

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



Aims of the talk

- Show that the generalization of text-to-sound indexing in corpora is an opportunity to integrate systematic segmentation and annotation of speech into formally-defined units
 - allowing systematic investigations e.g. on phonology/ phonetic properties
- Show how those units can be relevant for linguistic analysis
 - especially in relation to morphosyntactic parsing
 - and in confrontation with analyses based on virtual units (e.g. clause defined as predicate+arguments +adjuncts, semantic completeness/ self-sufficiency)

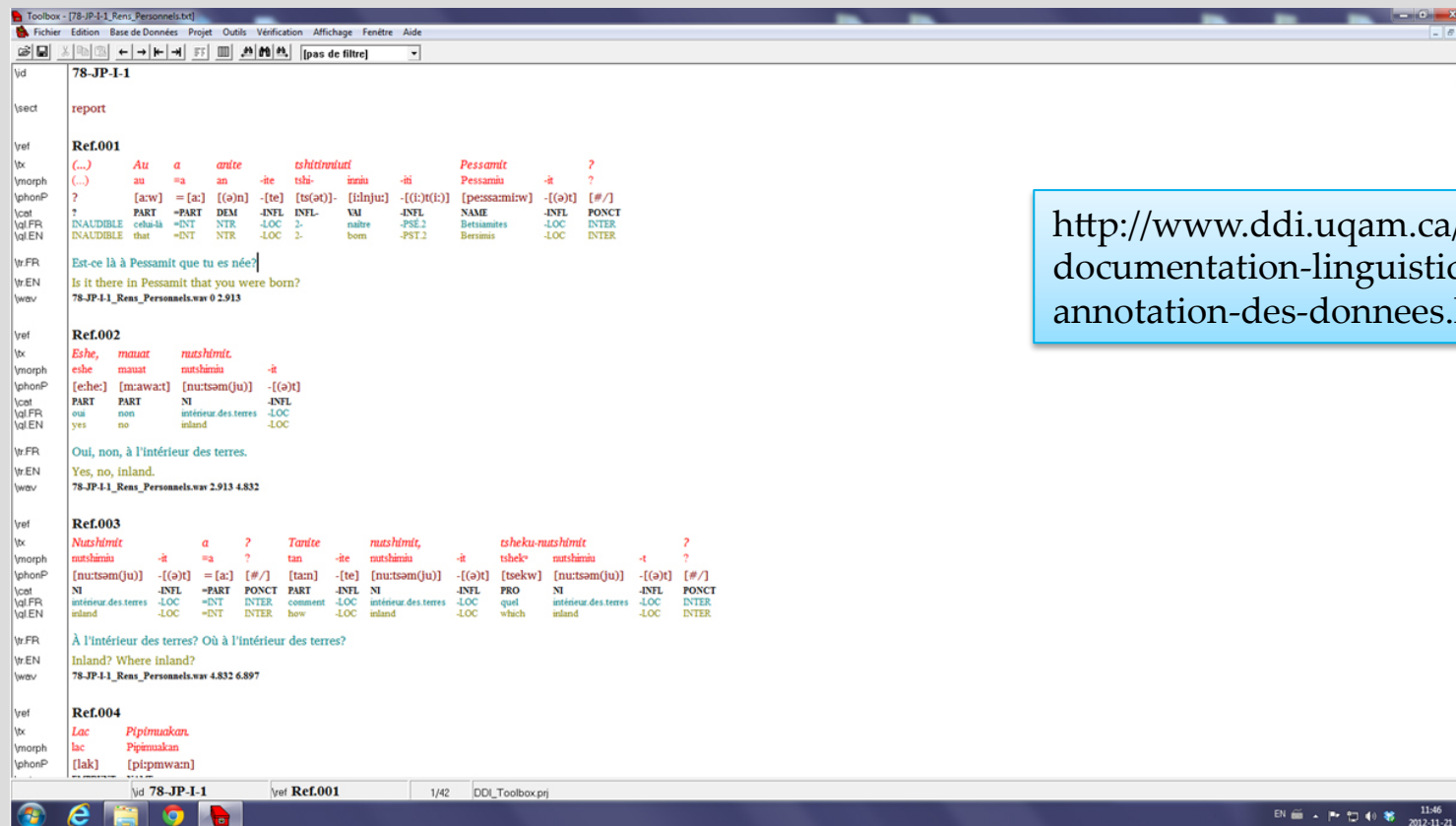
- Focus on clause/intonation unit mapping
 - show that it may well be an artefact of addition of analytic layer on preexisting analysis habits
- Propose an alternative type of (empirical) analysis based on notion of formal coding means
- Advocate the use of sound-indexed corpora as a basis for the analysis of lesser-described (and well-described) languages

Fieldwork on lesser-described languages

- Unfamiliar /unknown languages
- Analysis of the whole grammar of the language
 - phonology
 - morphology
 - syntax
 - information structure, pragmatics ...
- Methods
 - elicitation techniques
 - paradigms, complementary/contrastive distribution
 - native speaker judgements, tests
 - analysis of recorded narratives and conversations
 - grammatical contexts, situation
 - actual use of the language

Software for field linguists

- SIL International: <http://www-01.sil.org/computing/toolbox/information.htm>
 - Shoebox, Toolbox
 - FLex (Fieldworks)



