Prosodic segmentation and cross-linguistic comparison in CorpAfroAs and CorTypo: Corpus-driven and corpus-based approaches

Amina Mettouchi, Martine Vanhove

Prosodic segmentation and cross-linguistic comparison in CorpAfroAs and CorTypo: Corpus-driven and corpus-based approaches

Amina METTOUCHI (EPHE-PSL; LLACAN-CNRS)
Martine VANHOVE (LLACAN; CNRS – INALCO – EPHE)

Abstract
The paper addresses the issue of corpus-design in relation to research questions, for under-described languages. It shows how a corpus emerges from the methodology and *habitus* of its contributors, and how it is shaped by the technical tools used for data organization. It also underlines the ways in which a morphosyntactically-annotated corpus, segmented into intonation units, is amenable to a wide array of searches, both corpus-based and corpus-driven, and both formal and functional.
After a presentation of the annotation layout, and the segmentation choices that characterize the two projects, CorpAfroAs and CorTypo, scientific results are illustrated for two languages, Kabyle and Beja, and more marginally for Zaar, Juba Arabic and Modern Hebrew. They exemplify corpus-driven and corpus-based approaches of information structure and grammatical relations. Both types of approaches plead for an integrated view of prosody, closely interacting with syntax, semantics, phonology, information structure, and all levels of human communication and cognition. They also plead for a general endeavour to annotate as much as possible the large array of prosodic cues that are inseparable from speech processing and interaction dynamics.

Keywords: prosody; morphosyntax; information structure; reported speech; corpus; AfroAsiatic

1. Introduction

Since the 2000s, typology has taken on new directions: from generalizations based on descriptive grammars of diverse languages, it has opened up to text-based cross-linguistic comparisons. Parallel corpora of under-described languages, as advocated by Cysouw and Wälchli (2007), are not the only types of corpora liable to provide the kind of data that typology needs. But extremely diverse corpora are not easy to handle either. This paper expands on cross-linguistic corpora and databases that are non-parallel, but nevertheless sufficiently similar in terms of structure to allow such comparisons and generalizations: CorpAfroAs and CorTypo.[1]

Both projects, which will be described in Section 2, are based on field recordings, collected, transcribed and annotated by the language-specialists themselves. The spoken nature of the data was taken as central, and implying a segmentation based on prosodic rather than syntactic units. The fact that CorpAfroAs involved languages belonging to the same phylum, characterized by common morphosyntactic features, made it natural and desirable to integrate the articulation of prosody and morphosyntax in the layout and research questions underlying the CorpAfroAs project. While CorpAfroas is

[1] CorpAfroAs = *A Corpus of Spoken Afro-Asiatic languages*, ANR-06-CORP-0018 grant, 2007-2012. CorTypo = *Designing Spoken Corpora for Cross-Linguistic Research*, ANR-12-BSH2-0011 grant, 2013-2017. We thank our anonymous reviewers for their insightful comments, our colleagues in both projects (cf. footnote 4) for the great work we did together, the participants of the workshop where this paper was first presented for their rich discussions, the convenors and editors of this issue for their open and stimulating approach to the topic, and last but not least, all the speakers who contributed so gracefully and generously to our corpora, by giving us such beautiful recordings of their languages.

directly accessible as a searchable textual corpus, CorTypo, the second project referred to in this paper, is accessible as a searchable functional database, with the textual corpora in the background, from which all examples illustrating given functions are automatically retrieved and displayed for the end-user.

The paper presents studies and results, achieved through queries in the databases, language-internally for three of them (Sections 4.1.2; 4.1.3; 4.2.2) and cross-linguistically for the fourth one (Section 4.3). Despite the centrality of single-language studies on Kabyle and Beja (Sections 4.1 and 4.2), the developments in this paper are crucially relevant to the question of comparability, and contain reflections on the creation of cross-linguistic resources, with an emphasis on the possibility of combining language-internal annotations with comparability between languages (Section 3).

After the precise presentation of CorpAfroAs and CorTypo (Section 2) with their functionalities, which allow an end-user to conduct queries on the data, we explore the question of the interaction between prosody and morphosyntax, through four case-studies: one on grammatical relations, two on information structure, and one on reported speech (Section 4).

Throughout this paper, we address the issue of corpus-design in relation to research questions, for under-described languages. We show how a morphosyntactically-annotated corpus, segmented into intonation units, is amenable to a wide array of searches, both corpus-based and corpus-driven, and both formal and functional.

2. CorpAfroAs and CorTypo

The corpora that will be presented in this paper have been compiled in the framework of two research projects, CorpAfroAs[2] (2007-2012) and CorTypo[3] (2013-2017), funded by the French National Research Council ANR (Agence Nationale de la Recherche).[4]

CorpAfroAs, as its name indicates, contains only Afroasiatic languages, and is therefore a phylum-based cross-linguistic corpus. Each branch of the phylum is represented by one or several languages/corpora. Each language corpus was created within the scope of the CorpAfroAs project, from data collection to annotation and dissemination. The aim was to have one hour of morphosyntactically-annotated spontaneous spoken data, with one third of dialogue, and two thirds of monologue. Not all subcorpora have reached that threshold; information on each of them is accessible online on the CorpAfroAs website. The project was a pilot endeavour, aiming at making available the first online searchable prosodically segmented and morphosyntactically-annotated corpus of spoken AfroAsiatic languages, with extensive documentation of its structure and contents, thus paving the way for gradual enrichment of the corpus by external contributors[5] (and facilitating other endeavours for other languages/projects).

[2] https://corpafroas.huma-num.fr/

[3] https://cortypo.huma-num.fr/

[4] Members of the two projects (in bold those who participated in both projects, in roman the CorpAfroAs-only, and in italics the CorTypo-only participants) are: *Evangelia Adamou*, **Azeb Amha**, *Mourad Aouini*, Alexandrine Barontini, *Isabelle Bril,* **Bernard Caron**, Dominique Caubet (co-PI), **Christian Chanard** (co-PI)**, Bernard Comrie**, Huyen-Tô Dan-Rabier, *Zygmunt Frajzyngier* (co-PI), *Katharina Haude*, Shlomo Izre'el, *Marc Kemps-Snijders*, Cécile Lux, *Tahar Meddour*, **Il-Il Malibert-Yatziv,** Stefano Manfredi**, Amina Mettouchi** (Principal Investigator (PI)), Christophe Pereira, *Stéphane Robert*, *Paulette Roulon-Doko*, Graziano Savà, *Erin Shay*, Marie-Claude Simeone-Senelle, Mauro Tosco, **Martine Vanhove** (co-PI), Angeles Vicente, Coralie Villes, *Jeanne Zerner*.

[5] Recently, a file in Siwi was added to the Berber subcorpus by V. Schiattarella.

language	*duration (mn:sc)*	*words*	**mb cells or **morphemes*[6]
Berber			
Kabyle	49:11	7437	**23336
Tamasheq	10:38	1228	*2325
Chadic			
Hausa	11:55	11981	*14736
Zaar	61:20	10629	*13304
Creole (Arabic)			
Juba	45:55	9667	*12865
Cushitic			
Beja	56:35	5890	*12507
Gawwada	19:38	2394	*5651
Tsamakko	28 :18	2222	**5154
Omotic			
Wolaytta	13:33	1408	*2893
Semitic			
Arabic, Moroccan	83:17	12430	*23905
Arabic, Tripolitanian	27:23	3293	*5760
Modern Hebrew	63:40	7537	*14339

Table 1: The CorpAfroAs corpora (https://corpafroas.huma-num.fr/Archives/corpus.php)

CorTypo is larger in its representation, as it contains languages from various phyla. There is no attempt at representativeness, the corpus interfaced to the typological query database is minimum 30 minutes of annotated spoken data per language, most of the corpora following the CorpAfroAs annotation template.

[6] Morphosyntactic words were segmented into morphemes by each language specialist. Each specialist decided to put into each cell what was amenable to segmentation. The number of mb cells does not necessarily match the number of morphemes. All figures correspond to numbers of mb cells, except for Kabyle and Tsamakko, for which the actual number of morphemes is given.

Figure 1: CorTypo languages

Contrary to CorpAfroAs, the corpus was not conceived as the main deliverable: it is invisible to the end-user, remains in the background and feeds the typological database. This typological database (Figure 4) is the point of access for end-users, and provides them with the complete list of excerpts corresponding to their query, automatically retrieved, in real time, from the various language corpora.

The Database is accessible through various points of entry: by language, by functional domain, by functions (i) as labelled by the language contributor in CorTypo, and (ii) as they might be labelled within different theoretical frameworks or approaches via the 'keywords' point of entry. For instance in Kabyle, both 'agent-affecting subject' and 'agent-oriented change of state' functions (defined language-internally on the basis of their being different constructions) might be searched under the keyword 'causative', which is the label a general linguist would probably use to look for this kind of function in a typological database. It only contains constructions for two functional domains,[7] predication and reference, but those domains have been investigated for all the languages of the project. The principle is for the end-user to browse the database and conduct comparisons (which will be explained further in 2.1.) based on the empirical and language-internal organisation of the database: only the functions that are formally encoded in the language are displayed. The challenge (and the innovation) of the approach, based on Frajzyngier's Systems Interactions framework, is to provide a formally-grounded typological approach to language data, resolutely favouring the slow empirical building of established language-internal categories as a basis for further comparison, in a bottom-up perspective.

3. Corpus design and research questions

3.1. Cross-linguistic comparability

There are as many ways of putting together a corpus, as there are purposes for that corpus.

Both projects involved field linguists who were used to glossing textual data with such softwares as Shoebox, Toolbox, and later, Elan. As the main aim of the field

[7] A functional domain, in Frajzyngier's approach, is a system of forms that have at least one functional/semantic feature in common and that are in complementary distribution.

linguists involved in the projects was to analyse the grammar of the language they worked on, their data were usually annotated at the morpheme-level. This level of annotation naturally became the starting point of the first project.

At the time when the project was conceived, in 2004, there were very few corpora in lesser-described languages, and none of them had established a standard layout in view of automatic exploration. Most of the existing ones used annotation either for the creation of dictionaries, or, for the few that were already time-aligned,[8] in order to archive online data that could be understood by non-specialists of the language. Allowing non-specialists of a given language to read illustrative examples in publications was also the purpose of the standardized interlinear glossing system of the Leipzig Glossing Rules, developed in relation to the expansion of typological studies.

The CorpAfroAs project was conceived as a pilot endeavour whose purpose was to create an automatically searchable, time-aligned corpus in several AfroAsiatic languages, in view of allowing comparative investigations across those languages. Morphemic glossing was taken to a radically different level from previous usage: The aim was no longer to assist the reader in understanding the various grammatical components of a clause in a language they were not familiar with, or to allow the creation of dictionaries, but to reflect the grammatical organization of the language under study, through the systematic and consistent morphosyntactic annotation of whole transcribed texts. Contrary to isolated glossed examples in scientific articles, in which the categories that are not relevant for the demonstration at that point in the paper are not necessarily glossed, in CorpAfroAs all morphemes were systematically and consistently annotated, thus allowing a variety of queries.

The project involved data collection, transcription, annotation, in Toolbox or in Elan: CorpAfroAs is a created corpus, not a compiled one. Each language corpus was annotated according to its author's analysis of the language's grammatical features, with some degree of cross-corpora convergence based on the use of English terms for glossing, and the use of common abbreviations for the chosen categories. Typically and as an example, this involved using 'Perfective' instead of 'prétérit' or 'accompli' which are the terms traditionally used in francophone Berber studies. The term 'Perfective' was defined language-internally, with specific nuances depending on the language, but by consensus, each contributor of the corpus using the gloss 'Perfective' used the abbreviation PFV (not PERF, or PRFV, or whatever other abbreviation). Using common abbreviations (which doesn't entail common definitions) has been crucial for cross-linguistic corpus queries, and was supervised by our invited expert, Bernard Comrie. The elaboration of a common template using the same number and hierarchy of tiers (Table 2) was also key in making the subcorpora easily interoperable.

Choices of annotation template and segmentation (Table 2, Figure 2) were grounded in the main initial research question of the project, namely: How does morphosyntax interact with prosody in AfroAsiatic languages? This question drove and pollinated subsequent research, up to this day, as we are going to illustrate in this paper.

[8] See e.g. the *Pangloss* archive, originally developed by the LACITO research unit and now a joint multi-team project (https://pangloss.cnrs.fr/).

ref	identifier for the annotation unit (time-associated)		
tx	transcription in broad phonetics into phonological words (SA)		
	mot	intermediary tier with segmentation into morphosyntactic words (SS)	
		mb	morphophonological transcription into morphemes (SS)
		ge	morpheme-by-morpheme gloss of mb according to the Leipzig Glossing Rules, expanded within the project (SA)
		rx	part-of-speech and other information relevant for retrieval purposes (SA), with standardized glosses available on https://corpafroas.huma-num.fr/glosses.html
ft	free translation into English (SA)		
SA: symbolic association. SS: symbolic subdivision			

Table 2: Annotation template of CorpAfroAs

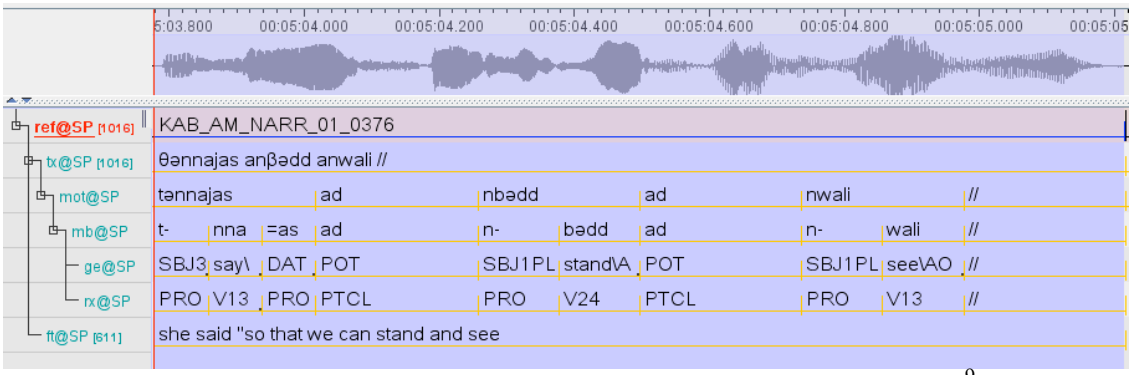


Figure 2: Annotation example in CorpAfroAs (Kabyle (Berber) subcorpus)[9]

The corpora were ultimately annotated in Elan-CorpA, a version of Elan[10] developed within the CorpAfroAs project (2007-2012), with an added internal parser linked to a lexicon, for semi-automatic interlinearization purposes, and a query language. Elan-CorpA was then interfaced to an online query tool, Elan Websearch (Figure 3).

