Corpus-Based Typology
(a qualitative/pilot approach)

Perspectives for cross-linguistic comparison

Amina Mettouchi
EPHE, Paris & CNRS LLACAN
mettouchi@vjf.cnrs.fr

Talk dedicated to Bernard Comrie
MPI closing conference 1-3 May 2015
Issues broached

• Corpora as language-specific resources
  – how to make them relevant for Typology
  – pilot projects (methodological)
• From CorpAfroAs to CorTypo
  – evolution of project-design
    • from phylum-internal cross-linguistic comparison
    • to empirical typology based on unrelated languages
• Corpus Annotation
  – A crucial issue for queries
  – Different annotations for different purposes
• Categories
  – Language-internal and cross-linguistic categories
  – comparative concepts, semantic maps...
  – the issue of similarity/comparability
Spontaneous Spoken Corpora

• Language-specific resources by nature
• Traditionally used for analysis of languages (+ elicitation, input for Grammars)
• Also within Documentation Projects
  – DOBES, ELDLP
  – Attempts to use them for Cross-linguistic comparison / Typology
    • Verb/Noun project (PI- F. Seifart)
    • Referentiality project (PI- C. Lehmann)
• Dedicated pilot projects (for cross-linguistic comparison)
  • CorpAfroAs, CorTypo (PI- A. Mettouchi)
  • ...
• Annotation
  – in view of broad uses: CorpAfroAs, CorTypo
  – for a particular project: Referentiality, Verb-Nouns
• Differences
  – in granularity
  – in types of segmentation units
  – in number and types of tiers
  – in theoretical/methodological approach
  – …
Cooñ Hieneqa hirohapa rúañqaga, wakañașeįka giwá rehii!

Hana first opens the window and (then) hollers for Frogy!

 Courtesy of Iren Hartmann Hoocako (Reference Project)
http://corpafronas.huma-num.fr/

https://www.benjamins.com/#catalog/books/scl.68/main

Table of Contents

Preface
Amina Mettouchi, Martine Vastove and Dominique Caubet

Part 1: Phonetics, phonology and prosody
Representation of speech in Corpafronas: Transcriptional strategies and prosodic units
Shihori Izerei and Amina Mettouchi

Tone and intonation
Bernard Caron

Part 2: Interfacing prosody, information structure and syntax
The intonation of topic and focus: Zaar (Nigeria), Tamasheq (Niger), Juba Arabic (South Sudan) and Tripoli Arabic (Libya)
Bernard Caron, Cécile Lux, Stefano Manfredi and Christophe Pereira

Quotative constructions and prosody in some Afroasiatic languages: Towards a typology
Il-il Meilbert and Martine Vastove

Part 3: Cross-linguistic comparability
Glossing in Semitic languages: A comparison of Moroccan Arabic and Modern Hebrew
Angélique Vicente, Il-il Meilbert and Alexandra Baronti

From the Lezgic Glossing Rules to the GE and RC lines
Bernard Comrie

Cross-linguistic comparability in Corpafronas
Amina Mettouchi, Giorgio Savia and Mauro Tosco

Functional domains and cross-linguistic comparability
Zygmunt Frączek and Amina Mettouchi

Part 4: Language contact
Language contact, borrowing and code-switching
Stefano Manfredi, Marie-Claude Simonne-Senelle and Mauro Tosco

Part 5: Information technology
ELAN-Corpafronas: Lexicon-aided annotation in ELAN
Christian Chassard

Language index
Subject index
Participants

• **PI**: Amina Mettouchi
• **Engineering team**: Christian Chanard, Coralie Villes, Huyen-Tô Dan-Rabier
• **Experts**: Bernard Comrie, Shlomo Izre’el
• **Contributors**: Azeb Amha, Alexandrine Barontini, Bernard Caron, Cécile Lux, Il-Il Malibert-Yatsiv, Stefano Manfredi, Amina Mettouchi, Christophe Pereira, Mauro Tosco, Graziano Savà, Marie-Claude Simeone-Senelle, Martine Vanhove, Angeles Vicente.
CorpAfroAs