The screenshot displays the Toolbox software interface, which is used for linguistic analysis. The main window shows a text editor with the following content:

```
78-JP-1-1
report

Ref.001
(...) Au a anite tshítivutui Pessamit ?
(...) au =a an -ite tshí- innú -it Pessamú -it ?
? [a:w] = [a:] [[a]n] [-te] [ts(ə)t]- [i:nju:] -[i:](f:)] [pessamit:w] -[i:](f:) [#/]
? PART =PART DEM -INFL INFL- VU -INFL NAME -INFL PONCT
InAUDIBLE cehi-lá =INT NTR -LOC 2- náire -PSE.2 Betsamites -LOC INTER
InAUDIBLE that =INT NTR -LOC 2- bom -PST.2 Betsamit -LOC INTER

WFR Est-ce là à Pessamit que tu es née?
WEN Is it there in Pessamit that you were born?
WAV 78-JP-1-1_Res_Personnels.wav 0.2913

Ref.002
Eshe, mauat nushimít.
eshe mauat nushimú -it
[e:he:] [m:awat:] [nushim(ju)] -[i:](f:)
PART PART NI -INFL
oui non intérieur des terres -LOC
yes no inland -LOC

WFR Oui, non, à l'intérieur des terres.
WEN Yes, no, inland.
WAV 78-JP-1-1_Res_Personnels.wav 2.913.4832

Ref.003
Nushimít a ? Tanite nushimít, tsheku-nushimít ?
nushimú -it =a ? tan -ite nushimú -it tshék- nushimú -it ?
[nushim(ju)] -[i:](f:) = [a:] [#/] [tan] -[te] [nushim(ju)] -[i:](f:) [tshék] [nushim(ju)] -[i:](f:) [#/]
NI -INFL =PART PONCT PART -INFL NI -INFL PRO NI -INFL PONCT
intérieur des terres -LOC =INT INTER comment -LOC intérieur des terres -LOC quel intérieur des terres -LOC INTER
inland -LOC =INT INTER how -LOC inland -LOC which inland -LOC INTER

WFR À l'intérieur des terres? Où à l'intérieur des terres?
WEN Inland? Where inland?
WAV 78-JP-1-1_Res_Personnels.wav 4.832.6897

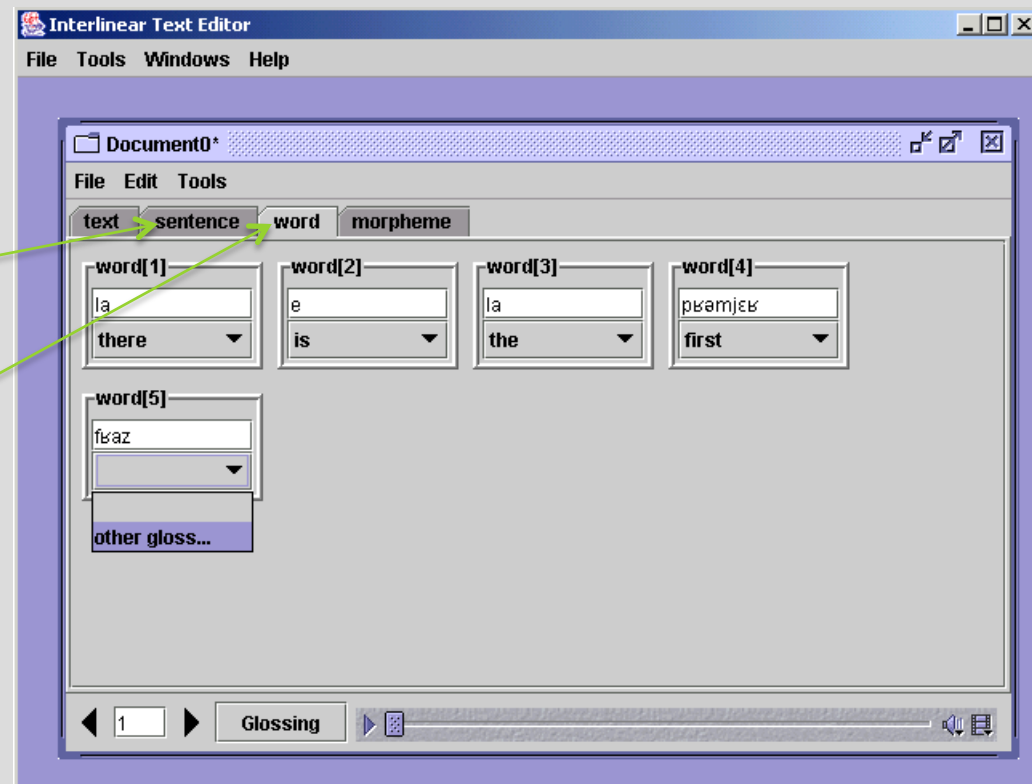
Ref.004
Lac Pipimwakan.
lac Pipimwakan.
[lak] [pipimwakan]
```

<http://www.ddi.uqam.ca/documentation-linguistique/l-annotation-des-donnees.html>

Indexing text to sound

- In addition to parsing
- Proposed levels of chunking: text, sentence, word, morpheme

Interlinear Text Editor
M. Jacobson
<http://michel.jacobson.free.fr/ITE/>



Transcriber: <http://trans.sourceforge.net/>

The screenshot displays the TranscriberAG application window. The title bar reads "TranscriberAG". The menu bar includes "File", "Edit", "Display", "Search", "Annotate", "Signal", "Speakers", "Window", and "Help". The toolbar contains various icons for file operations and playback. The left sidebar shows a "Filter" set to "Audio / Annotation files" and a tree view with "SYSTEM" (My Computer) and "MY SHORTCUTS" (Work AG, demoAG). The main workspace shows a transcription of an audio file named "PKBN_ENG_US_2009...". The transcription text is as follows:

