be delimited through special symbols, either placed directly in the list of tokens, or in a dedicated sentence feature containing the macrosyntactic annotation (and replaced in the text by standard punctuation signs; see Note 5).

In our corpora, the end of each IU is marked by a symbol indicating an IU boundary. In the corresponding CoNLL-U file, these may be integrated directly into the list of tokens and annotated as punctuation marks. In order to better identify the modality of each IU, we recommend introducing a separate boundary marker for each type of utterance. In our corpus of spoken Naija, the symbol // is used to mark the end of an assertion, ?// the end of a question, and !// the end of an order or an exclamation. The symbol &// is used to mark the end of an interrupted or incomplete utterance. However, more traditional symbols such as periods and question marks may also be used so long as they are used consistently as IU boundary markers.

(4) *I no pass all di subjects //*
'I didn't pass all of the subjects.'

(5) *wetin be Ponzi Scheme ?//*
'what is a Ponzi Scheme?'

(6) listen attentively !//
'listen attentively!'

(7) *e get things wey you &//*
'there are things that you…'

As demonstrated previously in examples (2) and (3), IUs can contain internal units with varying degrees of autonomy. All completed utterances must contain at least one (and typically only one) nucleus, a fully autonomous macrosyntactic unit capable of forming an acceptable utterance when spoken alone in the same discursive position (Pietrandrea & Kahane 2019). Examples (4-6) are each composed of one nucleus, but IUs can also contain multiple adnuclei located to the left or right peripheries of the nucleus. These are illocutionarily dependent on the nucleus and typically lack any marked microsyntactic relationship with the elements of the nucleus. These can be divided into prenuclei, located to the left of the nucleus, and postnuclei, located to the right. Like the IU boundaries, we recommend explicitly delimiting these units using a dedicated set of symbols. In our corpora, we favor using < to mark the ends of prenuclei (see ex. 8-9) and > to mark the beginning of postnuclei (ex. 10-11). However, more traditional symbols such as commas may also be used.

(8) *les chaises < il faut me les donner //*
'the chairs, you need to give them to me.'
(9) *en ce qui me concerne < j'aimerais enseigner dans un établissement public //*
'as for me < I would like to teach in a public school'
(10) *dat's what de use to do > some of dem o //*
'that's what they used to do, some of them.'
(11) *good evening > my daughter //*
'good evening, my daughter.'

Note that it is possible for an IU to contain strings of adnuclei

(12) *donc < alors < ça date de quand à peu près > ce fauteuil-là ?//*
'so, then, it dates from when approximately, this armchair?'
(13) *Osas < dis your wedding wey you dey prepare < me < I dey look forward to di ting o !//*
'Osas, that wedding of yours that you are preparing, me, I'm looking forward to it!'

These adnuclei are typically connected to the root of the utterance using the UD relations *dislocated, vocative, discourse,* and in some cases *parataxis*. These are discussed in Section 6.[7]

It is important to note that our segmentation generally avoids coordination between main verbs in a single IU, since each part can form an autonomous utterance. (13) shows a string of consecutive IUs each containing a single main verb.

(14) *vous oubliez vos produits habituels.*
et vous mettez à l'intérieur la boule magique, la boule de lavage avec vos vêtements.
et vous allez voir le résultat.
'you forget your usual products.
and you put inside the magic ball, the washing ball with your clothes.
and you will see the result.'

4. Paradigmatic lists

The annotation of paradigmatic structures is challenging in any analysis that marks heads, such as dependency- and X-bar-based phrase structure, because the idea of paradigm itself supposes that two or more elements jointly qualify as the head of a phrase. SUD and UD use the same set of basic relations to encode paradigmatic structures: The *conj* and *cc* relations for standard coordination, the *list* relation that is absent from the analysis of spoken data that we propose,[8] and finally the *orphan* and *reparandum* relations used respectively for ellipsis and disfluency.

