A case for " slow linguistics " *

Bernard Caron

A case for "slow linguistics"*

Bernard Caron

CNRS, IFRA-Nigeria, USR3336

Bernard.Caron@cnrs.fr

This paper pays tribute to the two people that were most influential in the author's professional activity in the past twenty years and who both passed away in November 2016: Russell Schuh, who advised him on the choice of the South-Bauchi group of Chadic languages as a topic after his thesis on Hausa, and Marvellous S. Davan, a Zaar speaker of Bauchi State, Nigeria, who had been his main informant and language assistant. The author takes this opportunity to defend his vision of corpus studies and "slow linguistics", inspired by "slow food" which emerged as an alternative to "fast food". It is illustrated by a passage from Schuh's Grammar of Miya *and by a discovery concerning tones in Zaar that Davan contributed to make. The author then introduces the concept of macrosyntax and its corresponding annotation system as a contribution to the development of corpus studies in African linguistics.*

1. Introduction

As a beginner Hausaist and Chadicist, I was introduced to Russell Schuh's work in the late 70's, when attending Claude Gouffé's lectures at the Ecole Pratique des Hautes Etudes. I had never met him, and I was surprised when I received a letter where he reacted to one of the very first articles I published on aspect in Hausa. My article was clumsy and fumbling, but Russell took the time to criticise it and offer

* In memoriam Marvellous S. Davan, R.I.P.

suggestions to improve my approach to the topic. I was touched by his generosity and honoured that he had taken the time to share his reflections with the beginner that I was.

After defending my "Thèse d'Etat" on a Hausa dialect of the Niger Republic, as I had the opportunity to work in Nigeria for a few years, I decided to undertake the description of a lesser known Chadic language. It was only natural that in 1990, before taking up my appointment with a French research institute in Ibadan, Nigeria, I should write to him and several other renowned specialists of Chadic linguistics to ask for their advice as to which Chadic language was higher on the list of languages in need of research. I received a long, detailed letter from Russell Schuh where he made a broad description of the state of the art in the description of the whole Chadic family, branch by branch, with the location of languages and the identification of those where the information we possessed needed to be complemented, and the classification problems that were involved. As far as Nigerian Chadic languages were concerned, one language group emerged as the most worthy of interest. This was the South Bauchi language dialect cluster, with Saya as its main representative, or rather Zaar, after the autonym that I decided to use. I followed Russell Schuh's advice, and this set me on the scientific path I have been following for the past 25 years. This has literally changed my life, and I never regretted it.

Linguistics is done by people (linguists), with people (speakers, called 'informants' in our jargon), for people (other linguists, and hopefully reaching beyond the circle of linguists). It is this human aspect I would like to emphasize in my appreciation of Russell Schuh's work, and in a few examples of my own work, which I hope does not compare too badly with his.

I shared with Russell Schuh an interest in language documentation, its stress on data and the need to interpret it beyond face value. The best summary of this community of approach is found in the preface to his *Grammar of Miya* (1998), where he explains what a grammar should be:

> "What is an optimal descriptive grammar? Ideally, it should state and illustrate every generalization and idiosyncrasy of every structure that exists in the language being described. […] The grammar should thus be organized in such a fashion and the description couched in such terminology that anyone with basic training in any tradition of linguistic theory and

> description could find the structures of interest and learn how they work without going beyond the description provided in the grammar itself. The grammar should be equally useful to the European structuralist working in the tradition of Andre Martinet or the American formal theorist working in the tradition of Noam Chomsky, to the semiotician concerned with the interrelations of signs or the typologist interested in cross-linguistic patterns, to the language area specialist or the general linguist. Finally, the grammar should be as comprehensible and valid in 100 years as it is today (which is not to say that the grammar might not be amended and expanded to incorporate facts unavailable to the compiler)." (Schuh 1998:xvii)

And he has succeeded beyond expectation! His *Grammar of Miya* (*op.cit.*) is my favourite grammar alongside (Huddleston & Pullum 2008). This has been an inspiration and one of the reason for introducing the concept of "slow linguistics".

2. The case for "slow linguistics"

"Slow linguistics" needs to be defended in the same way as people are promoting "slow food" to fight against the infamous fast food industry. Slow linguistics takes its time to identify and prepare good quality data, carefully analysed with methods that respect its nature and structure, producing results that will endure and be appreciated for a long time.

In that respect, Russell Schuh's grammar was not a speedy exercise in adding a language name to his list of trophies. He started his fieldwork in 1982 and the book finally came out in 1996. Of course, it was slowed down by numerous teaching and administrative duties. But the grammar benefited from this slow maturation. One of the two "paramount models" to his work which he mentions in his preface, R.C. Abraham's *The Language of the Hausa People*, was published in 1959, and he mentions it as unsurpassed despite its formal quirkiness. Likewise, Russell Schuh's work stands independent from those theoretical experiments that have a life expectancy of about five years. This is fortunate, when one thinks of the African students who have to bow to the theoretical whims of their supervisor when they take their PhD in Western universities: they stay trapped in that theory imposed on them, and inflict the framework on their own students once they go back to lecture in their countries. Although Russell Schuh's wish for a grammar that "should be as

comprehensible and valid in 100 years as it is today" (*op.cit*.) may be overoptimistic, his is a much wiser option. When you have such an aim in view, you are allowed to take your time!