report

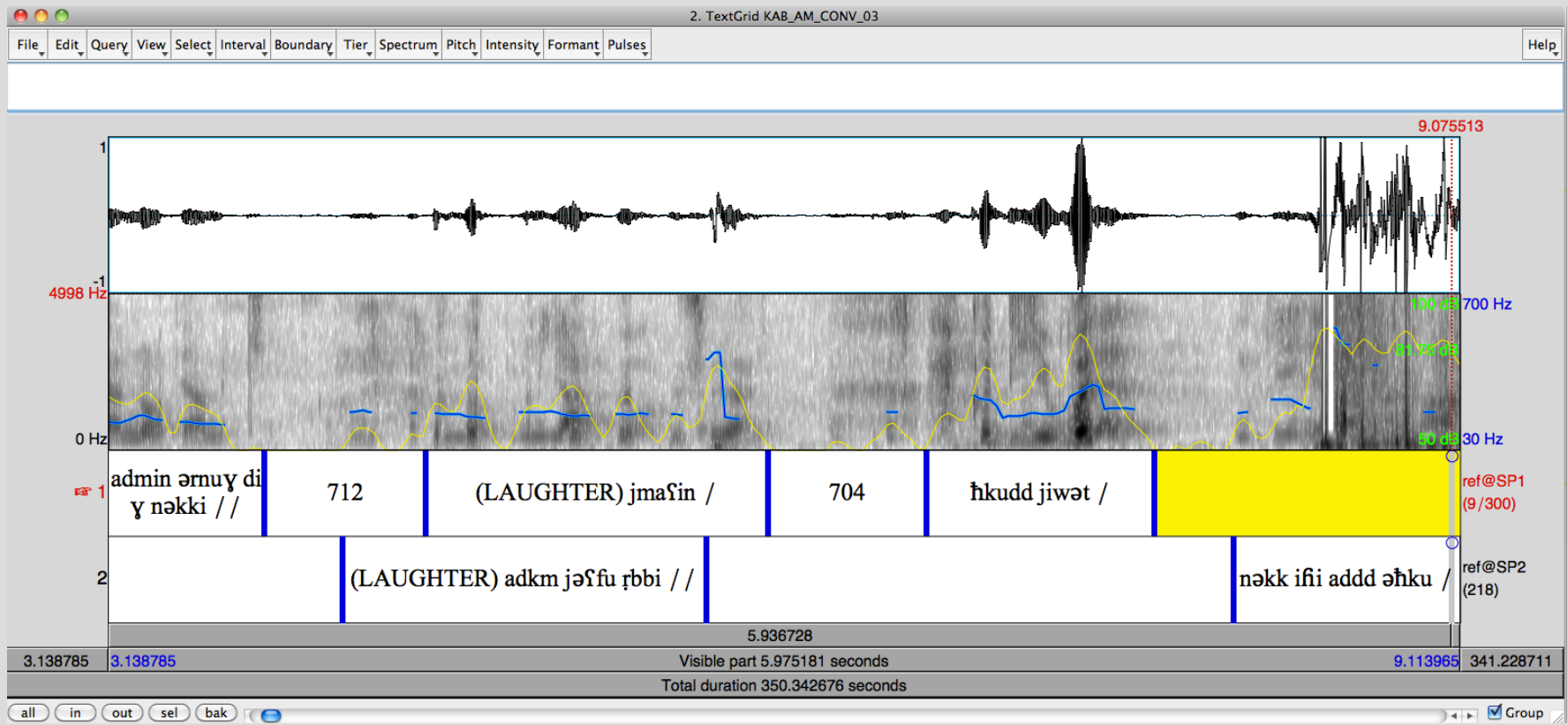
Williams Sarah

- o {b} this is Reporter's Notebook on the Voice of America . well now we had figures at least , South of the border to Mexico where drug related violence is increasingly worrying US officials . {b}
- o {b} US Secretary State Hillary Clinton will travel there next week to meet with Mexican officials . {b}
- o {b} we're joined now by VOA Houston correspondent Greg Flakus and we're joined by VOA congressional correspondent Deborah Tate . {b}
- o {b} Greg you've %er travelled in the the border region . w() what are conditions like there ?

(No speaker)

The playback controls at the bottom include buttons for "Playback" (Play, Stop, Loop), "Selection" (Auto play), "Tempo" (x1.00), "Zoom", and "Informations" (Cur. [23:51,382], Sel. [none], Total : 29:59,811). A volume control is set to +0.0 dB. The audio waveform and transcription text are visible in the playback area, with a time axis from 23:40 to 24:40. The version number "1.6.0" is shown in the bottom left corner.

Praat: <http://www.fon.hum.uva.nl/praat/>



Elan: <http://tla.mpi.nl/tools/tla-tools/elan/>

The screenshot displays the ELAN 3.9.1 software interface. The main window shows a video of a man and a woman sitting and talking. The interface includes a menu bar (File, Edit, Annotation, Tier, Type, Search, View, Options, Window, Help) and a toolbar with various playback and annotation tools. The 'Audio Recognizer' tab is active, showing parameters for 'Tag vowels (volume peaks of voiced timespans)'. The 'Parameters' section includes sliders for 'Pitch ceiling [Hz]' (set to 604.8), 'Intensity change [dB] to start/end a peak' (set to 2.0), and 'Minimum amplitude' (set to 0.1). The 'Progress' bar shows 'Ready'. Below the video, there are two waveforms: '#intensity/dB' and '#pitch/Hz'. The 'intensity/dB' waveform shows a peak at 38.9515 seconds with a value of 78.005. The 'pitch/Hz' waveform shows a peak at 504.576 Hz. Below the waveforms, there is a timeline with various annotation tiers: 'Event', 'Clause Transcri', 'Motion', 'Gesture #', and 'Gs Hand'. The 'Clause Transcri' tier shows the text: 'and so he climbs up a tree and he starts with the ladder | and he starts picking pears off the tree | and he puts the pears into an apron'. The 'Motion' tier shows 'motion' and 'non-motion' segments. The 'Gesture #' tier shows 'gestu', 'gestu', 'gesture 4', 'gestur', 'gesture 6', 'gesture 7', 'gesture', 'gesture 9', and 'gesture 1'. The 'Gs Hand' tier shows 'R', 'R', 'R', 'R', 'R', 'R', 'B', 'B', 'B'.

