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

Amina Mettouchi & Christian Chanard
Aim of the presentation

• How can query languages used in computer queries help formulate, test and falsify linguistic claims about morphosyntactic constructions?

• In what respect is the annotation/query process itself an instance of the scientific method?

• How does this way of annotating and querying data illustrate the importance of fruitful collaboration between linguists and computer analysts?
THE NATURE AND TREATMENT OF DATA

Mettouchi & Chanard 2019
Texts, recordings, corpora

- Well-described languages with a written tradition (e.g. POLIQARP) ([http://nkjp.pl/poliqarp](http://nkjp.pl/poliqarp))
  - Texts (orthographic) presented without a translation
  - corpora: tagging (& parsing)

---

**Context**

Nie... przecież już się zdecydowałam... ŁUKASZ A szczerze...? AGNIESZKA PO CHWILI. Boję się to razy bardziej niż przed klasówką z chemii...
Nie lubiłam chemii... ŁUKASZ Chciałem ci tylko przypomnieć, że _na_ studiach dziennikarskich chemia jest w bardzo okrojonym zakresie...
AGNIESZKA Żnów zaczynasz po swojemu... przypomina sobie Zresztą ty mnie tu zagadujesz, a za chwilę Kasia dostanie lody zamiast mleka. Agnieszka wybiera do salonu. Łukasz spogląda na Kasię, Kasia patrzy na Łukasza
Spoken corpora

- Major languages
  - orthography
  - no translation

Transcriber: http://trans.sourceforge.net/
Praat: http://www.fon.hum.uva.nl/praat/

- Indigenous languages, oral recordings
  - transcribed

  » type of transcription (phonetic, phonological, morpho-phonological, orthographic...)
  » units (word, constituent, discourse, ...)

Mettouchi & Chanard 2019
Role of translation (understand L1?); Provide a stylistic ‘equivalent’ to L1?

Translatability, alignment L1/L2 (units)

Type of annotation
- Semantic?
- Functional?
- Syntactic?
- Other?

Granularity
A corpus: what for?

Mettouchi & Chanard 2019

Elan: http://tla.mpi.nl/tools/tla-tools/elan/
• Reflecting the spoken nature of the data → segmented into intonation units
• Searchable at the level of morphosyntax → morphosyntactically annotated
  – Information Technology: development of Elan-CorpA = ELAN + addition of an internal parser linked to a lexicon, for semi-automatic interlinearization + a query language
• Yielded unexpected findings at the level of morphosyntax/prosody interface
  • http://corpafroas.huma-num.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
Queries in Elan-CorpA
Elan-CorpA WebSearch

http://corpafroas.huma-num.fr/Archives/help.html

- **The Search form**
  - **case sensitive**: uppercase and lowercase are not equivalent
  - **regular expression**: how the search targets and contexts must be interpreted (cf. bottom of the page)
  - **minimal duration**: search only in units of this minimal duration (0 = any duration)
  - **maximal duration**: search only in units of this maximal duration (0 = any duration)

- **The command line searching**: in this box, one can write a query in the specific *CorpA query language*. In the screenshot below, search is: look for the label OBL in the gloss tiers type (ge) fully aligned with whatever (.) in the morphem tiers type (mb)
- **The graphical searching interface**: it is the same graphical interface than ELAN’s multiple files, multiple layers one.

Target: searched sequence. Don’t forget to specify, at the right of the layer, the tier (or tier type) where you want to search for this (morpheme, word, gloss or category tiers...). (The screenshot shows how to express in this graphical interface, the same request as command line searching.

show me the morphemes (mb=.) that are labelled OBL (ge=OBL)

Mettouchi & Chanard 2019
Regular expressions

Regular expressions provide flexible means to match strings of text like 'beginning with', 'ending with', 'any from a list'... By default, the sequences in target or context boxes are searched inside the annotations. Then, for example, searching label 'PFV' will retrieve also the annotations labelled 'IPFV', 'IPFV.3SG.F'...

∀b is a mark for a word frontier (beginning, end, punctuations). ex: \bIPF\b = only the 'IPFs' inside an annotation (complex or not)

^ means beginning of the text. ex: ^N = all annotations beginning by N; ^ = all annotations that are suffixes (in this corpus, prefixes present a hyphen to the right, suffixes, a hyphen to the left)

$ end of text. ex: -$ = all annotations that are prefixes

. any single character. (if nothing after, it will be interpreted as 'any sequence of characters')

\ or [.] the character ','

? the previous character or no character. e.g.: 'gr?ave' will match 'gave' and 'grave'

[?] the character '?'

+ the previous character at least one time. e.g.: 'me+t' will match 'met', 'meet'...

[aeiou] one of these vowels. e.g.: '[aeiou]pe' will match all the words 'pape', 'pipe', 'pepe'...

[^ptk] any character but 'p', 't' or 'k'

[a-h] any letter between 'a' and 'h'

NOT() annotation not containing the text between parenthesis. ex: NOT(\.) in rx or ge = the plain annotations (not complex, i.e without '.')

Multiple layer search

You can refine your initial search by adding vertical constraints in the layer below. In this case, you will have to choose the type of constraint you want to impose to the targets.

- Fully aligned: both annotation cells must have the same temporal duration
- Inside: the upper cell must be a child of the bottom one (like 'ge' child of 'mb')
- Within: the upper cell must be a parent of the upper one (like 'mot' parent of 'mb')
- Overlap: there must be a temporal overlap between the upper cell and the bottom cell
How does ELAN handle queries?