- Phylum-internal cross-linguistic comparison
  - 1 hour (40mn monologal, 20mn dialogal) per language
  - 13 Afro-Asiatic languages
  - annotated on the same principles
    - same template for all the languages
    - functional annotation of morphemes
      - 1 morpheme = 1 gloss (ignore variation of readings)
        » cf. IPFV regardless of progressive/habitual interpretation
      - 1 lexeme = 1 gloss
        » cf. xdәm in KAB always ‘make’, regardless of contextual interpretation (make sthg vs. work)
    - part-of-speech information
    - free translation into English
    - same abbreviation for given gloss
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
TX: transcription

- segmentation into prosodic units
  - terminal vs. non-terminal boundaries
- transcription in ‘broad phonetics’, into phonological words
- conversations into separate actors
About TX

• No segmentation into syntactic units
• Phonological word : language-specific
• Possible queries given type of layout:
  – average length of Intonation Unit (ms)
  – number of phonological words per I.U.
  – length of pauses, speech/pause ratio…
  – frequency of some phonemes
  – …
• comparison with MOT tier
  (morphosyntactic words)
(1)  
```
\text{p`inted} \text{se`beta`thaz`e`min} \text{gesy'aren} / \\
\text{wwi-nt} = \text{dd} \quad \text{sb`a} \quad \text{tzdmin} \quad \text{n} \quad \text{jsyarn} / \\
\text{bring\_PFV-SBJ3PL.F = PROX} \quad \text{seven} \quad \text{bundle\_ANN.PL.F} \quad \text{GEN} \quad \text{firewood\_ANN.PL.M} / \\
\text{V14-PRO = PTCL} \quad \text{NUM} \quad \text{N.OV} \quad \text{PREP} \quad \text{N.OV} / \\
"they brought seven bundles of firewood"
```
GE and RX in **ELAN-CorpA**

Dev. by C. Chanard (LLACAN)

