Macrosyntactic corpus annotation. The case of Zaar

Bernard Caron

Macrosyntactic corpus annotation. The case of Zaar

Bernard CARON, CNRS, IFRA-Nigeria, USR 3336

Key terms: Oral Corpus Annotation, Macrosyntax, Universal Dependency Grammar, Identifying Clause, Topic

ABSTRACT

This paper argues for a minimal annotation representing in a simple and concise way the interface between information structure and syntax. The article uses the concept of macrosyntax, based on illocutionary units, for a new level of annotation using existing morphosyntactic tiers in Elan. One of the main assets of this system of annotation lies in the notion of piles it uses to represent the oral discursive flow and account for dysfluencies, discontinuities and ellipses. A pilot 15,000 words corpus has been annotated in Elan to run a preliminary study of the information structure of illocutionary components in Zaar, a Chadic language spoken in Nigeria. Their micro- and macro-syntactic properties are represented using Universal Dependencies Grammar.

1 Introduction

This article argues that morphosyntactic glossing of oral corpora is not sufficient for languages with little morphology. A minimal annotation system must be introduced to represent in a simple and concise way the interface between information structure and syntax. The article introduces the concept of macrosyntax, based on illocutionary units, for this specific level of annotation using morphosyntactic tiers in Elan. With the corresponding

annotation script, a pilot 90 min (15,000 words) corpus has been annotated for Zaar, a Chadic language spoken in Nigeria and a preliminary study of peripheries in this language has been done on this annotated corpus.

The article is organised as follows: after this introduction, the second section presents the Zaar language and the corpus collected; the third sections highlights three properties of oral data (dysfluencies, afterthoughts and coordination over turn-taking) that support the need for a different approach to the syntax of oral corpora. Section 4 introduces the concept of macrosyntax and illocutionary units as the foundation of this different approach. Section 5 shows how this approach enables us to annotate, analyse and represent a morphosyntactic ambiguity between left-dislocated topics and unmarked identifying clauses in Zaar. Section 6 introduces an attempt at tagging the Information Structure function of Illocutionary components in Elan. Finally, section 7 establishes a typology of illocutionary components in a 15Kw (15,000 word) oral corpus of Zaar.

2 Zaar and the Zaar corpus

Zaar, also known as Saya, is spoken by about 150,000 speakers in the South of Bauchi State (Nigeria), in the Tafawa Ɓalewa and Ɓogoro Local Government Areas. Together with 30 or so other related languages first identified by Shimizu (1978), Zaar forms a sub-branch of West Chadic languages named the South-Bauchi languages.[1] Apart from the dominant languages, i.e. English (official national language) and Hausa (dominant all over Northern half of Nigeria), South Bauchi languages are surrounded by Niger-Congo languages. Two isolates inside South-Bauchi languages are *Bankal* in the North and *Ɓoi* in the South. Most Zaar

[1] Newman (1990) classified South-Bauchi languages as the B3 sub-branch of West Chadic, but he now treats these languages as a third sub-branch (West-C) within West Chadic (Newman 2006; 2013).

people of the younger generation are Hausa-Zaar bilinguals. They are schooled in Hausa in primary school, before learning English. From a typological point of view, Zaar is a SVO language where TAM is conflated with the exponent of the subject function into a pre-verbal pronominal clitic, as in Example 1:

(1) *ka ɓəl łərtín //*

ka	*ɓəl*	*łərti*	*-ín*
2SG.FUT	dig	root	PROX

'You will dig this root.' // (Moral_Har_069)

This pre-verbal complex does not include the expression of focus to the extent that there is no restriction on which TAMs can be used in a sentence with a focused element. This same portmanteau morpheme can be omitted in sequential clauses – a phenomenon different from subordination, and appearing in narration to indicate consecutive events – and in Serial Verb Constructions. Zaar uses prepositions and the genitival modifier follows the noun it modifies. There is no case marking of object and subject. Zaar does not use relative pronouns, but has a relative subordinator *ɗan*, different from interrogative pronouns (see Caron 2005; 2015a).

The 90 min annotated corpus used for this paper was collected in the 1990's in the village of Tudun Wada, (Bauchi State, Ɓogoro Local Government Area) where the author worked regularly for almost 20 years and became part of the social life. The 11 files have been selected to balance genres (3 traditional animal tales; 3 free conversations; 5 extracts from an interview about Zaar history and culture), gender (5 men and 5 women), and age (from 20 to 75). They have been transcribed, using a phonological orthography marking tone and vowel length, and translated into Hausa by M.S. Davan, a trained and highly competent native speaker.[2] The translation into English, alignment and glossing were done with his help using the Elan-Cortypo programme (see Chanard 2014).

[2] Marvellous S. Davan was constantly involved in the annotation and revision of the transcription which spanned from 2005 when the corpus began, to 2015 when he came to Paris for one month to work with

3 Oral corpora and macrosyntax

Corpus studies in African linguistics must take into account an obvious fact which has methodological but also theoretical consequences: African languages (apart from Arabic and colonial languages such as English, French, etc.) have no written and grammatical tradition. They are oral languages. Oral corpora are greatly structured by features associated with performance: dysfluencies (hesitations, reformulations, pause fillers, aborted utterances) but also the stylistics of oral art, such as rhetorical repetitions, parallel constructions, etc. Grammatical frameworks are not equipped to account for the specificities of oral data. This argues in favour of a new descriptive paradigm and new methods specifically geared at describing oral data. Syntax most commonly takes the sentence as its defining object. However, in oral data, syntactic relations go beyond the sentence, and sometimes, beyond turn-taking. Likewise, dislocated elements, e.g. topics, do not necessarily have a proper syntactic function. A new framework, new tools for annotation, and new tools for syntactic representation, need to be devised so as to take those phenomena into account.

What is often considered as dysfluencies, bits of incomplete sentences, are actually the backbone of the communication process and reveal, when properly analysed, the complexity and intricate structure of this process. Let me give three examples of the specificities of oral corpora and how they can be annotated and represented. These are: dysfluencies, afterthoughts and coordination over turn-taking.

---

me on this paper. Mr Davan was able to use the competence and knowledge acquired in the process to publish *Ɓup Dzanyi Gwaa* in 2010, a book about Zaar history, religion and culture, entirely written in Zaar. Unfortunately, he died suddenly at the end of 2016, aged 40, leaving behind a family of two young children, and an unfinished project of developing an orthography of Zaar for his community. His death is a tragic loss to all of us.

3.1 Dysfluencies

A common configuration of oral corpora that needs to be accounted for concerns dysfluencies, as in (2). The hesitations of the speaker result in repetitions: kə́ | kə́, separated by a pause (#), and teː | teː, separated by a pause filler (ɣə́):

(2) a *Tôː kə́ # kə́ dû̌ teː ɣə́ teː gə̀ʃî tsə́n ŋ.*[3]

tôː	*kə́*	*kə́*	*dû̌*	*teː*	*ɣə́*
DM	2PL.AOR	2PL.AOR	beat	around	FILL

teː	*gə̀ʃî*	*tsə́n*	*kən*
around	downhill	like_this	COP2

'Well you... you would beat it towards er... downhill like this indeed.' (Bury_Har_052)

This type of dysfluencies is pervading oral performances, and has to be taken into account in our description of African languages. An easy solution would be to tap into the speakers' "competence" and ask them to rephrase the sentence, removing the "mistakes" so that it can

---

[3] Zaar is transcribed using the International Phonetic Alphabet, except for /j/ which is transcribed /y/. Vocalic phonemic length is marked after the vowel by single colon (ː). Phonetic length (in fillers, emphasis, etc.) is marked with three colons (ːːː). Phonemic tone is marked with diacritics: *á, à, â* and *ǎ* for High, Low, Falling and Rising respectively. Mid tone is left unmarked.
General Abbreviations: IC, Identifying Clause; IC1, Unmarked Identifying Clause; IC2, Marked Identifying Clause; IlC, Illocutionary Component; IlU, Illocutionary Unit; IS, Information Structure; IU, Intonation Unit; TAM, Tense-Aspect-Mood complex; UD: Universal Dependencies.
Abbreviations in morphosyntactic glossing: AOR, Aorist; ASS, Assertive; COMP, Complementiser; COND, Conditional; COP1, Copula #1; COP2, Copula #2; CPL, Completive; DEF, Definite; DIR, Directional; DIST, Distal; DM, Discourse Marker; EXCL, Exclamation; FILL, Pause Filler; FUT, Future; ICPL, Incompletive; INCH, Inchoative; INDF, Indefinite; ITER, Iterative; LOC, Locative; NEG, Negation; NMLZ, Nominaliser; OBJ, Object; PL, Plural; POS, Possessive; POSL, Possessive Link; PROX, Proximal; QLT, Qualitative; QUEST, Question; REM, Remote Past; RES, Resultative; SG, Singular; VRT, Virtual.
Function tags in UD representation: *advmod*, adverbial modifier; *aux:nsubj*, TAM and subject compound; *ccomp*, clausal complement; *conj:coord*, coordinated conjunct; *conj:dicto*, dysfluency, reformulation, elaboration; *cop*, copula; *csubj*, clausal subject; *dep*, unspecified dependency (e.g. predicative complement of *yi* 'be'); *det*, determiner; *discourse*, discourse element; *dislocated*, dislocated element; *dobj*, direct object; *iobj*, indirect object; *mark*, marker; *obl*, oblique; *punct*, punctuation.

fit into our descriptive frameworks. However, these so-called mistakes are traces of cognitive processes (reformulation, etc.) that are meaningful and need to be documented.

Such a need was integrated into the work that was initiated in the 1970's in France by Claire Blanche-Benveniste and the Groupe aixois de recherches en syntaxe (GARS) with the introduction of the concept of piles (*entassements*). Blanche-Benveniste *et al*. (1990) analysed dysfluencies and reformulations as a paradigm occupying one syntactic slot, and working like coordination structures. A visual representation using their annotation system is used for our Zaar example in (2b):

(2) b *tô:* *ká #*
ká dû *te: ɣá*
te: gàʃî tsán kən
'well you would…
you would beat (it) toward er
downhill like this'

This presents the advantage of reintegrating dysfluencies into syntax, and account for their role in discourse (reformulation, specification, elaboration of thought). The syntactic structure can now be represented e.g. by a dependency graph, as in Figure 1.

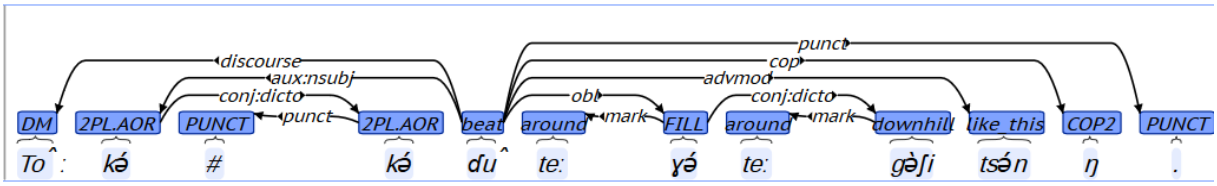


Figure 1: UD representation of Example 2

Formally, a dependency is a directional relation between two words, which is represented by an arrow: the origin of the arrow is called the governor and the target the dependent. Each dependency represents a government relation. In Figure 1, an arrow tagged *aux:nsubj* points from *dû*, 'beat' to *ká*, a 2PL.AOR TAM and Person marker: *ká* is the subject and auxiliary marker of *dû* and is governed by it.[4] A government unit (GU) is a maximal unit for

[4] See Note 3 for the list of UD function tags used in the paper.

government. A GU has a head, which is not governed, e.g. the verb *ɗu*, 'beat' in (2), and all the elements of the GU are dominated by this head. In other words, a GU is the maximal projection of a non-governed lexeme.