As shown in Blanche-Benveniste et al. (1990) and Gerdes & Kahane (2009), reformulation is hard to delimit on a spectrum from coordination to actual repairs, where a second phrase replaces the first. Are (15a) and (15b) coordinations or reformulations?

(15) a. *et puis après, ben, j'ai travaillé sur* ***les micro-processeurs, l'informatique***. (Rhap_D0005-9)
'and then, well, I worked on **microprocessors, computers**.'
b. *mais comme toujours, l'acte d'écrire peut prendre* ***différents masques, différentes valeurs***. (Rhap_2009-2)
'but as always, the act of writing can take on **different masks, different values**.'

[7] Fillers uttered by the listener, such as Fr. *mh* 'hum', are considered discourse markers attaching to the speaker's utterance. They have their own tree but carry the features AttachTo and Rel=discourse.

[8] *list* encodes a list containing only fragments such as address information or phone numbers.

Reformulations are not necessarily repairs but rather elaborations that are also frequently used by rhetorically skilled speakers.

(16) a. *from dat day < as I go meet am for shop < im con { **dey observe me** || **dey observe my attitude** } //* (ABJ_GWA_12_Accident_MG__123)
'From that day, as I went to meet him in the shop, he started **observing me, observing my attitude**.'
b. *mais le, le dis~, **le commissaire du district, le préfet du lieu** ne voulait pas de, de femme non mariée.* (Rhap_D2004-11)
'but the, the dis~, **the district commissioner, the prefect of the place** did not want any, unmarried woman.'
c. *euh, **deux petites phrases, deux vraies options** qui dessinent votre route, une route qui témoigne **d'une certaine, d'une bonne, d'une très bonne** conduite*. (Rhap-D2001-3)
'uh, **two small sentences, two real options** that draw your road, a road that testifies to **a certain, a good, a very good** conduct.'

We excluded the *reparandum* relation from the analysis of spoken data in our SUD annotation scheme, and instead use the concept of paradigmatic lists developed by Blanche-Benveniste et al. (1990) and produce a simple typology of paradigmatic lists (Kahane & Pietrandrea 2012) by annotating three sub-relations of the *conj* dependency:

- *conj:coord*: the standard coordination, where each conjunct denotes a different referent;
- *conj:dicto*: used for several denotations of the same referent, which are repetitions or reformulations of the same denotation, as in (16), including disfluencies as in (16b);
- *conj:appos:* used for double denotations of the same referent, as in (17).[9]

We decided not to distinguish reformulation and disfluencies because the line between voluntary and unvoluntary reformulation or correction is very difficult to draw.[10] In the transfer to UD annotation, *conj:dicto* is renamed *reparandum*, even if it is not exactly a standard *reparandum* annotation, because the first conjunct remains the head.[11]

(17) a. ***match nul, zéro partout** pour Lille au Mans.* (Rhap_M2006-72)
'**draw, zero all** for Lille at Le Mans.'
b. *et c'est dans cet esprit qu'est proposée par Szymanoski **l'ouverture du concert, son opus douze**.* (Rhap_D2012-47)
'and it is in this spirit that Szymanoski proposes the opening of the concert, his opus twelve.'
c. *eh bien, le secret, il est **là, dans ce petit bruit que l'on entend***. (Rhap_D2011-57)
'well, the secret is **there, in this small noise that we hear**.'

In the case of a partial answer to a *wh*-question, we also consider that the answer forms a *conj:appos* pile with the wh-word, encoded using the AttachTo and Rel features.

[9] UD uses the same *appos* relation for appositions that form paradigmatic lists and appositions where one unit modifies the other. We propose encoding the latter as a standard modification (*mod* in SUD or *nmod* in UD) or as subtype of modification (*mod:appos* in SUD or *nmod:appos* in UD) since the two constructions have different semantic and prosodic properties: in *conj:appos*, the two conjuncts are separate prosodic units, the apposed unit being in a different register, while in *mod:appos*, the two elements form one cohesive prosodic unit (*my brother John, le journal Le Monde*).