- **Annotation Lexicon** (excerpt) →
- **Glosses** (excerpt) →

<table>
<thead>
<tr>
<th>CTP</th>
<th>Centripetal</th>
</tr>
</thead>
<tbody>
<tr>
<td>CVB</td>
<td>Converb</td>
</tr>
<tr>
<td>DAT</td>
<td>Dative</td>
</tr>
<tr>
<td>DBL</td>
<td>Double</td>
</tr>
<tr>
<td>DECL</td>
<td>Declarative</td>
</tr>
<tr>
<td>DED</td>
<td>Deductive</td>
</tr>
<tr>
<td>DEF</td>
<td>Definite</td>
</tr>
<tr>
<td>DEICT</td>
<td>Deictic</td>
</tr>
<tr>
<td>DELAT</td>
<td>Deleative</td>
</tr>
<tr>
<td>DEO</td>
<td>Deonctic modality</td>
</tr>
<tr>
<td>DEPMID</td>
<td>Deponent middle voice</td>
</tr>
<tr>
<td>DEPREC</td>
<td>Depreciative</td>
</tr>
<tr>
<td>DIFF</td>
<td>Diffusive</td>
</tr>
<tr>
<td>DIM</td>
<td>Diminutive</td>
</tr>
<tr>
<td>DIR</td>
<td>Directional</td>
</tr>
<tr>
<td>DIST</td>
<td>Distal</td>
</tr>
<tr>
<td>DISTR</td>
<td>Distributive</td>
</tr>
<tr>
<td>DITR</td>
<td>Ditransitive</td>
</tr>
<tr>
<td>DS</td>
<td>Different Subject</td>
</tr>
<tr>
<td>DTR</td>
<td>Detransitivizer</td>
</tr>
<tr>
<td>DU</td>
<td>Dual</td>
</tr>
<tr>
<td>DUB</td>
<td>Dubitative mood</td>
</tr>
<tr>
<td>DUR</td>
<td>Durative</td>
</tr>
<tr>
<td>ELAT</td>
<td>Elative case</td>
</tr>
<tr>
<td>EMPH</td>
<td>Emphatic</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Parse</th>
<th>Parameters</th>
<th>Linked File</th>
</tr>
</thead>
<tbody>
<tr>
<td>.. Nr</td>
<td>Lexicon</td>
<td>Variant</td>
<td>Gloss</td>
</tr>
<tr>
<td>1...</td>
<td>tarast</td>
<td>edible_herb_ABS.SG.F</td>
<td>N.cov</td>
</tr>
<tr>
<td>1...</td>
<td>robbi</td>
<td>educate_AOR</td>
<td>V14</td>
</tr>
<tr>
<td>1...</td>
<td>trebbi</td>
<td>educate_IPFV</td>
<td>V14_PRFX_APHO</td>
</tr>
<tr>
<td>1...</td>
<td>timellalin</td>
<td>egg_ABS.F.PL</td>
<td>N.OV</td>
</tr>
<tr>
<td>1...</td>
<td>n̄h</td>
<td>eh</td>
<td>INTJ</td>
</tr>
<tr>
<td>1...</td>
<td>ja</td>
<td>either</td>
<td>CONJ</td>
</tr>
<tr>
<td>7...</td>
<td>tmenzut</td>
<td>elder_ANN.SG.F</td>
<td>N.OV</td>
</tr>
<tr>
<td>9...</td>
<td>tristi</td>