What makes Russell Schuh's work outstanding is that it is not merely descriptive. It is not just a list of facts randomly collected. The data is carefully analysed, from a systematic, diachronic and typological perspective, with a great amount of theoretical reflection. Let me take as an example his analysis of word order in his Miya grammar.

When studying the order of syntactic constituents in Miya, Russell Schuh does not take for granted the results of sentence elicitation. He takes into consideration not only syntactic (main vs. dependent) and morphological (TAM markers) elements but also information structure and genre (narrative vs. dialogue, reported speech, side comments, etc.). All this considered, Russell Schuh (1998:281-300) argues that although Miya independent clauses typically have SXV order, VXS order is possible and is probably the most "neutral" order. In his work, he meticulously studies the relationship between Information Structure and Grammatical Relations before doing a statistical study of word order in Miya and making a generalization concerning the evolution of word order in West-Chadic languages and stressing its importance for typological studies.

In Miya, constituents of X are focused *in situ*, and the subject appears then in initial position. Subjects are focused in initial position, with restrictions on TAM markers. Outside focused sentences, nominal subjects appear sentence-finally. Topicalized subjects are of course left-dislocated, with no resumptive pronoun, and no restriction on TAM verbal markers. He suggests that in Miya, "all independent main clauses with preverbal nominal subjects may, grammatically, actually have the form TOPIC-COMMENT rather than SUBJECT-PREDICATE" (1998:281). This would explain why statistically pre-verbal nominal subjects are dominant in narrations, and lead the linguist who often relies on narrations as a source of textual data to the wrong conclusions. This reconstruction of a basic constituent order against the face value of statistical evidence stresses the importance of genre and even style in the sampling of texts for linguistic studies. Of course, this takes time: the time for field work, sampling, recording, transcription and then annotation. Elicitation from a questionnaire is faster but may yield dubious results.

This type of in-depth, fine-grained analysis and annotation relies on a good understanding of the grammatical and lexical structure of the language that is documented. But as far as oral corpora are concerned, and more specifically dialogues, understanding the context, the common knowledge shared by the speakers, and even their personal history is often necessary for a proper analysis. As a consequence, much of the quality of the work we do depends on our ability to communicate with our informants.

First of all, the linguist must identify the person who possesses, beside immersion in the culture and a good competence in the language, some qualities that are not necessarily shared by everyone: communicative skills, mental agility, and an interest in abstract thinking. These qualities should hopefully match the linguist's own. When, added to this, the informant can write his own language, the linguist is blessed.

Russell Schuh, in the first chapter of his Miya grammar, gives credit to his own main informant, Vàziya Círòoma Tilde Miya, and stresses the importance of his help as regards the collection of texts.

> "[…] Vaziya's ability to write Miya was invaluable. Anyone who has ever tried to transcribe recorded texts knows the tedium and frustration involved, even from languages s/he knows fairly well much less a language which s/he does not speak. But with Vaziya's written version available, it simplified the task immensely because the basic flow of the text was already there, and only specific words and constructions needed clarification." (Schuh 1996:9)

In the same vein, I personally would like to pay homage to the second person who has been most influential in my work on Zaar, i.e. M.S. Davan, who sadly and unexpectedly passed away at the age of 40, not long after Russell Schuh. I had been working with him for almost twenty years, and he truly deserved the first name he had chosen for himself: "Marvellous". He was a natural linguist who was not given the time to use his gifts for the development of his culture and the defence of his language.

3. The discovery of a “marvellous” linguist and community

Around 1995, I had been working for 5 years on Zaar with Sunday M. Dariya as main assistant, when I felt I needed to find somebody with whom I could communicate better on the work, and who would be more interested in the language. Sunday was a competent speaker, but had no particular interest in understanding what I was doing. And his family obligations prevented him from giving me more of his time for the large scale corpus transcription and annotation that I wanted to start at the time.

In Sunday’s village where I had settled for my work, I advertised a position of assistant, offered to the many young adults who had completed their secondary school and who remained idle in the village. Very soon, five candidates volunteered. One morning, as I was teaching them the orthography I was using for transcription, I saw one young man in the audience reading over their shoulder, and pointing to them the mistakes they were making. I asked one of the applicants who was struggling with the task to leave him his seat, and it soon became obvious that that young man was by far the fastest learner. Within a couple of hours, he had mastered my orthography, which I quickly modified during the exercise, in order to make it easier for him to use. That’s how I met Marvellous S. Davan, or rather Gaba as he was then called, and started working with him.