Other tools on e.g.: http://liceu.uab.es/~joaquim/phonetics/fon_anal_acus/herram_anal_acus.html

- Features

- discourse-oriented (above the clause)

- Transcriber, Elan, Praat

- vs grammar-oriented (below the clause)

- Toolbox, ITE

- generalization of sound-to-text alignment

- ITE, Elan, Flex...

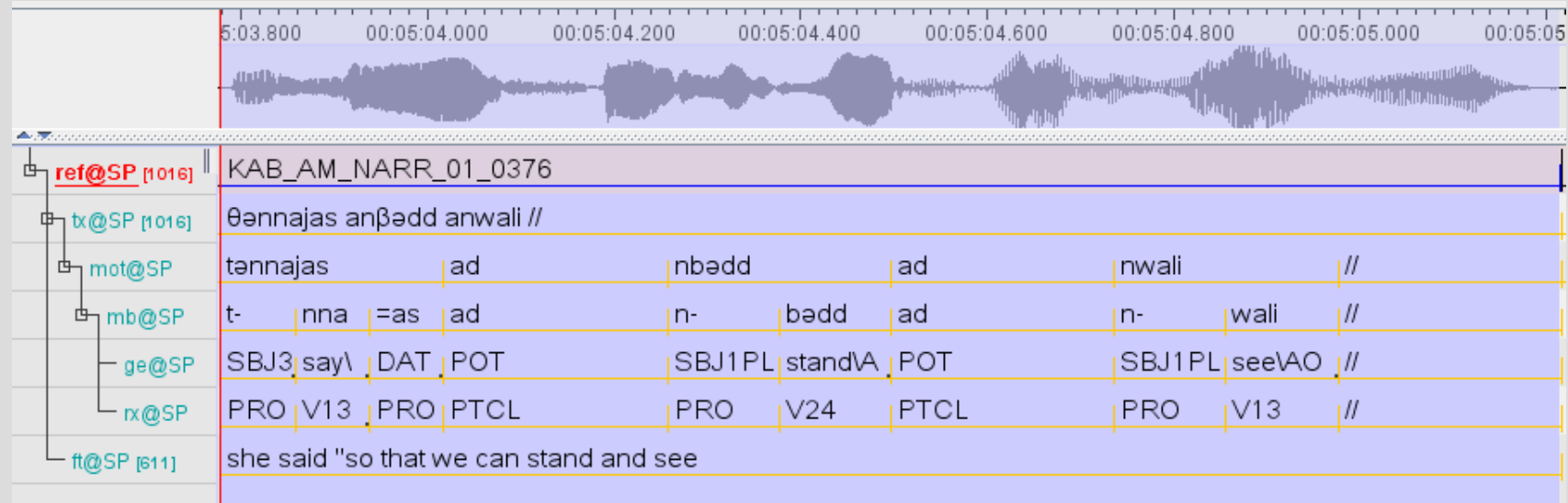
- Elan-CorpA: development of Elan within the CorpAfroAs Project:

- addition of an internal parser linked to a lexicon, for semi-automatic interlinearization

- <http://corpafroas.tge-adonis.fr/>



CorpAfroAs Layout



The various ELAN-CorpA annotation tiers (template available on CorpAfroAs website)

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)

ft free translation into English (SA)

SA: symbolic association. SS: symbolic subdivision

- Text-sound indexing → decisions as to how to segment the speech continuum
 - arbitrary basis (breath-groups, orthographic transcription) = arbitrary but consistent
 - intuition = no consistency, no explicit criteria
 - precise aims/perspective = explicit criteria: oriented, consistent and systematic
 - Opportunity for further discoveries

What kind of segmentation for my data ?

- What do I need for my analysis?
 - Information about
 - phonology,
 - morphology,
 - syntax,
 - semantics,
 - pragmatics...
- Texts not only as sample illustrations of the language, but as **corpora** ⇔ **searchability**

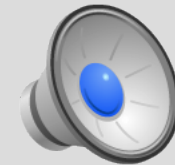
What are the relevant units of speech/language in my language?

Transcription

- Sound-indexing: all chunks can (/should?) be transcribed
- Lesser-described languages:
 - readability of transcript in relation to audio
 - possibility to work on phonetics/phonology

wwi-nt = dd sbfa tzdmin n jsyarn /

p^wintəd səβfaθzəðmin gəsyarən /



| | | | |
|-----|--|--------------------------|-------------------------------|
| (1) | p ^w intəd | [səβfʌθzəðmin | gəsy'arən] / |
| | wwi-nt = dd | sbfa tzdmin | n jsyarn / |
| | bring\PFV-SBJ3PL.F = PROX | seven bundle\ANN.PL.F | GEN firewood\ANN.PL.M / |
| | V14-PRO = PTCL | NUM N.OV | PREP N.OV / |
| | "they brought seven bundles of firewood" | | |

WORD

- Confusions
 - "(1) between a lexeme and its varying forms; (2) between an orthographic word [...] and other types of word; (3) between a unit primarily defined on grammatical criteria and one primarily defined on phonological criteria" (D&A 2002:6)
 - cf. Dixon & Aikhenvald (2002); Haspelmath (2011)
- Doing without « Word » ?
 - (Hockett (1944), Milewski (1951), Haspelmath (2011))
- Distinguishing between types of Words
 - "But is 'word' primarily a grammatical unit, with some phonological properties; or is it primarily a phonological unit, with some grammatical properties; or is it equally a unit in grammar and in phonology?" (D&A 2002: 9)

Possible criteria (D&A 2002:10)

- A **phonological word**
 - (a) *Segmental features* – internal syllabic and segmental structure; phonetic realisations in terms of this; word boundary phenomena; pause phenomena.
 - (b) *Prosodic features* – stress (or accent) and/or tone assignment; prosodic features such as nasalisation, retroflexion, vowel harmony.
 - (c) *Phonological rules* – some rules apply only within a phonological word; others (external sandhi rules) apply specifically across a phonological word boundary.
- A **grammatical word** consists of a number of grammatical elements which:
 - (a) always occur together, rather than scattered through the clause (the criterion of cohesiveness);
 - (b) occur in a fixed order;
 - (c) have a conventionalised coherence and meaning.