Each selected annotation in the research domain of the ELAN file is entered into the ‘annotations’ table of a PostgreSQL relational database, with the id of the tier it belongs to, its order in the tier and its begin and end times.

The ELAN queries are then transformed into SQL queries. Begin and end times will allow querying with vertical constraints and order in the tier will allow querying with horizontal constraints.

Mettouchi & Chanard 2019
Now we can search anything… 
… or can we?

- What you coded is what you get!
- «Basic/simple» question: size of the corpus
  - How many «words» in the corpus?
    - what is a word?
    - Is ‘word’ the best indicator of corpus size?

<table>
<thead>
<tr>
<th>File</th>
<th>word/sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONV01</td>
<td>2.848</td>
</tr>
<tr>
<td>NARR01</td>
<td>2.228</td>
</tr>
<tr>
<td>NARR02</td>
<td>2.375</td>
</tr>
<tr>
<td>NARR03</td>
<td>2.720</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Size</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>KAB_AM_aLISTOFGLOSSES.PDF</td>
<td>1.175 Ko</td>
<td></td>
</tr>
<tr>
<td>KAB_AM_CONV_01</td>
<td>1384 words</td>
<td>8:06 mn</td>
</tr>
<tr>
<td>KAB_AM_NARR_01</td>
<td>1803 words</td>
<td>13:29 min</td>
</tr>
<tr>
<td>KAB_AM_NARR_02_MIDGET</td>
<td>1748 words</td>
<td>12:16 mn</td>
</tr>
<tr>
<td>KAB_AM_NARR_03</td>
<td>2502 words</td>
<td>15:20 mn</td>
</tr>
</tbody>
</table>

Mettouchi & Chanard 2019
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

- metouchi & chanard 2019
So what is a ‘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)

– "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)
Criteria  (D&A 2002:10)

• **Phonological word**
  – *Segmental features* – internal syllabic and segmental structure; phonetic realisations in terms of this; word boundary phenomena; pause phenomena.
  – *Prosodic features* – stress (or accent) and/or tone assignment; prosodic features such as nasalisation, retroflexion, vowel harmony.
  – *Phonological rules* – some rules apply only within a phonological word; others (external sandhi rules) apply specifically across a phonological word boundary.

• **Grammatical word** consists of a number of grammatical elements which:
  – *always occur together*, rather than scattered through the clause (the criterion of cohesiveness);
  – *occur in a fixed order*;
  – *have a conventionalised coherence and meaning.*
In Kabyle

• Prosodic/phonological word
  – its delimitation is based on rules of schwa insertion, sandhi phenomena, pharyngealization, the presence of a word stress

• Grammatical word
  – delimited on the basis of morphemic status: a lexical or grammatical base and its affixes and clitics are considered a grammatical word

Mettouchi & Chanard 2019
PROSODIC-PHONOLOGICAL WORDS QUERY for number of phono-prosodic words, excluding (length of) pauses, boundary markers, breath intakes...

<table>
<thead>
<tr>
<th>Nr</th>
<th>Annotation</th>
<th>Temps de d...</th>
<th>temps de fin</th>
<th>Durée</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>BI-435 / θura?aminan /</td>
<td>00:00:00.585</td>
<td>00:00:01.021</td>
<td>00:00:00.436</td>
</tr>
<tr>
<td>3</td>
<td>θura?aminan /</td>
<td>00:00:01.021</td>
<td>00:00:01.637</td>
<td>00:00:00.616</td>
</tr>
<tr>
<td>4</td>
<td>amidawiy θamajahu /</td>
<td>00:00:01.637</td>
<td>00:00:02.589</td>
<td>00:00:00.952</td>
</tr>
<tr>
<td>5</td>
<td>θaseβ’ta θaqifin //</td>
<td>00:00:02.589</td>
<td>00:00:03.417</td>
<td>00:00:00.828</td>
</tr>
<tr>
<td>6</td>
<td>1314</td>
<td>00:00:03.417</td>
<td>00:00:04.731</td>
<td>00:00:01.314</td>
</tr>
<tr>
<td>7</td>
<td>amajahu //</td>
<td>00:00:04.731</td>
<td>00:00:05.610</td>
<td>00:00:00.879</td>
</tr>
<tr>
<td>8</td>
<td>519</td>
<td>00:00:05.610</td>
<td>00:00:06.130</td>
<td>00:00:00.520</td>
</tr>
<tr>
<td>9</td>
<td>BI-399</td>
<td>00:00:06.130</td>
<td>00:00:06.529</td>
<td>00:00:00.399</td>
</tr>
<tr>
<td>10</td>
<td>windd innanahu //</td>
<td>00:00:06.529</td>
<td>00:00:07.576</td>
<td>00:00:01.047</td>
</tr>
<tr>
<td>11</td>
<td>333</td>
<td>00:00:07.576</td>
<td>00:00:07.910</td>
<td>00:00:00.334</td>
</tr>
<tr>
<td>12</td>
<td>θamajahušiwi ασελfu /</td>
<td>00:00:07.910</td>
<td>00:00:09.419</td>
<td>00:00:01.509</td>
</tr>
<tr>
<td>13</td>
<td>ασσεβγ αθοσαरu //</td>
<td>00:00:09.419</td>
<td>00:00:10.500</td>
<td>00:00:01.081</td>
</tr>
<tr>
<td>14</td>
<td>697</td>
<td>00:00:10.500</td>
<td>00:00:11.198</td>
<td>00:00:00.698</td>
</tr>
<tr>
<td>15</td>
<td>BI-501</td>
<td>00:00:11.198</td>
<td>00:00:11.699</td>
<td>00:00:00.501</td>
</tr>
<tr>
<td>16</td>
<td>jippwass sęgussanəppwiw a?aminan /</td>
<td>00:00:11.699</td>
<td>00:00:13.571</td>
<td>00:00:01.872</td>
</tr>
</tbody>
</table>