[9] Glossing follows the Leipzig Glossing Rules. Additional abbreviations are:
ABSL: absolute state; ABS: absolutive pronominal paradigm; ANN: annexed state; AOR: aorist; APHO: apophony; CIRC: circumfix; CNS: consensual; CONJ: conjunction; COV: covert; DIR: directional; EXS: existential; GEM: derivation by germination; HESIT: hesitation; IRG: irregular; KIN: kinship; MID: middle; N.AC: action noun; N.P: proper noun; OV: overt; PFX: prefix; PNG: person, number, gender; POT: potential; PREP: preposition; PRO: pronoun; PTCL: particle; REAL: realis; RELSBJ: subject relativization affix; SBJ: subject pronominal paradigm; SING: singulative; TAM: tense, aspect, mood; V3a: verb class sp.; V + number: verb category; V%: different thematic vowels in 1st and 2nd SG than in other persons in PFV; V.DER: derived verbal form.

[10] Elan is a software created and developed by the MPI in Nijmegen https://archive.mpi.nl/tla/elan.

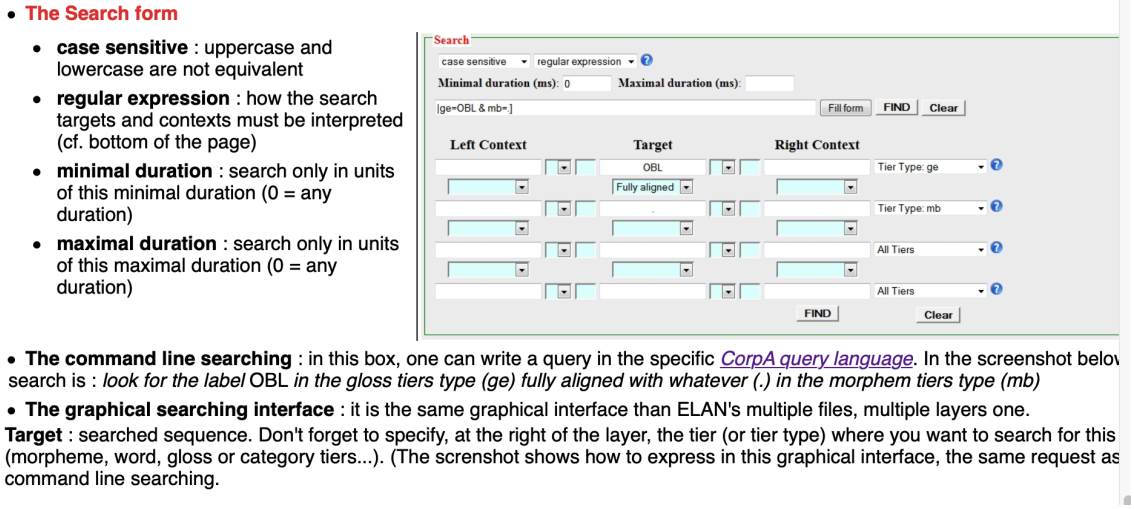


Figure 3: Elan Websearch (Chanard 2015)

The CorTypo project was born of the limits of cross-linguistic comparison in CorpAfroAs, which had been carried out directly from the corpus data, using grammatical sketches or lists of glosses as sources for definitions of the functions and categories used by each corpus author. Some searches were indeed possible without much mediation (e.g. "which languages have a postverbal vs. a preverbal negator, or both?" an automatic cross-linguistic query made possible by the fact that there is a standard abbreviation for "negator" (NEG), a standard one for "verb" (V), and that all files have the same template), but many others (e.g. "Does the language formally distinguish between negative existential and negative locative?") required the mediation of a grammatical description of some kind ("is there a dedicated negative existential/locative construction, if this is not straightforwardly marked either by a dedicated morpheme, or by an existential/locative predicate to which a standard negator is added?") before the data could be searched.

CorTypo therefore carried the comparative dimension further, and interfaced a full comparative database to a corpus whose template was basically the same as the CorpAfroAs one, for most languages.

The comparative database was inspired by the Systems Interactions framework developed by Frajzyngier (1999, 2004, 2016), and features an empirical approach, where all categories defined in each language are based on formal evidence of their grammatical encoding in that language. Given this bottom-up approach, the project could only be a pilot one, featuring a limited (but established) number of functional domains: Predication, and Reference. Predication refers to the way in which states and events are encoded by predicate types in the language, while Reference is the way in which the status of entities is grammatically encoded.

While CorpAfroAs contains exclusively Afro-Asiatic languages, the scope of CorTypo was not limited in terms of language families, and the languages were included on the basis of the willingness of the language specialists contacted to take part in a project involving radical empirical re-analysis of their data, in order to test an alternative model of cross-linguistic comparability, and its implementation in the form of a pilot database. Some of the languages (and contributors) in CorTypo are the same as the ones in CorpAfroAs.

The issue of cross-linguistic comparison has informed both projects from the start, and is broached in Mettouchi, Savà and Tosco (2015), as well as in Frajzyngier and

Mettouchi (2015). Empiricism was key in both projects, as well as bottom-up convergence rather than top-down annotation, but was pursued more radically in CorTypo than in CorpAfroAs.

In CorpAfroAs, the corpora were accompanied by lists of glosses, or full grammatical sketches including definitions of the categories and functions existing in the language. The purpose was to underline the fact that it was the end-user's responsibility, if she was a typologist looking for cross-linguistic generalizations, to decide that category X in language A and category Y in language B were subsumable under the same comparative concept (Haspelmath 2010) or cross-linguistic general category (Lazard 2006). In this way, a number of biases could be avoided when using the corpora for cross-language comparisons. And crucially, the authors of the corpora could use them for their own research, based on their own analysis.

In CorTypo, we took it upon ourselves to define the existing functional domains, functions and constructions in each language (as exemplified in Figure 7), based on verifiable formal evidence, in the same way that categories and constructions are established in the grammar of a language. As shown in Figure 4, the principle underlying the comparison is systemic: what are the type and number of domains in L1, L2, L3...? For each existing domain, what is the structure of the domain in L1, L2, L3…?, and for each language, what are the specific constructions and forms used to code the functions?

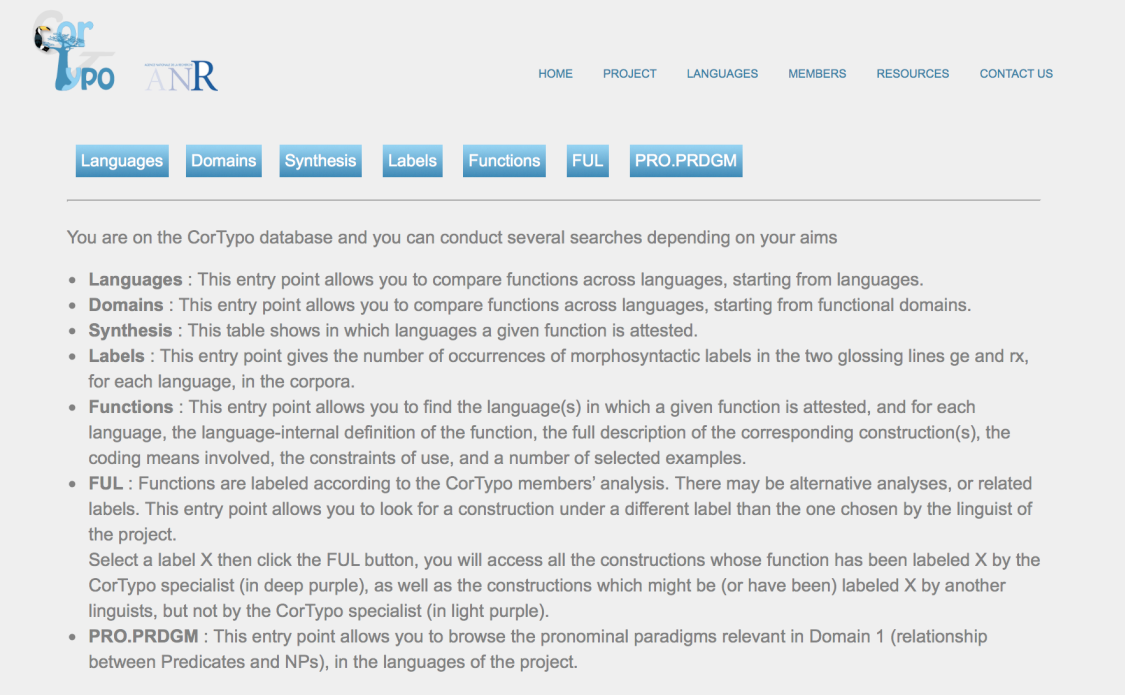


Figure 4: End-User interface for interrogation of the CorTypo database

For an entry point from the perspective of LANGUAGES, the end-user gets the type of result given in Figure 5, clicking on Phyla, and developing each branch and language. For each language, the available functional domains (Predication, and Reference) appear, and the end-user can then develop each domain, displaying the language-internal constructions that form this functional domain.

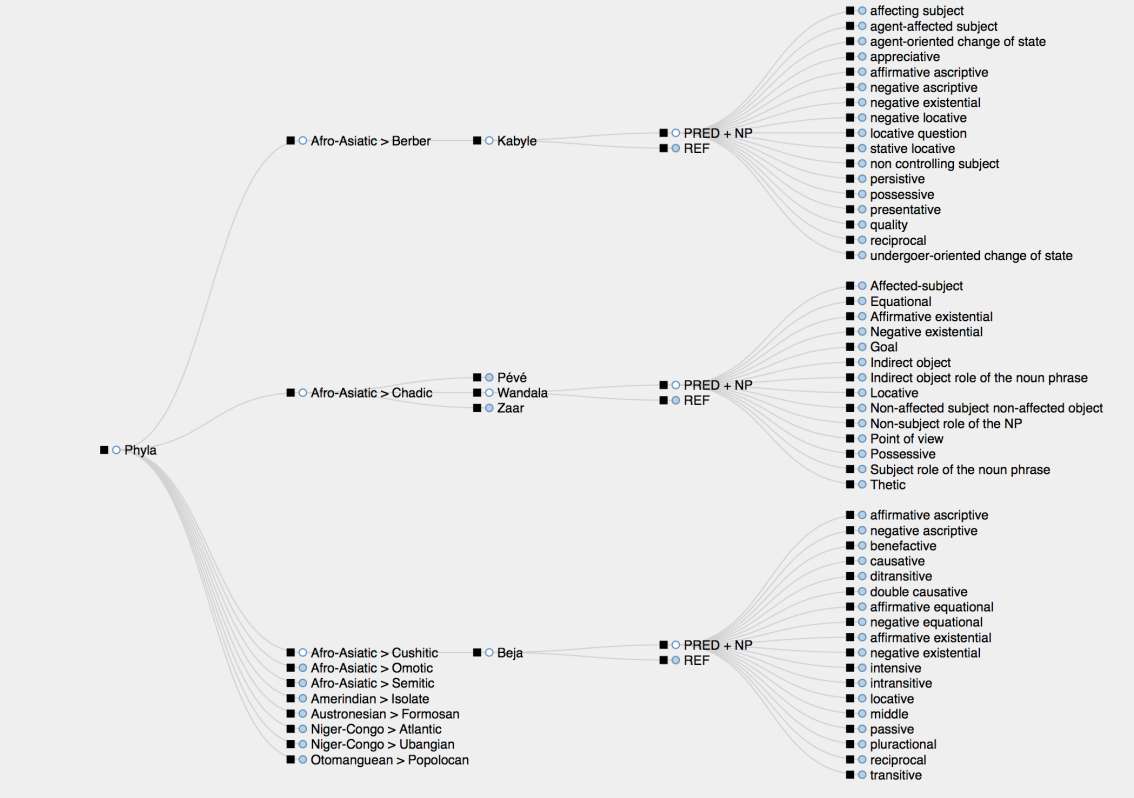


Figure 5: Unravelling the LANGUAGES entry point, to reach the synthetic display of the functional domain of PREDICATION[11] (abbreviated as PRED+NP) in each selected language.

As functions within each functional domain are defined on the basis of their formal encoding in the language, some functions may appear in the list of language A, and not in that of language B: Existential is a formally-specific construction in Amis, Movima, Gbaya, Wolaitta and Zaar as shown in Figure 6, not e.g. in Kabyle where it is expressed by a standard verbal predication. Kabyle, however, has a special non-verbal construction for the negative existential. The absence of a construction in the list may mean that (i) either the function is not grammaticalized (i.e. marked by a dedicated structure in the grammar) in the language, or that (ii) it is grammaticalized, but within another functional domain (as each domain is defined by its internal form-function cohesion, not by an a-priori list of universal functions).

For instance, transitivity or valency are not functions of the Domain of Predication in Kabyle or Chadic; in Kabyle for example, the construction "Agent-oriented change of state" is transitive, but this is one of its properties, not its function. The construction describes a change of state involving an agent and an undergoer, viewed as a telic action performed by the agent on the undergoer. It involves a labile verb and two arguments, one expressed by a bound subject pronoun, the other by an absolutive clitic pronoun, or by a noun in the absolute state. This construction underlines the fact that the change of state referred to by the labile verb is perceived

[11] The working definition for this domain is: the semantic relation between a (verbal or nominal or otherwise) predicate and its arguments/participants. It is not limited to verb valency or transitivity, nor is it strictly syntactic ('clause-type'), or defined based on truth-values ('proposition'), but is rather a group of relational functions hinging around a predicate.

from the perspective of the agent, which is the subject of the labile verb, the undergoer being its object. This construction aligns subject with agent, and object with undergoer.

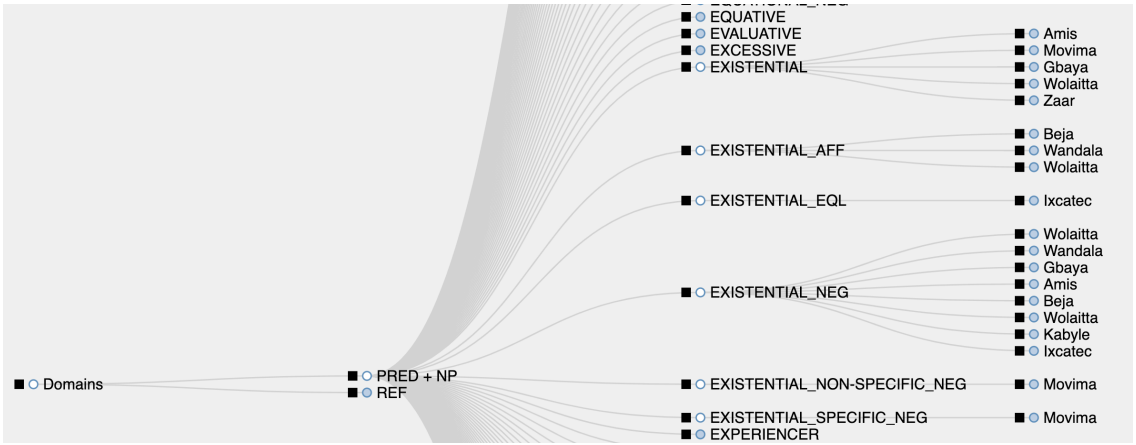


Figure 6: Unravelling the DOMAINS entry point, to reach the synthetic display of the languages in which existential constructions are found.