<td>electricity_ABS</td>
<td>N.COV</td>
</tr>
<tr>
<td>9...</td>
<td>tristi</td>
<td>electricity_ANN</td>
<td>N.COV</td>
</tr>
<tr>
<td>6...</td>
<td>taggarā</td>
<td>end_ABS.F.SG</td>
<td>N.COV</td>
</tr>
<tr>
<td>2...</td>
<td>annfliat</td>
<td>enormous</td>
<td>PHRASE</td>
</tr>
<tr>
<td>4...</td>
<td>kjam</td>
<td>kaj;...</td>
<td>enter_AOR</td>
</tr>
<tr>
<td>1...</td>
<td>kjam</td>
<td>enter_IPFV</td>
<td>V23</td>
</tr>
<tr>
<td>4...</td>
<td>kjam</td>
<td>enter_NEGPFV</td>
<td>V23</td>
</tr>
<tr>
<td>4...</td>
<td>kjam</td>
<td>enter_PFV</td>
<td>V23</td>
</tr>
<tr>
<td>8...</td>
<td>ssadhaj</td>
<td>entertain_CAUS.IPFV</td>
<td>V24_APHO</td>
</tr>
<tr>
<td>65...</td>
<td>irkal</td>
<td>entirely</td>
<td>ADV</td>
</tr>
<tr>
<td>5...</td>
<td>mon̄</td>
<td>escape_IPFV</td>
<td>V23</td>
</tr>
<tr>
<td>1...</td>
<td>ilumjān</td>
<td>European_ANN.PL.M</td>
<td>N.cov</td>
</tr>
<tr>
<td>1...</td>
<td>&lt;loro&gt;</td>
<td>euros</td>
<td>N.CSW.FRA</td>
</tr>
<tr>
<td>1...</td>
<td>&lt;même&gt;</td>
<td>even</td>
<td>CONJ.CSW.FRA</td>
</tr>
<tr>
<td>98...</td>
<td>ula</td>
<td>even</td>
<td>ADV</td>
</tr>
<tr>
<td>7...</td>
<td>yas</td>
<td>even_if</td>
<td>CONJ</td>
</tr>
<tr>
<td>1...</td>
<td>lammedit</td>
<td>evening</td>
<td>ADV</td>
</tr>
<tr>
<td>1...</td>
<td>mammedit</td>
<td>evening_ANN.SG.F</td>
<td>N.OV</td>
</tr>
<tr>
<td>4...</td>
<td>kulje</td>
<td>everything</td>
<td>INDF.PRO</td>
</tr>
<tr>
<td>1...</td>
<td>swaswa</td>
<td>exactly</td>
<td>ADV</td>
</tr>
<tr>
<td>1...</td>
<td>tli</td>
<td>exist_IPFV</td>
<td>V13_PRFX.A...</td>
</tr>
<tr>
<td>1...</td>
<td>lla</td>
<td>lli</td>
<td>exist_NEGPFV</td>
</tr>
<tr>
<td>1...</td>
<td>fəfəy</td>
<td>fəfəy</td>
<td>exit_AOR</td>
</tr>
<tr>
<td>6...</td>
<td>ssuffix</td>
<td>ssuffix</td>
<td>exit_CAUS.AOR</td>
</tr>
<tr>
<td>3...</td>
<td>l</td>
<td>fəfəy</td>
<td>exit_IPFV</td>
</tr>
<tr>
<td>58...</td>
<td>fəfəy</td>
<td>fəfəy</td>
<td>exit_IPFV</td>
</tr>
</tbody>
</table>
Glosses