Identifying words in texts

- Tuttle (2008) word-definition in Ahtna Athabaskan:
 - stem stress (prosodic)
 - consonant cluster identify boundary before verb stems, not words (phonological)
 - suffixes: restricted affix sets occur at right edge of words (morphological)

‘a purely prosodic word-recognition strategy is ineffective in identifying word edges in this language.(...) The impression is of an interactive strategy for word identification that builds on word-level morphology, lexical information *and* word-internal prosody’ (2008:24)

- This interactive strategy may not be limited to word-identification, but may be at play more generally in unit-identification, and speech processing in general

Other units (above the word)

(Fox 2000, Scheer 2011, ...)

- Grammatical
 - Phrase, Clause, Sentence (constituency)
 - Comp, X', X'' ... (X-bar)
 - Nucleus, Core, Clause, Sentence (RRG)
- Prosodic
 - Accentual phrase, intermediate intonational phrase, full intonational phrase (Beckman & Pierrehumbert)
 - Clitic group, phonological phrase, intonational phrase, utterance (Nespor & Vogel)
 - Minor phrase, major phrase, intonational phrase (Selkirk)
 - Intonation unit, utterance/paratone, period... (cf discussion in Izre'el & Mettouchi (submitted))

Prosodic hierarchy

- Pivotal status of the Word
- "The Phonological Word (or Prosodic Word) is located within the phonological hierarchy **between the constituents defined in purely phonological terms** (i.e., mora, syllable, foot) and **those that involve a mapping from syntactic structure** (i.e., clitic group, phonological phrase, intonational phrase, utterance)." (Vogel 2006: 531)
- "A headstone of Lexical Phonology is that there are two distinct computational systems that assess strings of morphemes (i.e. chunks up to the size of words) and strings of words (i.e. chunks of word size and larger). The former set of rules defines **lexical**, the latter **postlexical** phonology". (Scheer 2011)
 - lexical : cyclic, phase-based
 - prosodic: representational, prosodic constituency

Prosody-Syntax mapping (Selkirk 2009)

Prosodic hierarchy

- (utterance)
- intonational phrase
- phonological phrase
 - major (intermediate phrase)
 - minor (accentual phrase)
- Prosodic word
- Foot
- syllable

Syntactic hierarchy

- (sentence)
- clause
- Phrase
 - X''
 - X'
- Word (X^0)

- Mismatches are due to phonological markedness constraints (Selkirk 2009)
- Non-isomorphism : intonational chunks simply begin with every CP (not necessarily end with every CP) (Scheer 2011)

Functional Approach: Form/Function mapping

- Phonetic approach of the IU
- DuBois et al. (1992), Chafe (1994)'s criteria:
 - a coherent intonation contour (DB)
 - pause (DB, C)
 - pitch reset at IU boundaries (DB)
 - changes in F0 (overall decline of pitch contour) (C)
 - changes in duration (DB, C)
 - anacrusis
 - final lengthening
 - changes in intensity (loudness) (C)
 - changes in voice quality (creaky voice) (C)
 - speaker change (C)

- Functions of prosodic units of different types
 - determination of the 'idea unit' (Chafe)
 - information structure unit
 - discourse structure unit (subtopic, ...)
 - pragmatic unit (speech act, ...)
- Investigation of
 - the patterning of grammatical units / prosodic units
 - the factors (discourse, grammar, semantics...) shaping prosodic phrasing

Relationship between IU and Clause/Phrase

- Findings (e.g. Ross 2011, about Dalabon and Kayardild)
 - Clause: ‘a grammatical construction which includes a predicate, its core arguments and adjuncts, where the predicate need not be verbal and may be adjectival or nominal’ (2011:116)
 - Clause is more likely to correspond to IP than IP to clause: IPs are more commonly found to comprise part of a clause
 - Grammatical complexity and prosodic length are not factors in prosodic phrasing
 - Discourse/informational factors override grammatical phrasing (NP IPs, multi-verb IPs)

- Tao (1996), about Mandarin
 - Clause : ‘a verb plus its core arguments, with modifiers (e.g. locatives, adverbials, etc.) optionally present’ (1996:17)
 - Frequencies :
 - elliptical clausal IUs > NP IUs > Full clausal IUs
 - Discourse patterns : NP IUs have three major functions:
 - referential, interactional, and rhetorical
 - ‘grammatical exponents of the IU’ (‘speech units’)
= fundamental level of grammatical structure for Mandarin:
 - NP
 - VE (verb expressions)
 - XV (AV, VO/OV, SV/VS)

Discussion

- Common assumption among most studies: prosodic units
 - are either mapped from syntactic constituents
 - or constitute a separate hierarchy with their own functional value
- Also similar assumption in studies of local phenomena
 - mapping prominence / focus markers
 - mapping prominence / focused nouns
 - typically clefts (but there are counterexamples cf Mettouchi 2003)

- Interfacing prosodic units with syntactic, discourse, pragmatic, semantic, cognitive units
 - leaves aside a residue (often non-negligible, e.g. exact clause/IP mapping in Dalabon & Kayardild =
 - IP = 1 Full clause 22.8% (Kayarldid) – 33.4% (Dalabon)
 - Full clause = 1 IP 44.6% (Kayarldid) - 43.9 (Dalabon)
 - ascribed to
 - syntactic factors (NP IPs = right- or left- dislocation)
 - phonological markedness constraints
 - processing phenomena (length, complexity...)
 - short-term memory capacity
 - etc.