Selecting all existential constructions in the languages where they have been proved to exist by the language specialists, the end-user is taken to a list presenting the semantic definition, and the formal encoding of those constructions, as well as the formula allowing automatic retrieval of the corresponding examples in the appended corpus (Figure 7).

Kabyle (Mettouchi)
[PRED + NP] **negative existential**
Definition : Negative existential denies the existence of a referent, or a situation.
Construction : Negative Existential has the form ulaʃ, with NO clitic attached, preceded or followed by an NP in the absolute state, where the form ulaʃ is the negative existential predicate.\nulaʃ + NO absolutive pronoun or proper noun
Contrast : Negative existential predication is different from negative locative predication because in the former there must not be any absolutive clitic attached, nor proper noun.
Request : [ge=NEG.EXS & rx=PRED]{ge=1 & rx=1}[ge=NOT(ABS\b) & rx=NOT(PRO|NP)]

Wandala (Frajzyngier)
[PRED + NP] **Affirmative existential**
Definition : Asserts the existence of an entity
Construction : áŋkwà NP
Contrast : Contrasts with negative existential predication, contrasts with other verbless predications, viz. the equational, possessive, and locative predications.
Request : [ge=EXIST]

Wandala (Frajzyngier)
[PRED + NP] **Negative existential**
Definition : Denies existence of an entity
Construction : ɓákà or ɓáakà in clause-initial or clause-final position
Contrast : The predication is in contrast with the affirmative existential predication
Request : [mb=ɓáakà]ORmb=ɓákà]

Beja (Vanhove)
[PRED + NP] **affirmative existential**
Definition : a constuction which refers to the existence or presence of something/someone
Construction : Existential has the following forms: Noun (in nominative) + one of the three locative verbs: haj, faj or da ´be_there´\n(ge = NOM) + rx = N + ge = be_there + rx = V
Contrast : locative; existential
Request : [rx=\bN\b&ge=.]{rx=1&ge=1}[rx=PNG&ge=.]?{rx=1&ge=1}[rx=V&ge=be_there]

Beja (Vanhove)
[PRED + NP] **negative existential**
Definition : a constuction which refers to the non existence or absence of something/someone
Construction : Negative existential has the following forms: Noun (in nominative) + one of two locative verbs: haj or da ´be_there´
Contrast : negative equational, negative identification
Request : [ge = NEG]{ge<3}[ge = be_there]

Figure 7: Selecting all existential predications, and displaying their details in some languages.

By clicking on the query, which is a formulation of the construction using regular expressions and the abbreviations of the corpora, as well as its tiers and cells identifiers, the end-user can retrieve the corresponding corpus examples, and then expand their textual contexts (Figures 8 and 9).

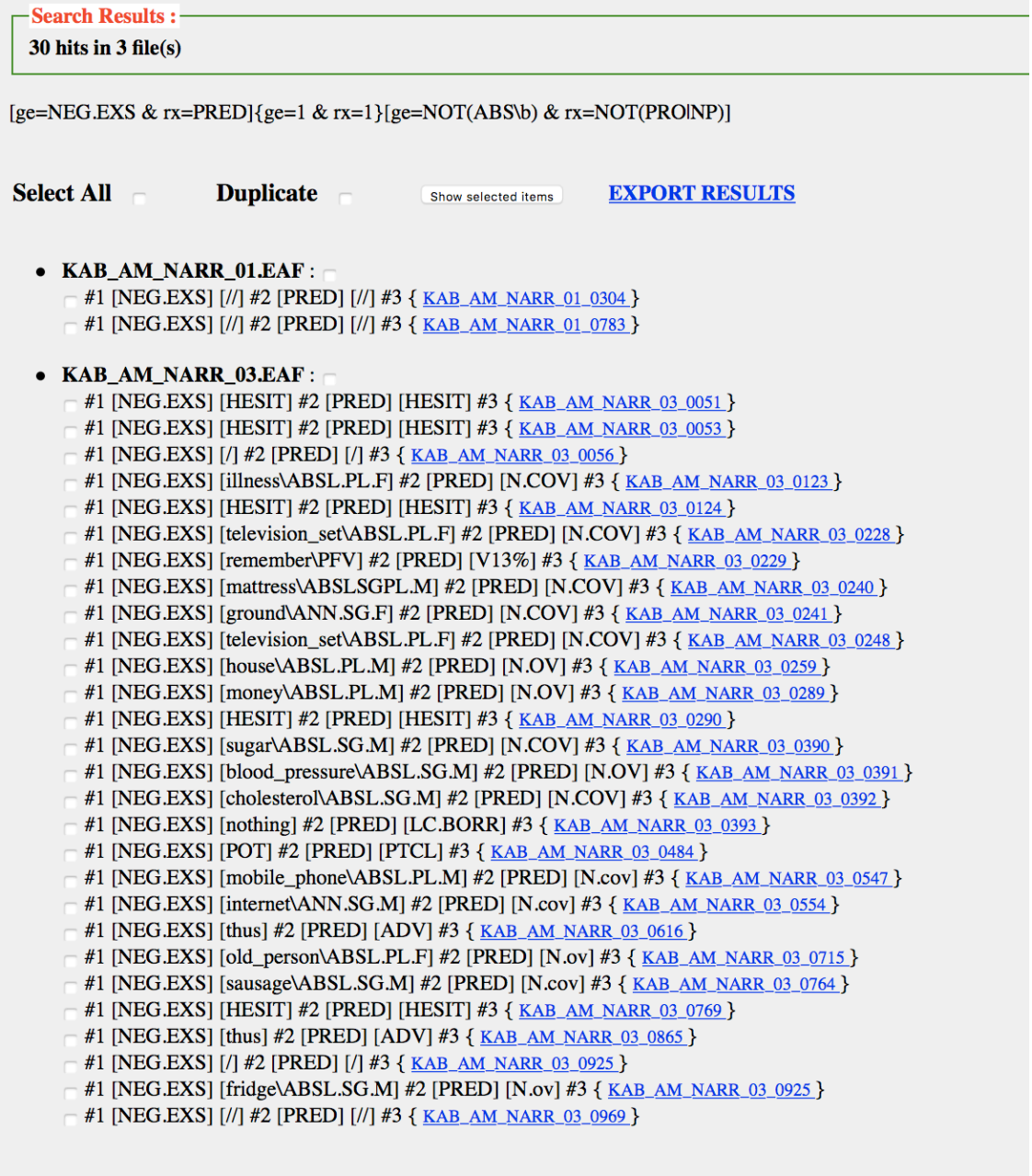


Figure 8: Automatically retrieving all the hits of the corpus (here an excerpt for Kabyle)

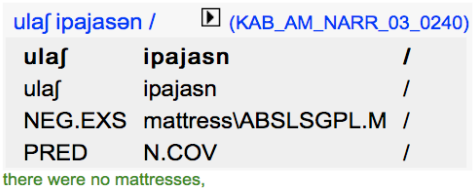


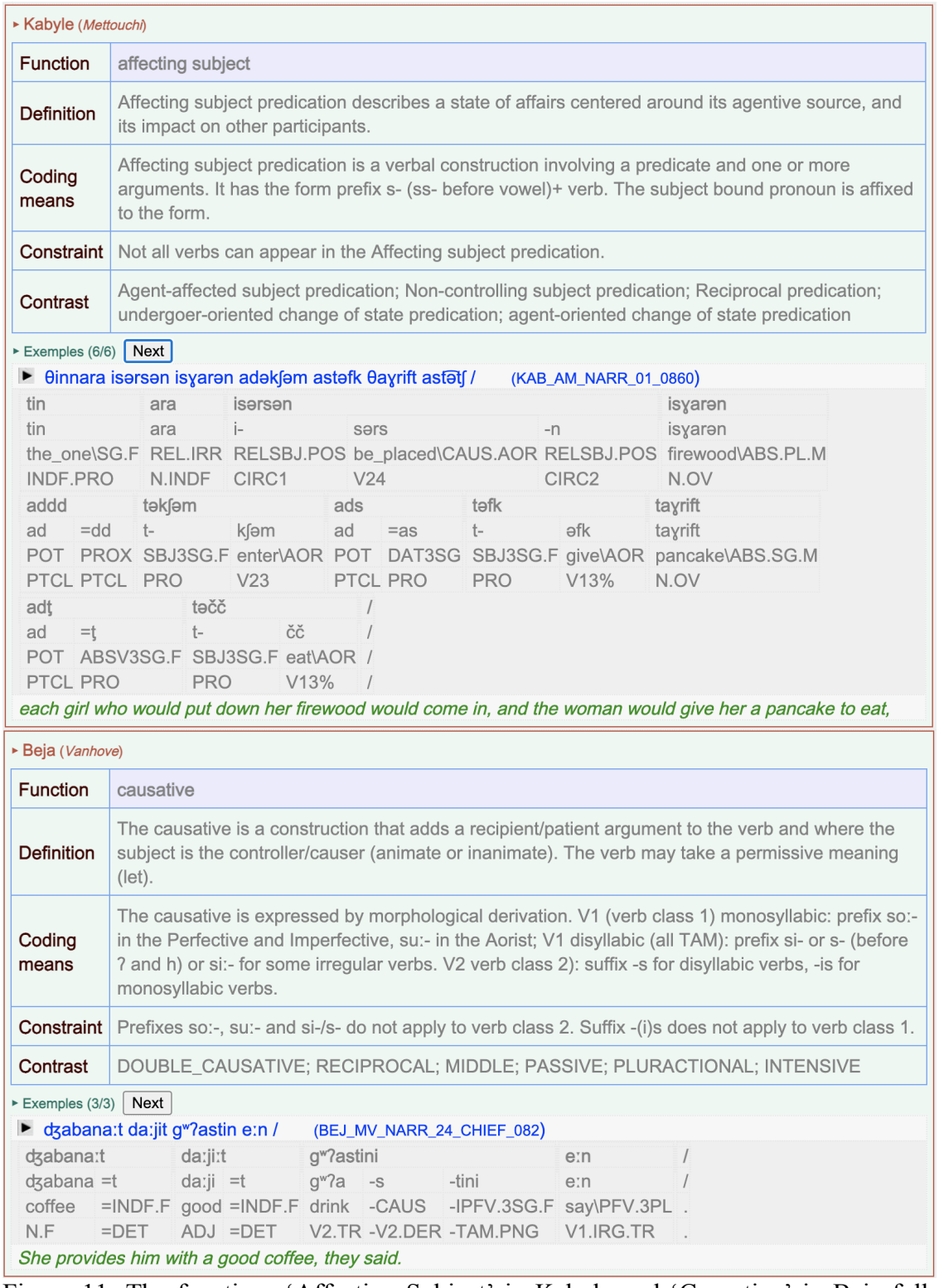
Figure 9: Corpus example (here in Kabyle)

For an entry point from the perspective of FUNCTIONS, the end-user gets the type of results given in Figure 6, clicking on Domains, and developing one of the two branches. Then, she can choose to expand some constructions, based on the hypothesis that they are indeed comparable, and check for herself if this is indeed the case. For example, by

clicking on “Causative” and “Affecting subject” (two labels used by the contributors of the database for their language corpus), she can notice that eight languages contain such functions (Figure 10) and explore this further.

Figure 10: Unravelling the FUNCTIONS entry point, to reach the list of functions one wishes to explore and compare (here various types of ‘Causative’ and ‘Affecting Subject’).

Then each of the results can be expanded to reveal definition, coding means, constraints of use, and possible contrasts with other forms within the same domain, for each language concerned as in Figure 11.



▸ Kabyle (*Mettouchi*)

Function	affecting subject
Definition	Affecting subject predication describes a state of affairs centered around its agentive source, and its impact on other participants.
Coding means	Affecting subject predication is a verbal construction involving a predicate and one or more arguments. It has the form prefix s- (ss- before vowel)+ verb. The subject bound pronoun is affixed to the form.
Constraint	Not all verbs can appear in the Affecting subject predication.
Contrast	Agent-affected subject predication; Non-controlling subject predication; Reciprocal predication; undergoer-oriented change of state predication; agent-oriented change of state predication

▸ Exemples (6/6) Next

▶ θinnara isərsən isɣarən adəkʃəm astəfk θaɣrift asʃəʃ / (KAB_AM_NARR_01_0860)

tin		ara	isərsən					isɣarən	
tin		ara	i-	sərs			-n	isɣarən	
the_one\SG.F		REL.IRR	RELSBJ.POS	be_placed\CAUS.AOR			RELSBJ.POS	firewood\ABS.PL.M	
INDF.PRO		N.INDF	CIRC1	V24			CIRC2	N.OV	
addd		təkʃəm		ads		təfk		taɣrift	
ad	=dd	t-	kʃəm	ad	=as	t-	əfk	taɣrift	
POT	PROX	SBJ3SG.F	enter\AOR	POT	DAT3SG	SBJ3SG.F	give\AOR	pancake\ABS.SG.M	
PTCL	PTCL	PRO	V23	PTCL	PRO	PRO	V13%	N.OV	
adt		təčč		/					
ad	=t	t-	čč	/					
POT	ABSV3SG.F	SBJ3SG.F	eat\AOR	/					
PTCL	PRO	PRO	V13%	/					

each girl who would put down her firewood would come in, and the woman would give her a pancake to eat,

▸ Beja (*Vanhove*)

Function	causative
Definition	The causative is a construction that adds a recipient/patient argument to the verb and where the subject is the controller/causer (animate or inanimate). The verb may take a permissive meaning (let).
Coding means	The causative is expressed by morphological derivation. V1 (verb class 1) monosyllabic: prefix so:- in the Perfective and Imperfective, su:- in the Aorist; V1 disyllabic (all TAM): prefix si- or s- (before ʔ and h) or si:- for some irregular verbs. V2 verb class 2): suffix -s for disyllabic verbs, -is for monosyllabic verbs.
Constraint	Prefixes so:-, su:- and si-/s- do not apply to verb class 2. Suffix -(i)s does not apply to verb class 1.
Contrast	DOUBLE_CAUSATIVE; RECIPROCAL; MIDDLE; PASSIVE; PLURACTIONAL; INTENSIVE

▸ Exemples (3/3) Next

▶ ʤabana:t da:jit gʷʔastin e:n / (BEJ_MV_NARR_24_CHIEF_082)

ʤabana:t		da:ji:t		gʷʔastini			e:n	/
ʤabana	=t	da:ji	=t	gʷʔa	-s	-tini	e:n	/
coffee	=INDF.F	good	=INDF.F	drink	-CAUS	-IPFV.3SG.F	say\PFV.3PL	.
N.F	=DET	ADJ	=DET	V2.TR	-V2.DER	-TAM.PNG	V1.IRG.TR	.