This syntactic analysis includes elements (e.g. dysfluencies and reformulations, tagged "*conj:dicto*"[5]) that are usually discarded, which justifies the name "macrosyntax", and the decision to take this all-encompassing GU as the basic unit of our description.

3.2 Afterthoughts

Afterthoughts are another example of the specificities of oral language data, as exemplified in (3):

(3) a *Tô: mə́ ŋgyǎ:r gya: gà:l ɓét ɗaŋ. Kó: gèri kó: ma:t.*

tô:	*mə́*	*ngyá:r*	*gya:*	*gà:l*	*ɓet*	*ɗaŋ*
DM	1PL.AOR	slaughter	PL	cow	all	too

kó:	*gèri*	*kó:*	*ma:t*
or	chicken	or	goat

'Well we slaughtered many cows too. Or hens, or goats.' (Cal_Sdy_032)

In this example, the first intonation unit finishes with the final adverbial adjunct *ɗaŋ*, 'too' and the end of the unit is marked with a terminal prosodic break. Then, as an afterthought, two nouns are added, forming with the direct object of the previous intonation unit a discontinuous chain of three coordinated direct objects (*gyá: gà:l*, 'cows; *gèri*, 'hens' and *ma:t*, 'goat') of the verb *ngyá:r* 'slaughter'. The afterthought forms a second intonation unit starting with a pitch reset and finishing with its own terminal prosodic break.

---

5 Conjunctions are treated asymetrically with the first conjunct as the head, and all the other conjuncts as dependants of this element.

As indicated in section 3.1, the GARS representation of dysfluencies with piles is used for coordination as well. The coordination operating over the final break of the first intonation unit can be represented as in (3b), where the coordinated elements are in a paradigmatic relationship, and inherit their syntagmatic function from the first element of the pile

(3)	b	*tôː mə́ ŋgyə̌ːr*	*gyaː gàːl* *kóː gèri* *kóː maːt*	*ɓét ɗaŋ*
		‘well we slaughtered	cows or hens or goats ’	too

This however poses a dilemma: if one follows the prosodic cues, the constituents of the second intonation unit are syntactic orphans without governor. If one follows the macrosyntactic structure, they are coordinated to *gàːl*, ‘cow’, and inherit their syntactic function from this link, but there is a discrepancy between the intonation and syntactic units. This militates in favour of an analysis called macrosyntax that can account for the phenomena characterising oral corpora. The macrosyntactic level describes the whole set of relations holding between all the segments that make up one and only one illocutionary act:

> *The accomplishment of an illocutionary act is the main property that a language event must have in order to be considered an utterance. […] From an operational point of view the utterance can be defined as the minimal linguistic unit such that it allows a pragmatic interpretation in the world.*
>
> *Cresti & Moneglia (2005: 16)*

Figure 2 represents this analysis in a UD graph.

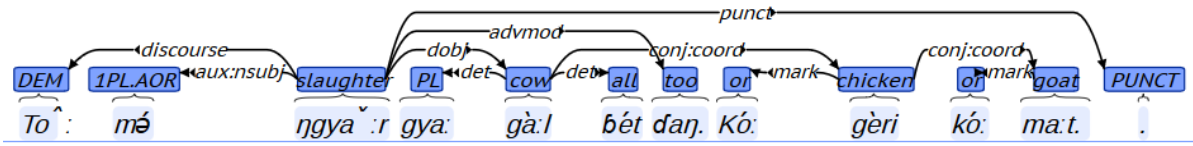


Figure 2: UD representation of Example 3

NB: the governing link that goes from *gàːl* ‘cow’, the first element of the pile, to *geri* ‘chicken’, the second one, cuts across the *advmod* link linking *ɗan* ‘too’ to its governor, *gyáːr* ‘slaughter’, the verbal root of the GU. This feature translates the anomalous position of the adverbial modifier *ɗan* which should occur in final position, and represents the anomalous syntactic structure of the afterthought.

3.3 Syntactic relations over turn-taking

Coordinated piles can occur across turn-taking and result in elliptic structures. But instead of considering those as either incomplete structures or structures where most of the elements have been omitted, they are represented as a special case of coordination across turn-taking, and another case of syntactic relation outside the sentence, hence the use of the term macrosyntax.

It is illustrated in (4) below, part of a passage where the first speaker [S1] is interviewed by [S2] about funeral rites. In this example, the nouns *gàt* ‘woman’ in (4a and e), and *ŋaː gàt* ‘girl’ in (4c) are part of the same pile that spreads over several turn-takings, and share the same syntactic properties as initially stated in (4a). This utterance in (4a) is divided in two parts: the nucleus *tá gìː tə gòs dôːʔ* ‘where will they bury her?’ and the pre-nucleus *tô gàt kən yáː mâs kúmá* ‘well if it is a woman that dies’, a conditional dependant clause whose subject *gàt*, ‘woman’ is identified by the copula *nə*, ‘be’. The identified element is coordinated over several turns of conversation without repeating the rest of the (4a) initial utterance.

(4) a [S1] *tô* ***gàt*** *kən yáː mâs kúmá tá gìː tə gòs dôːʔ*

tôː	*gàt*	*kən*	*yáː*	*məs*	*kúmá*
DM	woman	COP2	3SG.COND	die	too

tá	*giː*	*tə*	*gòs*	*dôː*
3PL.FUT	bury	3SG.OBJ	3SG.POS	where

‘Well if it is *a woman* that dies, they will bury her where?’

b [S2] *gədâ:* ***?***

gət *a:*
woman QUEST

'A woman?'

c [S1] ***kó: ŋa: gət.***

kó: *ŋa:* *gət*
or young woman

'Or a girl.'

d [S2] *ŋa: gət tá gì: ʃĭ ɓáɮən. Kápwâ:sən [...]*

ŋa: *gət* *tá* *gi:* *ʃĭ* *ɓáɮən* *káp* *wá:sən*
young woman 3PL.FUT bury 3PL.OBJ outside all 3PL.POS

'Girls, they would bury them outside. All of them. […]'

e [S1] ***tə́ gət*** *ɓét kó: ?*

tə́ *gət* *ɓét* *kó:*
with woman all or
And women generally? //

f [S2] *m̀: tə́ gət ɓét tá gì: ʃĭ dân.*

m̀: *tə́* *gət* *ɓét* *tá* *gi:* *ʃĭ* *dáni*
er with woman all 3PL.FUT bury 3PL.OBJ there
'Er and women generally, they would bury them there.'

(Bury_Sdy_20)

The elements coordinated across the turns of conversation are linked to the structure of the first question, and inherit their syntactic function from the first element of the pile, as represented in (4'):

(4') *gət* *kən yá: mâs kúmá […]*
kó: ŋa: gət
tə́ gət ɓét

' if it is women that die […]'
or girls
and women in general

This analysis and its accompanying annotation system elegantly underline the coherence of this large passage without postulating the existence of elements that have been deleted through ellipsis. Each element in (4c and e) is linked to the initial utterance (4a), and inherits its referential coordinates from this unit.

3.4 Parallel constructions

(5) *yà:ʃí mə́n::: malâ:ri má:::, lə̌pm za:r má:* ***tá tû:r gyá: dû:, tá tû:r gyá: náɣat***

yà:ʃí	*mə́n*	*malâ:r*	*-i*	*má:*	*lə̌pm*	*za:r*	*má:*
3PL	people	Malar	-INDF	too	moon.of	Zaar	too

tá	*tú:r*	*gyá:*	*dû:*	*tá*	*tú:r*	*gyá:*	*náɣat*
3PL.FUT	PL	cook	PL	beer	3PL.FUT	PL	food

'The people of Malar, (at) the Zaar festival, *they would brew beer, they would cook food*.' (Cal_Har_045)

The different macrosyntactic models acknowledge that sequences such as *tá tû:r gyá: dû:* and *tá tû:r gyá: náɣat* in (5) have to be considered as forming a cohesive unit at some level of linguistic description that should be accounted for.

4 Macrosyntactic corpus annotation

The elements that make up the specificity of oral corpora need to be annotated so as to be retrieved when analysing the corpora and incidentally in order to improve the training of automatic taggers and parsers. To do this, macrosyntactic units in Zaar are annotated with the script developed in the ANR Rhapsodie project (*Corpus de français parlé annoté pour la prosodie et la syntaxe*; Lacheret, Pietrandrea & Tchobanov 2014). This script has proved to be particularly efficient in dealing with the specificities of oral corpora, e.g. piles, dysfluencies, repetitions, discourse markers, overlaps, co-enunciation, false starts, self-repairs and

truncations. This method is data-driven, inductive (the relevant units are identified through annotation) and modular.

The macrosyntactic level describes the whole set of relations holding between all the segments that make up one and only one illocutionary act. The macrosyntactic punctuation marks macro-syntactic boundaries (i.e. illocutionary units and their main components: nuclei, pre nuclei and post nuclei, including discourse markers) and limits between pile layers (dysfluencies, reformulation, coordination).

4.1 Illocutionary Units and basic Illocutionary Components

Each text is segmented into a string of illocutionary units (IlU); each IlU is composed of 3 kinds of Illocutionary Components (IlC): a nucleus (obligatory), pre-nuclei (optional) and post nuclei (optional). Heuristically, in order to identify IlUs and IlCs, annotators rely on prosodic cues perceived while listening to the data that is annotated. Perceptively relevant prosodic cues enable them to identify terminal and non-terminal breaks, the former constituting the IlU limits. They are defined as follows (Cresti & Moneglia 2005:17):

a. Prosodic break*: perceptively relevant prosodic variation in the speech continuum such as to cause the parsing of the continuum into discrete prosodic units.*

b. Terminal prosodic break*: given a sequence of one or more prosodic units, a prosodic break is considered terminal if a competent speaker assigns to it, according to his perception, the quality of concluding the sequence.*

c. Non-terminal prosodic break*: given a sequence of one or more prosodic units, a prosodic break is considered non-terminal if a competent speaker assigns to it, according to his perception, the quality of being non-conclusive.*

The basic prosodic distinction in Zaar is between pre-nucleus units whose boundary is characterised by a level intoneme followed by an initial step-up (pitch reset) at the onset of the following unit; and final prosodic breaks signalling the end of an IlU by a fall. The final fall can be replaced by or combined with other intonemes (e.g. rise and high-rise) in case of emphasis or exclamation (see Caron 2015b:17).[6]

IlCs are annotated as follows: "<" follows a pre-nucleus and precedes a nucleus or another pre-nucleus; ">" precedes a post-nucleus and follows a nucleus or a previous post-nucleus; and "//" indicates the right boundary of an IlU.