Search:
- sequence of characters which is NOT a number, a slash, a hashtag, or a hyphen (breath intake: BI-123)
- ^$: beginning and end
- ^...$: sequence of characters that doesn't exist in the list
- ^...$ = only one character that doesn't exist in the list

Mettouchi & Chanard 2019
PROSODIC-PHONOLOGICAL WORDS QUERY for number of phono-prosodic words, excluding (length of) pauses, boundary markers, breath intakes...

Search: sequence of character which is NOT a number, a slash, a hashtag, or a hyphen (breath intake: BI-123)
^$ : beginning and end
[^...]+ : sequence of characters that doesn’t exist in the list...
[^...]= only one character that doesn’t exist in the list...

<table>
<thead>
<tr>
<th>Nr</th>
<th>Annotation</th>
<th>Temps de d...</th>
<th>temps de fin</th>
<th>Durée</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>BI-435</td>
<td>00:00:00.585</td>
<td>00:00:01.021</td>
<td>00:00:00.436</td>
</tr>
<tr>
<td>3</td>
<td>θura?aminan /</td>
<td>00:00:01.021</td>
<td>00:00:01.637</td>
<td>00:00:00.616</td>
</tr>
<tr>
<td>4</td>
<td>amidawiγ θamajahūts /</td>
<td>00:00:01.637</td>
<td>00:00:02.589</td>
<td>00:00:00.952</td>
</tr>
<tr>
<td>5</td>
<td>ənsəβa əsepʃin //</td>
<td>00:00:02.589</td>
<td>00:00:03.417</td>
<td>00:00:00.828</td>
</tr>
<tr>
<td>6</td>
<td>1314</td>
<td>00:00:03.417</td>
<td>00:00:04.731</td>
<td>00:00:01.314</td>
</tr>
<tr>
<td>7</td>
<td>amajahu //</td>
<td>00:00:04.731</td>
<td>00:00:05.610</td>
<td>00:00:00.879</td>
</tr>
<tr>
<td>8</td>
<td>519</td>
<td>00:00:05.610</td>
<td>00:00:06.130</td>
<td>00:00:00.520</td>
</tr>
<tr>
<td>9</td>
<td>BI-399</td>
<td>00:00:06.130</td>
<td>00:00:06.529</td>
<td>00:00:00.399</td>
</tr>
<tr>
<td>10</td>
<td>windd innanahu //</td>
<td>00:00:06.529</td>
<td>00:00:07.576</td>
<td>00:00:01.047</td>
</tr>
<tr>
<td>11</td>
<td>333</td>
<td>00:00:07.576</td>
<td>00:00:07.910</td>
<td>00:00:00.334</td>
</tr>
<tr>
<td>12</td>
<td>θamajahūtsiw aiseļhu /</td>
<td>00:00:07.910</td>
<td>00:00:09.419</td>
<td>00:00:01.509</td>
</tr>
<tr>
<td>13</td>
<td>aṭsfeγ afθusaru //</td>
<td>00:00:09.419</td>
<td>00:00:10.500</td>
<td>00:00:01.081</td>
</tr>
<tr>
<td>14</td>
<td>697</td>
<td>00:00:10.500</td>
<td>00:00:11.198</td>
<td>00:00:00.698</td>
</tr>
<tr>
<td>15</td>
<td>BI-501</td>
<td>00:00:11.198</td>
<td>00:00:11.699</td>
<td>00:00:00.501</td>
</tr>
<tr>
<td>16</td>
<td>jippwass segussaneppwin a?aminan /</td>
<td>00:00:11.699</td>
<td>00:00:13.571</td>
<td>00:00:01.872</td>
</tr>
</tbody>
</table>

Mettouchi & Chanard 2019
PROSODIC-PHONOLOGICAL WORDS QUERY for number of phono-prosodic words, excluding (length of) pauses, boundary markers, breath intakes...

An annotation on tier "motproso" that matches regular expression ^[^0-9-/#]+$
Comparability

- Kabyle (Berber - AA) vs Zaar (Chadic-AA)

<table>
<thead>
<tr>
<th></th>
<th>Kabyle Narr01</th>
<th>Zaar Narr01</th>
</tr>
</thead>
<tbody>
<tr>
<td>File duration</td>
<td>809 seconds</td>
<td>765 seconds</td>
</tr>
<tr>
<td>Words</td>
<td>1805</td>
<td>2052</td>
</tr>
<tr>
<td>Morphemes</td>
<td>6639</td>
<td>2617</td>
</tr>
<tr>
<td>Words/sec</td>
<td>2.2</td>
<td>2.7</td>
</tr>
<tr>
<td>Morphemes/sec</td>
<td>8.2</td>
<td>3.4</td>
</tr>
</tbody>
</table>

- Kabyle prosodic vs grammatical words in NARR01

<table>
<thead>
<tr>
<th>Narr01 Kabyle</th>
<th>809 seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phono-prosodic words</td>
<td>1152</td>
</tr>
<tr>
<td>Gramm. words</td>
<td>1805</td>
</tr>
<tr>
<td>PhoPrW/sec</td>
<td>1.4</td>
</tr>
<tr>
<td>GrW/sec</td>
<td>2.2</td>
</tr>
</tbody>
</table>