She provides him with a good coffee, they said.

Figure 11: The functions ‘Affecting Subject’ in Kabyle and ‘Causative’ in Beja fully developed with examples

Other searches are possible through the database, for instance with keywords, that cannot be detailed here for lack of space. A query on the keyword “Causative” would yield all the constructions for which the linguist responsible for the language in the

database has considered that this label might be or has been chosen as a label for this construction by other specialists of the language / by general linguists. It would yield the "Affecting subject" construction of Kabyle, which does not correspond to the usual definition in terms of valency-changing, or causer-causee pair (it contains for example verbs derived from onomatopoeia, referring to animal sounds), but would probably be categorized as "causative" by a general linguist. The Keywords allow the contributors of the database to have radically language-internal labels for their constructions, while offering other possible points of entry from different perspectives.

3.2. Transcription issues and prosodic segmentation

Because a number of languages (mostly in the Semitic and Berber branches) displayed a high degree of sandhi and specific surface phonetic realization of underlying forms, a decision was made to have two transcription lines: one that reproduced as faithfully as possible the spoken monologue or interaction, allowing the end-user to recognize the elements of the speech continuum (tx), and one that uncovered the underlying morphosyntactic structure (mot). This triggered reflections about the notion of word in each language: **tx** is dedicated to prosodic or phonological words (whose properties are based on such elements as accentuation, vowel harmony etc.), transcribed in broad IPA, whereas **mot** contains grammatical words (whose properties are based on delimitation of morphemes composing them, and their combination rules), transcribed morphophonologically (see Izre'el & Mettouchi 2015 for details).

The morphosyntactic transcription was then segmented into morphemes (**mb** tier), themselves glossed on two different lines: **ge** for categories and functions, as well as lexical information, **rx** for parts of speech and other relevant annotations.

Regarding time-alignment, the innovative feature of CorpAfroAs is that it treated segmentation not as a practical chunking issue, but as a theoretical one. It was crucial to CorpAfroAs that segmentation should be consistent within and across the languages of the Afroasiatic corpus, and based on a comparable, clearly defined unit. Building on previous research conducted on both well-described and lesser-described languages (Chafe 1994, Cruttenden 1997, among others), the intonation unit, a prosodically coherent contour, delimited by prosodic boundaries, was taken as the unit for segmentation. Two boundary types were manually annotated: Terminal, and non-terminal. In some subcorpora, a third type, truncated intonation unit, was added. Consistency of segmentation, which was based only on formal prosodic cues (mainly pause, final lengthening, initial rush, pitch reset), and done based on a combination of speaker perception and acoustic control with Praat, was supervised by our invited expert, Shlomo Izre'el. Not all languages, genres, speakers favoured exactly the same cues, but those four cues were the most frequently involved in intonation unit boundary marking (for details, see Izre'el & Mettouchi 2015). Although those cues also play a role in marking the boundaries of lower-level units (such as prosodic words, or phrases), it is possible to attribute them to the right hierarchical level in most cases because (i) intonation unit boundaries tend to be stronger than word or phrase boundaries, always involving more than one cue, and (ii) speakers/hearers perceive them comparatively, in relation to the surrounding stretch of speech, for a given speaker inside a given recording (this works best if segmentation is done by one speaker/hearer in real time, especially for narratives). Of course in some cases the decision can be hard to make (for a discussion on these aspects of segmentation, see Barth-Weingarten

(2011)), but overall, segmentation was relatively unproblematic. This aspect of corpus-building will be developed in the case-studies in Section 4.

Finally, it was decided that pauses above a 200ms threshold (100ms for some sub-corpora with faster speech-rates) would be annotated, with their duration indicated. In some corpora, their nature (silent vs. with a breath intake) was indicated. Dysfluencies (hesitations, false starts) were annotated as well.

3.3. Discussion

As is clear from those preliminaries, annotation and segmentation of the corpora, in CorpAfroAs and CorTypo, are grammatical, in the sense that they reflect the morphosyntactic organization of each language, and prosodic, in the sense that they reflect the way the information units are packaged during the flow of speech production. Corpus annotation is conceived as amenable to all investigations one can conduct in order to analyse the grammatical organization of a language. This allows for a wide number of research questions to be addressed.

Both corpus-based and corpus-driven research can be conducted on our data, and verifiability, falsifiability, transparency and contextualization are important aspects of both projects.

Finally, it must be underlined that even though both projects have comparative dimensions, they are primarily the work of language specialists annotating their own data in order to answer their own language-internal questions, and come up with a consistently and accurately annotated corpus.

The comparative dimension is an organizing principle linking the corpora together, it is not the essence of the annotation or the segmentation themselves, which remain specific to each language. We decided to use English as annotation and translation language, we used labels that are encountered in many language descriptions, such as ‘definite’, or ‘future’, the abbreviations for those labels were standardized (e.g. if a language had the category ‘definite’, it was annotated DEF, and not DFNT, but the definition of ‘definite’ was not imposed, each language specialist was invited to enter theirs in a “wiki” internal to CorpAfroAs, and to explicit them in their grammatical sketch or list of glosses. Later, within CorTypo, those definitions were showcased as essential elements in the pursuit of a cross-linguistic comparison anchored in language-internal categories. In CorpAfroAs the frontier between language-internal and cross-linguistic categories was not systematically underlined, and approaches using “comparative concepts” (Haspelmath 2010) were easily implementable by directly using corpus labels; for CorTypo, the stance was more radical, towards a resolutely language-internal, bottom-up approach. A more mainstream perspective, in terms of general, cross-linguistic, prototypical categories and the like, is made possible through the KEYWORDS entry point of the database, where usual terms (that each language specialist of the corpus identified as possible labels for the categories they identified) were entered as tags for each construction. In this way, what for Mettouchi in Kabyle is an “affecting subject” construction, can be retrieved through this entry point, by using the label “causative”, more often used in the literature.

All this underlines the fact that our annotated corpora are compatible with several approaches. This will be illustrated in the following studies of information structure and grammatical structures in Kabyle (Berber) and Beja (Cushitic), and in a comparative analysis of reported speech.

4. Case Studies in Prosodic Segmentation

4.1. Kabyle

4.1.1. Language profile and corpus

Kabyle (Berber, Afroasiatic) has about four million speakers in the north of Algeria. The variety represented in the corpus (Mettouchi 2012) is a Central-Western dialect, spoken in the village of Ait Ikhlef, close to the town of Bouzeguene.

In Kabyle (Mettouchi 2017), a minimal predication consists either of a verb and its bound personal pronoun, or of a non-verbal predicate. In addition, the clause may contain noun phrases, and prepositional phrases, as well as adverbs. Within noun phrases, modifiers follow the modified constituent. The language has two genders and two numbers, marked on adjectives, nouns, and pronominal affixes and clitics hosted by verbs, nouns and prepositions. It also has a binary morphological alternation marked on nouns, called states (Mettouchi 2014) (Table 3).

	Masculine		Feminine	
	Singular	Plural	Singular	Plural
Absolute	**a**-mɣar	**i**-mɣar-**n**	**t-a**-mɣar-**t**	**t-i**-mɣar-**in**
Annexed	**w**-mɣar	**j**-mɣar-**n**	**t**-mɣar-**t**	**t**-mɣar-**in**
(root *mɣar*, 'old person')				

Table 3: Gender, Number and State in Kabyle

The Kabyle corpus that is accessible online is composed of two folktales, a personal recount, and a three-party conversation. The different measurements are given in Table 4. Morphosyntactic words are defined as including the root and its affixes and grammatical clitics.

	DURATION (minutes)	MORPHEMES	WORDS (grammatical)	INTONATION UNITS	SILENT PAUSES (or breath intakes)
FOLKTALE 01	13:29	6639	1803	614	392
FOLKTALE 02	12:16	6044	1748	546	372
RECOUNT 03	15:20	7302	2502	794	365
Total Monologues	41:05	19 985	6053	1954	1129
CONVERSATION	8:06	3351	1384	680	-

Table 4: Size of the Kabyle corpus (CorpAfroAs)

Apart from state on nouns, the other important coding means for the issues broached in the next developments is the wealth of pronominal paradigms. Mettouchi (2017) lists eleven paradigms. In most paradigms, only the first person does not distinguish gender (Table 5).

	Singular		Plural	
	masculine	feminine	masculine	feminine
1	i-nɣa=**i**		i-nɣa=**aɣ**	
2	i-nɣa=**k**	i-nɣa=**km**	i-nɣa=**k^{w}n**	i-nɣa=**k^{w}nt**
3	i-nɣa=**t**	i-nɣa=**tt**	i-nɣa=**tn**	i-nɣa=**tnt**

Table 5: Absolutive (referential undergoer) pronouns in Western Kabyle (with *nɣa* 'kill')

4.1.2. Information Structure

There are several ways of studying information structure in a language, and cross-linguistically. It is frequent for instance to start from concepts (topic, focus and the like) and apply them to categories (argument focus, topicalisation of the object, etc.) or to a level of analysis (sentence topic, discourse topic). The question can then be for instance "how is object focus encoded in Language L (or languages L1, L2, L3...?)". In this perspective, for transparency and falsifiability, one needs not only a corpus, but also explicit functional definitions of such categories as topic and focus, which are (a) notoriously different across frameworks and approaches, (b) not necessarily marked in all languages, and (c) not necessarily forming the same systems of oppositions from one language to the next. Of course, it goes without saying that even for similar functions, the encoding forms will vary across languages. One advantage of this approach is however that it provides fast results, and is easy to apply to multiple languages, because it does not rely on forms but on a semantic/functional definition that one may consider cross-linguistic. Two huge challenges nevertheless remain: Are we sure we are studying the language itself and not the translation? And are we sure that those categories are indeed universal?

Among the other ways of working on information structure, there are empirical approaches based on form. This is the type favoured below. It consists in starting from formal configurations in the language under investigation, finding out what their functions are, and seeing how they form a system within one language. For cross-linguistic comparison, one has to conduct empirical investigations on each language, and then one can compare them. This can be done either on the basis of similar form-function pairs (focus by clefting), or on the basis of similar internal organization of the form-function pairs inside the domain (information structure), or of the relationships of that domain with other domains in the language (reference, TAM, etc.). This is the approach chosen in CorTypo, which was inspired by the Systems Interactions Framework (Frajzyngier 1999, 2004). The advantages of the approach (a more language-internal approach, more sensitive to the uniqueness of each language, and to differences across languages) are also its drawbacks (an approach that requires a full analysis of the domain before comparison can be undertaken, and does not necessarily provide comparable categories, since these emerge from the language data, and not from a neat prior definition. It is hence less amenable to vast and fast comparisons).

Of course those two approaches are not opposed perspectives, and there is a subset of shared assumptions and methods between them. It is nevertheless important to position oneself and state one's starting point, especially in relation to corpus analysis: The first approach implies a more corpus-based perspective, whereas the second is associated to a more corpus-driven one.

For the analysis of Kabyle, it is this second approach that was implemented, with a heuristic method based on the first author's native competence in the language, as well as preliminary investigations of the data determining which coding means were

good candidates for the encoding of functions pertaining to information structure. The data were then systematically investigated, using automated searches based on regular expressions. It is worth underlining that the basic annotation of the CorpAfroAs corpus only contains morphematic glossing and parts of speech; no information structure category was annotated, unless it was marked by a dedicated morpheme, which is not the case in Kabyle.

The categories involved in information structure constructions are nouns, independent and bound pronouns, verbs, relativizers and demonstratives (all annotated as such in the corpus). The coding means involved are word order, state, prosodic segmentation into intonation units, and boundary tone of those units. Those coding means were either annotated (state, IU boundary, boundary tone (final vs non-final), or automatically searchable (e.g. order of nouns relative to verbs).

Then the presence vs absence of a noun phrase relative to the verb[12] and relative to the prosodic boundary (which due to its necessary presence and perceptual salience is also a reference point (Mettouchi 2011, 2013, 2015 and 2018a)), were investigated. The following sequences were found using automatic queries:[13]

$[V_{sbj}]$
$[V_{sbj}\ N_{absl}]$
$[V_{sbj}\ N_{ann}]$
$[V_{sbj}\ N_{ann}\ N_{absl}]$
$[V_{sbj}\ N_{absl}\ N_{ann}]$
$[N_{absl}\ V_{sbj}]$
$[N_{absl}\ V_{sbj}\ N_{absl}]$
$[N_{absl}\ V_{sbj}\ N_{ann}]$
$\mathbf{N_{absl}}\ [V_{sbj}\ (N)\ (N)]$
$[V_{sbj}\ (N)\ (N)]\ \mathbf{N_{ann}}$

Then each sequence was investigated for its functional role, and some sequences were lumped together under one construction. For two or more sequences to be variants of a single construction, they have to express the same function. At the level of information structure, both sequences $[V_{sbj}]$ and $[V_{sbj}\ N_{absl}]$ carry the narrative or discourse forward, without change in topic or subtopic.[14] They are therefore variants of the same information structure construction, formally expressed as $[V_{sbj}\ (N_{absl})]$.[15]

[12] The verb is taken as a reference point for word order, due to its necessary presence, and its always bearing an overt subject bound pronoun. For criteria to establish a linguistic marker as a reference point, see Frajzyngier & Shay (2003: 60–62), and Mettouchi (2018a: 264).

[13] V_{sbj} refers to the obligatory combination of a verb and its subject affix; N_{ann} to a noun in the annexed state; N_{absl} to a noun in the absolute state. Square brackets [ ] represent prosodic boundaries.

[14] Topic and subtopic are understood here as discourse-based categories: the topic is what the larger text/episode is about (e.g. weaving), and this topic is developed into several related subtopics (e.g. shearing the sheep, washing and drying the wool, carding and spinning it, etc.).

[15] The parentheses indicate that in the domain of information structure, the sequence involving a noun in the absolute after the verb within the prosodic group of the verb $[V_{sbj}\ N_{absl}]$ and the sequence involving only the verb and its bound subject pronoun $[V_{sbj}]$ are variants of the same construction. The two sequences may however be different constructions in the domain of grammatical relations, or in the domain of reference: although both sequences express the same information structure function, at the level of grammatical relations, $[V_{sbj}]$ is an intransitive construction, whereas $[V_{sbj}\ N_{absl}]$ is a transitive construction with a nominal object.