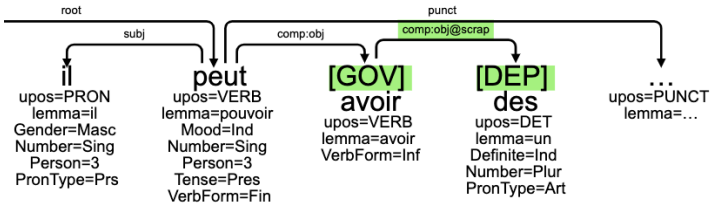
[10] The initial Rhapsodie annotation considered seven different cases of paradigmatic lists (Kahane & Pietrandrea 2020). The distinction between disfluencies and reformulation was based on purely formal criteria (repetition of a unit) to avoid poor inter-annotator agreement (Bawden et al. 2014).

[11] The notion of *reparandum* presupposes that the second conjunct repairs the first (Shriberg 1994). Following Blanche-Benveniste (1990), we instead consider that, in reformulations, even involuntary ones, information is cumulative, and nothing that has been said can be completely forgotten.

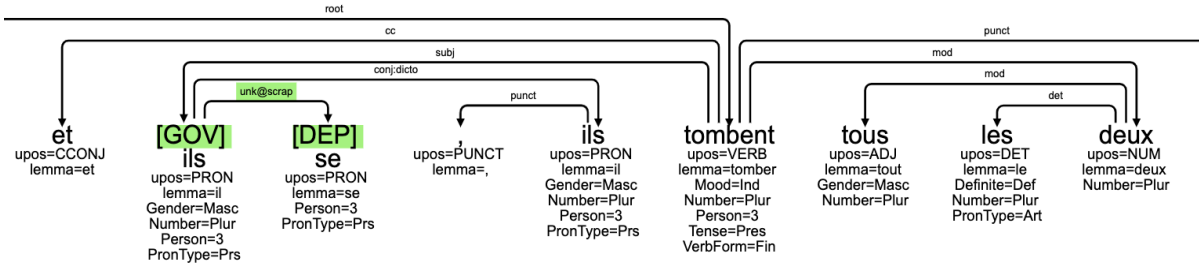
(18) $*L1 Magalie utilise* ***depuis combien de temps*** *notre boule magique?*
$L5 ***depuis deux mois***.
'$L1 **how long** has Magalie been using our magic eight ball?
$L5 **for two months**.'

5. Disfluencies

One kind of disfluency, often called repair, is observed when a unit, then called the reparandum, gets overridden by a new unit, the repair. As shown in section 4, this is a subtype of the common mechanism of listing, which is not necessarily disfluent and which we formalize with the relation *conj:dicto*. Another case of disfluency corresponds to incomplete units. In this case, UD chooses to promote one element as the head of the unit and to attach the other elements to it (Dobrovoljc & Nivre 2016). When this produces nonstandard relations, we propose adding a feature. In SUD, we add the feature *scrap* on these relations using @ as a separator (*@scrap*) This is converted into UD by inserting the Scrap=Yes feature on the dependent. Fig. 4 and 5 show some uses of *@scrap*. This feature is particularly useful for error mining: for instance, a relation between a verb and a determiner, as in Fig. 4, should not be allowed without a *@scrap*.



'it can have some …'
Figure 4. Incomplete object (comp:obj@scrap).

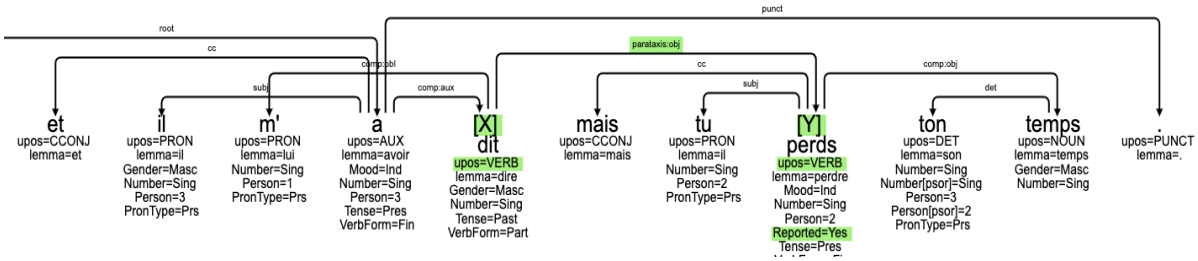


'and they REFL, they both fall'
Figure 5. Unknown relation inside a reparandum (unk@scrap).