- Reveals preconceptions about prosody
 - prosody as a different module from syntax, discourse...
 - prosody as encapsulation of speech, highlighting of / superimposition on pre-existing structures...
- Artefact of development of linguistics
 - analysis of written data, introspection: importance of 'langue' vs 'parole'
 - focus on units of competence vs units of performance: sentence, clause, phrase etc.
 - lack of proper means of recording, measure/analysis: late introduction of prosodic analysis
 - addition to theoretical models, to preexisting domains of analysis

Integrated approaches

- see however Chafe's definition of the 'idea unit' using three criteria: intonation, pause and syntax (all three needn't be present, nor need the presence of a single criterion necessarily signal an idea unit (1980:14)
- see also Mithun's analyzes of factors shaping word order strategies in Siouan, Caddoan and Iroquoian languages (1995)
 - morphological innovations
 - natural prosodic tendencies (pitch declination from most to less prominent in the intonation unit)
- also Morel & Danon-Boileau (1998) where the authors combine syntactic and prosodic factors to define the units of spoken French
- and others...

Proposal

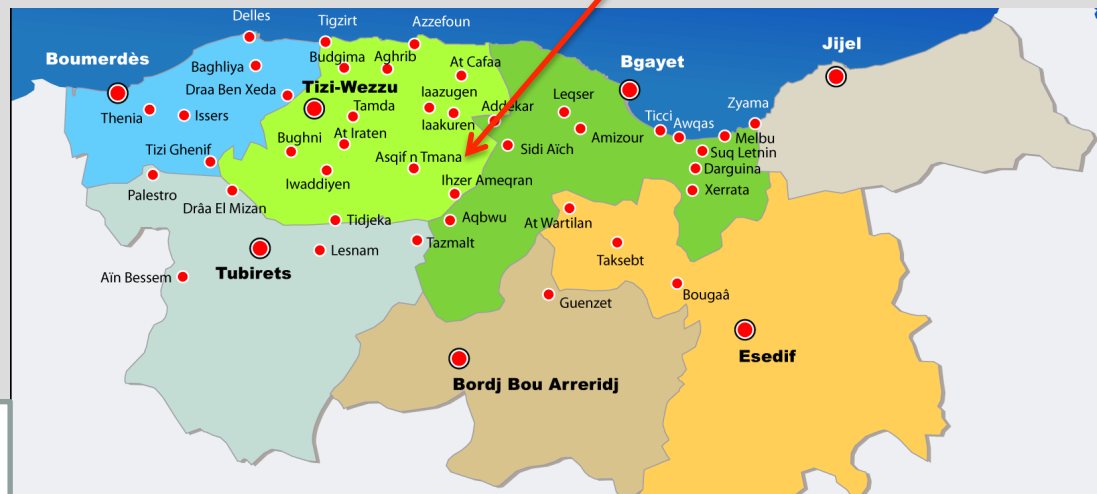
- Alternative solution to mapping approaches:
 - Avoid separation of supra-segmental and segmental by using the notion of **formal means** (Frajzyngier & Shay 2003).
 - All elements of speech, regardless of their nature, can be treated as forms, and investigated for their functions
- Study of Western Kabyle data involving simultaneously
 - prosodic boundaries (presence, absence)
 - linear ordering (with respect to a point of reference)
 - morphology

Kabyle (Berber, Afroasiatic)



28-35 languages
Number of
dialects unknown

Ait Ikhlef (At Idjer tribe)



≈ 25 000 km²
≈ 5 million population

1- Information structure and Grammatical relations

- The minimal verbal clause consists of the verb and its obligatory personal affix V_{subj}

n-ɸja

SBJ1PL-be_tired:PFV

'We are tired'

ɸja-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 ðamʃiʃbuðrar // (423) iθizəðyən /

| | | | | | | | | |
|-------------------|-----|--------------|------|-------------------|------|---------------------|-----------------------------------|-------------------|
| t-ufa | d | amʃiʃ | n | wəðrar | // | i=t | i-zdəy-n | / |
| SBJ3SG.F-find\PFV | COP | cat\ABS.SG.M | GEN | mountain\ANN.SG.M | // | REL.REAL=ABS V3SG.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,

wəxʃamni // (271) (BI-225) əntsa ðamʃiʃagi /

| | | | | | | | |
|--------------------|-------|----------|-----------|----------------------|----------------------|---------------|---|
| wəxʃam-nni // | nəttə | d | amʃiʃ-agi | / | i-sfa | taqəðʃit | / |
| 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,

itsffəy ikəss // (320)

| | | | | | | | | |
|--------------------|------------------------|----|------|--------------------|---------|------|----------------|----|
| i-ttffəy | i-kəss | // | ur | i-ttyim | ara | i | wəxʃam | // |
| SBJ3SG.M-exit\IPFV | SBJ3SG.M-graze\IPFV // | | NEG | SBJ3SG.M-stay\IPFV | POSTNEG | LOC | house\ANN.SG.M | // |
| PRO-V23.PFX | PRO-V24.LAB | // | PTCL | PRO-V24.PFX | N.INDF | PREP | N.OV | // |

he went out regularly to take it to pasture. He did not stay at home.

θʒa aruxʃamis kuləlɣirjəllana ʔamina // (418)

| | | | | | | | | |
|--------------------|----|------------------------|----------|----------|--------------------|----------|----------|-------|
| t-dla | ar | wəxʃam-is | kul | lxir | 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.COVS | PRO-V13% | PTCL | NP // |

She visited his house, it was full of good things, Amina."