The following constructions[16] (ex. 1-9) are part of the functional domain of Information Structure in Kabyle, and their function is defined:

(i) [V_{sbj} (N_{absl})] codes (sub-)topic continuation. It is by far the most frequent, and the less marked construction. The narration or conversation thread is just carried forward. This construction subsumes sequences:

[V_{sbj}] [V_{sbj} N_{absl}]	(1)	iqqaz iqqaz iqqaz iqqaz iqqaz iqqaz / (KAB_AM_NARR_01_0191) iqqaz iqqaz iqqaz iqqaz iqqaz iqqaz / i- qqaz i- qqaz i- qqaz i- qqaz i- qqaz i- qqaz / SBJ.3SG.M dig\IPFV SBJ.3SG.M dig\IPFV SBJ.3SG.M dig\IPFV SBJ.3SG.M dig\IPFV SBJ.3SG.M dig\IPFV SBJ.3SG.M dig\IPFV / PRO V23.GEM PRO V23.GEM PRO V23.GEM PRO V23.GEM PRO V23.GEM PRO V23.GEM / He dug and dug, iχðəməlβir / (KAB_AM_NARR_01_0192) ixdəm lbir / i- xdəm lbir / SBJ.3SG.M make\PFV well\ABSL.SG.M / PRO V23 N.COV / he made a well,

(ii) [V_{sbj} N_{ann} (N)] codes promotion to discourse topic of a new event or situation. This construction subsumes sequences:

[V_{sbj} N_{ann}] [V_{sbj} N_{ann} N_{absl}] [V_{sbj} N_{absl} N_{ann}]	(2)	θʕawəð θəswiʕθ // (KAB_AM_NARR_03_0541) tʕawəd təswiʕt // t- ʕawəd təswiʕt // SBJ.3SG.F change\PFV period\ANN.SG.F // PRO V14 N.OV // times have changed,
	(3)	aɣiddəɦiku səttsi θimuʃuɦia // (KAB_AM_NARR_03_0245) adaɣidd təɦiku sətti timuʃuɦia // ad =aɣ =dd t- ɦiku sətti timuʃuɦia // POT DAT.1PL PROX SBJ.3SG.F tell\AOR grandmother\ANN.SG.F tale\ABSL.PL.F // PTCL PRO PTCL PRO V13% N.KIN.COV N.OV // My grandma would tell us folktales.
	(4)	təsʕaθajazʕitˤ jemmanuʒa / (KAB_AM_NARR_02_Midget_594) təsʕa tajaẓiṭ jəmma Nuʒa / t- sʕa tajaẓiṭ jəmma Nuʒa / SBJ.3SG.F possess\PFV hen\ABSL.SG.F mother\ANN.SG.F Nuʒa / PRO V13% N.OV N.KIN.COV N.P / Mother Nuja has a hen,

When the verb is *ili* 'exist', then the construction additionally codes introduction of a new referent in view of making it salient, and promoting it to the role of protagonist in the following discourse.

[16] In order not to occupy too much space, examples are given without their context. In order to fully grasp their information structure values, see Mettouchi (2015, 2018), and/or find the constructions and their contexts in the online corpus (https://corpafroas.huma-num.fr/Archives/).

[V_{sbj} N_{ann}] [V_{sbj} N_{ann} N_{absl}] [V_{sbj} N_{absl} N_{ann}]	(5)	illajiwən / (KAB_AM_NARR_01_0016) illa jiwn / i- lla jiwn / SBJ.3SG.M exist\PFV one\ANN.SG.M / PRO V13% PRO.N.COV / there was a man,

(iii) [N V_{sbj} (N)] codes recapitulation of a salient preceding situation, so that the listener grasps the whole situation and its pragmatic importance for the current and following discourse. This backgrounding construction subsumes sequences:

[N_{abs} V_{sbj}] [N_{absl} V_{sbj} N_{absl}] [N_{absl} V_{sbj} N_{ann}]	(6)	θamtˤtˤuθ lləqβajəl θəsʕa nnif / (KAB_AM_NARR_03_0565) tamṭṭut n lqbajl təsʕa nnif / tamṭṭut n lqbajl t- sʕa nnif / woman\ABS GEN kabyle_tribe\ANN.PL.M SBJ.3SG.F possess\PFV pride\ABSL.SG.M / N.OV PREP N.COV PRO V13% N.cov / the Kabyle woman had a sense of honor,
	(7)	θaqjuntənni θssəʝləf / (KAB_AM_NARR_01_0753) taqjuntnni tssəglaf / taqjunt -nni t- ssəglaf / dog\ABSL.SG.F CNS SBJ.3SG.F bark\CAUS.IPFV / N.OV DEM PRO V24.APHO / the dog was barking, BI-408 (KAB_AM_NARR_01_0754) BI-408 azdduznni ʃ͡tsawiθuβəħri / (KAB_AM_NARR_01_0755) azdduznni jəṭṭawit ubəħri / azdduz -nni i- ṭṭawi =t ubəħri / big_stick\ABSL.SG.M CNS SBJ.3SG.M bring\IPFV ABS.3SG.M wind\ANN.SG.M / N.OV DEM PRO V14.PFX PRO N.OV / the wind moved the stick,

(iv) **N_{absl}** [V_{sbj} (N) (N)] codes (sub-)topic shift, a shift in perspective with respect to what is introduced in the previous discourse.

(8)

θənnajas ajargaz θurajəssikagi / (KAB_AM_NARR_01_0165)

tənnajas			aj	argaz	tura	jəssikagi			/
t-	nna	=as	a	argaz	tura	jəssi	-k	-agi	/
SBJ.3SG.F	say\PFV	DAT.3SG	VOC	man\ABSL.SG.M	now	daughter\ABSL.PL.F	KIN.2SG.M	PROXb	/
PRO	V13%	PRO	PTCL	N.OV	ADV	N.KIN.COV	PRO	AFFX	/

she said "my husband, now those daughters of yours,

524 (KAB_AM_NARR_01_0166)

524

uzdəɣɣara jiðsənt // (KAB_AM_NARR_01_0167)

ur	zəddɣəɣ		ara	jidsənt		//
ur	zəddɣ	-ɣ	ara	jid	-snt	//
NEG	dwell\IPFV	SBJ..1SG	POSTNEG	COM	PREP.3PL.F	//
PTCL	V23.GEM	PRO	N.INDF	PREP	PRO	//

I'm not living with them,

(v) [V_{sbj} (N) (N)] $\mathbf{N_{ann}}$ codes referent reactivation - because the referent has special importance in the narrative, and often also in order for it to become the topic of the following intonation units.

(9)

tufa ðamʃiʃbuðrar // (KAB_AM_NARR_01_0413)

tufa		d	amʃiʃ	n	wədrar	//
t-	ufa	d	amʃiʃ	n	wədrar	//
SBJ.3SG.F	find\PFV	COP	cat\ABSL.SG.M	GEN	mountain\ANN.SG.M	//
PRO	V13%	PRED	N.OV	PREP	N.OV	//

she found it was the Mountain Cat

423 (KAB_AM_NARR_01_0414)

423

iθizəðɣən / (KAB_AM_NARR_01_0415)

it		izədɣən			/
i	=t	i-	zdəɣ	-n	/
REL.REAL	ABS.3SG.M	RELSBJ.POS	dwall\PFV	RELSBJ.POS	/
DEMPRO	PRO	CIRC1	V23	CIRC2	/

who inhabited it,

wəχχamnni // (KAB_AM_NARR_01_0416)

wəxxamnni		//
wəxxam	-nni	//
house\ANN.SG.M	CNS	//
N.OV	DEM	//

the house.

To give an idea of the frequency of those constructions, the following counts (Table 6) were conducted on two monologues, out of more than a thousand intonation units.

Recording	Construction (i)	Construction (ii)	Construction (iii)	Construction (iv)	Construction (v)
Folktale Narr1	563 (**88,5%**)	39 (**6,2%**)	21 (**3,3%**)	7 (**1%**)	7 (**1%**)
Recount Narr3	451 (**81%**)	52 (**9,2%**)	32 (**6%**)	19 (**3,5%**)	2 (**0,3%**)

Table 6: Frequency of the constructions in two narratives of the corpus (Mettouchi 2015)

The various functions coded by those constructions do not neatly cover the array of functions usually investigated for the study of information structure, when the starting point is the category, not the language. But the procedure chosen here actually tells us a lot about the language, and the way it is organized, based on its coding potential (itself relying on the available coding means). For instance, Siwi (Eastern Berber) doesn't have the state distinction, which results in fewer possible formal combinations. As a result, some combinations (those involving two nominals), because they are constrained (the order of nouns relative to the verb is dedicated to the encoding of grammatical relations) are not available for information structure (Mettouchi & Schiattarella 2018).

For lack of space, other constructions cannot be discussed here, but the interested reader can read Mettouchi (in press), in which it is shown that the construction expressing narrow focus in Kabyle can be automatically retrieved in corpora (provided F0 peaks are annotated), based on a formal definition involving the interaction of morphology, word order, and prosody. Annotation of F0 peaks and open palm hand gestures has recently been added to one of the folktales in CorpAfroAs, and another folktale outside the online corpus, and this has resulted in a preliminary study of cultural gestures and their prosodic correlates (Ferré & Mettouchi 2020). Systematic annotation of F0 peaks will eventually be implemented in the rest of the online Kabyle corpus.

4.1.3. Grammatical relations

As clarified above, grammatical relations are not coded by dedicated morphemes on nouns in Kabyle.

- Grammatical relations are morphologically coded only on the subject pronoun paradigm (other paradigms code semantic roles)

- Nouns (whose function is coded by the interaction of the following formal means: Absence/presence of the noun, position relative to the predicate and the prosodic boundary, state marking, co-reference with a pronoun), overwhelmingly participate in the expression of information structure and referent activation or tracking - except for one specific construction, which is purely grammatical, the direct object (Mettouchi 2018b).

Those findings are based on systematic analyses of corpus data, and here again, the CorpAfroAs layout was crucial in the discoveries made, because its segmentation is prosodic, and its annotation is based on glossing that has been tested for consistency and accuracy through the process of semi-automatic annotation allowed by Elan-CorpA (Chanard 2015): All possible glosses are automatically suggested to the annotator each time she annotates a given morpheme, based on her previous annotation of what the software recognizes as the same sequence of characters. The annotator then chooses the right gloss in this particular instance, or creates a new one. This semi-automatic process

not only ensures consistency, but by automatically proposing all glosses that were entered in the annotation lexicon for the same string of characters, it attracts the annotator's attention on polyfunctionality, grammaticalization effects, as well as on homonymy, thus participating in the analysis of the language.

The example of the direct object shows how systematic analysis of such consistently annotated corpus data allows to propose, and test, a verifiable and falsifiable definition for Kabyle.

In Mettouchi (2018b), after establishing the grammatical or semantic role of pronominal paradigms, it is demonstrated that only one noun can appear within the prosodic group of the verb without being coreferent to a bound pronoun. It is in the absolute state, it follows the verb, either immediately, or separated from it by an adverb, a postverbal negator, and/or a noun in the annexed state. This characterization is considered to be the formal definition of direct objects in Kabyle. In semantic terms, the paper shows that the noun refers to an undergoer macrorole, and can be abstract or concrete, referential or non-referential, effected or affected.

It is possible to automatically retrieve those nouns in the corpus by launching the query: 'Inside the prosodic group of the verb, look for a noun in the absolute state immediately following the verb': QUERY: [rx = \bV & ge=. < mot=.] {rx < 3 & ge < 3 & mot=1} [rx=N & ge=ABSL < mot =.]

The kind of hits retrieved by this query are as shown in Figure 12:

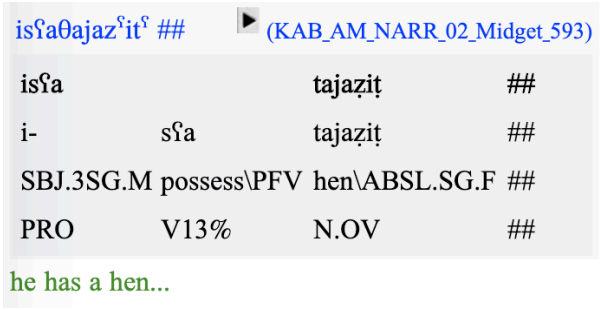


Figure 12: Retrieval of direct object

Manual investigations are necessary in order to check whether all and only the examples that meet the definition are retrieved. This investigation of the data aims at retrieving sequences that have the same function, but differ in form from the one that was initially retrieved, so that ultimately, a construction subsuming those sequences can be defined. This procedure is similar to the one detailed in Section 4.1.2 for information structure.

This might be done using a powerful algorithm that could retrieve all sequences resembling the one under investigation, but differing from it by one formal feature. Not having one at hand, multiple searches were done for all sequences that looked like counter-examples to the initial formal definition, which implied immediate proximity of the noun with respect to the verb. Those sequences were then integrated into the formulation of the retrieval instructions: 'inside the prosodic group of the verb, look for a noun in the absolute state immediately following the verb, or following <the verb followed by a noun in the annexed state (ANN in ge)> or following <the verb followed by an adverb (ADV in rx)> or following <the verb followed by a postverbal negator (POSTNEG in ge)>', which was translated into the query language as in Figure 13:

```
QUERY : [rx = \bV & ge=. < mot=.] {rx < 3 & ge < 3 & mot=1} [ rx=N & ge=ABSL < mot
=.]OR[rx = \bV & ge=. < mot=.] {rx < 3 & ge < 3 & mot=1} [ rx=N & ge=ANN < mot =.] {rx <
3 & ge < 3 & mot=1} [ rx=N & ge=ABSL < mot =.]OR[rx = \bV & ge=. < mot=.] {rx < 3 & ge <
3 & mot=1} [ rx=ADV & ge=. < mot =.] {rx < 3 & ge < 3 & mot=1} [ rx=N & ge=ABSL < mot
=.]OR[rx = \bV & ge=. < mot=.] {rx < 3 & ge < 3 & mot=1} [ rx=N & ge=POSTNEG < mot =.]
{rx < 3 & ge < 3 & mot=1} [ rx=N & ge=ABSL < mot =.]
```

Figure 13: Complex query

Then another series of counterexamples came up, this time involving nouns in the absolute state that were detached from the prosodic group of the verb and appeared in a separate Intonation Unit, but were nevertheless functionally equivalent to the more canonical structures listed above. The inference here is either (i) that the formal characterization "within the prosodic group of the verb" does not hold as a necessary feature of the definition of the Direct Object, or (ii) that there is something in those counterexamples that points to the fact that the noun should normally be inside the prosodic group of the verb. And indeed, there is prosodic evidence that the intervening prosodic boundary separating the verb from what is functionally a direct object should be overridden in its interpretation as clausal boundary, and that those sequences with the direct object appearing in the following intonation unit are departures from the default construction. Evidence for this is given by prosodic indications (hesitations or false starts on the one hand, overarching contour linking together several IUs on the other hand), which were interpreted as signals for the listener to override the boundary as a non-coding mark, and were considered as two linked units. This happens in the following cases:

- either because what caused the presence of the boundary was a cognitive disfluency (finding the right word, remembering the term correctly), as shown in Figure 14 (and the speaker signalled this by a filled pause or syllabic reduplication),

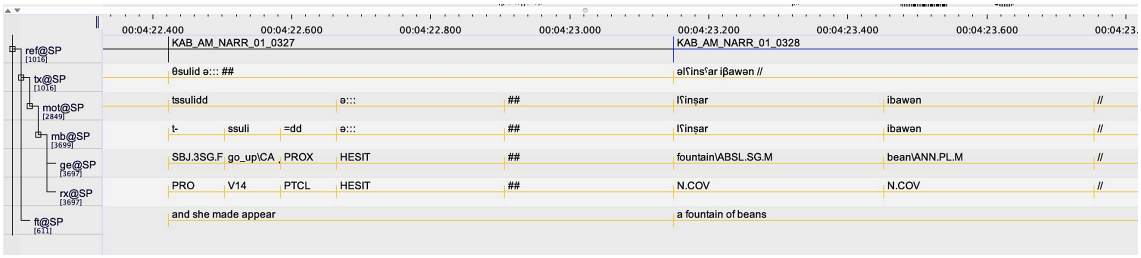


Figure 14: Example of disfluency

- or because the boundary was used as a stylistic device to underline each word of the clause separately, as shown in Figure 15.

