Segmenting spoken corpora in lesser-described languages:

New perspectives for the structural analysis of speech

Amina Mettouchi
mettouchi@vjf.cnrs.fr
Ecole Pratique des Hautes Etudes (Paris) & CNRS LLACAN
Kabyle (Berber, Afroasiatic)

28-35 languages
Number of dialects unknown

≈ 25 000 km²
≈ 5 million population

Ait Ikhlef (At Idjer tribe)
1- Information structure and Grammatical relations

- The minimal verbal clause consists of the verb and its obligatory personal affix $V_{subj}$
  
  n-ṣja  
  SBJ1PL-be_tired:PFV  
  ‘We are tired’
  
  ṣja-nt  
  be_tired:PFV-SBJ3PL.F  
  ‘They are tired’.

- Clitic pronouns may appear:
  - Absolutive (object with verbs, subject with some nominal predicates)
    n-wala=t  
    SBJ1PL-see:PFV=ABSV3SG.M  
    ‘We saw him/it’
  - Dative (only with verbs)
    n-fka=yas=t  
    SBJ1PL-give:PFV=DAT3SG=ABSV3SG.M  
    ‘We gave it to him/her’
• Nouns appear according to informational needs of speakers, under two forms: Annexed ($N_{ann}$) and Absolute ($N_{abs}$), and in various positions.

• State is not case, not grammatical role, not dependent marking (Mettouchi and Frajzyngier 2013)

• **ANN**: the noun provides the value for the variable of the function grammaticalized in the preceding constituent. Such functions are diverse.

• **ABS**: default form of the noun - as such, has no overall function of its own
Linear orders

• “For linear order to be a viable coding means there must be a point of reference, and this point of reference must be available to the hearer.” (Frajzyngier & Shay 2003:60-62)

• In Berber, the verb is a salient potential reference point (morphology).

• Another salient reference point is the prosodic boundary (discontinuity)
  • Main cues:
    • (1) final lengthening;
    • (2) initial rush;
    • (3) pitch reset;
    • (4) pause.
• Interaction of Morphological marking and Linear ordering of N and PP (with respect to V and Prosodic boundaries)
  • prosodic entity (IU boundary)
  • morphosyntactic words (noun, verb, preposition)
  • No other prior assumption concerning roles or functions of N or PP
Example

tufa ḏāmīʃjībuqḍar // (423) iḥīzādāyān //
t-ufa d āmīʃjī n wādḍar // i = t i-zdāy-n //
SBJ3SG.F-find\PFV COP cat\ABS.SG.M GEN mountain\ANN.SG.M // REL.REAL = ABSV3SG.M RELSBJ.POS-dwell\PFV-RELSBJ.POS /
PRO-V13% PRED N.OV PREP N.OV // DEMPRO-PRO CIRC1-V23-CIRC2 //
"she found it was the Mountain Cat who inhabited it,"

wɔًxɔxamnn // (271) (BI-225) anīṣa ɗamīʃjīagjī // isṣa ʔaqadįṭ //
wɔxam-nn // natta d amīʃ-jagi // i-iṣa taqadįṭ //
house\ANN.SG.M-CNS // IDP3SG.M COP cat\ABS.SG.M-PROXb // SBJ3SG.M-possess\PFV herd\ABS.SG.F //
N.OV-DEM // PRO PRED N.OV-AFFX // PRO-V13% N.OV //
the house. This cat had a herd,

ursday ikāss // (320) uwitsyimara guryxam // (434) (BI-339)
i-tṣufu i-kāss // ur i-ṭtyim ara i wɔxam //
SBJ3SG.M-exit\PFV SBJ3SG.M-graze\PFV // NEG SBJ3SG.M-stay\IPFV POSTNEG LOC house\ANN.SG.M //
he went out regularly to take it to pasture. He did not stay at home.