6. Paratactic constructions

Paratactic constructions are particularly frequent in spoken corpora. The UD *parataxis* relation covers all cases where a finite verbal construction is neither the root of an utterance, nor a governed clause. In our corpora, we have identified seven different situations that we propose to distinguish: *parataxis:obj, parataxis:discourse, parataxis:parenth, parataxis:insert, parataxis:dislocated, parataxis:mod, parataxis:conj*.

The first occurs when an illocutionary unit occupies a governed position. The most common situation is reported speech in the object position of a verb of saying. In such cases, UD uses the relation *parataxis*, which we propose replacing with *parataxis:obj*.



'and he told me but you are wasting your time.'

Figure 7. Reported speech.

We add the feature Reported=Yes on the head of the reported IU, because some of them are not directly governed, such as the second reported IU in (19), *ça pourrait être grave*, encoded as a separate sentence.[12]

(19) *après, elle a dit, on est aux affaires. ça pourrait être grave.*
'then she said, we're in business. it could be serious.'

Thanks to the feature Reported=Yes, a standard *comp:obj* relation can be used in SUD and should be preferred because reported speech can commute with a complementizer phrase and, in accordance with SUD principles, two units that are in the same commutation paradigm must have the same syntactic function.

Another case of an IU in a governed position is called a graft by Deulofeu (1999). A graft is an IU that is produced in a position where a noun phrase is expected, such as the IU *disons ma carrière pour simplifier* 'let's say my career to simplify', which occupies a subject position.

(20) *vous avez dit que, euh,* ***disons ma carrière pour simplifier,*** *témoigne de ma bonne conduite.*
'you said that, uh, **let's say my career to simplify**, shows my good behavior.'

In such a case we use a plain *subj* relation rather than a *parataxis* relation, but we add a feature Graft=Yes. Some graft constructions are lexicalized such as the construction (21) from SUD_French-ParisStories, where an IU is introduced by the idiom *en mode* 'in the mode', or the construction (22), where a question is apposed to the word *question* 'question'.

(21) *et là lui il l'a regardée en mode* ***euh mais ça va madame ?***
'and there he looked at her in the mode **uh but is it ok miss ?**'

(22) *alors on pourrait poser la question* ***les écrivains ont-ils existé ?***
'then one could ask the question **did the writers exist?**'

UD uses the relation *discourse* for discourse markers, except for verbal expressions, such as *I mean, I guess, you know,* etc., as in (23), as well as tag questions, where *parataxis* is used.[13] This is a category-based distinction that we believe is unwarranted.[14] In SUD, we decided to extend *discourse* for these cases, which is automatically converted into *parataxis:discourse* in UD. The relation has already been introduced in the UD spoken Slovenian treebank (Dobrovoljc & Nivre 2016).

[12] Another solution could be to add quotation marks during the transcription, but this might become unreadable as the reported speech can span over a whole set of illocutionary units.

[13] As shown by Kahane & Pietrandrea (2009), verbal discourse markers have properties that distinguish them from parentheses. They do not accept tense modifications and they cannot be modified. And they have a transitive verb without an overt object, but which takes its host as its object argument.

[14] Many distinctions for relations are category-based in UD, such as *nsubj* vs *csubj*, for nominal vs clausal subjects, *obj* vs *ccomp* for nominal vs clausal objects, or *amod* vs *advmod* vs *nmod*, for adjectival vs adverbial vs nominal or adpositional modifiers.