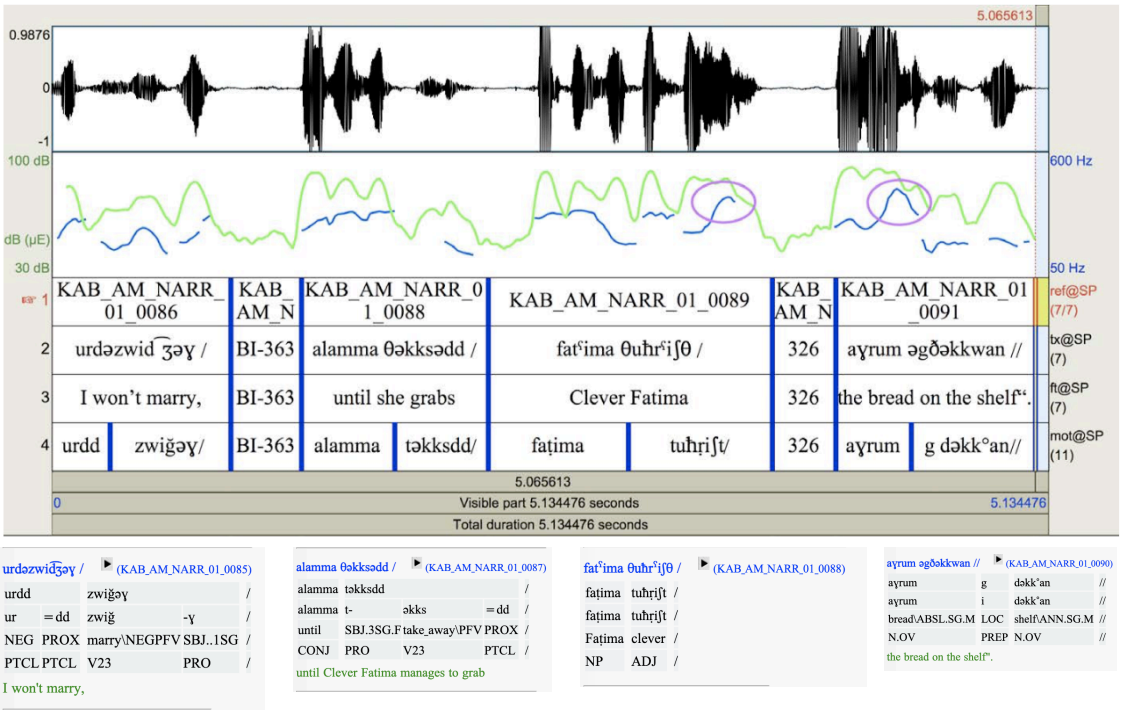


Figure 15: Example of tonal integration of verb and direct object across prosodic boundaries (HYPERLINK TO WAV FILE HERE)

In the first case (Figure 14), supporting evidence for the reinterpretation of the prosodic boundary as an unintended interruption was found in all instances, in the form of disfluencies: Each time the noun interpretable as Direct Object is separated from the verb by a prosodic boundary, a hesitation marker or a false start appears. It signals to the hearer that there should not normally be a boundary here, and instructs her not to interpret the chunking as significant in parsing terms. The sequence is then immediately resumed. The very fact that the separation between verb and direct object is bridged by a vocalized indication means that the prosodic group of the verb and the intonation unit containing the noun in the absolute are integrated, that a tight relationship holds between Verb and Direct Object. Which in turn explains why the typical construction has the Direct Object inside the prosodic group of the verb.

In the case of stylistic devices (Figure 15), such as highlighting (here realized by the use of intonation unit boundary cues where phrasal ones would be expected: "until she$_i$ grabs... Clever Fatima$_i$... the bread on the shelf!"), the interplay of almost identical intensity peaks at the beginning of each IU, and of a marked rising tone with continuative value at the end of each unit inform the addressee that the prosodic group of the verb is not to be processed as complete, but computed as part of a stylized arch-contour with a final fall, which re-creates an encompassing structure integrating together the verb and the direct object.[17]

Ultimately, it is possible to formulate a query that also includes those cases as in Figure 16:

[17] More details on this construction can be found in Mettouchi (2018b).

```
QUERY : [rx = \bV & ge=. < mot=.] {rx < 3 & ge < 3 & mot=1} [ rx=N & ge=ABSL < mot
=.]OR[rx = \bV & ge=. < mot=.] {rx < 3 & ge < 3 & mot=1} [ rx=N & ge=ANN < mot =.] {rx <
3 & ge < 3 & mot=1} [ rx=N & ge=ABSL < mot =.]OR[rx = \bV & ge=. < mot=.] {rx < 3 & ge <
3 & mot=1} [ rx=ADV & ge=. < mot =.] {rx < 3 & ge < 3 & mot=1} [ rx=N & ge=ABSL < mot
=.]OR[rx = \bV & ge=. < mot=.] {rx < 3 & ge < 3 & mot=1} [ rx=N & ge=POSTNEG < mot =.]
{rx < 3 & ge < 3 & mot=1} [ rx=N & ge=ABSL < mot =.]OR[rx = \bV & ge=. < mot=.] {rx < 3 &
ge < 3 & mot=1} [ ge = HESIT | # & rx=. < mot =.] {rx =1 & ge =1 & mot=1} [ rx=/ & ge=. <
mot =.]{rx =1 & ge = 1 & mot=1} [ rx=N & ge=ABSL < mot =.]OR[rx = \bV & ge=. < mot=.]
{rx < 3 & ge < 3 & mot=1} [ ge = FS & rx=. < mot =.]{rx =1 & ge = 1 & mot=1} [ rx=## &
ge=## < mot =##]{rx =1 & ge = 1 & mot=1} [ rx=N & ge=ABSL < mot =.]
```

Figure 16: Complete query for Direct Object

The approach adopted here is theoretically different from most treatments of the role of prosody in relation to grammar. Phenomena such as disfluencies, and stylistic devices that are usually ascribed to 'other levels' of language analysis are not discarded. Rather, prosodic cues are treated as elements of the fabric of language, just like morphological marks, linear ordering, and other formal coding means are. Prosody is not viewed as a separate module, and intonation units are not seen as a projection of other structural levels of grammar, or as a pragmatic unit with a single functional value (speech-act or other). These findings plead for an integrated view of prosody, closely interacting with syntax, semantics, phonology, information structure, and all levels of human communication and cognition, in a way that is best represented as a complex weaving of various threads, rather than a piling up of neatly stacked and hierarchically organized layers.

The consequence, which is not fully responded to yet, is that not only prosodic boundaries should be annotated in the corpus, but also peaks and probably contours. This is a challenge for under-described languages that should nevertheless be addressed.

4.2. Beja

4.2.1. Language profile and corpus

Beja, the sole language of the North-Cushitic branch of AfroAsiatic, is basically an SOV language, whose word order may be modified for pragmatic reasons.

In addition to TAM, verbs index person, number and gender of the subject, hence no overt nominal or pronominal subject or object is compulsory to form a complete utterance.

Beja is a marked nominative language, with three core cases, nominative, accusative and genitive, plus a (non-compulsory) vocative. The first two core cases are marked on determiners by long vowels sensitive to number (NOM.SG *u:*, PL *a:*; ACC.SG *o:*, PL *e:*). The genitive markers are suffixed to the dependent noun (SG *-i*, PL *-e:*). Independent pronouns have in addition dative and ablative sets, and enclitic pronouns an ablative one. Note that nominative and accusative markers (fused with number) surface only if the noun is definite and conforms to one of the two following templates: (i) Mono-syllabic nouns, e.g. NOM.SG. *u:=ka:m,* ACC.SG *o:=ka:m* 'the camel', NOM.PL *a:=kam*, ACC.PL *e:=kam* 'the camels'; (ii) bisyllabic nouns with a first *i* vowel, which is dropped after the article: *riba* 'mountain', *o:=rba* 'the mountain (ACC.SG)' (for details see Vanhove 2017).

The Beja 56:35 mn-long CorpAfroAs corpus (Vanhove 2014) consists of eighteen monologic semi-spontaneous narratives (folktales, anecdotes, religious stories, joke and personal narrative as detailed in Table 7).

	duration	Morphemes	words	IU + pausal Units	Pausal Units
Folktale 1	02:42	589	256	106	69
Folktale 2	05:28	1111	544	211	118
Folktale 3	02:04	540	252	112	77
Folktale 4	01:04	225	111	45	25
Folktale 5	02:22	652	289	109	52
Folktale 6	04:51	1114	545	223	117
Folktale 7	04:27	1049	489	202	97
Anecdote 1	03:49	823	392	160	96
Anecdote 2	01:10	232	112	51	30
Anecdote 3	01:16	280	136	50	30
Anecdote 4	01:05	225	117	48	25
Anecdote 5	01:51	437	197	76	43
Joke 1	00:33	128	58	21	11
Personal narrative 1	06:47	1560	709	277	152
Religious story 1	03:25	746	358	143	86
Religious story 2	02:06	419	190	82	54
Religious story 3	05:31	1131	552	232	149
Religious story 4	05:26	1246	583	234	132
Total	56:35	12507	5890	2382	1363

Table 7: Size of the Beja corpus (Corpafroas)

The 98mn-long CorTypo corpus (Vanhove 2017) consists of more genres: 1 conversation, 1 language play, 1 personal narrative, 1 religious story, 2 jokes, 2 pear stories, 3 procedural texts, 4 anecdotes, and 25 folktales. It contains 20,692 morphemes, 9969 words, 4152 intonation units, and 2421 pauses.
The number of words in the two corpora amounts to 15,859 and the number of morphemes to 33,199.[18]

4.2.2. Information Structure
Beja, as opposed to many Cushitic languages, is rather poor in morphosyntactic devices and functional particles for the expression of topics. There are no dedicated topic markers, but when a demonstrative precedes instead of following the noun it modifies, this sequence serves this purpose. Moreover, the obligatory marking of case means that Beja belongs to the type of languages where topics are fully integrated into the utterance (Maslova & Bernini 2006), a feature which does not make it easy to define form-function relationship at the relevant level of analysis, i.e at the pragmatic and/or the grammatical levels.

The research hypothesis was thus to explore if it was possible to disentangle the syntactic roles of core arguments and the pragmatic roles of NPs, i.e. for instance to

[18] They are grouped together in the CorpOrAn website (Vanhove 2020)

distinguish subject and topic, on the basis of word order and prosody, i.e. a corpus-based approach in the sense defined in Section 4.2.1.
What follows briefly sums up the question concerning the distinction between three categories of information structure, topics, antitopics and afterthoughts, and the grammatical function of subject.

First, a count of utterances containing a noun with an overt nominative marker was made in the corpora. 94% of the 605 utterances of this type correspond to the canonical linear order S(O)V. Among those, a majority of NPs_{NOM} occurs in the same intonation unit (IU) as their verbal predicate, as shown in Table 8.

	NP_{NOM} +V in same IU	NP_{NOM} +V in ≠ IUs	V+NP_{NOM}	Total
# NP_{NOM}	360	210	35	605
% of NP_{ACC}	59.5%	34.7%	5.8%	100%

Table 8: Intonation units and word order of NP_{NOM} + V

These counts were automatically retrieved by several queries on the rx and ge tiers about sequences of word classes in order to capture all possible combinations, including when pauses, adverbs, accusative NPs, adjectives, hesitations and false starts occur between NP and V: e.g. (i) an article in the nominative case, a noun and a verb; (ii) an article in the nominative case, a noun, a prosodic break (marked by / or // in the annotation system), a pause (whose duration is marked in milliseconds) and a verb; (iii) an article in the nominative case, a noun, an article in the accusative case, a noun, and a verb; etc. (see Figure 17).

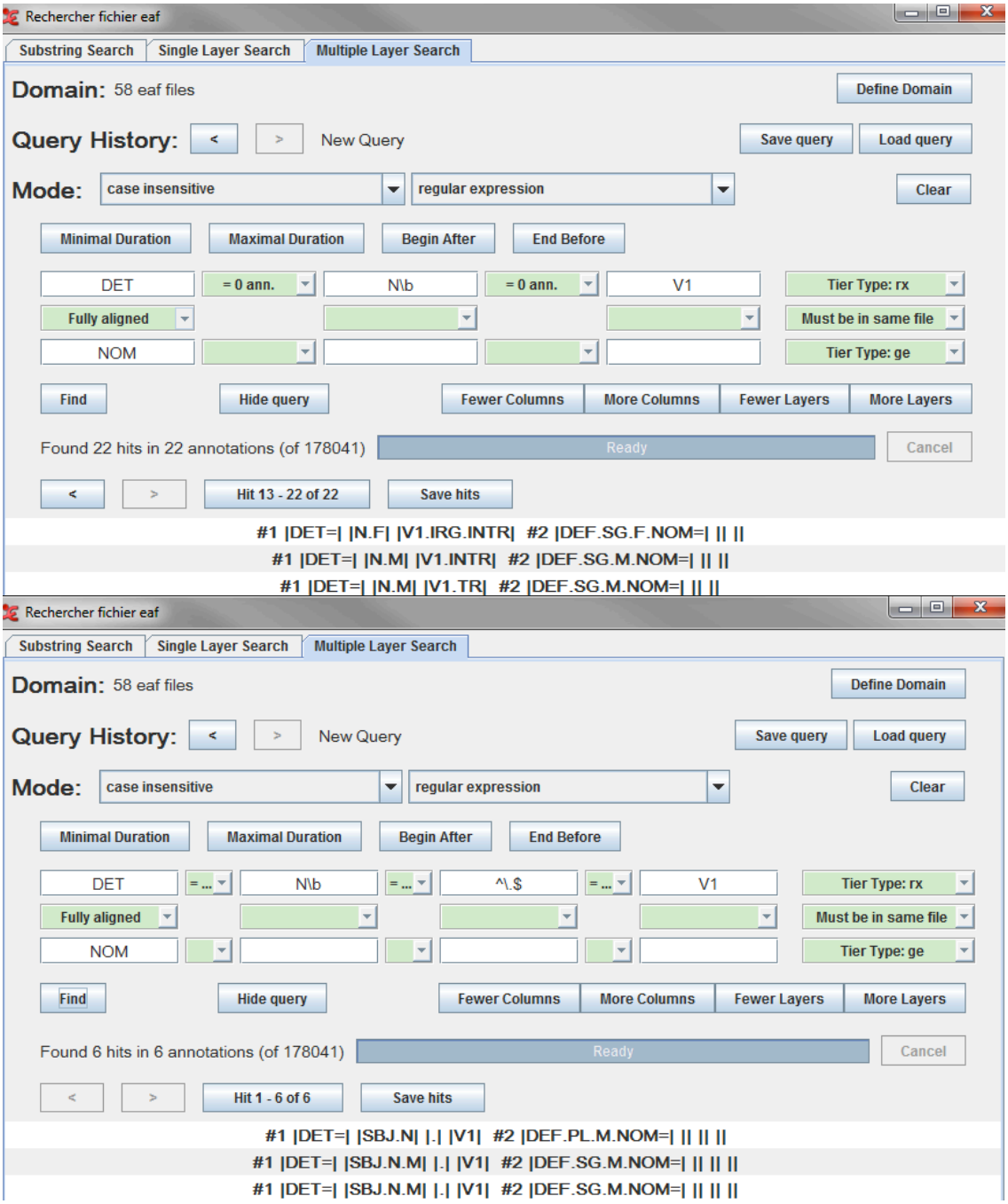


Figure 17: Examples of queries (i) and (iii) to retrieve the sequences of Table 8 in Beja

Due to the annotation system adopted for the corpora only the overtly marked definite NPs and the pronouns could be automatically retrieved. However, there were enough tokens to show clear tendencies concerning the interface between prosody and functions, which will be summarised below. In addition, since for long distances between NP and V, the queries could not be fine-grained enough, even with regular expressions, manual checking had to be done in order to exclude irrelevant sequences. The rest of the analysis had also to be done manually.