ʔaṣa aruyxamis kulāxīrjɔllana ṭamina // (418)
t-ḍla ar wɔxam-is kul lɔxir i-lla a Amina //
SBJ3SG.F-visit\PFV to house\ANN.SG.M-POSS3SG all good\ANN SBJ3SG.M-exist\PFV VOC Amina //
PRO-V13% PREP N.OV-PRO DET N.COVA PRO-V13% PTCL NP //
She visited his house, it was full of good things, Amina."
Constructions (subsuming structures in the data)

- \([V_{sbj} (N_{abs}) (PP) (PP)]\) which subsumes the following:
  - \([V_{sbj}]\)
  - \([V_{sbj} N_{abs}]\)
  - \([V_{sbj} PP]\)
  - \([V_{sbj} PP PP]\)
  - \([V_{sbj} N_{abs} PP]\)

- \([N_{abs} V_{sbj} (N) (PP)]\) which subsumes the following:
  - \([N_{abs} V_{sbj}]\)
  - \([N_{abs} V_{sbj} PP]\)
  - \([N_{abs} V_{sbj} N_{abs}]\)
  - \([N_{abs} V_{sbj} N_{ann}]\)

- \([V_{sbj} (N) N_{ann} (N) (PP)]\), which subsumes the following:
  - \([V_{sbj} N_{ann}]\)
  - \([V_{sbj} N_{ann} PP]\)
  - \([V_{sbj} N_{ann} N_{abs}]\)
  - \([V_{sbj} N_{ann} N_{abs} PP]\)
  - \([V_{sbj} N_{abs} N_{ann}]\)

- \(N_{abs} [V_{sbj} (N) (N)]\)
- \([V_{sbj} (N) (N)] N_{ann}\)

- \(PP [V_{sbj} (NP)]\)
- \([V_{sbj} (NP)] PP\)

NB: other structures involving other formal means than those listed in slide 34 belong to the functional domain of Information Structure. They are not studied here. Therefore this series of constructions is only a part of the domain.
Information structure functions

- \([V_{\text{sbj}} (N_{\text{abs}}) (PP) (PP)]\): Unmarked informational status: subtopic continuity
- \([V_{\text{sbj}} (N) N_{\text{ann}} (N) (PP)]\): Promotion to topical status of an event or state (sentence-focus): new episode
- \([N_{\text{abs}} V_{\text{sbj}} (N) (PP)]\): Recapitulation (for backgrounding) of salient preceding situation
- \(N_{\text{abs}} \rightarrow [V_{\text{sbj}} (N) (N)]\): contrast of the assertion with a previous presupposition from previous discourse.
- \([V_{\text{sbj}} (N) (N)] N_{\text{ann}}\): Promotion to topical status of a referent that had lost its (semi-)active status

- \(\text{PP} \rightarrow [V_{\text{sbj}} (NP)]\): temporal or spatial frame for the event/situation
- \([V_{\text{sbj}} (NP)] \text{PP}\): highlighting of the PP.
Grammatical relations

- Interaction of linear ordering with respect to V and PB, presence/absence of N, morphology codes information structure functions
- Grammatical relations are coded on bound pronouns (SBJ, ABSV, DAT)
- Are grammatical relations coded on nouns as well?
  - by agreement?
  - by morphology (ANN or ABS state)?
  - by linear ordering?
• “Agreement” alone does not mark SBJ
  – same features of gender/number as SBJ affix is not enough

(6) jə-kraŋ  igaŋ
SBJ3M.SG-plough:PFV  field:ABS.SG.M
‘He ploughed the field’ (and not ‘The field was ploughed’)

• ANN/ABS distinction alone does not mark SBJ or OBJ
  – N_{ann} can be coreferent to subject or object (or possessive or kinship) pronoun

(7) te-mmut  tqifit //
SBJ3SGF-die:PFV  child:ANN.SG.F
‘The girl died’

(8) t-ufa  d  amqif  n  wədrar //
SBJ3SGF-find:PFV  COP  cat:ABS.SG.M  GEN  mountain:ANN.SG.M //

i = t  izədyan /  wəxəam-ñni //
REL.REAL = ABSV3SG.M  inhabit:PFV:RELSBJ.POS /  house:ANN.SG.M-CNS //
‘She found it was the Mountain Cat who inhabited it, the house’
• Position alone does not mark SBJ or OBJ
  – nouns that are computable as subjects and objects can both precede or follow the verb, and if they follow the verb, there is no fixed ordering between them

(9) ayərdaj-nni jə-čča = t
rat.ABS.SG.M-CNS SBJ3M.SG-eat.PFV = ABSV3M.SG
‘He ate the rat’ or ‘The rat ate it’

(10) jə-swa wǝmsij ajfki
SBJ3M.SG-drink.PFV cat.ANN.SG.M milk.ABS.SG.M
‘The cat drank milk’

(11) jə-swa ajfki wǝmsij
SBJ3M.SG-drink.PFV milk.ABS.SG.M cat.ANN.SG.M
‘The cat drank milk’
How to interpret the facts?