That very first day, as he found it so easy to learn, I tried my luck at teaching him how to mark the tones of the language. I gave him a list of minimal pairs in monosyllabic and disyllabic words as reference patterns, and started dictating a new word list to him. It took him 30 minutes to master the system, and return a faultless transcription. I ended the session by handing over to him a 30 minute long recording of an interview of an old man who was narrating his biography and his experience as a worker in the tin mines of the Plateau State under the British colonial occupation. Within a week, he returned a neat transcription of the cassette, marked with tones, with an interlinear translation into Hausa. A few years later, I realised that he had actually corrected what the speaker said, removing the hesitations, and changing a word here and there when he thought the man had made a mistake. I asked him to completely redo the transcription which I needed for a work on intonation, without changing anything to what the speaker had said. The passages that he had not modified ten years before came back with the same

transcription, down to each single tone. He was a fast learner, a fast worker, and very precise and exact.

After working for years with Marvellous, and living within the Zaar community, a close relationship developed. I was asked to organise a cultural festival where 5 musical groups performed with their dancers. Sunday Dariya named one of his sons Bernard after me. I was turbaned "Sarkin Pada Tudun Wada Davan". I wanted to do something within the scope of my activity as a linguist that could be relevant to the community. I share with Russell Schuh the feeling that, if we need to gather data from the languages as part of our work as linguists, we need at the same time to produce "*output of interest and value to the speakers of those languages."*(Schuh n.d.). In a modest way that cannot compare with Russell Schuh's *Yobe Languages Research Project* on Bade, Bole, Duwai, Karekare, Maka, Ngamo and Ngizim, I have locally published a book on Zaar grammar, with a dictionary and collection of texts in Zaar, Hausa and English (Caron 2005). I conceived the book as a linguistic help for Zaar children to have access to English starting with texts whose cultural context was familiar to them. The book was launched in Bogoro, Bauchi State, with a small price tag thanks to the financial help of the French Embassy. The Zaar community had refused to have the book distributed freely as I had planned initially. "What is free has no value" they said.

What the Zaar community was keen on getting from my work was a Zaar translation of the Bible, which I politely declined. I directed them to the SIL people in Jos, but they did not agree on the terms of their collaboration.

After a good many years of work with me, Marvellous caught the virus, and decided to start his own work. I helped him to scout the Zaar area looking for old speakers to interview about their oral traditions. He wrote and typed in Zaar a summary of his findings, which he complemented with some proverbs, word lists, and a few passages from the Bible. He published the book in 2010 with the title *Ɓup Dzanyi Gwaay*, 'Improve yourself' (Davan 2010).

After that, Marvellous decided to teach the young children of the community to write and read in Zaar. For that, he set out to write a method, with a primer which he tested for some time in a local primary school. He registered for a BA in the Distance Learning Center that had just opened in Ɓogoro, Bauchi State. He also went to Jos to get some training in orthography development in a workshop

organised by the SIL. But all this beautiful project was nipped in the bud by his untimely death in November 2016 in Bauchi. Let him rest in peace.

As a conclusion to this homage to Marvellous, I would like to illustrate the pivotal role he has had, as an informant, in the development of my understanding and analysis of Zaar. I will take two examples, one from phonology, and one from intonation structure.

3.1. The mystery of the falling tone

A phonological problem had troubled me for more than two years before I met Marvellous. It concerns tonal verb classes and more precisely, monosyllabic CV verbs with a short vowel (*ɗu* 'to beat'; *nda* 'to enter'; *fu* 'to say'; *ɬə* 'to leave'; *ʃa* 'to drink'; *su* 'to return'; *ta* 'to climb'; *tu* 'to meet'). In the Perfective, I heard a Mid tone for all the verbs. Therefore, those verbs did not belong to the class of verbs with a lexical High tone, but rather to the lexical and morpho-phonological class of non-High verbs beginning with a voiceless obstruent.

Now, since this class of non-High verbs had a falling tone in the third person singular and plural of the Narrative, I was expecting the short list of CV verbs to behave in the same way, but it was difficult for me to check whether this was correct. The shortness of the vowel made the perception of the tone dicey, even after making my informants whistle as slowly as possible. Most of the time, however, I heard not a Mid tone but a High, sometimes with a rippling that I was tempted to interpret as a Falling tone. So, High or Falling tone? This was a pertinent question because it was possible that the falling modulation on a light syllable is systematically realized as high, in which case certain high tones could, in certain contexts, be hidden falling tones.