4.1.1 Nuclei

Nuclei bear the main prosodic prominence in the utterance. They are identified as the locus of the illocutionary force. Nuclei are usually governed by a tensed verb, as in (6), but not always. See e.g. (7) where a noun (*laː,* 'work') is governing the nucleus, and (8) where the whole nucleus of the second IlU is an exclamation (*kâːy* 'hey').

(6) *féːlêks < kyâːn máː < **káː rigá kə yisə́n tíː** "áy" //*

féːlêks	*kyáːni*	*máː*	*káː*	*rigá*
felix	2SG	even	2SG.CPL	already

kə	*yisə́ŋ*	*=tə*	*-íː*	*áy*
2SG.AOR	know	=3SG.OBJ	-RES	indeed

'Felix < you too < *you know him* eh. //' (Girls_B_092)

(7) *gíː < **ŋaː laː** ɓastə //*

gíː	*ŋaː*	*laː*	*ɓas*	*=tə*
DIST	small	work	at	3SG

'That's *easy* for him.' (*lit*. 'that < *small work* at him //') (Girls_B_094)

[6] See also Section 5.3 for a more detailed presentation of intonation patterns in Zaar.

(8) *kúní: à: məs ɓasəm sò:séy // myâ:n kúmá < **kâ:y** //*

kúni	*-í:*	*à:*	*məs*	*ɓas*	*=mə*	*sò:séy*	*myá:ni*	*kúmá*	*kâ:y*
boy	-DIST	3SG.CPL	die	at	1SG.OBJ	quite	1SG	also	EXCL

'That boy is dying for me. Myself, *I don't care*!'
(*lit.* 'as for me < *hey!*') (Girls_B_087)

4.1.2 Pre- and post-nuclei

Pre-nuclei include topics, left-dislocated adjuncts, IlU introducers and associated IlUs.

Example 9 shows an instance of a pre-nucleus topic and a post-nucleus associated IlU:

(9) ***ndà:ɗə̂m má:*** *< má łə́ yéltə > "**áy**" //*

ndà:ɗə̂m	*má:*	*má*	*łə*	*yel*	*=tə*	*áy*
Ndadəm	even	1PL.FUT	go	see	=3SG.OBJ	eh

'*Ndadem too* < I will go and see him > "*eh*". //' (Girls_A_005)

Post-nuclei include right-edged topics, afterthoughts, associated IlUs and backgrounded elements in marked identifying clauses (see Section 5).

4.2 IlU introducers

Illocutionary Unit introducers are conjunctions like 'but', 'then', 'since', etc. They specify the nature of the relation between the IlUs they introduce and other IlUs in the discourse, especially the preceding one; they have no syntactic dependency with any other elements in the IlU; they are blocked to the initial position of the IlU. Subordinating conjunctions can work as an IlU introducer when the clause that follows is not integrated into a higher IlU. In the macrosyntactic annotation, they are preceded by a circumflex accent.

(10) ***^sé: ^ɗan*** *tə́ măni //*

sé:	*ɗan*	*tə́*	*mán*	*-i*
then	then	3PL.AOR	come	-INDF

'*Then*, they came back.' (Mbrt_S1_114)

4.3 Associated Illocutionary Units

These are expressions like ‘well, you know, I think’, etc. They are not autonomous IlUs but they are not part of the dependency frame of the predicate in the nucleus.[7] Together with pause fillers, the most frequent associated IlU in Zaar is *tôː*, ‘well’. In the macrosyntactic annotation, they are written between quotes:

(11) *“**tôː**” < zəgì < ʧì gòs < dzàŋ gyòː //*

tôː	*zəgì*	*ʧì*	*gòs*	*dzàŋ*	*gyôː*
well	Ziggy	3SG.be	3SG.POS	day	which

‘ “*Well”* < Ziggy < as for him < when was that?’ (Boys-A_457)

4.4 Piling

Kahane & Pietrandrea (2012) propose a complete annotation and a functional tagging of piles. The format is: { — | — }. See (2c) below for the annotation of Example 2 in Section 3:

(2) c “*tôː”* ***{ kə́ | kə́ }*** *dû* ***{ teː ɣə́ | teː gəʃì }*** *tsə́n ŋ //*

‘ “Well” < *{ you’d | you’d }* beat (it) *{ toward er | downhill }* like this //’

Apart from dysfluencies and reformulations, piles note a syntactic relation where text segments occupy the same position in the dependency structure. They appear within IlUs, e.g. in coordination, as in (12). The sign “^”, e.g. in ^*kóː* --- ^*kóː*, ‘^or --- ^or’ identifies words working as conjunctions.

(12) a *wáːni { àː fin & | àː ŋgap ŋaː gədî wáːni maráy | ^kóː àː ʧì mə̂ːr | ^kóː á fì maːndə } //* [8]

wáːni	*àː*	*fi*	*àː*	*ngap*	*ŋaː*	*gət*	*-i*
that_person	3SG.CPL	do	3SG.CPL	catch	young	woman	INDF

[7] See Note 9 on microsyntax.

[8] The “&” indicates an aborted IlC.

wáːni	*mará*	*-íː*	*kóː*	*àː*	*ʧî*	*mâːr*
that_person	spoil	RES	or	3SG.CPL	eat	stealing

kóː	*á*	*ʃi*	*maːndə*
or	3SG.AOR	do	fight

‘So-and-So {has done & | has caught So-and-So’s daughter (and) spoilt her | ^or has stolen | ^or has gone into a fight}’ // (Rel_Har_188)

Following Blanche-Benveniste *et al.* (1990) this pile can be also represented as in (12b):

(12) b

wáːni	*àː ʃin &*
	àː ŋgap ŋaː gədî wáːni maráy
	^kóː àː ʧi mâːr
	^kóː á ʃi maːndə

‘So-and-So	has done &
	has caught So-and-So’s daughter and spoilt her
	^or has stolen
	^or has gone into a fight’

4.5 Non-alignment of Illocutionary Components and Governing

When all the constituents belonging to the dependency frame of the root of the IlU are contained inside the nucleus, IlUs and GUs are aligned. When they are not aligned and dependency links cross illocutionary boundaries, these are followed by a “+”, e.g. “//+”, “<+” and “>+”. In these cases, neither Intonation Unit (IU) boundaries nor turn-taking work as boundaries of syntactic dependencies. The annotation of piles is modified to mark the resulting discontinuous piling as follows: { --- |} {| --- }.

4.5.1 Piling across Intonation Unit boundaries

Example 3 of Section 3.2 on afterthoughts showed an example of piling across IU boundaries. It is annotated in (3c) to represent the piling that links the nouns in the afterthought to the direct object of the nucleus across the IU boundary:

(3) c *"tô:" má ŋgyă:r {gya: gà:l |} ɓét ɗaŋ //+ {| ^kó: gèri | ^kó: ma:t} //*

‘"Well" we slaughter plenty {cows |} too //+ {| ^or hens | ^or goats}. //’ (Cal_Sdy_ 032)

Another illustration is given in (13), where the “//+ sign” shows that *yâ:n nə myâ:n,* ‘if it’s me’ is a clausal adjunct added to the IlU as an afterthought.

(13) *"tô:" dzàŋ gí: ɣəŋ >+ tá fì mátaŋgáy //+ yâ:n nə myâ:n //*

tô:	*dzàŋ*	*gí:*	*kən*	*tá*	*fi*	*mátaŋ*	*káy*
DM	day	DIST	COP2	3PL.FUT	do	ritual_flogging	LOC

yâ:n	*nə*	*myá:ni*
if	COP1	1SG

‘"Well” *it’s that day*>+ they will do matang. //+ If it is me. //’ (Bury_Ha_201)

NB: (13) is an utterance with an identified structure, where the “>+” sign indicates that the post-nucleus unit is in a dependency relationship with an element of the nucleus.

4.5.2 Piling across turn-taking

In the so-called elliptic constructions seen in section 3.3 (see Example 4), coordination occurs across turn-taking. It is annotated for macrosyntax in (4’):

(4’) a [S1] *“tô” { **gàt |}** kən yá: mâs kúmá tá gì: tà gòs dò: ?//*

‘ “Well” and if it is { *a woman* |} that dies, they will bury her where ?//’

b [S2] *gàdà: ?//*

‘A woman ?//’

c [S1] *{| **^kó: ŋa: gàt** |} //*

‘{| ^*Or a girl* |} //’

d [S2] *ŋa: gàt tá gì: fí ɓáƔàŋ > kápwâ:sàŋ […] //*

‘Girls they would bury them outside > all of them.[…] //’

e [S1] {| ^***tá gàt ɓét*** *kó:* } //?

{| ^*And women generally* } ?//

f [S2] *“m̀ː” tə́ gə̀t ɓét< tá gìː ʃĩ ɗân //*
‘ “Er” and women generally < they would bury them there //’(Bury_Sdy_20)

4.5.3 Left-dislocated circumstantial adjuncts

Left-dislocated circumstantial adjuncts have the same prosodic profile as topics, but they have a different function. It is agreed that adverbials and other circumstantial adjuncts are frame-setters that limit the applicability of the main predication to a certain restricted domain (see Chafe 1976). Using the concept of common ground, Krifka & Musan (2012) establish a difference between contrastive topics and frame-setters which remains valid when contrasting frame-setters with aboutness topics:

> *With contrastive topics, the current common ground management contains the expectation that information about a more comprehensive, or distinct, entity is given; contrastive topics indicate that the topic of the sentence diverges from this expectation. With frame setters, the current common ground management contains the expectation that information of a different, e.g., more comprehensive, type is given, and the frame setter indicates that the information actually provided is restricted to the particular dimension specified. (Krifka & Musan 2012:32)*

In Zaar, this difference in the management of information is paralleled by a syntactic difference which confirms that topics and frame-setters belong to different functional levels: topics are pragmatic, belong to Information Structure, whereas frame-setters belong to the dependency frame of the verbal root. In (14) and (15) for example, no adverb (such as *ɗâni*, ‘there’) need modify the verbs *tu*, ‘meet’. This is done by the left-dislocated adjunct *ɗa gìp kìmsə́y*, ‘in Kimsə’.

(14) ***ɗa gìp kìmsáy*** **<+** *kə łə́ tutə //= kə fu:tə //*

ɗa	*gìp*	*kímsə*	*-í:*	*kə*	*łə*	*tu*	*=tə*
at	inside	Kimsə	-DIST	2SG.AOR	go	meet	3SG.OBJ

kə	*fu:*	*=tə*
2SG.AOR	tell	3SG.OBJ

'In Kimsə <+ you go meet him //= (and) tell him.' (Boys_B_188)

The same is true for (15) where the frame-setter *ɗaŋgənín tsə́n*, 'right now' is dependent on the verb *ɲom*, 'wrestle', and this has no temporal adjunct as a dependant modifier inside the nucleus.