Mettouchi & Chanard 2019
• How many « words » in the corpus?
  – what is a word?
    • there is no single definition of word => it’s important to know what we are referring to when counting
    • and especially when comparing counts across languages
  – Is ‘word’ the best indicator of a spoken corpus size/complexity? Of course duration is problematic because of variation in speech rates, but if we are interested in how much a corpus can yield in terms of linguistic information, better use e.g.
    • **Number of Intonation Units** (cf Chafe 1994: 63-69) = the IU is a cognitive processing unit, at most one active focus of consciousness per IU (‘one new idea’) (information density)
    • **Number of morphemes**
      – captures best grammatical+lexical information
      – allows comparability between morphologically complex (e.g. polysynthetic) vs morphologically simple languages
        » however, grammatical information is not limited to morphemes


Mettouchi & Chanard 2019
Scientific method

• This first, basic exploration already shows that
  – there is no such thing as « raw data »
  – even apparently noncontroversial units such as ‘word’ are already quite problematic
  – any data analysis should provide transparent information about the ‘data’
  – comparability can only be achieved if the compared data are carefully defined

- Characterizations (observations, definitions, and measurements of the subject of inquiry)
- Hypotheses (theoretical, hypothetical explanations of observations and measurements of the subject)
- Predictions (inductive and deductive reasoning from the hypothesis or theory)
- Experiments (tests of all of the above)
  1. Define a question
  2. Gather information and resources (observe)
  3. Form an explanatory hypothesis
  4. Test the hypothesis by performing an experiment and collecting data in a reproducible manner
  5. Analyze the data
  6. Interpret the data and draw conclusions that serve as a starting point for new hypothesis

Mettouchi & Chanard 2019
STEP 2

MORPHOSYNTACTIC QUERIES AS A WINDOW ONTO THE SCIENTIFIC METHOD
Possible questions involving a grammatical entity

• « How often are objects non-referential in my corpus » ?
• « How often are referents introduced as objects » ?
• « What is the most frequent word order » ?
• However, in order to answer that question, one has to tag ‘Objects ’ (for instance) in a corpus
  – therefore identify them


Mettouchi & Chanard 2019
How do I recognize verbs in the corpus?

• In Kabyle, straightforward
  – always has PNG affix (except for imperative 2.SG)
  – may not be so in other languages

• tagged in the corpus
  – in RX as V+number for verb class/type
**direct object (DO)** An *object traditionally seen as identifying someone or something directly involved in an action or process: e.g. *my books* in *I might leave my books to the library*, where it is distinguished from the *indirect object to the library*. Hence, in particular, the object typically next to the verb in English, one marked by the accusative case in German, and so on. (MATTHEWS, 2007, p. 106)

**object (O)** 1. An element in the basic sentence construction of a language such as English which characteristically represents someone or something, other than that represented by the *subject (1)*, that is involved in an action, process, etc. referred to. E.g. *him* in *I met him*; both *her* and *aflower* (respectively the *indirect object and the *direct object*) in *I will give her aflower*; also, on the assumption that it is syntactically the same element, *that I did in I said that I did*. 2. An element seen as standing in a similar relation to a preposition: e.g. *Washington* in *from Washington*. 3. Any element, in any type of language, which characteristically includes the semantic role of *patient*. Cf subject (3): thus, in typological studies, a language may be classified as an *SVO* language simply because that is the commonest order, in texts, of agent, verb, and patient. (MATTHEWS, 2007, p. 272)
patient (P) 1. Noun phrase or the equivalent that identifies an individual etc. undergoing some process or targeted by some action. E.g. the house is a patient in I painted the house; Mary in I kissed Mary. 2. Thence of a syntactic role which is characteristically that of a patient. E.g. a direct object in English tends to be a patient, especially a patient rather than an *agent.

Therefore direct objects and elements in other languages which are in this respect equivalent to them may be called, in general, patients.

The sense is that of Latin patiens, ‘suffering’ or ‘undergoing’. Abbreviated to P especially in cross-linguistic studies, where opp. A for *agent (2); also opp. S (3). (MATTHEWS, 2007, p. 290)
Fuzzy annotation ...

<table>
<thead>
<tr>
<th>KAB_AM_NARR_02_Midget_717</th>
</tr>
</thead>
<tbody>
<tr>
<td>jëbb°asënduq //</td>
</tr>
<tr>
<td>jëwwi</td>
</tr>
<tr>
<td>i-</td>
</tr>
<tr>
<td>SBJ.3SG.M</td>
</tr>
<tr>
<td>PRO</td>
</tr>
<tr>
<td>V14</td>
</tr>
<tr>
<td>OBJECT</td>
</tr>
</tbody>
</table>

he brought a chest

that apple he gave it to those women

... will yield fuzzy results

Mettouchi & Chanard 2019
The scientific method

Hypothesis

Tests

Results

Annotation

Primary Data

Implement

YES

ANNOTATED DATA

Mettouchi & Chanard 2019
‘Object’ (?) bound pronoun

- Several pronominal paradigms in Kabyle
  - One of them: **undergoer** of verbal active transitive,
  - and **sole argument** of some non-verbal constructions

---

```
ibbwint / (KAB_AM_NARR_01_0726)
jewwitent /
i- wwi =tnt /
SBJ.3SG.M bring\PFV ABS.3PL.F /
PRO V14 PRO /
```

he took them,

---

```
ikkerd jufad mussisulafigent / (KAB_AM_NARR_01_0901)

ikkerd
i- kkr =dd
SBJ.3SG.M stand_up\PFV PROX

jufadd
i- ufa =dd
SBJ.3SG.M find\PFV PROX

jessis
jessi -s
daughter\PL KIN.3SG

ulajfigent
ulafig =itent /
PRO V14 PRO /
N.KIN.COV PRO
```

the father woke up and found that his daughters were no longer there,
• If object of verbal and sole argument (≈ subject) of non-verbal predicate then cannot be an ‘object’

• Cannot be non-referential
  – cf negative existential (& other contexts)
  – ➽ ABS PRO = REFERENTIAL UNDERGOER

Mettouchi & Chanard 2019
What about the other candidates?