Could I find assistance from parallels in other parts of the language? No noun in Zaar has this syllabic structure. The only grammatical morpheme bearing the same uncertainty as to whether it is a high tone or a falling tone was the Remote Past morpheme (*ta*). Its tonal behaviour is identical with that of the Recent Past (*naː*), which in certain contexts takes the form *nâː* with a clearly perceptible falling tone because it is carried by a long vowel. The parallel between the two past tense morphemes strongly suggests the existence of such a Falling tone over the Remote Past of *ta*, that is, over a monosyllable ending in a short vowel. Is this tone "simplified" in the form of a High

tone or did it keep its form, which meant that I could not hear it clearly? I could not resolve this problem just with my own ears.

A couple of years later, I devoted several days of work to the problem using CECIL, a now obsolete acoustic lab developed by SIL which could be carried to the field. But to no avail: the acoustic data was insufficient to get a pitch track that I could interpret.

And that's where Marvellous solved the problem for me in his first 30 minute long transcription. In the texts, all those monosyllabic CV verbs with a short vowel, as well as the *ta* TAM marker, were transcribed by Marvellous with a falling tone in the relevant contexts. I soon realised that Marvellous systematically transcribed the phonological value of tone, as realised after phonotactic rules had applied. I only needed to reconstruct the rules and the lexical values from his transcriptions.

I gradually came to rely totally on his intuitions. At first I doubted some transcriptions that were not consistent, thinking he had made mistakes, but soon realised that they were phenomena that had escaped my attention. Following this, my analyses became more precise, more detailed and covered more and more complex data. Marvellous's explanations of context, situations, and background knowledge of Zaar culture and village history became essential when I started doing some fine-grained corpus annotation, especially on Information Structure.

3.2. Clefts sans 'it' sans 'be', sans everything.

In Zaar, the possibility of dropping the copula in specifying copular clauses (Huddleston & Pullum 2008:1416 ff.), sometimes has the result of producing transcriptions with clefted sentences that look like left-dislocated topics. The only thing that differentiates them is intonation. Such a case is found in (1) where the speaker (Marvellous himself) talked about a game of football where he scored the sixth goal:

(1)

lim-ês má: MYAˆN mətá łya.[1]
six-DEF even 1SG 1SG.REM drink

'The sixth even, it is me who scored.' (Boys-A_407) 2

This game of football had become a bone of contention: Marvellous had been accused of not playing well. In (1) he defends himself by saying that he was the one who scored the sixth goal. *limês*, 'the sixth (goal)', is topicalised, and the subject *myâ:n* is focused in the form of an independent pronoun. The corresponding "neutral" sentence (without cleft or left-dislocation), which could appear in a narration of sequential events, would read like (2):

(2)

mətá łya lim-ês má:.
1SG.REM drink six-DEF even

'I even scored the sixth.'

A copula (either of the invariable particles *nə* 'COP1'; or *kən* 'COP2') is usually present to specify the clefted element. An equivalent of (1) would' then be:

---

1 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 (:). Phonemic tone is marked with diacritics: á, à, â and ă for High, Low, Falling and Rising respectively. Mid tone is left unmarked. The following abbreviations are used in the text and in morphosyntactic transcriptions: AOR, Aorist; COND, Conditional; COP, Copula; DEF, Definite; DM, Discourse Marker; FILL, Pause Filler; FUT, Future; INCH, Inchoative; NEG, Negative; OBJ, Object; PL, Plural; POS, Possessive; PROX, Proximal; QUEST, Question; REM, Remote Past; RES, Resultative; SG, Singular; TAM: Tense, Aspect and Mood; VRT, Virtual. By convention, in Universal Dependencies syntax, the dependency links are tagged in lowercase, e.g. advmod, adverbial modification; conj:dicto, dysfluency; dobj, direct object; nsubj, nominal subject; obl:comp, oblique complement; punct, punctuation; svc, serial verb construction.

2 When a reference is given for an example, it corresponds to my unpublished annotated Zaar corpus. Unreferenced examples are reconstructed for the purpose of the paper.

(3)

lim-ês máː nə MYAˆːn *mətá łya.*
six-DEF even COP1 1SG 1SG.REM drink

'The sixth itself, it is ME who scored (it).'

In a different context, if the speaker was listing the names of the players who scored the different goals, e.g. 'Justin scored the first one, Gaba the second...', the function of the 1SG independent pronoun would be changed to that of the topic: 'As for me, I scored the sixth one,' as in (4a):

(4a)

myâːn, mətá łya lim-ês máː.
1SG 1SG.REM drink six-DEF even

'As for me, I scored the sixth itself.'

In this example, the topic is pronounced with a suspensive intonation, and followed by a pause with a pitch reset (downstep) while the main prosodic prominence falls on the predicate *łya limês,* 'score the sixth'. With no pause or stepdown, and the main prosodic prominence falling on *myâːn 'I/me',* (4a) would now become (4b) where the first element of the sentence is focused. (4b) is a prosodically marked cleft, without identifying copula, without "it" pronoun, and without relativization:

(4a) *myâːn* = Topic
myâːn, mətá łya limês máː.
1SG 1SG.REM drink six-DEF even

'(as for) me, I scored the sixth itself.'