- Coreference (‘agreement’) alone, linear order alone, or morphology alone cannot unambiguously code Subject or Object functions on nouns,
- Are grammatical relations only marked on pronouns?
- Or can they be computed also on nouns, but through the interplay of formal means?
- Actually, it is the interaction of the three which allows the coding of subject and object roles on nouns.

Note that the existence of grammatical relations SBJ and OBJ is established in Kabyle through behavioral tests (relativization etc.), and the existence of a dedicated pronominal paradigm for SBJ. The question here is to see whether grammatical relations are marked on **nouns** as well as on pronominal affixes.
Outside the prosodic group containing the Verb

• **Before** the prosodic boundaries of this group, no transparent coding of SBJ or OBJ
• Same position, same state (co-reference ≠ ‘agreement’)

(12) ʿllafi  i-barək  mmi-s  /
God  SBJ3SG.M-give_luck:AOR  son:ABS.SGM-KIN3SG /

atan  i-ḡḡadd  såttə  //
PRST  SBJ3SG.M-leave:PVF=PROX  six  //

‘God bless her son, here he is with six children’.

(13) ajtma  /  t-uṛ-ḍ = asn = idd  /
brother:ABS.PL  /  SBJ2-take:PFV-SBJ2SG = DAT3PL.M = PROX /

‘My brothers, you bought them things’
Outside the prosodic group containing the Verb

- **After** the prosodic boundaries of this group, **no transparent coding of SBJ or OBJ**
- Same position, same state (co-reference ≠ ‘agreement’)

(14) tə-mmut  təmṭṭut-is /  wəmyar-nni //
SBJ3SG.F-die:PFV  woman:ANN.SG.F-KIN3SG /  old_man.ANN.SG.M-CNS //
‘His wife died, that man’

(15) ad = dd  ḥku-ɣ  /  amk i  /  ḥsif-nt  zik  /
POT = PROX  tell:AOR-SBJ1SG /  how REL-REAL /  live:IPFV-SBJ3PL.F  long_agono /
lixalat  n  lqbajl-nnəɣ /
woman:ANN.PL.F  GEN  kabyle_tribe:ANN.PL-POSS1PL/
‘I will tell how they lived in the old days, the Kabyle women’.
Within the prosodic group containing the Verb

- \([V_{\text{sbj}} (N_{\text{abs}}) (PP) (PP)]\): topic continuity
- The only noun which may appear is in the absolute state
- The combination between position after the verb and absolute state (within the PGV) unambiguously codes the object

(16) i-rūh
    ar wādrar        a Amina /
SBJ3SG.M-go:PFV  to mountain:ANN.SG.M VOC Amina /

‘He went to the mountain, Amina,

i-qqaz          i-qqaz          i-qqaz          i-qqaz
he dug and dug,

i-qqaz          i-qqaz /        i-xdēm          lбир /        ann[j]tilat //
SBJ3SG.M-dig:IPFV SBJ3SG.M-dig:IPFV / SBJ3SG.M-make:PFV well:ABS.SG.M / enormous //

he made a well, an enormous one.’
Within the prosodic group containing the Verb

• \([V_{\text{sbj}} \quad N_{\text{ann}} \quad (N) \quad (PP)]\) : topicalization of event or state

• Only one NP in the annexed state can occur. It is always coreferent to the subject affix.

• Annexed state + position (following (immediately or not) the verb within the prosodic group of the verb) = transparent coding of subject

(17) SP1: <ça fait> t-mmut = as taqaʃt / i Zahiwa Taʃliṭ //
               it.is   SBJ3SG.F-die:PFV = DAT3SG girl:ANN.SG.F / DAT ZahiwaTaʃliṭ

'So she lost a girl, Zahwa Taalits ?'