- **Noun coreferent to an absolutive pronoun**
  
  Mettouchi & Chanard 2019

- **Noun directly following the verb**

<table>
<thead>
<tr>
<th>NOUN</th>
<th>COORDS</th>
<th>DEM</th>
<th>PRO</th>
<th>PRO</th>
<th>PRO</th>
<th>V14</th>
<th>DEM</th>
<th>PRO</th>
<th>N.OV</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

He gave the apples to his wives.

Mettouchi & Chanard 2019
Nominal coreferent to a PRO

• Coreference: same Number and Gender in Noun and in Pronoun
  – SG.F ⇐ 3.SG.F
  – SG.M ⇐ 3.SG.M
  – PL.F ⇐ 3.PL.F
  – PL.M ⇐ 3.PL.M

• Various situations in the corpus
  – \( N_{\text{absl}} \) before the verb+ABS\(_{\text{pro}}\) in the same IU
  – \( N_{\text{absl}} \) before the verb+ABS\(_{\text{pro}}\) in a ≠ IU
  – \( N_{\text{ann}} \) after the verb+ABS\(_{\text{pro}}\) in a ≠ IU
Coreference with ABS.PRO

- **NOUN in ABSL before Verb**

  \[
  \begin{align*}
  \text{asenduqagi} & / (KAB\_AM\_NARR\_02\_Midget\_783) \\
  \text{asenduq} & \\
  \text{boxABS.L.G.M PROX1} & \\
  \text{N.OV} & \text{AFFX} \\
  \end{align*}
  \]

  this box,

- **Noun in ANN after Verb**

  \[
  \begin{align*}
  \text{tufa damj} & / (KAB\_AM\_NARR\_01\_0413) \\
  \text{tufa} & \text{d amjj} \ n \ \text{wadrar} \\
  \text{SBJ.3SG.F findV} & \text{COP catABS.L.G.M} \ \text{GEN mountainANN.S.G.M} \\
  \text{PRO} & \text{V13%} \ \text{PRD N.OV} \ \text{PREP N.OV} \\
  \end{align*}
  \]

  she found it was the Mountain Cat

- **Very different constructions**

  \[
  \begin{align*}
  \text{att\'awi\'t} & \ \text{i}jammak // (KAB\_AM\_NARR\_02\_Midget\_784) \\
  \text{adt} & \text{tawid} \ \text{i} \ \text{jammak} \\
  \text{POT.3SG.M} & \text{SBJ.2G} \ \text{bringV} \ \text{SBJ.2G} \ \text{DAT} \ \text{motherANN.S.G.F KIN.2G.M} \\
  \text{PTCL.PRO} & \text{CIRC1V14} \ \text{CIRC2} \ \text{DEMPRO N.KIN.COV} \ \text{PRO} \\
  \end{align*}
  \]

  you will take it to your mother

  \[
  \begin{align*}
  \text{ibiza\'yen} & / (KAB\_AM\_NARR\_01\_0415) \\
  \text{it} & \text{izadyan} \\
  \text{REL.REAL.3SG.M} & \text{RELSBJ.POS dwellV} \ \text{RELSBJ.POS} \\
  \text{DEMPRO.PRO} & \text{CIRC1} \ \text{V23} \ \text{CIRC2} \\
  \end{align*}
  \]

  who inhabited it

  \[
  \begin{align*}
  \text{wax\'amnn} & / (KAB\_AM\_NARR\_01\_0416) \\
  \text{wax\'amnn} & \\
  \text{wax\'am} & \text{-nni} \\
  \text{houseANN.S.G.M} & \text{N.OV} \ \text{DEMPRO} \\
  \end{align*}
  \]

  the house.

Mettouchi & Chanard 2019
• NOUN in ABSL before Verb (and prosodic boundary), with same Gender-Number on pronoun

My brothers, this box,

you bought them things,

you will take it to your mother

• Share information structure function – subtopic shift (Mettouchi 2015)

• **NOUN in ANN after Verb (and prosodic boundary), with same Gender-Number on pronoun**

```
tufa ḍamṭibudrar // (KAB_AM_NARR_01_0413)
tufa  d  amṭif  n  wadrar //
t-  ufa  d  amṭif  n  wadrar //
SBJ.3SG.F.findPFV COP cat\ABSL.SG.M GEN mountain\ANN.SG.M //
PRO   V13%  PRED N.OV   PREP N.OV //
```