(4b) MYAˆːN = Focus
MYAˆːN *mətá łya limês máː.*
1SG 1SG.REM drink six-DEF even

'(it is) me > who (lit. I) scored the sixth itself.'

The difference between these two constructions lies in the change of place of the main prosodic prominence. It falls on *łya limês*, 'score the sixth' in (4a) and *myâːn* in (4b). The main prosodic prominence indicates what functions as the illocutionary nucleus and

syntactic root in the utterance[3]. In a focused construction, the clefted element marked by the main prosodic prominence and/or a copula supersedes the verbal predicate as the syntactic root of the utterance. Consequently, the target of the illocutionary act moves away from the verbal predicate to specify the element of the predicative relation that is clefted. As a result, the verbal predication itself is backgrounded.

I personally relied on Marvellous's context-based explanations and paraphrases for my analyses, and looked for intonation cues to substantiate his interpretations. My own approach to the relation between the linguist and the informant may be extreme, but it is essential in the work of linguists who are not native speakers of the language they study. However, I still find it difficult to convince my colleagues of the existence of ambiguities such as the one between a topic and focus interpretation of (4a) and (4b), where the only difference is expressed by intonation.

As I became more and more involved in the study of information structure, in a bottom-up methodology based on corpus analysis, the need for an annotation system that would enable me to retrieve extensive and relevant data from large corpora became more and more urgent. That's where the concept of macrosyntax and its annotation scheme provides a powerful tool to study the interface between information structure, prosody and syntax. Once the corpus is annotated for macrosyntax, tagged and parsed, it can be queried for a study of the relative role of morphology, syntax, and prosody in establishing the relation between sound and meaning. The annotation process and its interpretation of the meaning of the data is at the foundation of corpus linguistics. This methodology has helped me to account for the specificity of oral corpora, which is at the centre of my current work in linguistics.

4. Macrosyntax and the specificity of oral corpora

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

[3] In Dependency Grammar, the root of the utterance is the single non-governed lexical item that operates as the syntactic head of the Government Unit. (See section 4.1 for a short presentation of Dependency Relations).

languages such as English, French, etc.) have no written and grammatical tradition. They are oral languages. Oral corpora are greatly structured by the features associated with performance: dysfluencies on the one hand (hesitations, pause fillers, aborted utterances) but on the other hand, the stylistics of oral art performance, such as rhetorical repetitions, parallel constructions, etc.

Descriptive grammatical frameworks, which on the whole remain heavily indebted to prescriptive grammars of European languages, are not equipped to account for the specificities of oral data. Dysfluencies for example, which are often considered as 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.

This argues in favour of a new descriptive paradigm and 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. 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. This means taking the Illocutionary Unit as the basic unit of representation, and going beyond the limits of sentential syntax to found a new syntax called "macrosyntax".

4.1. Macrosyntax and Macrosyntactic annotation

In this new approach to corpus annotation, the Illocutionary Unit is taken as the basic unit of representation. The Illocutionary Unit can be compared to Cresti & Moneglia's utterance, which they define in reference to Austin's theory of speech acts (Austin 1962):

> "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)

The Illocutionary Unit is not necessarily congruent with intonation units, and is defined as comprising all the elements bearing a syntactic relationship with the syntactic root of the unit. This

includes peripheral elements that hold a discursive relationship with the root, such as left- and right- dislocated elements, parentheses, etc. By doing so, the model lays the ground for an all-inclusive model of syntax called 'macrosyntax'. Macrosyntax subsumes 'microsyntax' which describes the relation between a head and its complements, adjuncts, determiners or modifiers.

In other words, the macrosyntactic level describes the whole set of relations holding between all the segments that make up one and only one illocutionary act (Cresti & Moneglia 2005). A macrosyntactic punctuation marking Illocutionary Constituents and their relations has been developed in the Rhapsodie Project (RP) for French (Lacheret, Pietrandrea & Tchobanov 2014). It 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 (disfluencies, reformulation, coordination[4]).

Illocutionary Constituents 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 Illocutionary Unit.

In (5) the pre-nucleus *Ndà:də̃m má:* is a left-dislocated topic, separated from the nucleus by "<". In (6) the post-nucleus *sarkinpá:da* is a vocative separated from the nucleus by ">".

(5)

Ndà:də̃m má: < má łə yel=tə áy //
Ndadəm even < 1PL.FUT go see=3SG.OBJ eh //

'Ndadəm < I will go and see him //' (SI_06_Girls_A_005)

(6)

á ɓân-í: ŋă:n > sarkinpá:da //
3SG.AOR finish-RES COP2.VRT.NEG2 > Sarkin_Pada //

'Is it finished now > Sarkin Pada ?//' (SI_07_Women_A_114)'

[4] See section 4.2. for an illustration of the use of the concept of "pile" to account for dysfluencies in oral corpora.