Constructions (subsuming structures in the data)

- $[V_{\text{subj}} (N_{\text{abs}}) (PP) (PP)]$ which subsumes the following:
 - $[V_{\text{subj}}]$
 - $[V_{\text{subj}} N_{\text{abs}}]$
 - $[V_{\text{subj}} PP]$
 - $[V_{\text{subj}} PP PP]$
 - $[V_{\text{subj}} N_{\text{abs}} PP]$
- $[N_{\text{abs}} V_{\text{subj}} (N) (PP)]$ which subsumes the following
 - $[N_{\text{abs}} V_{\text{subj}}]$
 - $[N_{\text{abs}} V_{\text{subj}} PP]$
 - $[N_{\text{abs}} V_{\text{subj}} N_{\text{abs}}]$
 - $[N_{\text{abs}} V_{\text{subj}} N_{\text{ann}}]$
- $[V_{\text{subj}} (N) N_{\text{ann}} (N) (PP)]$, which subsumes the following:
 - $[V_{\text{subj}} N_{\text{ann}}]$
 - $[V_{\text{subj}} N_{\text{ann}} PP]$
 - $[V_{\text{subj}} N_{\text{ann}} N_{\text{abs}}]$
 - $[V_{\text{subj}} N_{\text{ann}} N_{\text{abs}} PP]$
 - $[V_{\text{subj}} N_{\text{abs}} N_{\text{ann}}]$
- $N_{\text{abs}} [V_{\text{subj}} (N) (N)]$
- $[V_{\text{subj}} (N) (N)] N_{\text{ann}}$
- **PP** $[V_{\text{subj}} (NP)]$
- $[V_{\text{subj}} (NP)]$ **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_{subj} (N_{abs}) (PP) (PP)]: Unmarked informational status: subtopic continuity
- [V_{subj} (N) N_{ann} (N) (PP)]: Promotion to topical status of an event or state (sentence-focus): new episode
- [N_{abs} V_{subj} (N) (PP)] : Recapitulation (for backgrounding) of salient preceding situation
- N_{abs} \nearrow [V_{subj} (N) (N)]: contrast of the assertion with a previous presupposition from previous discourse.
- [V_{subj} (N) (N)] N_{ann} : Promotion to topical status of a referent that had lost its (semi-)active status
- $PP \nearrow$ [V_{subj} (NP)]: temporal or spatial frame for the event/situation
- [V_{subj} (NP)] 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-krəz igər
 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) tə-mmut tqfjft //
 SBJ3SGF-die:PFV child:ANN.SG.F
 ‘The girl died’

(8) t-ufa d amfjɪ n wədrar //
 SBJ3SG.F-find:PFV COP cat.ABS.SG.M GEN mountain.ANN.SG.M //

i = t izədyən / wəxxam-nni //
 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 wəmʃiʃ ajfki
 SBJ3M.SG-drink.PFV cat.ANN.SG.M milk.ABS.SG.M
 ‘The cat drank milk’

(11) jə-swa ajfki wəmʃiʃ
 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) **lla**fi i-barək **mmi-s** /
God SBJ3SG.M-give_luck:AOR son:ABS.SGM-KIN3SG /

atan i-ğğadd sətta //
PRST SBJ3SG.M-leave:PVF = PROX six //
‘God bless her son, here he is with six children’.

(13) **ajtma** / t-uy-đ = 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* *təmt̪tut-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* *hku-γ* / *amk i* / *ʔfiʃ-nt* *zik* /
POT=PROX tell:AOR-SBJ1SG / how REL.REAL / live:IPFV-SBJ3PL.F long_ago /

lxalat *n* *lqbajl-nnəy* /
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_{subj} (N_{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) $i\text{-}\dot{r}u\dot{h}$ ar wədrar a Amina /
 SBJ3SG.M-go:PFV to mountain:ANN.SG.M VOC Amina /
 'He went to the mountain, Amina,

| | | | |
|-------------------|-------------------|-------------------|-------------------|
| $i\text{-}qqaz$ | $i\text{-}qqaz$ | $i\text{-}qqaz$ | $i\text{-}qqaz$ |
| SBJ3SG.M-dig:IPFV | SBJ3SG.M-dig:IPFV | SBJ3SG.M-dig:IPFV | SBJ3SG.M-dig:IPFV |
| he dug and dug, | | | |

| | | | | |
|-----------------------------------|---------------------|-------------------|-----------------|----------------------|
| $i\text{-}qqaz$ | $i\text{-}qqaz$ / | $i\text{-}xdəm$ | $l\text{bir}$ / | $ann\dot{s}tilat$ // |
| 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_{subj} N_{ann} (N) (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 təqjift / i Zafiwa Taflit //
it_is SBJ3SG.F-die:PFV = DAT3SG girl:ANN.SG.F / DAT ZafiwaTaflit

‘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_{abs} V_{subj} (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_{abs} = **subject**

(18) **taqjunt-nni** **t-ssəglaf** /
dog:ABS.SG.F-CNS **SBJ3SG.F-bark:CAUS.IPFV** /
'The dog was barking'

(19) **tiqfifin** **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) *jiwət* *tə-fka=jas* *nnəfɕ //*
 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