• Consistent abbreviations
  – language-internal definitions

example: Absolute and Annexed states in Kabyle
(A. Mettouchi, working document)
Queries

• For the corpus author:
  – test hypotheses
  – discover new constructions etc.
    • unmediated access to the corpus
• For the end-user
  – language specialist:
    • similar queries as author
    • mediation through list of glosses
  – phylum specialist, or typologist: necessary mediation
    • cross-linguistic queries
    • mediation through grammars or grammatical sketches + list of glosses
Beja grammatical sketch

Table of Contents

1. INTRODUCTION ........................................................................................................... 3
2. MORPHOLOGY ............................................................................................................... 8
3. SYNTAX ........................................................................................................................... 9
3.1. NOMINAL MORPHOLOGY ......................................................................................... 9
3.2. VERBAL MORPHOLOGY ............................................................................................. 9
4. CONCLUSION .................................................................................................................. 12
<table>
<thead>
<tr>
<th>OK</th>
<th>Language</th>
<th>Title</th>
<th>Author</th>
<th>Identifier</th>
<th>Size</th>
<th>Duration</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Beja</td>
<td>The shelter and the lovers</td>
<td>Martine VANHOVE</td>
<td>BEJ_MV_NARR_01_SHELTER</td>
<td>256 words</td>
<td>00:02:42</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Beja</td>
<td>The farmer and the djinn</td>
<td>Martine VANHOVE</td>
<td>BEJ_MV_NARR_02_FARMER</td>
<td>544 words</td>
<td>00:03:28</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Beja</td>
<td>The camel race</td>
<td>Martine VANHOVE</td>
<td>BEJ_MV_NARR_03_CAMEL</td>
<td>392 words</td>
<td>00:03:49</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Beja</td>
<td>The cat-djinn</td>
<td>Martine VANHOVE</td>
<td>BEJ_MV_NARR_04_DJINN</td>
<td>252 words</td>
<td>00:02:04</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Beja</td>
<td>Lost in Eritrea</td>
<td>Martine VANHOVE</td>
<td>BEJ_MV_NARR_05_ERITREA</td>
<td>709 words</td>
<td>00:06:47</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Beja</td>
<td>The foreigner, the Beja and the leopard</td>
<td>Martine VANHOVE</td>
<td>BEJ_MV_NARR_06_FOREIGNER</td>
<td>112 words</td>
<td>00:01:10</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Beja</td>
<td>Cold at the pilgrimage</td>
<td>Martine VANHOVE</td>
<td>BEJ_MV_NARR_07_COLD</td>
<td>136 words</td>
<td>00:01:16</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Beja</td>
<td>The drunkard who became a Muslim and a saint</td>
<td>Martine VANHOVE</td>
<td>BEJ_MV_NARR_08_DRUNKARD</td>
<td>358 words</td>
<td>00:03:25</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Beja</td>
<td>The jewel and the monster</td>
<td>Martine VANHOVE</td>
<td>BEJ_MV_NARR_09_JEWEL</td>
<td>111 words</td>
<td>00:01:04</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Beja</td>
<td>The rabbit and the camel</td>
<td>Martine VANHOVE</td>
<td>BEJ_MV_NARR_10_RABBIT</td>
<td>117 words</td>
<td>00:01:05</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Beja</td>
<td>Muna's coffee</td>
<td>Martine VANHOVE</td>
<td>BEJ_MV_NARR_11_COFFEE</td>
<td>58 words</td>
<td>00:00:33</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Beja</td>
<td>The boy-eater witches</td>
<td>Martine VANHOVE</td>
<td>BEJ_MV_NARR_12_WITCH</td>
<td>289 words</td>
<td>00:02:22</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Beja</td>
<td>The dream by the grave</td>
<td>Martine VANHOVE</td>
<td>BEJ_MV_NARR_13_GRAVE</td>
<td>190 words</td>
<td>00:02:06</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Beja</td>
<td>Sijadok the Christian and the Qadi</td>
<td>Martine VANHOVE</td>
<td>BEJ_MV_NARR_14_SIADOK</td>
<td>552 words</td>
<td>00:05:31</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Beja</td>
<td>An old man kills a leopard</td>
<td>Martine VANHOVE</td>
<td>BEJ_MV_NARR_15_LEOPARD</td>
<td>197 words</td>
<td>00:01:51</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Beja</td>
<td>The Prophet, the fox and the crow</td>
<td>Martine VANHOVE</td>
<td>BEJ_MV_NARR_16_PROPHET_FOX</td>
<td>545 words</td>
<td>00:04:51</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Beja</td>
<td>The shoemaker and the fairies</td>
<td>Martine VANHOVE</td>
<td>BEJ_MV_NARR_17_SHOEMAKER</td>
<td>489 words</td>
<td>00:04:27</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Beja</td>
<td>Adam and the devil</td>
<td>Martine VANHOVE</td>
<td>BEJ_MV_NARR_18_ADAM_DEVIL</td>
<td>583 words</td>
<td>00:05:26</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Hausa</td>
<td>Hausa grammatical sketch</td>
<td>Bernard CARON</td>
<td>HAU_BC_GRAMMATICALSKETCH.PDF</td>
<td>563 Kb</td>
<td>00:05:26</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Hausa</td>
<td>Hausa Conversation 1 (Boys)</td>
<td>Bernard CARON</td>
<td>HAU_BC_CONV_01</td>
<td>2410 words</td>
<td>00:15:40</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Hausa</td>
<td>Hausa Conversation 2 (Boys)</td>
<td>Bernard CARON</td>
<td>HAU_BC_CONV_02</td>
<td>1638 words</td>
<td>00:10:18</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Hausa</td>
<td>Hausa Conversation 3 (Girls)</td>
<td>Bernard CARON</td>
<td>HAU_BC_CONV_03</td>
<td>1326 words</td>
<td>00:08:48</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Hausa</td>
<td>Hausa Conversation 4 (Men)</td>
<td>Bernard CARON</td>
<td>HAU_BC_CONV_04</td>
<td>4530 words</td>
<td>00:28:47</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Hausa</td>
<td>Hausa Narration 1 (Women)</td>
<td>Bernard CARON</td>
<td>HAU_BC_NARR_01</td>
<td>850 words</td>
<td>00:05:26</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Hausa</td>
<td>Hausa Narration 2 (Women)</td>
<td>Bernard CARON</td>
<td>HAU_BC_NARR_02</td>
<td>611 words</td>
<td>00:03:28</td>
<td>🔗</td>
</tr>
<tr>
<td></td>
<td>Hausa</td>
<td>Hausa Narration 3 (Girls)</td>
<td>Bernard CARON</td>
<td>HAU_BC_NARR_03</td>
<td>146 words</td>
<td>00:00:58</td>
<td>🔗</td>
</tr>
</tbody>
</table>
### Search Domain

