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

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/
Williams Sarah

- (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) US Secretary State Hillary Clinton will travel there next week to meet with Mexican officials.
- (b) we’re joined now by VOA Houston correspondent Greg Flakus and we’re joined by VOA congressional correspondent Deborah Tate.
- (b) Greg you’ve %er travelled in the the border region. w() what are conditions like there?

(No speaker)
Praat: http://www.fon.hum.uva.nl/praat/
Elan: http://tla.mpi.nl/tools/tla-tools/elan/

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/
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 /

\[\text{p'\text{int}e\text{d} s\text{e\beta}s\text{a\theta\epsilon\delta}\min \ g\text{es\gamma'r\epsilon\eta}} /\]

(1) p\text{int}\text{ed} [s\text{e\beta}s\text{a\theta\epsilon\delta}\min \ g\text{es\gamma'r\epsilon\eta}] /
wwi-nt = dd sbFa tzdmin n jsyarn /
bring\text{\textbackslash PFV-SBJ3PL.F} = \text{PROX} seven bundle\text{\textbackslash ANN.PL.F} GEN firewood\text{\textbackslash ANN.PL.M} /
V14-\text{PRO} = \text{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