(23) ***you know*** *gote na kind of delicacy weh dem dey prepare //*
you know gote is a kind of delicacy which they prepare.'

The third case is parenthetical clauses or parentheses, which are real illocutionary units that could be autonomous, but which are inserted in the middle of another illocutionary unit, as *my mama dey dat time* 'my mum was alive at the time' in the Naija sentence (24) or *je suis désolée* 'I'm sorry' in the French sentence (25). We note them as *parataxis:parenth*.

(24) *sometimes (**my mama dey dat time** //) she go help me carry small //*
'Sometimes, **my mum was alive at the time**, she would help me to carry a little.'

(25) *sauf que le clip, enfin, **je suis désolée**, mais il faisait pas très professionnel, hein*.
(ParisStories_2020_concoursInstagram_49)
'except that the clip, well, **I'm sorry**, but it was not very professional, eh'

The fourth situation is that of incises or inserts, which are distinguished from parenthetical ones by the fact that they are not saturated and could not form an independent statement (Bonami & Godard 2008). They belong to a more formal register (news, sermons, stories, etc.). In French, they are also characterized by the inversion of the subject. We annotate them as *parataxis:insert*. An insert can be found at the end of an illocutionary unit.

(26) *vraiment, **dit Job**, la vie de l'homme sur terre est une corvée*. (Rhap_M2003-33)
'truly, **says Job**, man's life on earth is a chore.'

(27) *alors, que va faire maintenant le président ? **se demande Le Progrès à Lyon***.
(Rhap_D2013-45)
'So what will the president do now? **wonders Le Progrès in Lyon**.'

The fifth case is that of verbal constructions in the prenucleus position where they function as dislocations. In example (28), *j'ai des copines actuellement* 'I have friends at present' is clearly not an illocutionary unit (it is not asserted by the speaker that she has friends). It is a prenucleus, which can switch with a nominal phrase like *mes copines* 'my friends'. For the time being, we use *dislocated* for these cases in SUD and UD, but we consider revisiting it as *parataxis:dislocated* in UD.

(28) ***j'ai des copines euh actuellement*** *euh, je m'entends super bien avec*.
(ParisStories_2019_experienceFac_37)
'**I have friends uh at present** uh, I get along great with.'

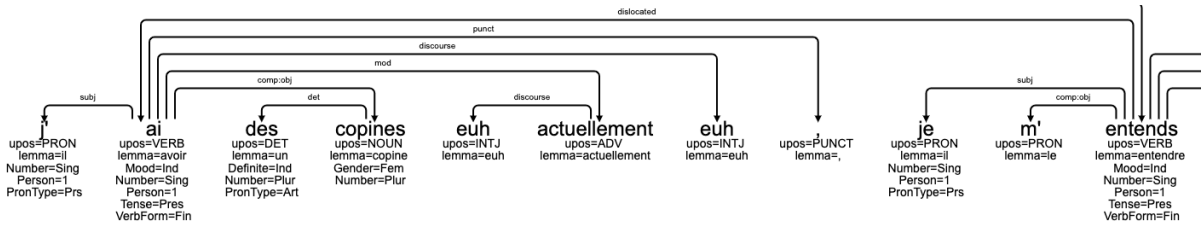


Figure 8. Parataxis:dislocated.