97 file(s)

### Concordances and Lists

List of tier type: [mot] & [?]
Order: [alphabetically] & [by frequency]
Concordances in: [mot] & [?]
Concordances in: [CONCORDANCE]
Context length (chars): [50]

### Search

Case sensitive: [on] & [off]
Regular expression: [on] & [off]
Minimal duration (ms): [0]
Maximal duration (ms): [ ]

<table>
<thead>
<tr>
<th>Left Context</th>
<th>Target</th>
<th>Right Context</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NEG. EKS</td>
<td>All Tiers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>All Tiers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>All Tiers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>All Tiers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>All Tiers</td>
</tr>
</tbody>
</table>

[FIND]
They told their little sister, "now, what are we going to eat?"

They said, "Let's eat Clever Fatima!"
Cross-Linguistic Comparison in CorpAfroAs

- Directional extensions (Mettouchi, Savà & Tosco 2015), attach to verbs
  - Hausa, Zaar (Chadic), Kabyle, Tamasheq (Berber), Gawwada, Ts’amakko (Cushitic)
  - Grammatical sketches & Lists of Glosses
    - DIR (directional, precisely ‘ventive’) in Hausa, CTP (centripetal) in Zaar
    - PROX (proximal) & DIST (distal) in Kabyle & Tamasheq
    - ASS (assertive) & DAT (dative) in Ts’amakko and CTP (centripetal) & CFG (centrifugal) in Gawwada
Examples

As we had left,

we were cutting wood,

she would go to the teacher,

she would tell him
Proportions

• Hausa and Zaar: only one directional extension (Hi tone +-o; -ɗi)
• Kabyle and Tamasheq: two extensions
  – Tamasheq: DIST= 40%, PROX=60%
  – Kabyle: DIST= 0.1%, PROX=99.9%
    • same diachronic origin =in/=n vs. =du /dd
• Ts’amakko and Gawwada:
  – Ts’amakko: DAT= 3%; ASS= 97%
  – Gawwada: CTP= 21%; CFG= 79%
    • same diachronic origin =nu vs. =na
Main types of verbs

- Hausa and Zaar
  - motion and handling verbs,
- Kabyle
  - verbs of motion, position and handling,
  - verbs of saying
  - all sorts of other verbs except statives
- Tamasheq
  - motion verbs and verbs of saying
- Gawwada and Ts’amakko
  - verbs of motion and handling,
  - verbs of saying
  - other verbs (human activities)
Corpus queries as suggestions, preliminary investigations for more detailed cross-linguistic/typological research.

- Grammaticalisation path from motion verbs
  - ρ diffusion to potentially all types of dynamic verbs (cf. Kabyle) – with modal values, via:
    - handling verbs
    - verbs of saying
- Link between deictic motion and speaker-stance?
  - movement towards speaker/addressee in Tamasheq, and Gawwada
  - + viewpoint of the speaker/ addressee in Kabyle
- etc.
Achievements and Limitations

• Achievements
  – 13 hours of prosodically-segmented, and morphosyntactically-annotated spoken data in 13 lesser-described languages
  – common annotation scheme & template
  – 40% dialogal, 60% monologal

• Limitations
  – No clear-cut distinction between language-internal categories & cross-linguistic ones
  – Necessary mediation through grammatical sketches, grammars...
CorTypo: Designing spoken corpora for cross-linguistic research

CorTypo is a project that was active from 2013 to 2017. It aimed to design spoken corpora for cross-linguistic research. The project was financed by the Agence Nationale pour la Recherche (ANR), and it was initially prepared in 2011 and submitted in January 2012.

Principal Investigator: Amina Meltouchi
Director of Studies at EPHE (Ecole Pratique des Hautes Etudes), member of the CNRS laboratory LLACAN
Professional webpage (CV, publications): http://aminameltouchi.linguista.net

Aim of the project

The aim of the CorTypo project is the elaboration of an innovative system of linguistic annotation of natural language corpora in lesser-described spoken languages.

Innovative nature of the project

1. An annotation of sound-indexed texts that is based on the formal means existing in a given language, including prosodic means, linear orders, and...
Participants

- **PI**: Amina Mettouchi
- **Engineering team**: Christian Chanard, Mourad Aouini, Tahar Meddour
- **Scientific Managers**: Zygmunt Frajzyngier (Functional database), Bernard Comrie (Table of Categories), Martine Vanhove (Corpus)
- **Contributors**: Evangelia Adamou, Azeb Amha, Isabelle Bril, Bernard Caron, Denis Creissels, Zygmunt Frajzyngier, Katharina Haude, Il-Il Malibert-Yatsiv, Stefano Manfredi, Amina Mettouchi, Nicolas Quint, Stéphane Robert, Paulette Roulon-Doko, Yvonne Treis, Martine Vanhove.
- **Administrative coordinator**: Jeanne Zerner
- **Invited IT expert**: Marc Kemps-Snijders
• 14 languages (different phyla)
• A pilot corpus, fully L-annotated
  – same template as CorpAfroAs,
  – with additional functionalities in Elan (allowing retrieval of complex constructions & reference-tracking)