(15) *"tô:"* ***ɗaŋgənín tsə́n*** *<+ tá ɲôm tə́ kádî //*

tô:	*ɗangəní*	*tsəní*	*tá*	*ɲom tə́*	*kádî*
well	now	like_this	3PL.FUT	take with	dog

' "well" *right now* <+ he will wrestle with Dog. //'
(Hyena_S1_282)

The adjunct is dependent on the verb it modifies. It contributes to the semantic component of the nucleus by restricting the circumstantial scope of its referential value. When a circumstantial adjunct appears in pre-nucleus position (left-dislocation, PR-Adv), it keeps a direct dependency relation with the verb, and no clitic or resumptive element is needed. This is marked in our macrosyntactic punctuation by the plus sign added to the chevron (<+). Another tests is given by their ability to be cleft. Clausal adjuncts, whether circumstantial or conditional share the prosodic and syntactic properties of adverbial adjuncts, and are annotated in the same way (<+).

(16) ***ɗam mə łə́ tulí:*** **<+** *sé: mə łə́ tu::: è: gàri gón //=*

ɗan	*mə*	*łə*	*tul*	*-í:*	*sé:*	*mə*	*łə*	*tu*	*è:*	*gàri*	*gón*
as	1SG.AOR	go	arrive	RES	then	1SG.AOR	go	reach	fill	town	QLT

'As I arrived <+ then I reached er... a village. //=' (Boys_A_151)

(17) ***yá: măni*** *<+ wò łyan wàhála > áy //*

yá:	*man*	*-i*	*wò*	*łya*	*-ni*	*wàhála*	*áy*
3SG.COND	come	-INDF	3SG.FUT	drink	-INCH	suffering	eh

'If she comes <+ she will suffer > indeed. //' (Boys_B_289)

5 Left-dislocation and marked identifying clauses in Zaar

I have argued in the preceding section that macrosyntactic corpus annotation is necessary to account for specific phenomena pertaining to oral corpora. In this section I argue that macrosyntactic analysis, annotation and representation of illocutionary units provide the tools to disambiguate utterances where morphosyntactic tagging alone cannot differentiate between e.g. left-dislocated topics and marked identifying clauses, also known as "it-clefts" in English syntax.

5.1 Topics

Topics are pre-nucleus IlCs that introduce a referent, selected out of the on-going conversation, or of the common knowledge of the speakers. These referents provide the necessary pragmatic information for the illocutionary act carried by the following nucleus. Topics do not enter in a *microsyntactic* relationship with the verbs of the nucleus.[9] When a topic is in a pragmatic relation with the verb, and this relation is part of the dependency frame of the verb (the element could be an argument of the verb), the syntactic relation must be realised as a clitic so that the valency of the verb is saturated. In (18), the topic *gə̀ːrí raːs,* 'old locust-bean tree' is co-referential with the adverb *ɗànì*, 'there', an adjunct of the verb *gìː*, 'bury'. In (19), a clitic (the direct object pronoun *ʃĩ*, 'them') saturates the dependency frame of the verb *gìː*, 'bury'.

[9] Microsyntax concerns dependency between a head and elements instantiating its dependency frame, e.g. the complements of a verb. Microsyntax is a subdomain of macrosyntax.

(18) *gyá: gà~ ra:sán tsán < tá gì: ʃí* ***ɗân*** *//*

gyá:	*gà:rí*	*ra:s*	*-ín*	*tsaní*	*tá*	*gi:*	*ʃí*	*ɗáni*
PL	old	locust_bean	PROX	like_this	3PL.FUT	bury	3PL.OBJ	there

'These old locust-bean trees like this < they would bury them *there*. //' (Bury_Har_109)

(19) *ŋa: gàt < tá gì:* ***ʃí*** *ɓáɮàŋ //+ kápwâ:sàŋ //*

ŋa:	*gàt*	*tá*	*gi:*	*ʃí*	*ɓáɮàŋ*	*káp*	*wá:sàŋ*
young	woman	3PL.FUT	bury	3PL.OBJ	outside	all	3PL.POS

'Girls < they would bury *them* outside //+ all of them. //' (Bury_Har_103)

This points to the fact that topics, separated from the nucleus by a prosodic break, do not enter in a microsyntactic relation with the verb or the predicate of the nucleus, and their relation to the nucleus is not microsyntactic but pragmatic. In other words, the repetition of the topic or the presence of a resumptive pronoun confirms the non-compositional nature of topics, working as a syntactic island (see Cresti & Moneglia 2005:34–38). Their relation to the nucleus is pragmatic, and is best defined by the notion of "aboutness" (see Sperber & Wilson 1986; Cresti 2012). The information function of the topic is to identify, through linguistic means, the domain of relevance for the illocutionary force carried by the nucleus, its pragmatic domain of identification. This is conveyed by the name "aboutness topic" commonly used to refer to this construction (see Krifka & Musan 2012; Schultze-Berndt 2013; Simard 2014).

5.2 Marked identifying clauses in Zaar

Marked identifying clauses are a subclass of identifying clauses as defined by Halliday (1967) in his series of articles on *Transitivity and Theme in English*. They translate in English as what is usually called "it-clefts" and are exemplified in (20).[10]

(20) *máː* ***nə zəgì*** *átâ wul vèːsòː*

máː	*nə*	*zəgì*	*átâ*	*wul*	*vìː*	*-ês*	*-oː*
even	COP1	Ziggy	3SG.REM	say	mouth	-DEF	-ASS

Actually, *it is Ziggy* who said it.

Halliday explains that any clause such as *John saw the play* can be organised into a "cleft sentence" with equative form through the nominalisation of one set of its elements, e.g. *what John saw was the play*. The former, without the nominalisation, is non-identifying and the second is identifying. The identifying clause adds the further information that one of the participants is definable by participation in the process. In an identifying clause, it is always the nominalization which is "to be identified". A further division is made by Halliday between Marked and Unmarked Identifying clauses:

> *There is thus an association of variable – value with theme – rheme similar to that of identified – identifier with given – new: in the unmarked case, the identified is given, the identifier new, and the variable is theme, the value rheme. […] in a sense a theme is a variable to which a value is to be assigned. But as always the speaker may exploit the contrastive possibility of not*

---

[10] Marked identifying clauses are often described under the name *it-cleft*, after Higgins (1973), referring to their morphosyntactic exponents in English. Cf. for example Huddleston & Pullum (2008:1414) where *it-clefts* are defined as "a bi-clausal copulative construction consisting of an impersonal pronoun (the cleft pronoun), a copular verb, the informationally prominent phrase (the cleft phrase) and an embedded relative clause (the cleft clause)". However, Zaar does not have an impersonal pronoun corresponding to the 'it' of English "it-clefts", and the copula can be omitted. This is the reason why I opted for Halliday's semantic approach, as less language-specific.

mapping the variable on to the theme; hence to the unmarked, operative

[Type(1) what John saw was the play] corresponds a marked, receptive form

[Type(2) it was the play that John saw]. (Halliday 1967:228)

I propose to name Type (1) UNMARKED IDENTIFYING CLAUSE (IC1); and Type (2) MARKED IDENTIFYING CLAUSE (IC2). Typical ICs in Zaar are exemplified below, starting from the non-IC (21) where the root of the utterance is the verb *wul* 'say':

(21) *^kəndá zəgì átâ wul vèːs //*

kəndá	*zəgì*	*átâ*	*wul*	*vìː*	*-ês*
then	Ziggy	3SG.REM	say	mouth	-DEF

'^Then Ziggy spoke //' (lit. 'said the speech')

In the corresponding IC1 in (22) the utterance in (21) is split. The subject 'Ziggy' identifies the variable 'the man who spoke'. The identification is marked by the non-verbal copula *nə* 'be'. The resulting Identifying Clause is called *Unmarked* because the order Identified/Identifier corresponds to the unmarked order Theme/Rheme (see Halliday *ibid.*).

(22) *"máː" dàːsóːdă̰ːtâ wul vèːs <+ nə zəgyòː //*

máː	*dàːsóːda*	*átâ*	*wul*	*vìː*	*-ês*	*nə*	*zəgì*	*-oː*
even	the_one_who	3SG.REM	say	mouth	-DEF	COP1	Ziggy	-ASS

Actually, the one who spoke is Ziggy. (Boys-A_455)

In the corresponding Marked Identifying Clause (IC2) in (20) repeated below, the order Theme/Rheme (Identified/Identifier) is reversed, and the Rheme ('Ziggy', the Identifier) comes first. IC2 is pragmatically marked through rhematisation.

(20) *"máː" **nə zəgì** >+ átâ wul vèːsòː*

máː	*nə*	*zəgì*	*átâ*	*wul*	*vìː*	*-ês*	*-oː*
even	COP1	Ziggy	3SG.REM	say	mouth	-DEF	-ASS

"Actually" *it is Ziggy* >+ who said it //

In (20) and (22), the copula *nə* ‘be’ does not have an expletive pronoun. This is the rule for the two copulas used in Zaar for identification with the meaning ‘(it) is X’: *nə* X (COP1, in (23)) and X *kən* (COP2, the most frequent, in (24)):

(23) ***nə łərtín*** *>+ ka ɓəl fá: //*

nə	*łərti*	*-ín*	*ka*	*ɓəl*	*fá:*
COP1	root	PROX	2SG.FUT	dig	indeed

‘(It) is this root >+ (that) you will dig indeed. //’ (Moral_Har_069)

(24) *“tô:” tə́ yîsə́ŋə́y tu* ***kyâ:ŋ*** *>+ mbwá:tə //*

tô:	*tə́*	*yisə́ŋ*	*-í:*	*tu*	*kyá:ni*	*kən*	*mbwa:*	*=tə*
DM	3PL.AOR	know	RES	COMP	2SG	COP2	shoot	3SG.OBJ

‘ “Well” they know that *(it) is you* >+ (who) shot it. //’ (Hunt_Har_047a)

Contrary to what was observed with topics, no resumptive pronoun appears with the verb or nominal predicate of the Identified. In (24), *mbwá: tə*, ‘shoot it’ has no subject clitic standing for the variable identified by *kyâ:n* ‘you’, nor does any pronoun stand for ***łərtín***, ‘this root’ in (23). This shows that the Identifier is still in a microsyntactic relation with the predicate of the Identified, hence the annotation with a “+’ added to the chevron.

Zaar can even omit the copula altogether. In (25), no copula is used for the IC2 structure. This example is taken from a conversation where two young girls complain that they stayed idle at home the previous week-end. They promise themselves that this Sunday, they will not merely go out, but ‘(it is) *everywhere’* (that) they will go.