She found it was the Mountain Cat

423  (KAB_AM_NARR_01_0414)
423

```
iṭiḏadyn   // (KAB_AM_NARR_01_0415)
it   izadyn   / /
 i   =t   zdāy  -n / /
REL.REAL.ABS.3SG.M RELSBJ.POS dwall\IPFV RELSBJ.POS /
DEMPRO PRO  CIRC1   V23   CIRC2 /
```

who inhabited it,

```
waxxamnni // (KAB_AM_NARR_01_0416)
waxxamnni   //
waxxam  -nni //
house\ANN.SG.M.CNS //
N.OV   DEM //
```

the house.

```
damadaṭ# ### (KAB_AM_NARR_01_0444)
təddmd  aṭ#  ###
t-  ddəm  =dd  aṭ#  ###
SBJ.3SG.F.grasp\PFV PROX  XXX  ###
PRO   V23   PTCL  XXX  ###
```

She took

```
θa# əsamtnən itsəmmultan  akkən  / (KAB_AM_NARR_01_0445)
ta#  tasamtnən  itsəmmut  akkən  /
ta#  tasamta  -nni  -t  tsəmmut  akk  -ən  /
XXX  pillow\ABSL.SG.F.CNS  SBJ.3SG.M use\as\pillow\IPFV  thus  DIST  /
XXX  N.OV   DEM PRO   V24.PFX.APHO  ADV.AFFX  /
```

the pillow on which he slept,

```
wamṭifjannī   // (KAB_AM_NARR_01_0446)
wamṭifjnī   //
wamṭifj  -nni   /
cat\ANN.SG.M.CNS   /
N.OV   DEM   /
```

the cat.

• **Share information structure function – referent reactivation (Mettouchi 2015)**

Mettouchi & Chanard 2019
Nominal coreferent to ABS PRO

- various formal configurations
  - $N_{\text{absi}}$ before the verb+ABS$_{\text{pro}}$ in the same IU
  - $N_{\text{absi}}$ before the verb+ABS$_{\text{pro}}$ in a $\neq$ IU
  - $N_{\text{ann}}$ after the verb+ABS$_{\text{pro}}$ in a $\neq$ IU
    - can we label all ‘object’ on the basis of them having a common referent which is an undergoer?

- Regardless of the type of coreferent pronoun (and the putative grammatical relation), the following constructions have the following IS functions
  - Noun$_i$/pro$_i$-PGV $\rightarrow$ recapitulation of previously introduced situation
  - Noun$_i$/pro$_i$-PGV $\rightarrow$ contrastive comments
  - pro$_i$-PGV / Noun$_i$ $\rightarrow$ reactivation of semi-active or inactive referent

Mettouchi & Chanard 2019
Analytic decisions

• Label the pronoun ABS (absolutive), with the semantic value ‘referential undergoer’
  – because it can be both 'object' and ‘subject’
• Not label nouns coreferent to ABS pronouns ‘object’, because formal means point to other functions for the constructions, at Information Structure level AND there’s no ‘object’ case
  – regardless of the nature of the pronoun (SBJ, ABS, DAT, POSS, KIN, PREP …)
    • leave them simply labelled by their morphological marks : State (absolute or annexed) + Gender and Number
OTHER NOMINALS

- NOUN in ABSL directly following the verb

Select All _DUPLICATE_  Show selected items  EXPORT RESULTS

- KAB_AM_CONV_01.EAF:

  - #1 [V13%] [N.KIN.COV] #2 [leave\PFV] [grandfather\ABSL.SG.M] #3 [iggadd] [jaddi] { KAB_AM_CONV_01_SP1_095 }
  - #1 [V14] [N.KIN.COV] #2 [take\PFV] [grandmother\ABSL.SG.F] #3 [juy] [sa\ddi] { KAB_AM_CONV_01_SP1_070 }
  - #1 [V14] [N.KIN.COV] #2 [take\PFV] [grandmother\ABSL.SG.F] #3 [juy] [sa\ddi] { KAB_AM_CONV_01_SP1_072 }
  - #1 [V14] [N.KIN.COV] #2 [take\PFV] [mother\ABSL.SG.F] #3 [juydd] [janna] { KAB_AM_CONV_01_SP1_076 }
  - #1 [V14] [N.KIN.COV] #2 [take\PFV] [father\ABSL.SG.M] #3 [juy] [3\ddi] { KAB_AM_CONV_01_SP1_112 }
  - #1 [V14] [N.KIN.COV] #2 [take\PFV] [mother\ABSL.SG.F] #3 [juy] [sa\ddi] { KAB_AM_CONV_01_SP1_116 }

jaddmottatsaffaht / (KAB_AM_NARR_02_Midget_029)

- jaddmadd  
ta\jaffaht /
- i-  
ddem =dd  ta\jaffaht /
- SBJ.3SG.M\grasp\PFV PROX apple\ABS\SG.F /
- PRO  V23  PTCL  N.OV /

He took an apple,

Mettouchi & Chanard 2019
The scientific method

- Hypothesis
- Tests
- Results
- Primary Data
- Implement

Annotation

ANOTATED DATA

YES

NO

Mettouchi & Chanard 2019
Query: \( N_{\text{abs}} \) following \( V \) immediately

Exactly similar function with three other structures:

- \( V \) Noun\(_{\text{Ann}} \) Noun\(_{\text{Abs}} \)
- \( V \) ADV Noun\(_{\text{Abs}} \)
- \( V \) POSTNEG Noun\(_{\text{Abs}} \)

he told him that his brothers had seen a carpet at Jemma Nuja’s place,

Mettouchi & Chanard 2019
$N_{\text{abs}}$ following $V$ (separated by $N_{\text{ann}}$)

Mettouchi & Chanard 2019
Exactly similar function with three other structures:

- V Noun<sup>Ann</sup> Noun<sub>Abs</sub>
- V ADV Noun<sub>Abs</sub>
- V POSTNEG Noun<sub>Abs</sub>