Contrary to Kabyle (Mettouchi 2018a, and above, Section 4.1.2), or to Modern Hebrew and Anal Naga (Ozerov this volume) it was not driven primarily by the forms

and constructions, but by a theoretical starting point, namely the definition of topic proposed by Gundel (1985) and her 'Topic-Familiarity Condition':

> givenness in the relational sense is correlated with topichood by definition. Second, givenness in one referential sense, that of assumed familiarity, appears to be a necessary precondition for felicitous topichood if the topic is to fulfill its function of relating a sentence to the discourse context in which it is used. (Gundel 1988: 213)

This stance was favoured by the fact that the corpora are semi-spontaneous and not interactional (with one marginal exception), thus not easily liable to interpretations such as those proposed by Ozerov (this volume) within the Interactional Linguistics framework. Nevertheless it does not rule out further studies from a bottom-up perspective.

In the absence of dedicated particles or constructions, and since the canonical word order is SOV, word order and prosodic contours are the only cues left for Beja (as in other languages of this type, see e.g. Hedberg and Sosa 2008 for English topics).

The prosodic analysis was done with PRAAT. It turned out that falling contours are typical of a large majority (295/360, i.e. 82%) of the NPs_{NOM} occurring in the same IU as V in the canonical word order,[19] while a high or rising contour (149/210, i.e. 71%) was typical when NPs_{NOM} are separated by a prosodic boundary (Fig. 12 and 13). The contour difference, together with the presence or absence of a prosodic boundary, were considered as indicative of the syntactic vs. pragmatic functions of NPs_{NOM} in initial position. To supplement the prosodic analysis, a semantic and pragmatic analysis of discourse sequences was also conducted. The results show that NPs_{NOM} systematically surface when a new referent at the beginning of a sequence is introduced, and that they are never separated from the verbs by a prosodic boundary (except when hesitations and false starts occur). These cases of NPs_{NOM} were analysed as syntactic subjects, introducing a new referent. Conversely, NPs_{NOM} followed by a prosodic break, be it minor, major, with or without a pause, turned out to be either instances of contrastive topic (ex.10, Fig. 18) or of selective topic (ex. 11, Fig. 19), pragmatic functions.

NP_{NOM} = contrastive topic
(Previous context: when it becomes a man, he says: "Oh, you, who are you?")

(10) *u:n* *ani* / *261* *ʤa:n-ta:ji=b=i*
PROX.SG.M.NOM 1SG.NOM . djinn-SING=INDF.M.ACC=COP.1SG
'As for me, I am a djinn.' (BEJ_MV_NARR_02_farmer_220-222)

[19] Most of the exceptions are explained by hesitations and difficulties in speech processing. Still, there are some cases where this explanation does not hold, and for which further research is needed.

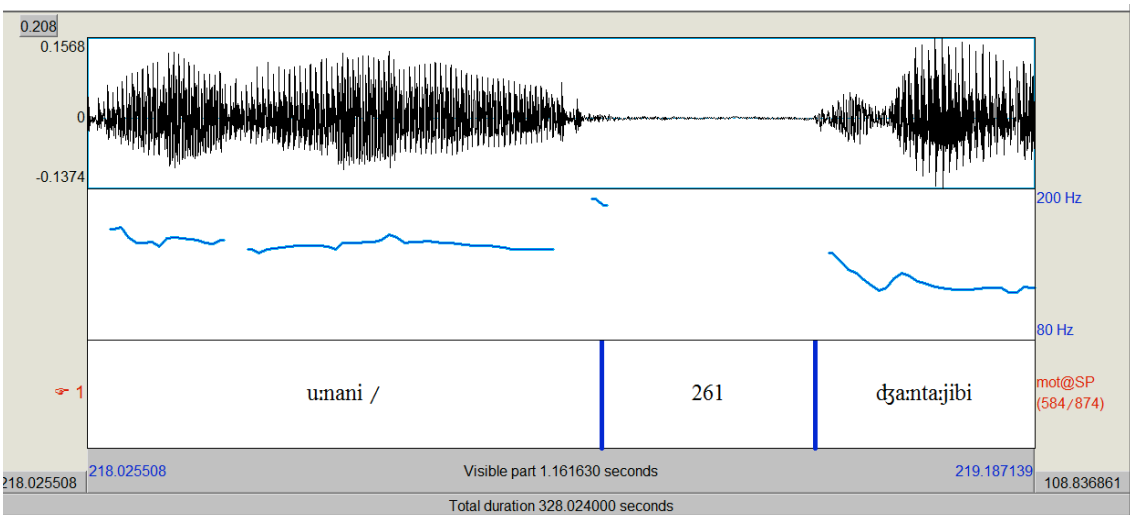


Figure 18: NP_{NOM} = contrastive topic

NP_{NOM} = selective topic

(11) *me:k-i=t* *misu:s* */ 312* *ti=ʃabaka* *ʔabik-a:=t*
donkey-GEN=INDF.F die\N.AC . DEF.F=net take-CVB.MNR=INDF.F

ha:j *ti-t-farʔi* *e:n* *//*
COM 3SG.F-MID-go_out\IPFV say\PFV.3PL .
'The corpse of a donkey is taken out by taking it with the net, they said.'
(BEJ_MV_NARR_02_farmer_179-181)

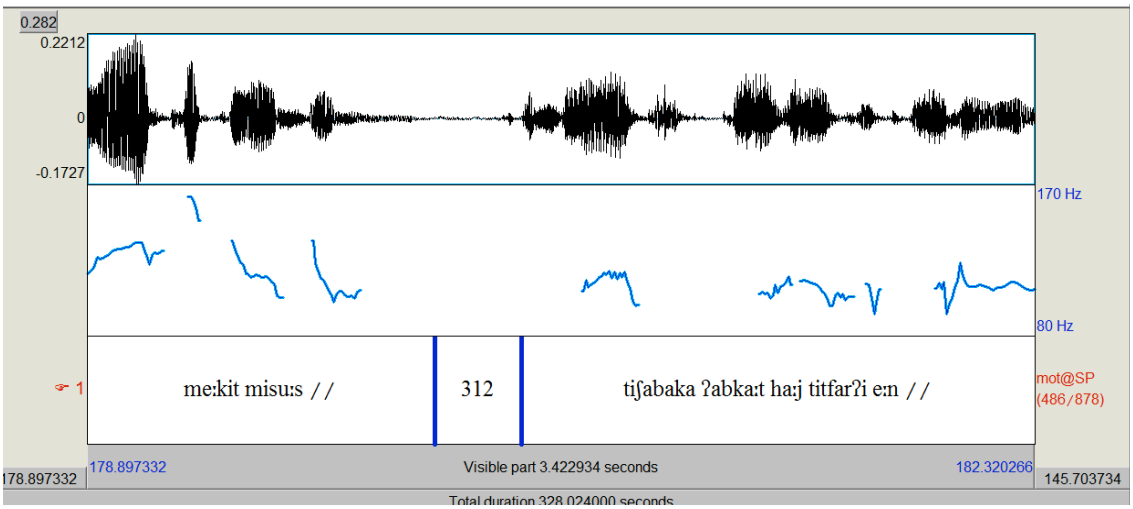


Figure 19: NP_{NOM} = selective topic

Regarding antitopics and afterthoughts, which both occur post-verbally, the same kind of manual analyses had to be done. The analysis of antitopics was based on Chafe's (1994: 176) definition of a referential antitopic, as a pragmatic category that functions to "confirm established information". For afterthoughts, it was Cresti and Moneglia's (2005) approach that was followed. They regard them as extrasentential or rhematic, and as non-activated referents.

In Beja, the distinction between the two categories is prosodically marked: Antitopics have a mid or low and rather flat contour, while afterthoughts have a mid or

high bell contour. It may be the case that antitopics tend to occur within the same IU as the verb, while afterthoughts usually occur after a prosodic break, but more examples from a larger corpus are needed to confirm this hypothesis.

The same kind of analyses was conducted for NP_{ACC} and gave similar results where prosody and word order play a key-role in disentangling pragmatic roles from syntactic ones. Details are found in Vanhove (submitted).

4.3. Reported speech in Beja, Zaar, Juba Arabic and Modern Hebrew

The study of the relationship between grammatical relations and prosody is illustrated here at the clause level. It concerns reported discourse, and was conducted for four languages of the CorpAfroAs database: Beja, Zaar (Chadic), Juba Arabic (Creole) and Modern Hebrew (Semitic) (Malibert & Vanhove 2015). The analysis was limited to speech reports marked by quotative frames which contain the most basic speech verb of each language, i.e. 'say' verbs, excluding verbs such as 'ask', 'demand', 'shout', and excluding reported speech without a quotative verb, impossible to retrieve automatically, as well as those containing just a complementizer.

The investigation was focused on the relationship between prosodic contours and reported speech. The descriptive tools and analysis of Genetti (2011) for direct speech report in Dolakha Newar (Tibeto-Burman) were used as a starting point and adapted to the annotation system of CorpAfroAs. If the examples were all automatically extracted from the CorpAfroAs corpus, the prosodic analysis in relation to the quotative frames of the reported speech and their right and left contexts, i.e. their prosodic integration cline, had to be done in PRAAT, since prosody, apart from minor and major prosodic breaks and pauses is not annotated. The analysis was extended to indirect reported speech in the three languages where it exists (Zaar, Juba Arabic, Modern Hebrew), since contradictory claims had been made concerning their prosodic features in the literature. The results for Beja are briefly summarized below.

The first observation was that the direct speech reports are rarely set off from the quotative verb by intonation-unit boundaries. Their prosodic integration within the same IU as the quotative verb represents the vast majority (almost 90% of the 317 examples). The quotative verb may belong (i) to the same IU as the whole quoted speech (90 examples); (ii) if the reported speech is split into several IUs, to the last IU of the quoted speech (175 examples); (iii) to an internal IU (12 examples). Quite often the quotative verb cliticizes to the speech report, is uttered in a very rapid tempo and uttered in such a low pitch that it does not show on the pitch trace provided by PRAAT. The 3SG.M Perfective *ini* 'he said' may even be phonetically reduced to a single vowel, often devoiced.

Example (12) is typical of the first category of prosodic integration. It is set off from the previous and next IUs by medium-length pauses, and includes the quotative verb in the same prosodic unit (Figure 20):

(12) *a:ladʒ-an=ho:b* *u:=jha:m* *d=he:* *far-ija*
tease-PFV.1SG=when DEF.SG.M.NOM=leopard DIR=1SG.ACC jump-PFV.3SG.M

ini //
say\PFV.3SG.M
'When I teased it, the leopard jumped on me, he said.'
(BEJ_MV_NARR_15_leopard_051)

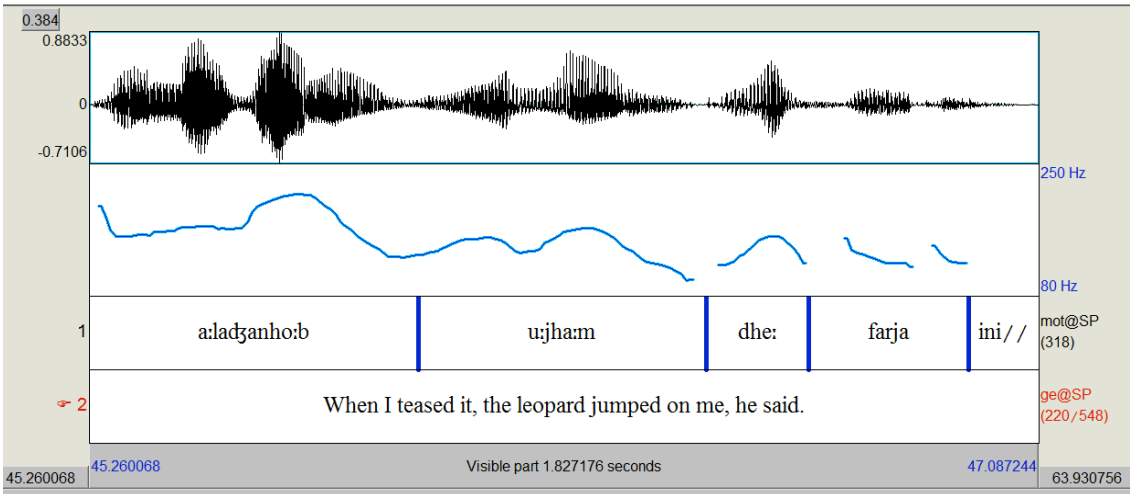


Figure 20: Prosodic integration of quotative verb in Beja (Cushitic)

In only 40 examples does the direct speech report occur in a different IU, set off from the quotative verb by a pause. This happens in three linguistics contexts: When the quoted speech consists of an exclamative utterance, when it contains an Imperative verb form or an onomatopoeia.

As for the onset of the speech report, it is most often set off from the previous context (98% of the 308 examples), indicating that the prosodic break is a marker of the onset of a quotation.