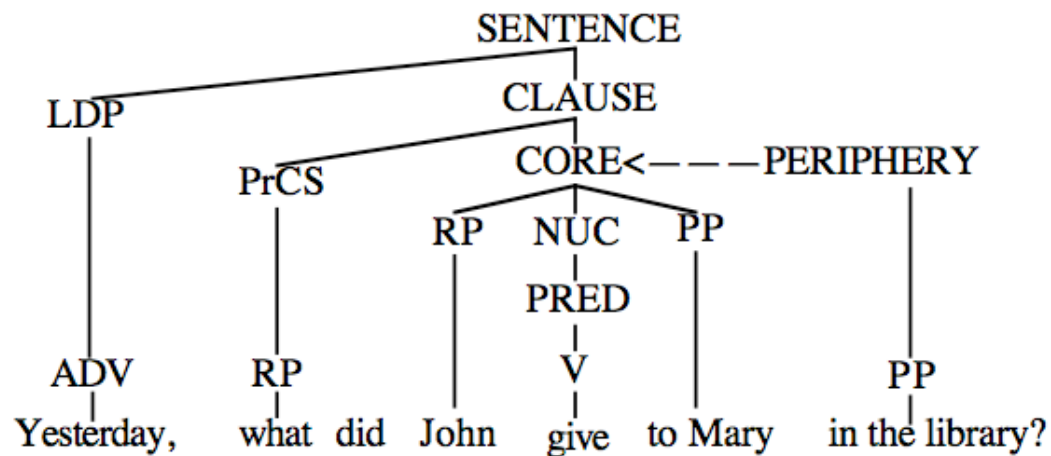


Figure 1: The layered structure of the clause²

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-nna=jas / ma t-sla-mt=dd / i wəzdduz-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-đ i jəmqarqar i uŋəbbuđ-iw /
 when2 REL.IRR = PROX SBJ2-hear\AOR-SBJ2SG DAT toad\ANN.PL.M LOC belly\ANN.SG.M-POSS1.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)**

(23) azəkka-nni / dg jid / t-nna = as /
 tomorrow-CNS / ASSOC.LOC night\ANN.SG.M / SBJ3SG.F-say\PFV = DAT3SG /
 ADV-DEM / PREP N.OV / PRO-V13% = PRO /

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

nni-y = ak = dd / ħafa tuččin ara ttw-əčč-nt //
 say\PFV-SBJ1SG = DAT2SG.M = PROX / only eating REL.IRR PASS-eat\AOR-SBJ3PL.F //
 V13%-PRO = PRO = PTCL / ADV N.V N.INDF V13%-PRO //

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 tuḥriṣt t-əkks = dd ayrum g dəkʷan //
 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-ssəwwa i jəstma-s /
 SBJ3SG.F-cook\CAUS.PFV DAT sister\PL-KIN3SG /
 ‘She cooked for her sisters’

- after the Prosodic Boundary, only
 - if PREP = Locative or Directional and verb = motion or position or existential (as opposed to activity etc.)

(24) nəy ad = ʧ i-ğğ ak°ija //
 or POT = ABSV3SG.F SBJ3SG.M-leave\AOR absolutely //

i təlməst n wəbrid //
 LOC middle\ANN.SG.F GEN way\ANN.MSG.M //

‘or if he will leave her completely, in the middle of the road.’

- or if PREP = Dative and Verb followed by dative clitic

(27) SP1 : <ça fait> t-mmut = as təqʃiʃt / i Zafiwa Tafliṭ //
 it_is SBJ3SG.F-die:PFV = DAT3SG girl:ANN.SG.F / DAT Zafiwa Tafliṭ

‘So she lost a girl, Zahwa Taalits?’ (lit: so a daughter died on her, (on) Zahwa Taalits ?)

- 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

- Beckman, M.E. & J.B. Pierrehumbert (1986) Intonational Structure in Japanese and English. *Phonology Yearbook* 3, 255-309.
- Chafe, W. (1980) The deployment of consciousness in the production of a narrative. In W. Chafe (ed), *The pear stories: cognitive, cultural and linguistic aspects of narrative production*. Norwood: Ablex.
- Chafe, W. (1994) *Discourse, consciousness and time*. Chicago & London: The University of Chicago Press.
- Dixon, R.M.W. & A. Aikhenvald (2002) *Word*. CUP: Cambridge.
- DuBois J.W., S. Schuetze-Coburn, S. Cumming & D. Paolino (1992) *Discourse transcription* (vol4). California: University of California.
- Frajzyngier, Z. & E. Shay (2003) Explaining language structure through systems interaction. Amsterdam/Philadelphia: Benjamins.
- Haspelmath, M. (2011) The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45/1. 31-80.
- Hockett, C.F. (1944) Review of *Linguistic Interludes* and *Morphology: the descriptive analysis of words* (1944 editions), both by E.A. Nida, in *Language* 20. 252-255.
- 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.

- Mettouchi, A. (2003) Contrastive Focalization on clefts in Taqbaylit Berber, in *Proceedings of IP2003 Interfaces Prosodiques*, Nantes 27-29 mars 2003, pp. 143-148.
- Mettouchi, A. and Z. Frajzyngier (2013) A previously unrecognized typological category: The state distinction in Kabyle. *Linguistic Typology* 17 (1) 2013: 30-59.
- Milewski, T. (1951) The conception of the word in the languages of North American natives. *Lingua Posnaniensis* 3. 248-268.
- Mithun, M. (1995) Morphological and prosodic forces shaping word order. In P. Downing & M. Noonan (eds), *Word Order in Discourse*. Amsterdam-Philadelphia: John Benjamins.
- Ross, B. (2011) *Prosody and Grammar in Dalabon and Kayardild*. PhD thesis. University of Melbourne.
- Scheer, T. (2011) Chunk definition in phonology: prosodic constituency vs. phase structure. In Maria Bloch-Trojnar & Anna Bloch-Rozmej (eds), *Modules and Interfaces*. Lublin: Wydawnictwo KUL. 221-253.
- Selkirk, E. (2009) On Clause and Intonational Phrase in Japanese: The Syntactic Grounding of Prosodic Constituent Structure. *Gengo Kenkyu* 136. 000-000. <http://people.umass.edu/selkirk/pdf/Selkirk2009GKprfs2.pdf>
- Tao, H. (1996) *Units in Mandarin conversation: prosody, discourse and grammar*. Amsterdam-Philadelphia: John Benjamins.
- Tuttle, S. (2008) Phonetics and word definition in Ahtna Athabascan. *Linguistics* 46-2, 439-470.
- Vogel, I. (2006) Phonological Words. In *Encyclopedia of Language and Linguistics*, 2nd Edition, Keith Brown (ed), 531-534. Oxford: Elsevier

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.