Mettouchi & Chanard 2019
\( N_{abs} \) following \( V \) (separated by ADV)

Search Results:
15 hits in 4 file(s)

\[ \text{rx = \text{ADV} & ge=. < mot=} \{ \text{rx < 3 & ge < 3 & mot=1} \} [\text{rx=ADV} & ge=. < mot=} \{ \text{rx < 3 & ge < 3 & mot=1} \} [\text{rx=\text{ADV} & ge=. < mot=} \{ \text{rx < 3 & ge < 3 & mot=1} \} [\text{rx=\text{ADV} & ge=. < mot=} \{ \text{rx < 3 & ge < 3 & mot=1} \} [\text{rx=\text{ADV} & ge=. < mot=} \{ \text{rx < 3 & ge < 3 & mot=1} \}

Select All  □  Duplicate  □  Show selected items  □  EXPORT RESULTS

KAB_AM_NARR_01.EAF:

- #1 [V23] [ADV] [N.OV] #2 [take_away\PFV] [thus] [bread\ABSL.SG.M] #3 [rakks] [akka] [ayrum] \{ KAB_AM_NARR_01_0098 \}
- #1 [V24] [ADV] [N.OV] #2 [be_placed\PFV.CAUS] [thus] [bread\ABSL.SG.M] #3 [isors] [akka] [ayrum] \{ KAB_AM_NARR_01_0119 \}
- #1 [V24] [ADV] [N.OV] #2 [throw\PFV] [thus] [rope\ABSL.SG.M] #3 [to\qaggaras] [akka] [ammar] \{ KAB_AM_NARR_01_0454 \}
- #1 [V14] [ADV] [N.OV] #2 [address\IPFV] [thus] [cat\ABSL.SG.M] #3[to\tal] [akka] [am\ifni] \{ KAB_AM_NARR_01_0596 \}
- #1 [V24] [ADV] [N.OV] #2 [be_tied\PFV] [thus] [dog\ABSL.SG.F] #3 [iqqan] [akka] [taqjint] \{ KAB_AM_NARR_01_0732 \}
- #1 [V14] [ADV] [N.OV] #2 [bring\PFV] [thus] [firewood\ABSL.PL.M] #3 [w\wint] [akcan] [isy\ran] \{ KAB_AM_NARR_01_0808 \}

\( \text{akka am\ifni} \) // (KAB_AM_NARR_01_0596)

- tak\if\n
- t\- tak\- a

- SBJ.3SG.F address\IPFV thus PROXa

- PRO V14.PFX ADV.DEM PRO N.OV DEM //

Mettouchi & Chanard 2019
Exactly similar function with three other structures:
- V Noun_{Ann} Noun_{Abs}
- V ADV Noun_{Abs}
- V POSTNEG Noun_{Abs}

he didn't add a gun to his horse.
N\textsubscript{abs} following V (separated by POSTNEG)

\[
\text{Search Results: } \\
\text{2 hits in 4 file(s)} \\
\]

\[
(rx = \text{\textbackslash V} & \text{ge}. < \text{mot} = ] \{ rx < 3 & \text{ge} < 3 & \text{mot} = 1 \} \{ rx = \text{N} \& \text{ge} = \text{POSTNEG} < \text{mot} = ] \{ rx < 3 & \text{ge} < 3 & \text{mot} = 1 \} \{ rx = \text{N} \& \text{ge} = \text{ABSL} < \text{mot} = ] \\
\]

- KAB\_AM\_NARR\_02\_MIDGET.EAF:
  - #1 [V13%][N.INDF][N.OV] #2 [come\:NEGPFV][POSTNEG][gun\:ABSL\:SG\:F] #3 [sddi][ara][tamughaltis] \{ KAB\_AM\_NARR\_02\_Midget\_081 \}

- KAB\_AM\_NARR\_03\_EAF:
  - #1 [V13%][N.INDF][N.COV] #2 [possess\:NEGPFV][POSTNEG][right\:ABSL\:SG\:M] #3 [ta\:sic][ara][<kroit>] \{ KAB\_AM\_NARR\_03\_0127 \}

\texttt{ursidd\ sddi\ ara\ tamughaltis\ d\ u\text{"u\textsc{diwis}}} //

<table>
<thead>
<tr>
<th>ursidd</th>
<th>sddi</th>
<th>ara</th>
<th>tamughaltis</th>
<th>d</th>
<th>u\text{&quot;u\textsc{diwis}}</th>
</tr>
</thead>
<tbody>
<tr>
<td>ur =-as\ =dd</td>
<td>s- =dd</td>
<td>ara</td>
<td>tamughalt</td>
<td>-is</td>
<td>d</td>
</tr>
</tbody>
</table>

\text{NEG\ DAT.3SG\ PROX CAUS\ come\ NEGPFV POSTNEG\ gun\:ABSL\:SG\:F\ POSS.3SG\ ASSOC\ horse\ANN\:SG\:M\ POSS.3SG\ //}

\text{PTCL\ PRO\ PTCL\ AFFX\ V13\%\ N.INDF\ N.OV\ PRO\ PREP\ N.OV\ PRO\ //}

\text{he\ didn't\ add\ a\ gun\ to\ his\ horse.}

Mettouchi & Chanard 2019
Query: \( N_{\text{abs}} \) following \( V \) inside the same IU

Exactly similar function with disfluencies:
- \( V \ [xxx] \ HESIT / \) Noun\(_{\text{Abs}}\)
- \( V \ [xxx] \ FS (## \ or /) \ Noun_{\text{Abs}}\)

