Query languages and linguistics categories: annotating and searching spoken electronic corpora in indigenous languages

Amina Mettouchi (EPHE, PSL & CNRS LLACAn)
Christian Chanard (CNRS LLACAn)

Computer science and Linguistics have had regular interactions ever since the exponential development of information technology. Their relationship can be one of application (of IT methods to linguistic data), but more fundamentally, they are related together by the fact that at the core of programming, there is language. This study will explore the topic of the conference under that angle: how can query languages used in computer queries (Fig.1) help formulate, test and falsify linguistic claims about morphosyntactic constructions?

For such questions to be investigated, it is essential to work on annotated electronic corpora. Their contribution to linguistics is in the systematicity of data treatment: for instance, viewed from a computer programming perspective, a "word" is a string of characters surrounded with a blank space at each end, and therefore, for a linguist transcribing recordings into a computer programme such as ELAN, the potential for accurate statistics of word counts relies from the start on decisions concerning word boundaries, which is one of the first steps in the analysis of an undescribed language. And every step of the transcription and annotation process of spoken data is a potential locus for fertile interaction of programming and linguistic analysis.

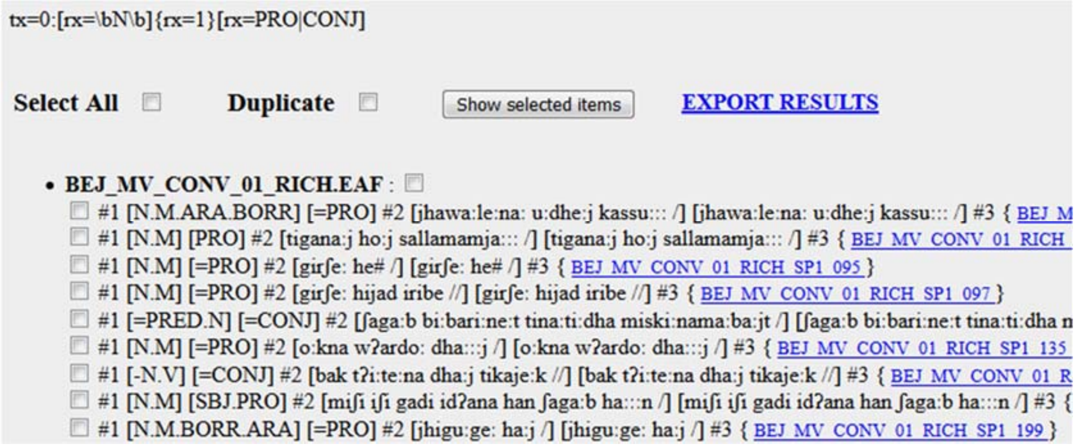
This will be demonstrated through the investigation of electronic spoken corpora from several indigenous languages (CorpAfroAs http://corpafroas.huma-num.fr & CorTypo http://cortypo.huma-num.fr), whose annotation template (Fig.2) allows the formulation of complex queries to answer such questions as e.g. "How often do nominal subjects co-occur with nominal objects in the same intonation unit?".

What is crucial however in this perspective, is for annotation to be not only systematic, but also transparent, and as non-aprioristic as possible. Indeed, a query will only retrieve results on the basis of annotation, so that defining what a nominal subject/object and a prosodic boundary are language-internally, is absolutely crucial for the results of the query to be valid and non-trivial. The talk will explore the way the annotation/query process itself, as an instance of the scientific method (Fig.3), participates in the analysis, yielding explicit, testable and falsifiable definitions, such as the following for Kabyle (Berber) nominal direct object "a noun in the absolute state, immediately following the verb, or following <the verb followed by a noun in the annexed state> or following <the verb followed by an adverb> or following <the verb followed by a postverbal negator>, within the prosodic group of the verb", a definition which in turn needs transparent definition and consistent annotation of such annotations/categories as Verb, Noun, Absolute/Annexed state, Adverb, and Postverbal Negator.

This incremental procedure allows the testing and refining of hypotheses in view of the grammatical analysis of the language, while also questioning the process of categorization itself, and ultimately, the nature of linguistic data, and the specificity of linguistics as a science, namely the fact that its metalinguistic toolkit is linguistic in nature. Beyond the material discussed here, this reflexive nature of linguistics makes it relevant to many fields in the humanities, such as philosophy, human cognition, and the anthropology and sociology of science.

(Figures on page 2)

Figure 1: Query language 'search N followed by PRO or CONJ in the same unit'
[rx=\bN\b < tx=.] {rx=1 & tx=0} [rx=PRO|CONJ <tx=.] or
tx=0:[rx=\bN\b]{rx=1} [rx=PRO|CONJ]

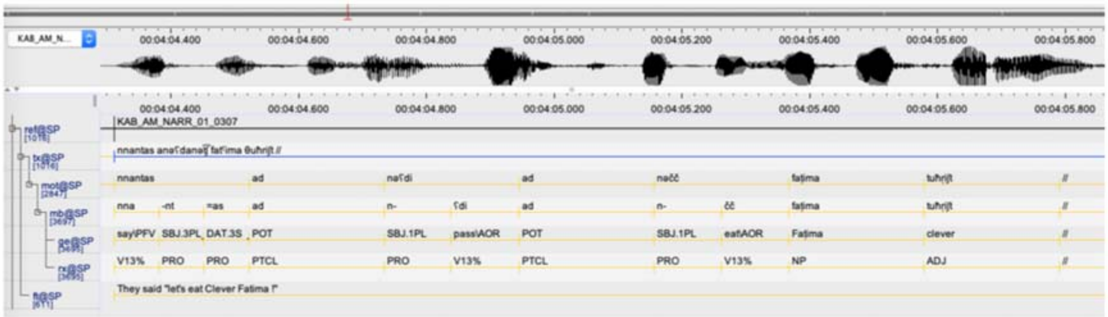


tx=0:[rx=\bN\b]{rx=1}[rx=PRO|CONJ]