<table>
<thead>
<tr>
<th>Language</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movima</td>
<td>Amerindian isolate</td>
</tr>
<tr>
<td>Ixcatec</td>
<td>Oto-Manguean</td>
</tr>
<tr>
<td>Amis</td>
<td>Austronesian</td>
</tr>
<tr>
<td>Gbaya</td>
<td>Ubangian</td>
</tr>
<tr>
<td>Wolof</td>
<td>Atlantic</td>
</tr>
<tr>
<td>Koalib</td>
<td>Heiban</td>
</tr>
<tr>
<td>Kabyle</td>
<td>Berber</td>
</tr>
<tr>
<td>Wandala</td>
<td>Chadic</td>
</tr>
<tr>
<td>Zaar</td>
<td>Chadic</td>
</tr>
<tr>
<td>Beja</td>
<td>Cushitic</td>
</tr>
<tr>
<td>Kambaata</td>
<td>Cushitic</td>
</tr>
<tr>
<td>Wolaitta</td>
<td>Omotic</td>
</tr>
<tr>
<td>Baggara Arabic</td>
<td>Semitic</td>
</tr>
<tr>
<td>Hebrew</td>
<td>Semitic</td>
</tr>
</tbody>
</table>
• Deliverables
  – A pilot corpus
  – Interfaced with a comparative database
  – And a table of categories (gloss, abbreviation and full language-internal definitions of categories + cross-linguistic definitions (supervised by B. Comrie))

• The comparative database
  – bottom-up empirically-built functional domains
  – inspired by Frajzyngier’s approach
    (Frajzyngier & Mycielski 1998, Frajzyngier & Shay 2003, Frajzyngier 2013, Frajzyngier & Mettouche 2015, …))
  – comparison among Functional domains
    • «The functional domain is a set of functions that all share one semantic characteristic, and the forms that realize them are in complementary distribution within a relevant constituent.” (Z.F.)
Similarities and differences (F&M 2015)

• Languages are similar/different
  – Different functional domains encoded in the grammar
    • L1 Aspect, L2 Tense, L3 Aspect+Tense...
  – Different structures of a given F.D.
    • L1 Three Aspects, L2 Five Aspects ...
  – Different grammaticalized meanings
    • Perfective in a language with a Negative Perfective (Kabyle) is different from Perfective in a language with no Negative Perfective (Shilha, Siwi...)
Predications