Macrosyntactic structures can be represented using the Universal Dependencies framework[5]. See (Figure 1) below which represents the dependency relationships in (5):

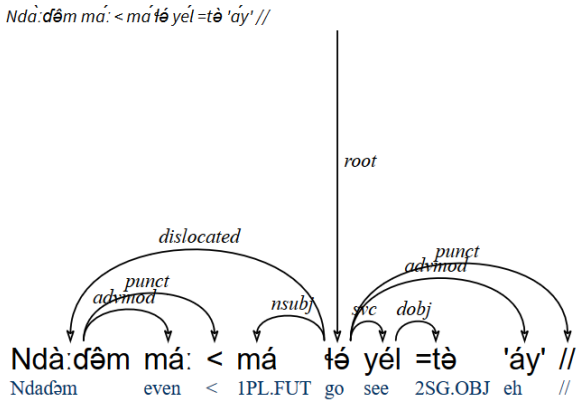

Figure 1: Syntactic representation of Example (5)

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 Fig.(1) an arrow tagged '*nsubj*' (nominal subject) points from *ɬə́*, 'go' to *má*, 1PL.FUT TAM and Person marker: *má* is the subject of *ɬə́* and is governed by it. A government unit (GU) is a maximal unit for government. A GU has a head, which is not governed, and all the elements of the GU are dominated by this head. We call "root" the head of the Illocutionary Unit. The root of (5) is ɬə, 'go' in Figure (1). In other words, a GU is the maximal projection of a non-governed lexeme. In our analysis, the GU is the Illocutionary Unit, i.e. the maximum macrosyntactic unit.

The following section examines 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. The corresponding annotation symbols will be presented and commented when they are introduced in the examples.

---

[5] See the Universal Dependencies website (http://universaldependencies.org/) for a detailed presentation of the theoretical framework and illustration by treebanks corpora in numerous languages.

4.2. Dysfluencies

A common configuration of oral corpora that needs to be accounted for concerns dysfluencies, as in (7), an example taken from Zaar. The hesitations of the speaker result in the repetition of *ká*, separated by a pause (*#*), and *teː*, separated by a pause filler (*ɣə́*):

(7a)

Tôː ká # ká dû teː ɣá teː gə̀ʃi tsə́n kən.
DM 2PL.AOR # 2PL.AOR beat around FILL around downhill like_this COP2

'So, you... you would beat it towards er... towards the East like this indeed.' (Bury_Har_052)

This type of dysfluency pervades 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 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) (Blanche-Benveniste et al. 1990). Their group was particularly innovative in their stress on documentation and oral corpora analysis. They developed a method to annotate dysfluencies of this type by turning them into a paradigm which has the same syntactic structure as coordination or apposition. They created the concept of "pile" (*empilement* in French) to describe the introduction of this paradigmatic dimension into syntax.

Though coordination and apposition, elements build a paradigm in which each of them fills the same syntactic function as the first element of the paradigm. A visual representation using the GARS annotation shows clearly the paradigmatic relationship between the coordinated elements of (8):

(8a)

ka ɓəl-ni gyáː łərtí giní, tə́ łərtí gín, tə́ łərtí gín.
2SG.FUT dig-INCH PL root PROX and root PROX and root PROX

You will dig these roots, and this root, and this root.
INT_05_Morals_SP1_117

(8b)

ka ɓəlni *gyá: łərtí giní*
tə́ łərtí gín
tə́ łərtí gín

‘you will dig these roots
and this root
and this root.’

In macrosyntactic annotation, piles are delimited by braces: { … }. The elements constituting the piles are separated by pipes: { ___ | ___ }. Various types of pipes annotate different types of piles, e.g. “|c” which annotates coordination: { ___ |c ___ }:

(8c)

ka ɓəl-ni { gyá: łərtí giní |c tə́ łərtí gín |c tə́ łərtí gín } //
2SG.FUT dig-INCH { PL root PROX |c and root PROX |c and root PROX} //

You will dig { these roots |c and this root |c and this root } //

The corresponding syntactic representation is shown in (Figure 2), where the dependency tag of coordination is ‘conj:coord’:

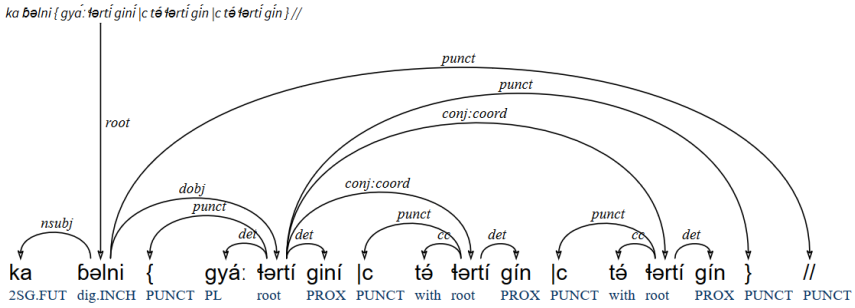

Figure 2: Syntactic Representation of Example (8)