(25) *“â:” < dzàŋ là:dì má: <+* ***kakáp*** *>+ má gè:wàyéy //*

â:	*dzàŋ*	*lá:dì*	*má:*	*kakáp*	*má*	*ge:wayé*	*-í:*
ah	day	Sunday	even	everywhere	1PL.FUT	walk_around	RES

‘ “Ah” < on Sunday indeed <+ *(it is) everywhere* >+ (that) we will stroll. //’ (Girls_A_010)

When the copula is omitted, the absence of morphological marking on the Identifier (e.g. *kakáp* ‘everywhere’ in (23)) can result in ambiguity between an IC2 and a compound IlU with

a pre-nucleus topic. See, e.g. (26) and (27) where there is no morphological change between *kyâːn* 'you' in (26), where it is a topic, and in (27), where it is an Identifier:

(26) *Féːlêks < kyâːn máː <* ***káː rigá kə yisə́n tíː*** *"áy" //*

Féːlêks	*kyáːni*	*máː*	*káː*	*rigáː*	*kə*	*yisən*	*tə*	*-íː*
Felix	2SG	even	2SG.CPL	precede	2SG.AOR	know	2SG.OBJ	-res

'Felix < you yourself < *you know him* "eh". ' (Girls_B_092)

(27) *"wókèː"* ***kyâːn*** *>+ kyáː ʧaːtəɣáy > ŋǎːn //*

wókèː	*kyáːni*	*kyáː*	*ʧaː*	*=tə*	*káy*	*ŋǎːn*
ok	2SG	2SG.ICPL	put	3SG.OBJ	LOC	quest

'"OK" *it's you* >+ who made him do it > isn't it?' (Hyena_S1_319)

In (27), the personal pronoun *kyâːn* 'you' is rhematised, and bears a prosodic prominence, whereas in (26) the pronoun *kyâːn* is topicalised (through a continuative prosodic contour) and the predicate *káː rigá kə yisə́n tíː* 'you know him' bears the prosodic prominence. In both cases, the pronoun *kyâːn* precedes the predicate, and neither the pronoun nor the predicate carry a morphological exponent of their rhematic or thematic status. The change in the rhematic status of the illocutionary components is expressed only through intonation, through the change in the main prosodic prominence from the predicate to the Identifier in the IC2 structure.

5.3 The prosody of topic and identifying clauses in Zaar

Let us see a brief illustration of the intonation patterns characterizing those utterances and start with the "neutral" declarative sentence in (28), and its corresponding pitch track in Figure 3.[11]

[11] The examples of intonation patterns described in this paper are quoted from Caron (2015b), Caron et al. (2015), a detailed study of tone, intonation and information structure in Zaar.

(28) *á lə̌ːrmí ŋáːwôs mə́ndî mà jèlí oː //*

á	*lə́ːr*	*=mí*	*ŋaː*	*=wôs*	*mə́n*	*-dî*
3SG.AOR	bring	=1PL.OBJ	son	=3SG.POS	ben	-DIR

mà	*jel*	*-i*	*-oː*
1PL.AOR	see	-INDF	-ASS

'He has brought his son for us to see.' (SAY_BC_CONV_02_SP2_029)

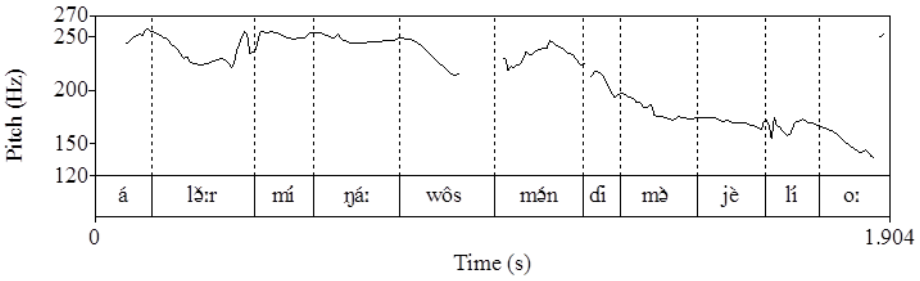


Figure 3: Pitch track of Example 28

This "neutral" intonation pattern is characterised in Zaar by a combination of declination and a final fall. This intonation pattern obtains for all types of sentences: assertions (both positive and negative), Wh-Questions and Yes/No-Questions.

By contrast, topics are pre-nucleus units characterised by various prosodic cues separating them from the comment. The two main cues that are always present are: suspension of declination, followed by a pause. These two exponents can be reinforced by a lengthening of the last segment of the topic, pitch reset and/or change of register. An example is provided in (29) and its pitch track in Figure 4, where the second topic, *ɮàmdî gòsdîːːː* 'the place where she goes', is separated by a change to a much lower register. The first topic *mə̌ːm móːmi kúma* 'as for Momi's mother' is a topic specified by the discourse particle *kuma* 'as for'.

(29) *mə̌ːm móːmi kúmá < ɮàmdî gòsdîːːː < ʧáː fini gòs <+ koyarwa makaranta //*

maːm	*ká*	*móːmi*	*kúmá*	*ɮam*	*-dî*	*gòs*	*-dî*
mum	POS	Momi	too	return	-DIR	3SG.POS	-DIR

ʧáː	*fi*	*-ni*	*gòs*	*koyarwa*	*makaranta*
3SG.ICPL	do	-INCH	3SG.POS	teaching	school

'As for Momi's mother < the place where she goes (lit. her going) < what she does <+ is teaching children in school.' (SAY_BC_CONV_02_SP1_023-6)

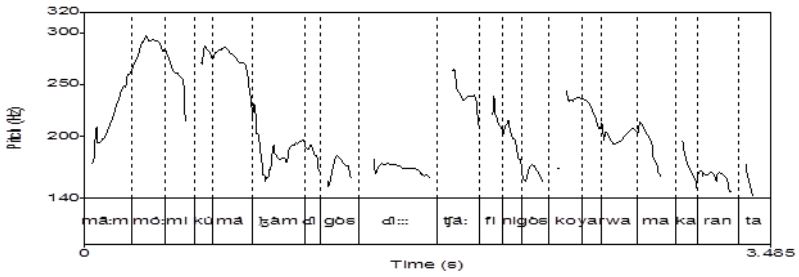


Figure 4: Pitch track of Example 29

As for IC2s, they constitute a single intonation unit with no break, and are characterised by a fall from a main stress falling on the identifier element, as in (30), and its pitch track represented in Figure 5:

(30) ***tákwâːràs ŋ*** *>+ átá mán tum //*

tákwâːràs =kən átâ man tu =mə
Takwaras =COP2 3SG.REM come meet =1SG.OBJ

[…] *it's Takwaras* >+ who came to meet me […:] //
(SAY_BC_CONV_03_SP1_695)

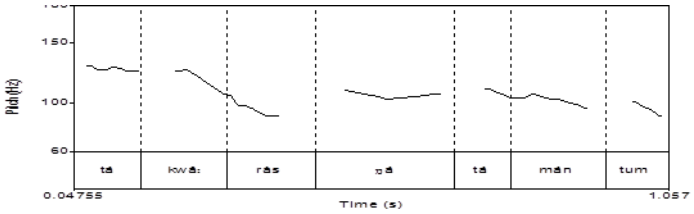


Figure 5: Pitch track of Example 30

In the corpus, out of a total of about 1,400 utterances, 586 have been tagged as compound utterances with pre- or post-nuclei, while 571 have been tagged as simple (thetic, all-new) and 108 have been tagged as marked Identifying Clauses (IC2).

5.4 Syntactic representation

The differences in the properties explored in the previous section can be neatly represented using dependency graphs, as developed e.g. in the Universal Dependencies Grammar project (de Marneffe, Dozat *et al*. 2014) and the annotating tool *Arborator* (Gerdes 2013). The

tagging of peripheries, discourse markers, etc. has been adapted to account for the properties described in Zaar (see Note 3).

5.4.1 *Topic*

Topics are represented as peripheric to the root, e.g. in (31) the IlCs *la:* ‘work’ and *mə:ríwôpm* ‘our children’. The two topics have been labelled as *dislocated* in the graph (see Figure 6).

(31) *la: < mə:ríwôpm < ʧî gwà:sàŋ <+ tə́ la: hń //*[12]

‘As for work < our children < they themselves <+ (they) don't have any work.’ (Wom_B_221)

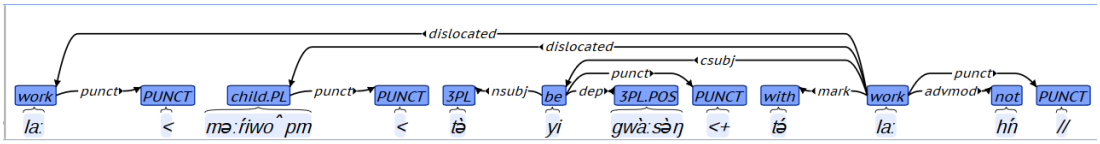


Figure 6: UD representation of Example 31

5.4.2 Marked Identifying Clause (IC2)

In UD analysis, the identifier of e.g. (32) is the root of the graph, and the identified variable (*tàtàyá: fû:mí ʧî:*, ‘(that) they used to tell us like this’) is a dependent of the identifier (see Figure 7).

(32) *“tô:” < **gí:** >+ tàtàyá: fû:mí ʧî: //*

‘ “Well” < *it is this* >+ (that) they used to tell us like this. //’ (lit. THIS, they told us like this.) Moral_Har_088

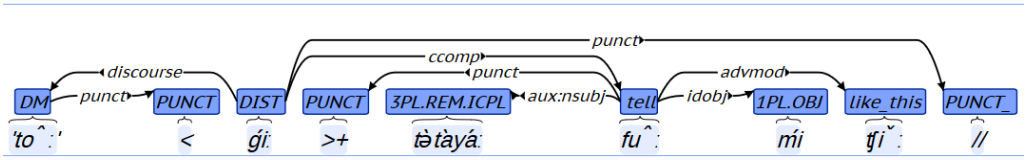


Figure 7: UD representation of Example 32

When the copulas *nə* or *kən* are used, they are represented as dependents of the lexical predicate, i.e. the Identifier, as in (33) and (34) below (see Figures 8 and 9).

12 *ʧî* is analysed as the fusion of the 3PL subject pronoun *tə* plus the defective verb *yi* ‘be’.

(33) ***nə ɬərtín*** *>+ ka ɓəl > fá: //*

‘(It) is this root >+ (that) you will dig > indeed. //’ (Moral_Har_069)

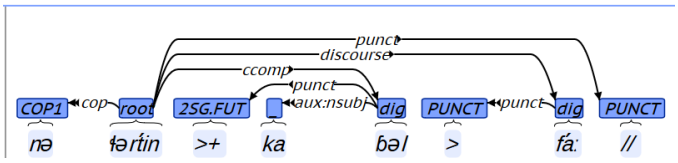


Figure 8: UD representation of Example 33

(34) *“tô:” < **yá:ni kən** >+ wò fi wuki gín dányâ:lín //*

‘ “Well” < it is THIS >+ (that) will make this very medicine. //’ (Moral_Har_076)

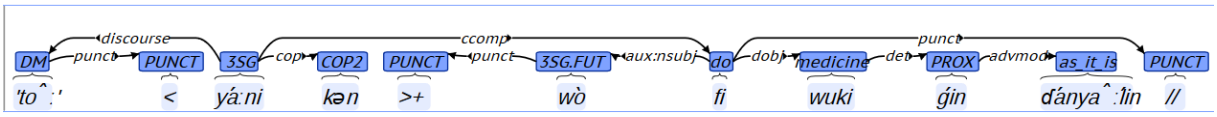


Figure 9: UD representation of Example 34

This analysis follows the UD general principle that only words with lexical content can be governors. The optionality of non-verbal copulas in Zaar as in many languages (e.g. Russian) reinforces this rule (de Marneffe, Ginter, et al. 2014).