The prosodic integration of overt subjects and recipient addressees of the quotative verb was also investigated but the number of tokens was too low to draw any conclusion.

As for the other languages of the sample, direct and indirect speech reports do not seem to differ greatly in terms of prosody. Nevertheless the prosodic integration cline of speech reports varies from language to language, according to different criteria. This led to preliminary hypotheses for a cross-linguistic study of the interface between prosody and speech report. They still need to be further tested empirically on other languages and on larger corpora. The hypotheses concern the interface between morphosyntax and intonation units:

1. In the languages without a complementizer, the prosodic integration of speech reports within the same IU as the quotative frame tends to be very high. (It concerns the end of speech reports in SOV languages, and their onsets in SVO language; there is no VSO language in the sample). Consequently, in SOV languages the onset of the speech report is systematically set off from the previous intonation unit. This is a clear prosodic cue, marking the beginning of the speech report. In SVO languages it is the end of the speech report which is set off from the next IU. This may be a first step towards a grammaticalization of the quotative verb into a complementizer (see Güldemann 2008).
2. Conversely, in the languages with a complementizer, whatever their word orders, speech reports tend to be less integrated within the quotative frame.
3. Non-clitic complementizers tend to be prosodically integrated within the quotative frame, not within the speech report.

5. Discussion

The four case studies presented in Section 4 demonstrate that corpora annotated with an entirely similar template do not hamper the possibility of a large range of methodological and theoretical approaches of the data, be they corpus-driven or corpus-based. For Kabyle the starting point of the analyses are the coding means (including prosody) of the language, clearly defined and contrasted within one linguistic system. It allowed not only precise descriptions of several phenomena, but also comparative investigations, even if not as on similar scale as when working with predefined comparative concepts (in the sense of Haspelmath 2010) as is often the case in typological studies. Without excluding also a bottom-up approach and language-internal definitions, the research questions for Beja were driven by the absence of certain morphosyntactic coding means to question the role of prosody in the identification of predefined comparative concepts and syntactico-semantic constructions.

Despite those differences, both types of approaches plead for an integrated view of prosody, closely interacting with syntax, semantics, phonology, information structure, and all levels of human communication and cognition. They also plead for a general endeavour to annotate as much as possible the large array of prosodic cues that are inseparable from speech processing and interaction dynamics. Without a notation of prosodic boundaries based on acoustic and perceptual cues rather than on syntactic or pragmatic or semantic assumptions, and without precise transcription of hesitations, false starts and pauses, it would not have been possible to conduct the investigations presented in this paper.

Comparison with approaches presented in this volume show similarities with Ozerov's perspective, which promotes expansion of the data sample of each study to include phenomena that only partially resemble the originally defined concept, rather than assuming a restrictive definition for the studied concept and selecting examples that fit it. Indeed, for both Vanhove and Mettouchi, back and forth movement between data and hypotheses is essential, and allow refinement of the initial claim (this is particularly clear in Mettouchi's analysis of the Direct Object construction in Kabyle, with the original 2018 paper retracing all stages of the demonstration step by step).

One difference is that Mettouchi considers prosodic forms as basic coding means at the same level as morphosyntactic ones, whereas Ozerov considers prosody as a set of pragmatic or interactional cues superimposed onto a syntactic structure (called "basic structure"). Consequently, Mettouchi also decomposes the syntactic sequence into more fine-grained cues (word-order, state marking, etc.), and weaves them very tightly with prosodic cues (prosodic boundary, contour, dysfluencies, peaks), which are not considered as more interactional or pragmatic than the sequence itself. For Mettouchi, all cues are forms, and it is the written bias which reinforces the separation into layers between morphosyntax and prosody. In this sense, and in the use of combinations of forms, Mettouchi's approach appears more radical in its assumptions about the nature of constructions than Ozerov's. It also allows automatic identification of sequences in the corpus (e.g. "noun in the annexed state directly following the verb"with "noun", "verb" being annotated in the corpus), whereas Ozerov's starting point is the manual identification of syntactic "interrogatives" or "left-dislocations". But those differences are secondary, and there is clear convergence on the importance of "questioning the meta-language used in the study design" (Ozerov), and doubts of the kind "Are we sure we are studying the language itself and not the translation? And are we sure that those categories are indeed universal?" (Mettouchi).

We also share with Haig and Schnell (this volume) a focus on fine-grained annotation of the corpus, and engagement with descriptive and analytical issues relevant to each individual language sampled, although the GRAID and RefIND systems use comparative categories and typological definitions. As indicated by the authors, the combination of GRAID/RefIND with other tiers of annotation of a more language-internal nature allows the conduct of complex queries that capture the expected variation within the multilingual corpora. Another important similarity between CorpAfroAs/CorTypo and MULTICAST is the fact that each corpus is annotated by a specialist of that language.

References

Barth-Weingarten, Dagmar. 2011. The fuzziness of intonation units: Some theoretical considerations and a practical solution. *LiSt - Interaction and Linguistic Structures* 51. 1-22, May 2011. http://www.inlist.uni-bayreuth.de/issues/51/InLiSt51.pdf. Accessed on 20/05/2021.

Chanard, Christian. 2015. ELAN-CorpA: Lexicon-aided annotation in ELAN. In Amina Mettouchi, Martine Vanhove & Dominique Caubet (eds.), *Corpus-based Studies of lesser-described Languages: The CorpAfroAs Corpus of spoken AfroAsiatic*, 311-332. Amsterdam, Philadelphia: Benjamins (Studies in Corpus Linguistics 68).

CorpAfroAs: *The CorpAfroAs Corpus of Spoken AfroAsiatic Languages*. http://dx.doi.org/10.1075/scl.68.website. Accessed on 12/04/2020.

Corporan: The LLACAN fully-annotated and searchable spoken corpus in lesser-described languages, https://corporan.huma-num.fr/Archives/corpus.php. Accessed on 11/14/2020.

CorTypo: *Designing Spoken Corpora for Cross-Linguistic Research*, https://cortypo.humanum.fr/. Accessed on 12/04/2020.

Chafe, Wallace. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: The University of Chicago Press.

Cresti, Emanuela & Moneglia, Massimo (eds). 2005. *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*, DVD + Vol. Amsterdam: Benjamins.

Cruttenden, Alan. 1997. *Intonation*. Second edition. (Cambridge Textbooks in Linguistics.) Cambridge: Cambridge University Press.

Cysouw, Michael & Bernhard Wälchli. 2007. Parallel texts: Using translational equivalents in linguistic typology. *STUF - Sprachtypologie und Universalienforschung* 60(2). 95–99. DOI: 10.1524/ stuf.2007.60.2.95.

Ferré, Gaëlle & Amina Mettouchi. 2020. A Cultural Study of Open-Palm Hand Gestures and their Prosodic Correlates. In *Proceedings of 10th International Conference on Speech Prosody 2020*: 25-28 May 2020, paper 21 (Online), Tokyo, Japan. 285-289. https://www.isca-speech.org/archive/SpeechProsody_2020/pdfs/21.pdf. Accessed on 12/04/2020.

Frajzyngier, Zygmunt. 1999. Domains of point of view and coreferentiality: System interaction approach to the study of reflexives. In Zygmunt Frajzyngier & Traci Curl (eds.), *Reflexives: Forms and Functions*, 125–152. Amsterdam, Philadelphia: Benjamins.

Frajzyngier, Zygmunt & Erin Shay. 2003. *Explaining Language Structure through Systems Interaction*. (Typological Studies in Language, 55). Amsterdam, Philadelphia: Benjamins.
Frajzyngier, Zygmunt. 2004. Principle of functional transparency in language structure and language change. In Zygmunt Frajzyngier, Adam Hodges & David S. Rood (eds.), *Linguistic diversity and language theories*, 259–283. Amsterdam, Philadelphia: Benjamins.
Frajzyngier, Zygmunt & Amina Mettouchi. 2015. Functional domains and cross-linguistic comparability. In Amina Mettouchi, Martine Vanhove & Dominique Caubet (eds.), *Corpus-based Studies of lesser-described Languages: The CorpAfroAs Corpus of spoken AfroAsiatic*, 257-279. Amsterdam, Philadelphia: John Benjamins (Studies in Corpus Linguistics 68).
Genetti, Carol. 2011. Direct speech reports and the cline of prosodic integration in Dolakha Newar. *Himalayan Linguistics. Special issue in memory of Michael Noonan and David Watters* 10(1). 55-71.
Güldemann, Tom. 2008. *Quotative Indexes in African Languages: A Synchronic and Diachronic Survey*. Berlin, New York: Mouton De Gruyter.
Gundel, Jeanette K. 1985. Shared knowledge and topicality. *Journal of Pragmatics* 9. 83–107.
Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663-687.
Hedberg, Nancy & Juan M. Sosa. 2008. The prosody of topic and focus in spontaneous English dialogue. In Chungmin Lee, Matthew Gordon & Daniel Buring (eds.), *Topic and focus: Cross-linguistic perspectives on meaning and intonation*, 101-120. Dordrecht: Springer.
Izre'el, Shlomo & Amina Mettouchi. 2015. Representation of speech in CorpAfroAs: Transcriptional strategies and prosodic units. In Amina Mettouchi, Martine Vanhove & Dominique Caubet (eds.), *Corpus-based Studies of lesser-described Languages: The CorpAfroAs Corpus of spoken AfroAsiatic*, 13-41. Amsterdam, Philadelphia: Benjamins (Studies in Corpus Linguistics 68).
Lazard, Gilbert. 2006. *La quête des invariants interlangues. La linguistique est-elle une sciences?* Paris: Champion.
Malibert, Il-Il & Martine Vanhove. 2015. Quotative constructions and prosody in some Afroasiatic languages: Towards a typology. In Amina Mettouchi, Martine Vanhove & Dominique Caubet (eds.), *Corpus-based Studies of lesser-described Languages: The CorpAfroAs Corpus of spoken AfroAsiatic*, 117-169. Amsterdam, Philadelphia: Benjamins (Studies in Corpus Linguistics 68).
Maslova, Elena & Bernini, Giuliano. 2006. Sentence topics in the languages of Europe and beyond. In Giuliano Bernini & Marcia L. Schwartz (eds.), *Pragmatic Organization of Discourse in the Languages of Europe: EUROTYP*, 67–120. Berlin: Mouton de Gruyter.
Mettouchi, Amina. 2011. Linear orders and prosodic groups in Kabyle (Berber). Paper presented at the *14th International Conference of AfroAsiatic linguistics*. Turin (Italy) 15-18 June 2011.
Mettouchi, Amina. 2012. Kabyle Corpus. In *The CorpAfroAs Corpus of Spoken AfroAsiatic Languages*, Amina Mettouchi & Christian Chanard (eds.). http://dx.doi.org/10.1075/scl.68.website. Accessed on 12/04/2020.

Mettouchi, Amina. 2013. Segmenting spoken corpora in lesser-described languages: New perspectives for the structural analysis of speech", Plenary talk at the *46th Annual Meeting of the Societas Linguistica Europaea, Split (Croatia) 18-21 September 2013*. http://llacan.vjf.cnrs.fr/pers/mettouchi/pub/SLE-2013-Split-Mettouchi-Plenary-Part11.pdf & http://llacan.vjf.cnrs.fr/pers/mettouchi/pub/SLE-2013-Split-Mettouchi-Plenary-Part21.pdf. Accessed on 12/04/2020.

Mettouchi, Amina. 2014. Foundations for a typology of the annexed/absolute state systems in Berber. *STUF - Sprachtypologie und Universalienforschung* 67(1). 47-61.

Mettouchi, Amina. 2015. Aspect-Mood and discourse in Kabyle (Berber) spoken narratives. In *Beyond Aspect: The expression of discourse functions in African languages,* Doris Payne & Shahar Shirtz (eds). Amsterdam, Philadelphia: John Benjamins, pp. 117-144.

Mettouchi, Amina. 2017. Predication in Kabyle (Berber). In Amina Mettouchi, Zygmunt Frajzyngier & Christian Chanard (eds), *Corpus-based cross-linguistic studies on Predication (CorTypo*), https://cortypo.huma-num.fr/Archives/publications/CorTypo%20-%20Kabyle%20(Berber)_PRED.pdf. Accessed on 12/04/2020.

Mettouchi, Amina. 2018a. The Interaction of state, prosody and linear order in Kabyle (Berber): Grammatical relations and information structure. In Mauro Tosco (ed), *Afroasiatic: Data and Perspectives*, CILT, 261–285. Amsterdam, Philadelphia: Benjamins.

Mettouchi, Amina. 2018b. Prosodic segmentation and grammatical relations: The direct object in Kabyle (Berber). *Revista de Estudos da Linguagem*, [S.l.], July 2018. ISSN 2237-2083. http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/13049. Accessed on 12/04/2020.

Mettouchi, Amina. in press. From a corpus-based to a corpus-driven definition of clefts in Kabyle (Berber): Morphosyntax and Prosody. In Enrique Palancar & Martine Vanhove (eds.), *Clefts and other related focus constructions, special issue of Faits de Langues* 52(1) - 2021.

Mettouchi, Amina, Martine Vanhove & Dominique Caubet (eds.). 2015. *Corpus-based Studies of lesser-described Languages: The CorpAfroAs Corpus of spoken AfroAsiatic*. Amsterdam, Philadelphia: Benjamins (Studies in Corpus Linguistics 68).

Mettouchi, Amina, Graziano Savà & Mauro Tosco. 2015. Cross-linguistic comparability in CorpAfroAs. In Amina Mettouchi, Martine Vanhove & Dominique Caubet (eds.), *Corpus-based Studies of lesser-described Languages: The CorpAfroAs Corpus of spoken AfroAsiatic*, 221-256. Amsterdam, Philadelphia: Benjamins (Studies in Corpus Linguistics 68).

Mettouchi, Amina & Valentina Schiattarella. 2018. The influence of the State distinction on word order and information structure in Kabyle and Siwi (Berber). In Evangelia Adamou, Katharina Haude & Martine Vanhove (eds), *Information Structure in lesser-described languages: Studies in prosody and Syntax*, 265-296. Cambridge: Cambridge University Press.

Vanhove, Martine. 2014. The Beja Corpus. In Amina Mettouchi & Christian Chanard (eds.), *The CorpAfroAs Corpus of Spoken AfroAsiatic Languages* http://dx.doi.org/10.1075/scl.68.website. Accessed on 11/14/2020.

Vanhove Martine. 2017. *Version démo du Corpus Bedja, projet CorTypo*. http://cortypo.huma-num.fr/Archives/pred5.php. Accessed on 11/14/2020.

Vanhove Martine. 2020. *Corpus de bedja*. https://corporan.huma-num.fr/Archives/corpus.php.
Vanhove, Martine. submitted. Information Structuring in Beja (North-Cushitic). In *Proceedings of the 47th Annual Meeting of the North Atlantic Conference on Afroasiatic Linguistics (NACAL 47)* (Provisional Title). Villejuif: Publications LACITO.