SP2: t-mmut = as tmənzut //
SBJ3SG.F-die:PFV = DAT3SG eldest:ANN.SG.F //
'Her eldest daughter died (on her)’
Within the prosodic group containing the Verb

- \([N_{\text{abs}} V_{\text{sbj}} (N) (PP)]\): recapitulation of event or state for backgrounding
- If and only if there is no bound pronoun (other than SBJ-affix) in the prosodic group of the verb, \(N_{\text{abs}} = \text{subject}\)

(18) taqjunt-nni
dog:ABS.SG.F-CNS
‘The dog was barking’

(19) tiqśiśin
ur mmut-nt ara /
girl:ABS.PL.F NEG die:PFV-SBJ3PL.F POSTNEG /
‘The girls didn't die’.
• If there are one or two bound pronouns the grammatical relation is no longer transparent:

(20) jiwəṭ  tə-fka = jas  nnəfs //
one:F  SBJ3SG.F-give:PFV = DAT3SG  half //
‘One woman gave him/her half (an apple), or ‘She gave half (an apple) to one woman.

• ⇒ In the position before the verb, the noun can be transparently coded as subject if and only if there are no pronouns cliticized to the verb, only the subject affix.
Summary

• a noun is a **nominal subject** if and only if, **within the prosodic group of the verb**:
  – the verb has **no bound pronouns** other than the subject affix AND the noun occurs **before the verb** (therefore necessarily in the absolute state);
  – the noun occurs **after the verb** (immediately or not) AND is in the **annexed** state.

• a noun is a **nominal object** if and only if, **within the prosodic group of the verb**, the noun occurs **after the verb** (immediately or not) AND is in the **absolute** state.

• Several dimensions **interact**: **prosody** (boundary), **syntax** (position) and **morphology** (verbal morphology, state) = in Kabyle, **nominal subjects and objects** can only be unambiguously computed **within the prosodic group containing the verb**.
Implications

- **Information structure** values are coded through the interplay of formal means belonging to different domains of grammar.
- **Grammatical relations** are coded on nouns only in some information structure constructions, thanks to the same interplay of several formal means:
  - topic continuation (Object)
  - ‘sentence-focus’ constructions, for topicalization of event or state, or recapitulation of event or state for backgrounding (Subject and Object)
• This shows that nominal coding of grammatical relations in Kabyle is a **by-product** of information structure constraints – themselves shaped by the features of Kabyle (indexation of arguments on the verb, V-initial language,...)

• Different perspective from first ascribing grammatical roles to each Noun (on what basis? translation into target language?), and then looking at mapping between Grammatical Relations and Information Structure.
2- Roles of Prepositional Phrases

- PP: usually studied with respect to their role as arguments or adjuncts

Figure 1: The layered structure of the clause

- e.g. layered structure of the clause in Role and Reference Grammar (semantically-based model of the clause)
Does the Argument/Adjunct distinction map onto prosodic segmentation?

- Inclusion in the prosodic group containing the verb is not linked to argument vs adjunct status

| (21) | i-naa = jas | ma | t-sla-mt = dd | i | wazdzu = agi |
| SBJ3SG.M-say \ PFV = DAT3SG | if | SBJ2-hear \ PFV-SBJ2PL.F = PROX | DAT | big stick \ ANN.SG.M-PROXb |
| PRO-V13% = PRO | CONJ | CIRC1-V13%-CIRC2 = PTCL | DEMPRO | N.OV-AFFX |
| 'he said, if you hear this stick' |

| (22) | imi | ara = dd | t-sl-ʤ | i | jomqarqar | i | usuqabbaq-w |
| when2 | REL.IRR = PROX | SBJ2-hear \ AOR-SBJ2SG | DAT | toad \ ANN.PL.M | LOC | belly \ ANN.SG.M-POS1.SG |
| CONJ | N.INDF = PTCL | CIRC1-V13%-CIRC2 | DEMPRO | N.OV | PREP | N.OV-AFFX |
| 'when you hear the toads in my belly' |

- What is the motivation of position inside/outside the prosodic group containing the verb?
• Look for preposition + noun
  – before/after V
  – before/after prosodic boundary

• The only PP before the Verb are
  – also before prosodic boundary
  – are LOCATIVE
  – function = frame (adverbial)

The following day, at night, the woman told her husband

I told you, I won't accept anything else than their being eaten (either they get eaten (by wild beasts in the forest) or...)
• All other PP occur after the Verb (and postverbal nouns if any)
  • within the Prosodic Group of the Verb
  • regardless of the type of PREP and the type of Verb (DEFAULT position)