6 Macrosyntax and Information Structure annotation in Elan

For retrieval purposes, a further step in corpus annotation was attempted for this work on Zaar by tagging in Elan the Information Structure (IS) function of the IlCs. Information Structure tagging was done with a new module of ElanCorpA that is being developed by M. Aouini & C. Chanard at Llacan, as part of the Cortypo programme.[13] This module is a new type of annotation, based on the annotation tiers that already exist in the CorAfroAs / Cortypo format. This new functionality in Elan is meant to create annotations on a dependent tier that cover

[13] The Cortypo programme directed by A. Mettouchi (http://cortypo.huma-num.fr/index_fr.html) is a follow-up of the CorpAfroAs programme (see Mettouchi, Vanhove & Caubet 2012; Chanard 2014).

non-contiguous annotations of the parent tier. For a given annotated file in the “classical” Elan format, extra annotations can be created as new lines in two sets of tables: Groups and Links. Individual groups and links in the table can then be highlighted in the annotation tiers when selected, and the corresponding passage in the sound file can be played. The file can be searched, with multiple criteria including tier annotations, table, and distances in terms of alignment, annotation and time span. These tables can be sorted by types, names or annotations, which has a great heuristic value and opens new possibilities for structural annotations (whether informational or syntactic) in Elan. Figure 10 shows a screenshot of Example 13 annotated for IS with Elan and the Links and Groups module.

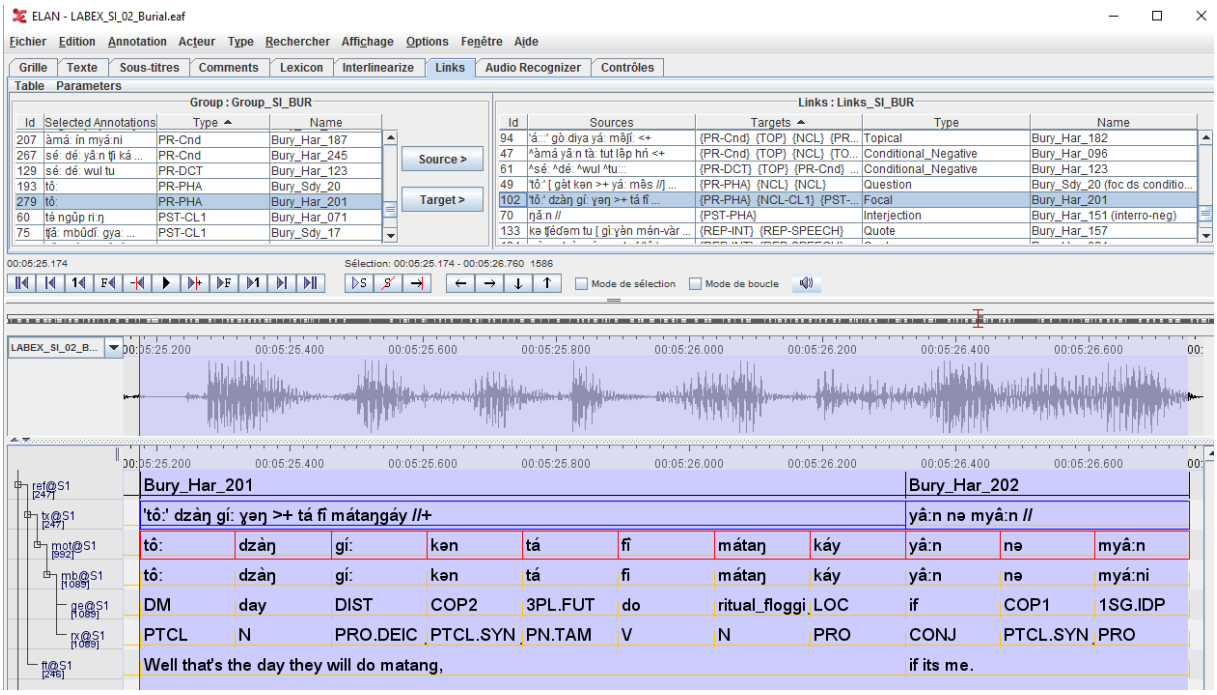
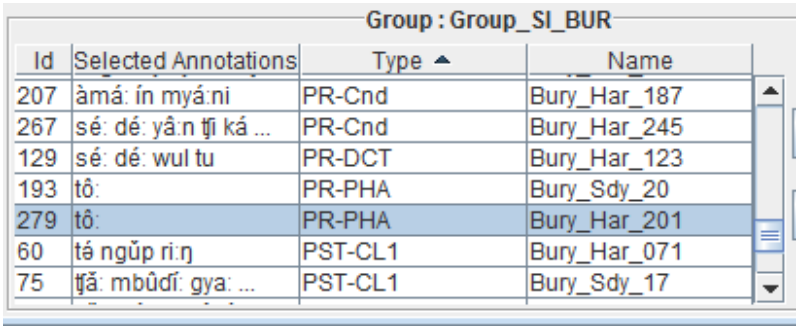


Figure 10 : Screenshot of Example 13

In the first table (called Groups, top left of the screen), to create a group, the annotator selects a set of annotations in any of the existing tiers, gives this group a name and a type that can be selected in a controlled vocabulary. These sets consist of a single or several annotations that can be selected from one or several tiers, and can be discontinuous.

For this work, I used the Groups table to identify sets of words that make up Illocutionary Components (IlC: Nuclei, Pre-nuclei, Post-nuclei and In-nuclei), and tagged them with their function (Type) and reference number (Name) (see Figure 11).



Group : Group_SI_BUR

Id	Selected Annotations	Type	Name
207	àmá: ín myá:ni	PR-Cnd	Bury_Har_187
267	sé: dé: yâ:n ʧi ká ...	PR-Cnd	Bury_Har_245
129	sé: dé: wul tu	PR-DCT	Bury_Har_123
193	tô:	PR-PHA	Bury_Sdy_20
279	tô:	PR-PHA	Bury_Har_201
60	tə́ ngŭp ri:ŋ	PST-CL1	Bury_Har_071
75	ʧă: mbûdî: gya: ...	PST-CL1	Bury_Sdy_17

Figure 11. Groups table sorted by Type
with the phatic pre-nucleus of Example 13 selected

In the second table (called Links, top right of Figure 10), the annotator creates links between two sets of annotations built on the same principle as groups. One set is called the Source, and the other set is called the Targets. The links created are given a Name and a Type in the same way as for groups. The sources or the targets can also be taken from the Groups table. In this case, the sets selected from the Groups table can be viewed either by showing the annotations in the tiers, or the types and names given to the groups in the Groups table.

For this work, I have used the Links table to tag the Illocutionary Units. For better readability and convenience sake, the table shows the full text on the text tier as the source of the links and the IlCs (groups) tagged in the Groups table as targets. I have used the "type" column of the Groups table for a temporary, rule of thumb functional tagging of IlUs, indicating whether they contain e.g. questions, conditionals, rhetorical devices such as parallel IlUs, etc. (see Figure 12).

Links : Links_SI_BUR

Id	Sources	Targets ▲	Type	Name
94	'á:::' gò:diya yá: mə̂jí: <+	{PR-Cnd} {TOP} {NCL} {PR...	Topical	Bury_Har_182
47	^àmá yâ:n tà: tut lə̂p hń <+	{PR-Cnd} {TOP} {NCL} {TO...	Conditional_Negative	Bury_Har_096
61	^sé: ^dé: ^wul ^tu:::	{PR-DCT} {TOP} {PR-Cnd} ...	Conditional_Negative	Bury_Har_123
49	'tô:' [gə̀t kən >+ yá: mə̂s //] ...	{PR-PHA} {NCL} {NCL}	Question	Bury_Sdy_20 (foc ds conditio...
102	'tô:' dzàŋ gí: ɣəŋ >+ tá fi ...	{PR-PHA} {NCL-CL1} {PST-...	Focal	Bury_Har_201
70	ŋă:n //	{PST-PHA}	Interjection	Bury_Har_151 (interro-neg)
133	kə ʧéɗəm tu [gì:yèn mán-vàr ...	{REP-INT} {REP-SPEECH}	Quote	Bury_Har_157

Figure 12. Links sorted by types showing Ex (13) in IlU classification

The corresponding annotation can be viewed in the tiers below when selected, and the corresponding sound segment can be played via the media player in Elan (see Figure 13).

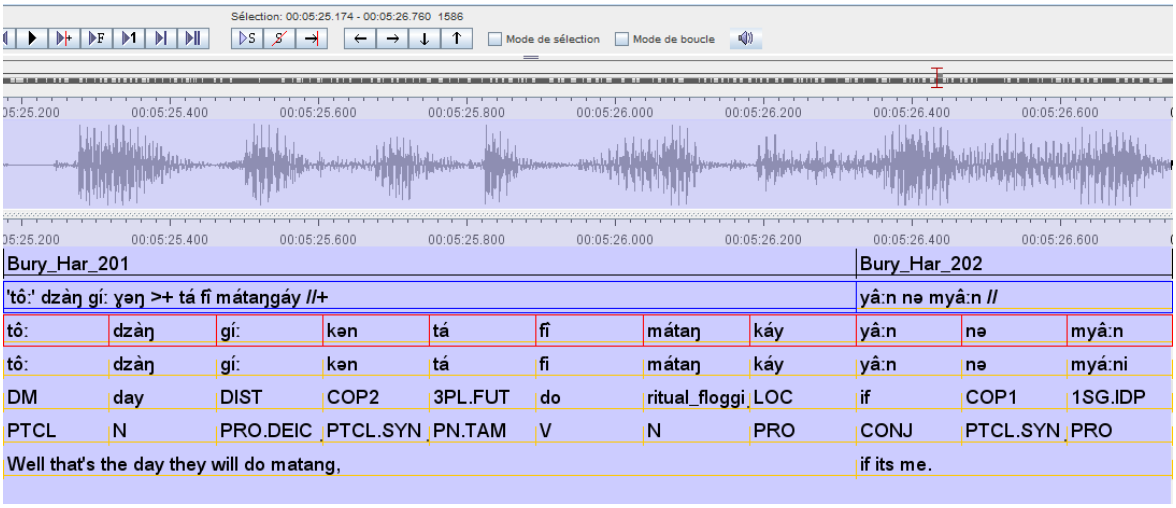


Figure 13. Annotation and sound file segment corresponding to Example 13

The labels tagging the macrosyntactic constituents are both structural and functional, as seen in Table 1.