The same paradigmatic relationship characterizing coordination is applied to dysfluencies in (7a), and are represented in the GARS annotation in (7b):

(7b)

tôː *ká #*
ká *dû* *teː ɣə́*
teː gàʃì tsán kən

‘well you would…
you would beat (it) toward er
toward the East like this’

In macrosyntactic annotation, the piles built by dysfluencies are marked by double pipes, as in (7c)

(7c)

Tôː { ká || # ká } dû { teː || ɣə́ teː } gàʃì tsán kən // [6]
DM { 2PL.AOR || # 2PL.AOR } beat { around || FILL around }
East like_this COP2 //

So, { you || you } would beat it { towards || er... towards} the East like this indeed.’ (Bury_Har_052)

The syntactic representation of dysfluencies is shown in Figure (3) where the dependency tag of dysfluencies is ‘conj:dicto’:

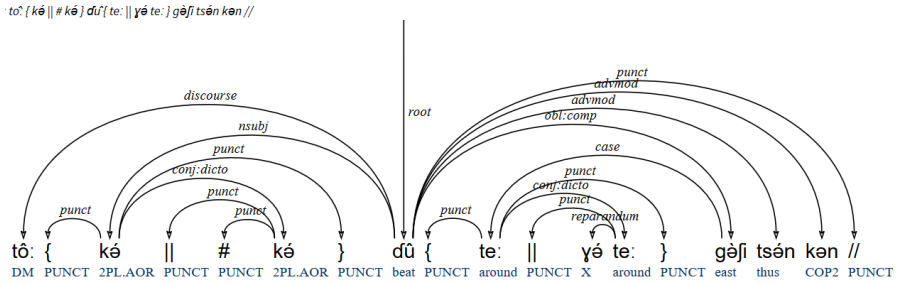

Figure 3: Syntactic representation of Example (7)

This concept of ‘piles’ which covers coordination, apposition and dysfluencies, presents the advantage of reintegrating dysfluencies into syntax, and account for their role in discourse (reformulation, specification, elaboration of thought).

6 In full macrosyntactic annotation, *tôː* and *ɣə* are called “Associated Illocutionary Units” and are surrounded by quotes.

4.3. Afterthoughts

Afterthoughts are another example of the specificities of oral language data, as exemplified in (9a):

(9a)

Tôː mə́ ŋgyə̌ːr gyaː gàːl ɓét ɗaŋ. Kóː gèri kóː maːt.
DM 1PL.AOR slaughter PL cow all too or chicken or goat

'Well we slaughtered many cows too. Or hens, or goats.'
(Cal_Har_ 032)

In this example, the first intonation unit finishes with the adverbial adjunct *ɓét ɗaŋ,* 'plenty too' and the end of the unit is marked with a terminal prosodic break. Then, as an afterthought, two nouns are added, forming a discontinuous chain of three coordinated direct objects (*gyáː gàːl*, 'cows'; *gèri*, 'hens' and *maːt*, 'goat') of the verb *ngyáːr* 'slaughter.' The afterthought forms a second intonation unit starting with a pitch reset and finishing with its own terminal prosodic break.

The coordination operating over the final break of the first intonation unit can be represented as in (9b), where the coordinated elements are in a paradigmatic relationship, and inherit their syntagmatic function from the first element of the pile:

(9b)

tôː mə́ ŋgyə̌ːr	*gyaː gàːl, ɓét ɗaŋ.*
	kóː gèri
	kóː maːt.
'well we slaughtered	cows, many too.
	or hens
	or goats.'

This poses a dilemma: if one follows the intonational clues, the constituents of second intonation unit are syntactic orphans without governor. If one follows the syntactic 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. Macrosyntax shows a way out of this dilemma by allowing syntactic relations (e.g. piling such as coordination, whether disjunctive or not) across prosodic boundaries. In (9), the coordination of the NPs in the first intonation unit (*gyà: gaːl* 'cows') and in the afterthought (*kóː gèri kóː maːt*, 'or chicken, or goat') takes place across the final prosodic boundary and over the utterance-final adverbial phrase *ɓét ɗaŋ*, 'too'.

In macro-syntactic annotation, the final prosodic boundary across which the coordination operates is marked with a plus: //+. The disjunctive coordination is annotated with the symbols:
{ ___ |c } ... {|c___ } as in (9c)[7]:

(9c)

tôː má ŋgyǎːr { *gyaː gàːl* |c} *ɓét ɗaŋ* //+ {|c *kóː gèri* |c *kóː maːt* } //
DM 1PL.AOR slaughter { PL cow |c} all too //+ {|c or chicken |c or goat } //

'well we slaughtered { many cows |c} too //+{|c or hens |c or goats} //' (Cal_Har_ 032)

4.4. Syntactic relations over turn-taking

Piling through coordination can also occur across turn-taking and result in elliptic structures. Instead of considering those as either incomplete structures or structures where most of the elements have been omitted, they can be considered as a special case of coordination across turn-taking.