Reference

Tense-Aspect-Mood

current query interface

<table>
<thead>
<tr>
<th>Affective</th>
<th>The subject is nominative-marked, the property is either expressed by an adjective or a relativised verb to which the equational/absolute copula is suffixed.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Casative</td>
<td>Morphological derivation, suffix -t (the suffix sometimes merges with the stem-final consonant) or -t(tae), the distribution of -(ta) vs. -(t) is partly grammatically, partly lexically determined and thus not entirely predictable; the case/controlling is marked by the NOM case, the case is marked by the ACC or DAT case.</td>
</tr>
<tr>
<td>Distributive</td>
<td>Morphological derivation consisting of three morphemes, the middle morpheme, the passive morpheme and the causative morpheme.</td>
</tr>
<tr>
<td>Equational</td>
<td>The subject is nominative-marked, the class/group to which the referent belongs is expressed by a noun to which the equational/absolute copula (COPO) is suffixed.</td>
</tr>
<tr>
<td>Existential</td>
<td>Existential copular verb yoo - COPI) with nominative subject; location expressed in a locative or oblique case adjacent for location, instrumental/comitative adjunct for companions and dative for possessors.</td>
</tr>
<tr>
<td>Identification</td>
<td>The subject is nominative-marked, the unique referent with which the subject referent is equated is followed by a non-verbal copula -VV- (COPI).</td>
</tr>
<tr>
<td>Middle</td>
<td>Morphological derivation, two allomorphs (i) a glottal suffix realised as a glottal stop if the stem-final single consonant is a sonant, realised as glottalisation if the stem-final single consonant is an obstruent, (ii) an aq form found after stem-final geminate consonants/clusters.</td>
</tr>
<tr>
<td>Passive</td>
<td>Morphological derivation, suffix -am.</td>
</tr>
<tr>
<td>Reciprocal</td>
<td>Morphological derivation consisting of two morphemes, the middle morpheme -t/-aq plus the passive morpheme.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Affective</th>
<th>Affected subject predications are an outcome of the combination of the verbs of class 3 with the exponent of the function suffix -t. Affected subject predications is always intensive.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affirmative existential</td>
<td>dagga NP</td>
</tr>
<tr>
<td>Equational</td>
<td>NP (Propositional)</td>
</tr>
<tr>
<td>Goal</td>
<td>Two constructions code the predications. When class 1 verbs (intransitive or transitive verbs) the predications is marked by the vowel deletion on the verb preceding the noun phrase. The noun phrase that follows the verb represents the direct object. Third person singular object process cannot occur on the verb. With class 2 verbs (non-affiliated subjects, non-affiliated objects and class 3 (affected subject) verbs, the predications is marked by the suffix e (cone the high tone) added to the verb.</td>
</tr>
<tr>
<td>Indirect object</td>
<td>Y + (object process) + 3p, ssg, object process u. With verbs other than inherently indirect object, the third person singular indirect object is marked by the sequence u + a. For inherently indirect object verbs, the third person singular is marked only by the pronoun. The marking on the verb is required regardless of whether there is or not a nominal indirect object.</td>
</tr>
<tr>
<td>Indirect object role of the noun phrase</td>
<td>In the indirect object predications, marked by the third person object suffix u, a noun phrase is marked for the indirect object through the preposition y -u.</td>
</tr>
<tr>
<td>Locative</td>
<td>If the predications is inherently locative the locative predications is coded by juxtaposition of the predications and the locative subject. If the predications is inherently locative, the locative predications is marked by the predications u complemented by the NP. The noun phrase is marked for the locative function by a variety of prepositions.</td>
</tr>
<tr>
<td>Negative existential</td>
<td>bis or fiiis in clause-final or clause-final position</td>
</tr>
<tr>
<td>Negated subject non-affiliated object</td>
<td>V2 a NP</td>
</tr>
<tr>
<td>Object role of the noun phrase</td>
<td>When separated from the verb by other constituents, vowel deletion on the constituent preceding the NP indicates that the NP is the direct object. When following the verb of class 1, vowel reduction on the verb indicates that the noun phrase is the direct object.</td>
</tr>
</tbody>
</table>
Formal queries allow the retrieval of relevant corpus data in the languages of CorTypo
BAGGARA ARABIC – sem.

Positive existential predication  \( \mathsf{fî = PRO.a, agreement} \)

Negative existential predication  \( \mathsf{má = fî = PRO.a} \)

Look for EXS in ge and/or COP in rx, at a distance of one "rx" (left) cell from a PROa

KABYLE – berb.

Negative existential

Negative Existential has the form \( \overline{ulaf} \), preceded or followed by an NP in the absolute state, where the form \( \overline{ulaf} \) is the negative existential predicate.

Negative existential denies the existence of a referent, or a situation.

\[ \{ \text{ge = NEG.EXS} \& \text{rx = PRED} \} \& \{ \text{ge = 1} \& \text{rx = 1} \} \]

\[ \{ \text{ge = NOT(ABSV)} \& \text{rx = PRO} \} \overline{ulaf} \mathsf{+ NO absolutive pronoun} \]

IXCATEC – oto-mang.

existential and equational  \( \mathsf{sf NP} \)

Look for EXIST in ge and PRED in rx, AND . in ge and N in rx

34
Comparing Languages

• In CorpAfroAs
  – homogeneised labels (PFV for perfective, completed, ‘accompli’…)
  – L-defined in the grammatical sketches
  – comparison within a phylum/families
  – open comparison
  • which categories are « comparable » is to be decided by the end-user depending on their theoretical position
    – Formal categories (Newmeyer 2007)
    – ‘Comparative concepts’ (Haspelmath 2007)
    – etc…
Comparing Languages

• In CorTypo
  – Labels fully language-internal
    • defined for each language in the Table of Categories
      – categories’ definitions can be compared
  – New functionalities in Elan-CorpA (constructions)
  – Separation between L-internal (corpus) and cross-linguistic (database) levels
  – Database providing domains, functions and forms + queries for retrieval of corpus data
  – Comparison guided through the database
    • which categories are comparable is prepared by the project members (through functional domains)
    • But other databases could be interfaced with the corpus (project open to further developments)
References


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)