	Aligned constituents	*Non-aligned constituents*
Pre-Nucleus	PR-ALL : Allocutive, Vocative PR-DCT: Discourse connector PR-EXP: Expressive 7 PR-PHA: Phatic PR-TOP: Left-edged Topic	PR-Adv: Left-dislocated adverbial adjunct PR –Cls: Left-dislocated clausal adjunct PR –Cnd: Left-dislocated conditional adjunct PR-IC1: Pre-nucleus section of IC1
Nucleus	NCL	NCL-IC2 (Nucleus of IC2) NCL-IC1 (Nucleus of IC1)
Post-Nucleus	PST-ALL: Allocutive, Vocative	APX: Nucleus Appendix (Afterthought)

	PST-DCT: Discourse connector PST-EXP: Expressive PST-TOP: Right-edged Topic	PST-IC2: Post-nucleus section of IC2
In-Nucleus	GFT: Graft ; PAR : Parenthesis	

Table 1: Group Types tagging macrosyntactic constituents

Using this tag set, I have been able to test a tentative typology of pre- and post-nucleus units on 11 annotated files (90 minutes; 15,000 words). I was able to extract the list of illocutionary constituents, and check the consistency of the annotation. The aim of this type of extraction is to look for regularities in the marking of the units, in syntax, morphology, intonation and reference tracking, i.e. do a basic bottom-up research. It is clear that the relevancy of the results is dependent on the tagging, which is based on my intuition and understanding of the language, checked and controlled by informants. Of course, this bottom-up stance is not devoid of any theoretical bias, but the exhaustivity of the annotation will (and already has) lead me to a revision of my analyses and some of the labels used for tagging. This labile method must strike a balance between rapidity of annotation (a process which can be very time consuming) and how fine-grained our analysis needs to be. To be fully labile, the tagging system must anticipate the need for regular revisions, e.g. automatic conversion and collapsing of categories.

The next section is devoted to a typology of peripheries, i.e. pre-and post-nucleus components of IlUs, as retrieved in the corpus using the "groups" table sorted by type (see Figure 11).

8 Typology of pre- and post- nuclei

When micro- and macro-syntactic dependencies are aligned, the boundaries of the nucleus correspond to the microsyntactic dependency unit of the verb/predicate carrying the illocutionary act, and include all the elements governed by this head. All the dialogical units

are aligned (*viz* outside the government of the nucleus head). The aligned textual units are: Discursive links (PR-DCT) and Topics (TOP and ANT). As for non-aligned units, the pre-nucleus governed constituents comprise left-dislocated adjuncts (PR-Adv, PR-Cls, PR-Cnd) and the pre-nucleus constituent of IC1 (PR-IC1). The post-nucleus governed constituents are the nucleus appendix (APX, e.g. afterthoughts) and the post-nucleus constituent of IC2s (PST-IC2).

8.1 Aligned peripheries

Aligned peripheries are divided into two classes which are respectively dedicated to different types of information functions: a) the textual construction of the utterance (textual peripheries, e.g. Topic, Appendix, Locutive Introducer); b) its communicative support (dialogical peripheries, e.g. Phatic, Allocutive, Expressive, etc.; see Cresti 1999:15). The only textual periphery that is not part of the dependency frame of the head of the nucleus is the Topic (TOP) and it appears massively in pre-nucleus position: only 2 examples of post-nucleus topics (also called right-edged topics, or antitopics: ANT) are found in the corpus, against 611 cases of TOP. Topics are illustrated in (33) for left-edged Topics (TOP) and in (34) for right-edged Topics (ANT). Right-edged topics are characterised by a low tone, flat contour, and follow a non-final prosodic break.

8.1.1 Topic

(33) ***tsátŋgân dàn má:*** *< myá: yel [nə lă: bàptàk ɓasmí //] //*

tsátn	*-kánì*	*dàn*	*má:*	*myá:*	*yel*
sit	-NMLZ	house	even	1SG.ICPL	see

nə	*la:*	*ká*	*bàptàk*	*ɓas*	*=mí*
COP1	work	POSL	useless	LOC	1PL.OBJ

'Sitting home < I see [it is useless for us. //] //' (Girls_B_035)

(34) *gòpm < kó:dzàŋgyò: <+ mìyá la: káwêy >* ***myà:ní::: gút za:r //***

gòpm	*kó:dzàngyó:*	*mìká*	*la:*	*káwêy*
1PL.POS	everyday	1PL.CONT	work	merely

myà:ní	*gudì*	*za:r*
1PL	woman.PL	human

'We < everyday <+ we do nothing but work > *we Zaar women.* //' (Wom_A_169)

8.1.2 Dialogical constituents

Dialogical constituents are used to establish, maintain or qualify the illocutionary act. They occur before, or after the nucleus. They are surrounded by inverted commas in the transcription. It is possible to distinguish the following types:

- *Phatic* (PR-PHA & PST-PHA), dedicated to control the communicative channel, ensuring its maintenance. They are either fillers (e.g. *er*…, *mm*…), discourse punctuators (*OK*, *well*, Zaar *tô:*), marks of agreement with the speaker (*uh*, Zaar *m:*, *è:*), etc.
- *Allocutive* (PR-ALL & PST-ALL), specifying to whom the message is directed, keeping their attention (Vocative, *you know*, *you see*) or introducing evidential modality (*I think*, etc.).
- *Expressive* (PR-EXP & PST-EXP), giving an emotional strength to the illocutionary act.
- *Connective* (PR-DCT & PST-DCT), linking different parts of the discourse (utterances within a turn, or across turns) maintaining some explicative, causal, temporal or concessive values. Most of them occur in pre-nucleus position.

8.1.2.1 Phatic

In (35), three cases of phatic units are exemplified: *tô:*, 'well' and *yâwwà:*, 'OK' as PR-PHA, and the TAG *ŋă:n*, 'no?' as PST-PHA.

(35) *"**tòː**" < yâːn < ʧăː ʧĭm tə ɣá vìː válti tu ʃĕro > **"ŋăːn"** // **"yâwwàː"** < **"tôː"** < átâ yi ʧĭk //*

tôː	*yâːn*	*ʧăː*	*ʧĭm*	*tə*	*ká*	*vìː*	*válti*	*tu*
DM	3SG	3PL.ICPL	call	3SG.OBJ	at	speech	muslim	COMP

ʃĕro	*ŋăːn*	*yâwwàː*	*tôː*	*átâ*	*yi*	*ʧĭk*
flogging	QUEST	ok	DM	3SG.REM	be	thus

' "*Well*" < this < they call it in Hausa 'shoro' > "*no*"? // "*OK*" < "*well*" < that's how it used to be. //' (Bury_Har_149)

8.1.2.2 Allocutive

Vocatives are examples of allocutives that can appear either before (PR-ALL, in [25]) or after the nucleus (PST-ALL, in (36) where Afo is a proper noun):

(36) *"ká" < "**Àfóː"** < ká: ye yáddiyóːɗam myáː súː sú kámʃâk > "kwăː" //*

ká	*Àfóː*	*ká:*	*yel*	*yáddiyóːɗan*
disapproval	Afo	2SG.CPL	see	how

myáː	*súː*	*sú*	*kámʃâk*	*kwăː*
1SG.ICPL	like	PL	Kamshak	DM

'"What" < "*Afo"* < you saw how I like Kamshak > "anyway". //' (Girls_B_073)

In (37), the speaker is protesting, using a yes/no rhetorical question, ending in a vocative, i.e. 'Afo', the name of the co-speaker.

(37) *mə káp Ngasaː > "**Àfóː"** ?!//*

mə	*kap*	*Ngas*	*-aː*	*Àfóː*
1SG.AOR	take	Angas	-VRT	Afo

'We should marry Angas people > "*Afo"* ?!//' (Girls_B_104)

In (38) *ká: yisə́ŋ*, 'you know', shows another way of maintaining the communicative channel, with an associated IlU:

*(38) ^dóŋ < "**ká: yisə́ŋ"** < farko máː<+ ɗaŋ kámʃâk tà wu tu ʧáː sûːm <+ mətá wultə tu baːbù //*

dón	*ká:*	*yisə́ŋ*	*farko*	*máː*	*ɗan*	*kámʃâk*	*tà*
because	2SG.CPL	know	beginning	even	as	Kamshak	REM

ʧáː	*súː*	*=mə*	*mətá*	*wul*	*=tə*	*tu*	*baːbù*

3SG.CPL love 1SG.OBJ 1SG.REM say 3SG.OBJ COMP no

‘“Because” < “*you know”* < in the beginning <+ as Kamchak said he loved me <+ I told him no way //’ (Girls_B_147)

8.1.2.3 Expressive

Exclamations in Pre-Nucleus position (e.g. PR-EXP; *kâ:y*, ‘hey’ in (39)).

(39) [Sp1] *sàké:dî ʧá: ndará //*

[Sp2] ***“kâ:y”*** **<** *ʧá: poləmgáy sò:séy //*

sàkê:t	*-i*	*ʧá:*	*ndará*
skirt	-INDF	3SG.CPL	be_proper

kâ:y	*ʧá:*	*pol*	*=mə*	*káy*	*sò:séy*
eh	3SG.CPL	please	1SG.OBJ	LOC	quite

[Sp1] ‘The skirt is nice. //’
[Sp2] ‘ “*Hey*” < I really like it. //’ (Girls_B_069)

8.1.2.4 Connective

In (38), *don*, ‘because’, is an initial discursive link (PR-CNT) working as a connective. It is annotated with a circumflex accent.

8.2 Non-aligned peripheries

Non-aligned peripheries are elements entering in a microsyntactic relation with the root of the IlU, and as such, pertain to the textual construction of the utterance.

8.2.1 Pre-Nucleus (<+)

The pre-nucleus governed constituents are left-dislocated adjuncts (PR-Adv, PR-Cls, PR-Cnd) and the pre-nucleus constituent of Unmarked Identifying Clauses (PR-IC1).

- PR-Adv, or left-dislocated adverbial adjunct

(40) *“tò:” <* ***dzàŋ là:dì*** **<+** *má łĭ: kində > “bá:” //*

tô:	*dzàŋ*	*lá:dì*	*má*	*łə*	*-í:*	*kində*	*bá:*
well	day	Tuesday	1PL.FUT	go	RES	Kində	NEG1

‘ “Well” < *on Tuesday* <+ we’ll go to Kində > “no”. //’ (Girls_A_001)

- PR-Cls, or left-dislocated clausal adjuncts

(41) *“tòː” < **kyàː gìː tíː** <+ “tôː” ká ɮəːríː ʧĭp //*

tôː	*kyăː*	*giː*	*tə*	*-íː*	*tôː*	*ká*	*ɮáːr*	*-íː*	*ʧĭp*
DM	2PL.ICPL	bury	3SG.OBJ	RES	DM	2PL.FUT	stay	RES	quietly

‘ “Well” < (after) you had buried him <+ “well” < you would sit still. //’ (Bury_Har_046)

Correlative conditionals (i.e. conditionals with a temporal meaning: ‘if (=when, =each time that) … then…’) are analysed just like ordinary adjuncts:

(42) ***yáː yelmə̆ŋ*** *<+ ʧáː fitə wusúŋə̆ŋ > “éy” //*

yáː	*yel*	*=mə*	*hḿ*	*ʧáː*	*fi*	*=tə*
3SG.COND	see	1SG.OBJ	NEG2	3SG.CPL	DO	3SG.OBJ

wusúŋ	*hḿ*	*éy*
be_nice	NEG2	indeed

‘If/when he does not see me <+ he is not happy > “hey”. //’ (Girls_B_077)

- PR-Cnd, or left-dislocated conditionals

(43) ***yâːn hali ɗa kàm*** *<+ má dîːbí //*

yâːn	*hali*	*ɗa*	*kàm*	*má*	*dîːp*	*-i*
if	chance	COP3	indeed	1PL.FUT	buy	-INDF

‘If there is a chance <+ we will buy it. //’ (Girls_B_056)

- PR-IC1, or pre-nucleus section of Unmarked Identifying Clauses (IC1)

(44) *^àmáː < **mə́n yóːɗan ʧăː fi*** <+ nə mə́n mársəŋ //

àmáː	*mə́n*	*yóːɗan*	*ʧăː*	*fi*	*nə*	*mə́n*	*mársəŋ*
but	people	which	3PL.ICPL	do	COP1	people	Marsang

‘^But < *the people who do it* <+ are the people of Marsang. //’ (Cal_Har_010)

8.2.2 Post-Nucleus (>+)

The post-nucleus governed constituents are the nucleus appendixes (APX, e.g. afterthoughts) and the post-nucleus constituent of Marked Identifying Clauses (PST-IC2).