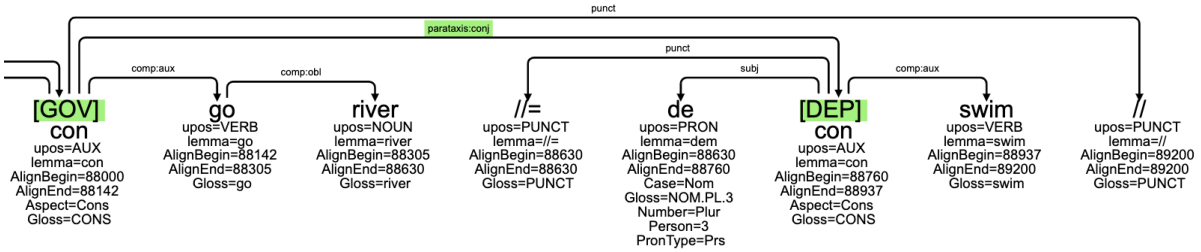
The sixth case of paratactic construction is the case of clausal modifiers that are not introduced by a subordinating conjunction. As in the previous case, the first clause *creche no dey my office*, lit. 'daycare is not (in) my office', is not asserted by the speaker and is a clear prenucleus. For such cases, we use *parataxis:mod* or simply *mod* in SUD (equivalent to UD *advcl*).

(29) *creche no dey my office now < e for no dey easy //*
'If I hadn't had a daycare in my office, it wouldn't have been easy.'

UD makes a seventh use of the *parataxis* relation, to link juxtaposed propositions, as in (19), taken from the spoken part of UD_English-GUM.

(30) *God, I didn't spend the night, that's what makes me so mad, I'm grounded for nothing.*

Our IU segmentation avoids this case most of the time; thus (30) can be perfectly split into three autonomous assertions. Nevertheless, in cases where we do not separate two juxtaposed or coordinated statements, we propose using *parataxis:conj*. In SUD_Naija-NSC, this relation has been used for sequences of parallel IUs.



'then they went to the river, and they swam.'[15]

Figure 9. Parataxis:conj.

So far, we have not found any examples of parataxis which do not fit into one of the seven cases considered, and as we have shown, quite a few can be done away with in SUD (e.g. *parataxis:obj, parataxis:discourse* and *parataxis:mod*).

7. Conclusion

The goal of this paper was to present the additional features we have introduced in our spoken treebanks and to give some proposals for future developments of spoken treebanks, in order to converge on a common set of features. This concerns both practical recommendations for text-sound alignment (# sound_url, BeginAlign, EndAlign), and theoretical propositions concerning text segmentation into sentences, paradigmatic lists, and paratactic constructions.

Our set of relations and tags are meant to be extended to the annotation of constructions typical of spoken texts in other languages as well. This can only be verified by the elaboration of more spoken treebanks for typologically different languages. We hope that the phenomena presented in this paper will motivate other linguists to work on spoken language treebanks and that this paper can serve as a first guide in this endeavour.

Acknowledgements

We would like to thank our reviewers for their valuable remarks. Several people have participated in the development of our annotation schemes for spoken language. We are particularly grateful to Marine Courtin, Vanessa Gaudray-Bouju, Menel Mahamdi, and Mariam Nakhlé.

References

Batchelor C. (2019). Universal dependencies for Scottish Gaelic: syntax. In *Proceedings of CLTW2019 at Machine Translation Summit XVII.*

Bawden R., Botalla M.-A., Gerdes K., Kahane S. (2014) Correcting and Validating Syntactic Dependency in the Spoken French Treebank Rhapsodie. In *Proceedings of the 9th Language Resources and Evaluation Conference* (*LREC*).

[15] The auxiliary *con* marks the consecutivity of two events.

Beliao J., Lacheret A., & Kahane S. (2015) Interface intono-syntaxique en français parlé : compter quoi, compter comment, compter pourquoi ?. In S. Loiseau (ed), *La fréquence textuelle* [*Langage,* 197], 129-153, Larousse.

Berrendonner A. (1990). Pour une macro-syntaxe. In *Données orales et théorie linguistique* [*Travaux de linguistique (Gent)*, 21], 25-36.

Blanche-Benveniste C. (1990). *Le français parlé (études grammaticales)*. Editions du CNRS.

Bonami O., & Godard D. (2008). On the syntax of direct quotation in French. In *Proceedings of the 15th International Conference on HPSG*, CSLI Publications, 358-377.