Select All □ **Duplicate** □ Show selected items EXPORT RESULTS

- **BEJ_MV_CONV_01_RICH.EAF** : □
  - □ #1 [N.M.ARA.BORR] [=PRO] #2 [jhawa:le:na: u:dhe:j kassu::: /] [jhawa:le:na: u:dhe:j kassu::: /] #3 { BEJ_M
  - □ #1 [N.M] [PRO] #2 [tigana:j ho:j sallamamja::: /] [tigana:j ho:j sallamamja::: /] #3 { BEJ_MV_CONV_01_RICH_
  - □ #1 [N.M] [=PRO] #2 [girʃe: he# /] [girʃe: he# /] #3 { BEJ_MV_CONV_01_RICH_SP1_095 }
  - □ #1 [N.M] [=PRO] #2 [girʃe: hijad iribe //] [girʃe: hijad iribe //] #3 { BEJ_MV_CONV_01_RICH_SP1_097 }
  - □ #1 [=PRED.N] [=CONJ] #2 [ʃaga:b bi:bari:ne:t tina:ti:dha miski:nama:ba:jt /] [ʃaga:b bi:bari:ne:t tina:ti:dha n
  - □ #1 [N.M] [=PRO] #2 [o:kna wʔardo: dha:::j /] [o:kna wʔardo: dha:::j /] #3 { BEJ_MV_CONV_01_RICH_SP1_135
  - □ #1 [-N.V] [=CONJ] #2 [bak tʔi:te:na dha:j tikaje:k //] [bak tʔi:te:na dha:j tikaje:k //] #3 { BEJ_MV_CONV_01_R
  - □ #1 [N.M] [SBJ.PRO] #2 [miʃi iʃi gadi idʔana han ʃaga:b ha:::n /] [miʃi iʃi gadi idʔana han ʃaga:b ha:::n /] #3 {
  - □ #1 [N.M.BORR.ARA] [=PRO] #2 [jhigu:ge: ha:j /] [jhigu:ge: ha:j /] #3 { BEJ_MV_CONV_01_RICH_SP1_199 }

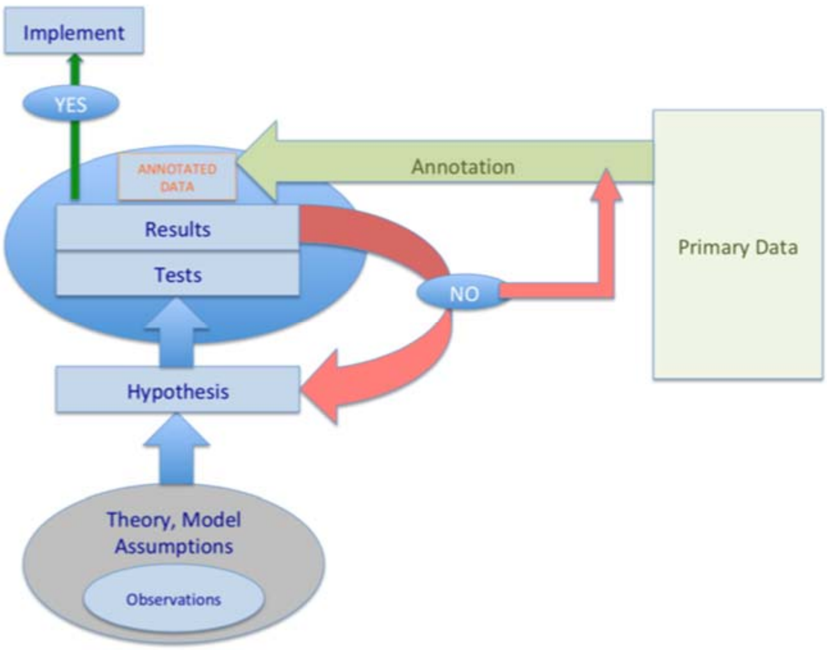
Figure 2: Annotation template in CorpAfroAs



ref@SP [1016]	KAB_AM_NARR_01_0307											
tx@SP [1016]	nnantas anaʕdanač faṭima θuħrijt //											
mot@SP [2847]	nnantas			ad	naʕdi		ad	načč		faṭima	tuħrijt	//
mb@SP [3697]	nna	-nt	=as	ad	n-	ʕdi	ad	n-	čč	faṭima	tuħrijt	//
ge@SP [3695]	say\PFV	SBJ.3PL	DAT.3S	POT	SBJ.1PL	pass\AOR	POT	SBJ.1PL	eat\AOR	Fatima	clever	//
rx@SP [3695]	V13%	PRO	PRO	PTCL	PRO	V13%	PTCL	PRO	V13%	NP	ADJ	//
ft@SP [611]	They said "let's eat Clever Fatima !"											

ref identifier for the annotation unit (time-associated)
tx transcription in broad IPA into phonological words
mot intermediary tier with segmentation into morphosyntactic words
mb morphophonological transcription into morphemes
ge morpheme-by-morpheme gloss of mb
rx part-of-speech and other information relevant for retrieval purposes
ft free translation into English

Figure 3: The Scientific Method in Corpus Annotation



Implement
YES
ANNOTATED DATA
Annotation
Primary Data
Results
Tests
NO
Hypothesis
Theory, Model
Assumptions
Observations