- PST-IC2

IC2s constitute a single intonation unit. In these constructions, the illocutionary nucleus is not on the verb of the clause, which follows in the post-nucleus situation, but on the Identifier. In the following examples, the nucleus is bolded, and the ">+" sign that follows the nucleus indicates that there is a microsyntactic relation with what follows.

(45) *^dò:mín < sáŋwa:rí* < ***sé: dà:ʃì yá: môr lǔ:y*** >+ *əndá ʧà:yi vər tə sáŋwa:rês //*

dò:mín	*sáŋwa:rí*	*sé:*	*da:*	*-ês*	*yá:*	*mor*
because	chief_priest	only_if	person	DEF	3SG.COND	do_a_little

lu:	*-í:*	*kəndá*	*ʧà:yí*	*vər*	*tə*	*sáŋwa:rí*	*-ês*
get_old	RES	then	3PL.ICPL.ITER	give	3SG.OBJ	chief_priest	DEF

'^Because < a chief < *(it's) only when a man is a bit old* >+ ^then they make him a chief. //' (Rel_Har_008)

(46) *"tò:"* < ***gí:*** >+ *ká mân ?//*

tô:	*gí:*	*ká*	*mán*
well	DIST	2PL.AOR	come

'"Well" < *that* >+ you have come (for)? //' (= "well", is it that what you have come for?) (Girls_A_090)

Afterthoughts, which are elaborations or correction of the illocutionary act of the nucleus, are expressed in a different IlU. They are preceded by a final intonation break and a pitch reset, and they receive a falling contour. In (47), the afterthought is bolded:

(47) *móɣʃì makaranta < ma ʤyă:ŋo: //+* ***^sé: tà ŋál kálá:sò:*** *//*

mókʃì	*makaranta*	*ma*	*ʤyá:*	*hń*	*-o:*
courting	school	1SG.FUT	be_able	NEG2	ASS

sé:	*tà*	*ŋal*	*kálâ:s*	*-o:*
unless	3SG.AOR	look_for	class	ASS

'Dating in school < I couldn't do it //+ ^*unless he changed class. //*' (Girls_A_076)

Likewise, *kápwâ:sàŋ*, 'all of them', a part of the long example (4d) repeated in (17), is an appendix added to the nucleus as an afterthought after a final break.

9 Conclusion

In this paper analysing peripheries in relation with syntax and information structure in Zaar, a Chadic language spoken in Nigeria, I have argued that a minimal annotation representing in a simple and concise way the interface between information structure and syntax is essential to retrieve meaningful data. The article uses the concept of macrosyntax, based on illocutionary units, for this new level of annotation using existing morphosyntactic tiers in Elan. With the corresponding annotation script, a pilot 90 min (15,000 words) corpus has been annotated and a preliminary study of peripheries in this language has been done on this annotated corpus. I have argued that, although topics and frame-setters share the same intonation pattern, their syntactic properties call for a specific syntactic representation for which I have used a system adapted from Universal Dependencies Grammar. Some concluding comments can be done concerning the system introduced in this paper to annotate the information structure of Zaar, and how this structure is patterned in the language. I have chosen this punctuation, and developed a corresponding set of tags bearing in mind that it should be as theory-neutral as possible in order to implement a genuine bottom-up methodology, with a heuristic aim in mind, and hoping that the results can be used for typological comparisons. Another quality of this system of annotation is related to the fact that the notion of piling, accounting for coordination, can easily and intuitively be extended to dysfluencies, discontinuities and ellipses, and is perfectly adapted to the restitution of the oral flow. Despite the apparent accidents, interruptions and ellipses, the restitution of the piles proves that meaning, syntax and information progress and develop like the fugues and counterpoints of a musical score, which a description limited to the boundaries of a canonical grammatical sentence has been unable to account for. Finally, in the way Zaar shapes sound into meaning with the help of intonation, syntax and semantics, it appears that the left periphery is dominant and ICs are a device that is all the more meaningful as it is sparsely used. The three components of Zaar

Illocutionary Units come forth with a clear specialisation: the pre-nucleus establishes the frame/ground/site around the speaker's point of view; the nucleus carries the action/opinion, etc. in relation to the site; the post-nucleus seeks the hearer's approval, reactions or comments.

REFERENCES

Blanche-Benveniste, Claire, Mireille Bilger, Christine Rouget, Karel van den Eynde & Piet Mertens. 1990. *Le français parlé: études grammaticales*. Paris: CNRS.

Caron, Bernard. 2005. *Za:r (Dictionary, grammar, texts)*. Ibadan (Nigeria): IFRA.

Caron, Bernard. 2015a. Zaar Grammatical Sketch. In Amina Mettouchi, Martine Vanhove & Dominique Caubet (eds.), *Corpus-based Studies of Lesser-described Languages. The CorpAfroAs corpus of spoken AfroAsiatic languages*. Amsterdam/Philadelphia: John Benjamins. <https://halshs.archives-ouvertes.fr/halshs-00647526v3>

Caron, Bernard. 2015b. Tone and Intonation. In Amina Mettouchi, Martine Vanhove & Dominique Caubet (eds.), *Corpus-based Studies of Lesser-described Languages. The CorpAfroAs corpus of spoken AfroAsiatic languages*, 43–60. Amsterdam/Philadelphia: John Benjamins.

Caron, Bernard, Cécile Lux, Stefano Manfredi & Christophe Pereira. 2015. The intonation of topic and focus: Zaar (Nigeria), Tamasheq (Niger), Juba Arabic (South Sudan) and Tripoli (Libya). In Amina Mettouchi, Martine Vanhove & Dominique Caubet (eds.), *Corpus-based Studies of Lesser-described Languages. The CorpAfroAs corpus of spoken AfroAsiatic languages*, 63–115. Amsterdam/Philadelphia: Benjamins.

Chafe, Wallace L. 1976. Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View. In Charles N. Li & Sandra A. Thompson (eds), *Subject and Topic*, 25–56. New York/San Francisco/London: Academic Press.

Chanard, Christian. 2014. *ELAN-CorpA-V4.7.3*. <http://llacan.vjf.cnrs.fr/res_ELAN-CorpA.php>

Cresti, Emanuela. 2012. The definition of Focus in Language into Act Theory (LACT). In Heliana Mello, Alessandro Panunzi & Tommaso Raso (eds), *Pragmatics and Prosody: Illocution, Modality, Attitude, Information Patterning and Speech Annotation*, 39–82. Florence: Firenze University Press.

Cresti, Emanuela & Massimo Moneglia (eds.). 2005. *C-ORAL-ROM: integrated reference corpora for spoken Romance languages*. [Studies in Corpus Linguistics 15]. Amsterdam/Philadelphia: Benjamins.

Gerdes, Kim. 2013. Collaborative Dependency Annotation. *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, 88–97. Prague: Matfyzpress.

Halliday, M. A. K. 1967. Notes on Transitivity and Theme in English: Part 2. *Journal of Linguistics* 3(2). 199–244.

Higgins, Francis Roger. 1973. The pseudo-cleft construction in English. Massachusetts Institute of Technology Thesis. <http://dspace.mit.edu/handle/1721.1/12988> (14 April, 2016).

Huddleston, Rodney & Geoffrey K. Pullum. 2008. *The Cambridge Grammar of the English Language*. 2nd ed. Cambridge: Cambridge University Press.

Kahane, Sylvain & Paola Pietrandrea. 2012. La typologie des entassements en français. vol. 1. SHS Web of Conferences. <http://www.shs-conferences.org/>

Krifka, Manfred & Renate Musan. 2012. Information structure: Overview and linguistic issues. In Manfred Krifka & Renate Musan (eds.), *The Expression of Information Structure*, 1–43. Berlin, Boston: De Gruyter Mouton. <http://www.degruyter.com/view/product/177467> (2 May, 2016).

Lacheret, Anne, Paola Pietrandrea & Atanas Tchobanov. 2014. Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French. <http://hal.upmc.fr/hal-00968959/document> (23 March, 2016).

Marneffe, Marie-Catherine de, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joachim Nivre & Christopher D. Manning. 2014. Universal Stanford Dependencies: a Cross-Linguistic Typology. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

Marneffe, Marie-Catherine de, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Joakim Nivre, et al. 2014. Universal Dependencies. Online documentation (Version 1). <http://universaldependencies.org/> (16 June, 2016).

Mettouchi, Amina, Martine Vanhove & Dominique Caubet (eds.). 2012. *The CorpAfroAs Corpus. ANR CorpAfroAs: a Corpus for Afro-Asiatic languages*. <http://corpafroas.huma-num.fr/>

Newman, Paul. 1990. *Nominal and Verbal Plurality in Chadic*. Berlin: Walter de Gruyter.

Newman, Paul. 2006. Comparative Chadic revisited. In Paul Newman & Larry M. Hyman (eds.), *West African linguistics: papers in honor of Russell G. Schuh*, 188–202. [Studies in African Linguistics: Supplements 11]. Columbus, Ohio: Published by the Department of Linguistics and the Center for African Studies, Ohio State University.

Newman, Paul. 2013. *The Chadic Family: Classification and Name Index*. (Electronic Publication). Mega-Chad Research Network / Réseau Méga-Tchad. <http://lah.soas.ac.uk/projects/megachad/misc.html> (23 February, 2014).

Schultze-Berndt, Eva. 2013. About the shifty notion of contrast. Identifying subtypes of topics in corpus data of two Australian languages. Oral presentation at *Labex TCA-ISGR, Llacan* (12/11/2013), Villejuif.

Shimizu, Kiyoshi. 1978. *The Southern Bauchi Group of Chadic Languages: A Survey Report*. (Africana Marburgensia Special Issue 2). Marburg.

Simard, Candide. 2014. Another look at right-detached NPs. In Aicha Belkadi, Kakia Chatsiou & Kirsty Rowan (eds.), *Proceedings of Conference on Language Documentation and Linguistic Theory 4*. London: SOAS. <www.hrelp.org/eprints/ldlt4_17.pdf>

Sperber, Dan & Deirdre Wilson. 1986. *Relevance: communication and cognition*. Oxford: Blackwell.