Braggaar A. & van der Goot R. (2021). Challenges in Annotating and Parsing Spoken, Code-switched, Frisian-Dutch Data. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*

Caron, B, Courtin, M., Gerdes, K., & Kahane, S. (2019) A Surface-Syntactic UD Treebank for Naija. In *Proceedings of the 17th international conference on Treebanks and Linguistic Theories (TLT)*, SyntaxFest.

Cresti E. (1995). Speech act units and informational units. In E. Fava (ed), *Speech Acts and Linguistic Research,* Proceedings of the Workshop. Center for Cognitive Science, SUNY at Buffalo, 89-107.

Deulofeu, J. (1999). *Recherches sur les formes de la prédication dans les énoncés assertifs en français contemporain (le cas des énoncés introduits par le morphème que)*. Thèse d'état, Université Paris 3.

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

Dister A., Goldman J.-P., & Marlet R. (2019) Orthographic and phonetic transcriptions of Rhapsodie recording. In Lacheret-Dujour et al. (2019), 21-34.

Dobrovoljc, K. & Nivre, J. (2016). The Universal Dependencies treebank of spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC),* 1566-1573.

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 Universal Dependencies Workshop (UDW)*, EMNLP.

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

Gerdes K. & Kahane S. (2009) Speaking in piles: Paradigmatic annotation of French spoken corpus. In *Proceedings of the 5th Corpus Linguistics Conference*, Liverpool, http://ucrel.lancs.ac.uk/publications/cl2009.

Kahane S., & Lacheret A. (2019) Syntax and prosody mapping: What and how? The case of intonational periods and illocutionary units. In Lacheret-Dujour et al. (2019), 339-363.

Kahane S., Pietrandrea P. (2009) Les parenthétiques comme « Unités Illocutoires Associées » : une perspective macrosyntaxique, in M. Avanzi & J. Glikman (eds), *Les Verbes Parenthétiques : Hypotaxe, Parataxe ou Parenthèse ?* [*Linx*, 61], 49-70.

Kahane S., Vanhove M., & Ziane R. (2021) A morph-based and a word-based treebank for Beja. In *Proceeding of the 20th Proceedings of the 18th international conference on Treebanks and Linguistic Theories (TLT)*, SyntaxFest.

Lacheret A., Kahane S., Beliao J., Dister A., Gerdes K., Goldman J.-P., Obin N., Pietrandrea P., & Tchobanov A. (2014) Rhapsodie: un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. In *Actes du 4ème congrès mondial de linguistique française* (*CMLF*), *SHS Web of Conferences*, vol. 8, EDP Sciences, 2675-2689.

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

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Øvrelid L., Kåsen A., Hagen K., Nøklestad A., Solberg P. E., & Johannessen J. B. (2018). The LIA treebank of spoken Norwegian dialects. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC),* 4482-4488.

Patejuk A. & Przepiórkowski A. (2018). *From Lexical Functional Grammar to Enhanced Universal Dependencies: Linguistically informed treebanks of Polish*. Institute of Computer Science, Polish Academy of Sciences, Warsaw. Downloadable from http://nlp.ipipan.waw.pl/Bib/pat:prz:18:book.pdf.

Pietrandrea P. & Kahane S. (2019) Macrosyntactic annotation. In Lacheret-Dujour et al. (2019), 97-125.

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

Pretkalniņa L., Rituma L., Saulīte B. (2018) Deriving enhanced Universal Dependencies from a hybrid dependency-constituency treebank. In *Proceedings of the 21sh International Conference Text, Speech, and Dialogue*, LNCS, Vol. 11107, Springer Link, 95-105.

Shriberg E. E. (1994). *Preliminaries to a theory of speech disfluencies*. Doctoral dissertation, University of California, Berkeley.

Tyers, F. M. and Mishchenkova, K. (2020) Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW)*, 195-204.

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.

Wong T., Gerdes K., Leung H. & Lee J. S. (2017) Quantitative comparative syntax on the Cantonese-Mandarin parallel dependency treebank. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, 266-275.

Zeldes, A. (2017) The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, 51(3), 581–612.