This is illustrated in (10) below, which is part of a passage where the first speaker [S1] is interviewed by [S2] about funeral rites. In this example, the nouns *gàt* 'woman' in (10a) and (10e), and *ŋaː gàt* 'girl' in (10c) are part of the same pile that spreads over several turn-takings, and share the same syntactic properties as initially stated in (10a).

The utterance in (10a) is divided in two parts: the nucleus *tá gìː tà gòs dɔ̀ːʔ* 'where will they bury her?' and the pre-nucleus *tô GÀT kàn yáː mâs kúmá* 'well if it is A WOMAN that dies', a conditional dependent clause whose subject *gàt*, 'woman' is clefted. The clefted element is coordinated over several turns of conversation without repeating the rest of the (10a) initial sentence.

[7] See Caron (2017) for a more detailed presentation of macrosyntax and the annotation of Zaar.

(10a) [S1]
tô ***GƏT*** *kàn yá: mâs kúmá tá gì: tà gòs dɔ̀:?*
DM woman COP 3SG.COND die too 3PL.FUT bury 3SG.OBJ 3SG.POS where

'Well and if it is a woman that dies, they will bury her where?'

(10b) [S2]
gədɗ*-à: ?*
woman-QUEST

'A woman?'

(10c) [S1]
kó: ŋa: gàt.
or young woman

'Or a girl.'

(10d) [S2]
ŋa: gàt tá gì: ʃĩ ɓáɮəŋ > kápwâ:səŋ […]
young woman 3PL.FUT bury 3PL.OBJ outside all 3PL.POS

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

(10e) [S1]
tá gàt *ɓét kó: ?*
with woman all or

'And women too or what?

(10f) [S2]
m̀: tá gàt ɓét tá gì: ʃĩ dân.
er with woman all 3PL.FUT bury 3PL.OBJ there

'Er and women too, they would bury them there.'(Bury_Har_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: *'{ gàt* |c *kó: ŋa: gàt* |c *tá gàt ɓét } kən yá: mâs […],* ' if it is { women |c or girls |c and women in general } that die […]').

Likewise, the noun in S2's echo-question (*gədâ:*, women?') is part of this coordinated pile too, and inherits the same function as the coordinated elements in S1's turns. (10b) is equivalent to (10b'):

(10b') [S2]
gə̀t (kən yá: mâs) a: […],
woman COP2 3SG.COND die QUEST

'(if it is) a woman (that dies) eh?'

This analysis and its accompanying annotation system elegantly underline the coherence of this large passage without postulating the existence of elements deleted through ellipsis. Each element in (10b), (10c) and (10e) is linked to the previous utterance of the speaker, and inherits its referential coordinates from this unit.

5. Conclusion

As a conclusion to this account of my experience of field linguistics in all its aspects, which I started in Nigeria with Russell Schuh's initial impulse, and continued with his constant encouragement, I would like to stress that corpus linguistics has entered a new era. With the aid of computers, what I call "slow linguistics" can now be done on a large scale while gaining in quality and saving on resources, thus getting the best of both worlds. New computer programmes, using algorithms commonly referred to as "deep learning" programmes, are used to produce automatic taggers, parsers, phonetizers and sound-text aligners. Those are beginning to be developed for under-resourced languages. I am planning to experiment some of these on Zaar as a pilot language and then, why not expand these methods to the study of other languages in a future Centre for the Study of Nigerian Languages?

References

Abraham, Roy Clive. 1959. *Hausa literature and the Hausa sound system*. London: Univ. of London Pr.

Austin, John Langshaw. 1962. *How to do things with words: the William James lectures, delivered at Harvard Univ. in 1955.* Oxford: Clarendon Press.

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. 2017. Macrosyntactic corpus annotation. The case of Zaar. Ms. 45p.

gCresti, Emanuela & Massimo Moneglia (eds.). 2005. *C-ORAL-ROM: integrated reference corpora for spoken Romance languages*. (Studies in Corpus Linguistics v. 15). Amsterdam ; Philadelphia, PA: J. Benjamins.

Davan, M. S. 2010. *Ɓup Dzanyi Gwaa*. Ibadan (Nigeria): Ifra-Nigéria.

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

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).

Schuh, Russell G. n.d. Yobe languages Research Project. http://aflang.linguistics.ucla.edu/Yobe/yobe.html (19 August, 2017).

Schuh, Russell G. 1998. *A Grammar of Miya*. Berkeley & Los Angeles: University of California Press.