(25) faṭima  tuḫrijt    t-ṣ̣aːks=dd     aṣirum     g    dakkān     //
Faṭima   clever       SBJ3SG.F-take_away\PFV = PROX   bread\ABS.SG.M   LOC   shelf\ANN.SG.M   //
‘Clever Fatima grabbed the bread from the shelf’

(26) t-ṣawwa       i   jastma-s     /
SBJ3SG.F-cook\CAUS.PFV    DAT  sister\PL-KIN3SG   /
‘She cooked for her sisters’
• after the Prosodic Boundary, only
  • if \text{PREP} = \text{Locative or Directional} and \text{verb} = \text{motion or position or existential (as opposed to activity etc.)}

\begin{align*}
(24) \text{noy} & \quad \text{ad} = \ddagger \quad \text{i-gg} \quad \text{ak'ija} \quad // \\
\text{or} & \quad \text{POT=AHSV3SG.F} \quad \text{SBJ3SG.M-leave\AOR} \quad \text{absolutely} // \\
\text{i talmast} & \quad \text{n wərid} \quad // \\
\text{LOC middle\ANN.SG.F} & \quad \text{GEN way\ANN.MSG.M} \quad // \\
\text{‘or if he will leave her completely, in the middle of the road.’}
\end{align*}

• or if \text{PREP} = \text{Dative} and \text{Verb} followed by dative clitic

\begin{align*}
(27) \text{SP1 : < ça fait >} & \quad \text{t-mmut = as} \quad \text{taqjf} \quad / \quad \text{i} \quad \text{Zaïwa Taflit} \quad // \\
\text{it is} & \quad \text{SBJ3SG.F-die:PFV = DAT3SG} \quad \text{girl:ANN.SG.F} \quad \text{DAT} \quad \text{Zaïwa Taflit} \\
\text{‘So she lost a girl, Zahwa Taalits?’} & \quad (\text{lit: so a daughter died on her, (on) Zahwa Taalits ?})
\end{align*}
• PP is after the boundary of the prosodic group containing the verb only if the PP expands on a feature already present in the preceding constituent
  • reminiscent of the annexed state’s function (ANN=“the noun provides the value for the variable of the function grammaticalized in the preceding constituent”)

• Other explanation : only if PP is part of the syntactic/semantic valency of the verb
  • only arguments can be separated from their verb!
Conclusion

• Empirical study of what is actually coded in the language allows to discover fundamental properties of that language
• Those discoveries would not have been possible without independent coding of formal means, and their annotation (automatic searches)
• Structures are probably not duplicated/mapped from one component of the grammar to another
Perspectives

• An analysis in terms of formal means, regardless of the nature of those means, may allow more integration of the various components of speech/language (including gestures),
• Analytic grammars of lesser-described (and well-described) languages would be enriched by the systematic definition of their units of speech and language,
• Segmented and annotated corpora of naturally-occurring speech are extremely useful for language analysis.
Thank you

Don’t sit and wait. Get out there, feel life. Touch the sun, and immerse in the sea.
Jalāl ad-Dīn Rūmī (1207-1273)
References

• Izre’el & Mettouchi (submitted) Representation of Speech in CorpAfroAs: Transcriptional Strategies and Prosodic Units. In Mettouchi, A., M. Vanhove and D. Caubet (eds), *Corpus-based Studies of lesser-described Languages: the CorpAfroAs Corpus of spoken AfroAsiatic Languages*. Amsterdam-Philadelphia: John Benjamins.
Acknowledgements

• The work presented here benefitted from discussions with colleagues at University of Colorado, Boulder and Max Planck Institute, Leipzig during my research stays as guest of both institutions. Many thanks in particular to Z. Frajzyngier and B. Comrie.
• My thanks also to Maya Hickman who helped me understand the basics of the psycholinguistic and cognitive background (and theories) of language production and processing.
• Various stages of results on Information Structure and Grammatical Relations were presented at 14th Italian Meeting of Afroasiatic Linguistics (June 2011); UC Boulder (Oct 2011); MPI Leipzig (Aug 2012); SLE 2012 (Sept 2012); ISSLaC Bielefeld (May 2013): many thanks to the audiences for their feedback.