Exactly similar function with high rise:
- \( V \ [xxx] \ / Noun_{\text{Abs}}\)

Mettouchi & Chanard 2019
N_{abs} following V in a different intonation unit (separated by HESIT)

\[
\text{rx = } \text{bV} \quad \& \quad \text{ge = } \text{<mot=1} \quad \{ \text{rx = 3} \quad \& \quad \text{ge = 3} \quad \& \quad \text{mot=1} \} \quad \{ \text{ge = HESIT} \quad \& \quad \text{rx = } \text{<mot=1} \} \quad \{ \text{rx = 1} \quad \& \quad \text{ge = 1} \quad \& \quad \text{mot=1} \} \quad \{ \text{rx = N} \quad \& \quad \text{ge=ABSL} \quad \text{<mot=1} \}
\]

Search Results:
4 hits in 4 file(s)

- **KAB_AM_NARR_02_MIDGET.EAF**: 
  - #1 [V13%] [HESIT] [] [N.OV] #2 [give\AOR] [HESIT] [] [bird\ABSL.PL.M] #3 [fको] [a:::] ['] [if्र]axa} {KAB_AM_NARR_02_Midget_176}

- **KAB_AM_NARR_03.EAF**: 
  - #1 [V14] [HESIT] [] [N.COY] #2 [bring\AOR] [HESIT] [] [beam\ABSL.PL.M] #3 [तविल] [a:::] ['] [ibawau} {KAB_AM_NARR_03_0079}
  - #1 [V23] [HESIT] [] [N.OV] #2 [make\AOR] [HESIT] [] [threshing\_area\ABSL.SG.M] #3 [काठमा] [a:::] ['] [enmar} {KAB_AM_NARR_03_0102}
  - #1 [V24] [HESIT] [] [N.KIN.COY] #2 [go\AOR] [HESIT] [] [grandmother\ABSL.SG.F] #3 [तीह] [a:::] ['] [setṭijagi} {KAB_AM_NARR_03_0797}

\[\text{awnoā}k̩̃y\quad \text{a:::} / (KAB_AM_NARR_02_Midget_176)\]
\[\text{ifr̩}axa / (KAB_AM_NARR_02_Midget_177)\]

- awnoāk̩̃y
- adwān
- ad
- POT
- DAT
- 2.PL.M
- give\AOR
- SBJ.1SG
- HESIT
- PTCL
- PRO
- V13%
- PRO
- HESIT
- and I'll give you
- N.OV
- AFFX
- those birds.

Mettouchi & Chanard 2019
Query: $N_{abs}$ following $V$ inside the same IU

Exactly similar function with disfluencies:
- $V$ [xxx] HESIT (/) Noun$_{Abs}$
- $V$ [xxx] FS (## or /) Noun$_{Abs}$

Exactly similar function with high rise:
- $V$ [xxx] / Noun$_{Abs}$

Example:
- Mettouchi & Chanard 2019
Query: \( N_{\text{abs}} \) following \( V \) inside the same IU

Exactly similar function with disfluencies:
- \( V \) [xxx] HESIT (/) Noun\(_{\text{Abs}}\)
- \( V \) [xxx] FS (## or /) Noun\(_{\text{Abs}}\)

Exactly similar function with high rise:
- \( V \) [xxx] / Noun\(_{\text{Abs}}\)

Mettouchi & Chanard 2019
For the last case (high boundary tone on preceding IU), intonation has to be automatically annotated → in progress.
Accurate Query

QUERY: [rx = \bV & ge=. < mot=.] {rx < 3 & ge < 3 & mot=1} [ rx=N & ge=ABSL < mot =.] OR [rx = \bV & ge=. < mot=.] {rx < 3 & ge < 3 & mot=1} [ rx=N & ge=ANN < mot =.] {rx < 3 & ge < 3 & mot=1} [ rx=ADV & ge=. < mot =.] {rx < 3 & ge < 3 & mot=1} [ rx=N & ge=ABSL < mot =.] OR [rx = \bV & ge=. < mot=.] {rx < 3 & ge < 3 & mot=1} [ rx=N & ge=POSTNEG < mot =.] {rx < 3 & ge < 3 & mot=1} [ rx=N & ge=ABSL < mot =.] OR [rx = \bV & ge=. < mot=.] {rx < 3 & ge < 3 & mot=1} [ rx=## & ge=## < mot =##] {rx=1 & ge = 1 & mot=1} [ rx=N & ge=ABSL < mot =.]

Mettouchi & Chanard 2019
SYNTHESIS

• Definition of the Direct Object
  – a noun in the absolute state, directly following the verb in the same intonation unit, or possibly separated from it by a noun in the annexed state (=nominal subject), an adverb, a postverbal negator
  – its tight relationship with the verb is marked by the fact that if it exceptionally occurs in a separate intonation unit:
    • it is due to a problem in speech production, signalled by HESIT or FS
    • it is an instance of stylistic highlighting (extra-high boundary tone)

• Other candidates for ‘object’ have been dismissed
  – referential undergoer (absolutive pronoun)
  – nouns coreferent to absolutive pronoun

• Use that definition for the re-annotation of the corpus
  – and then launch a query using that annotation coupled with others (depending on the questions we are asking ourselves)

Mettouchi & Chanard 2019
• For a corpus to yield non-trivial and non-aprioristic results, annotation has to be rigorous
  – Annotation should be a heuristic process, not an aprioristic one
  – Definitions should be provided for all categories (especially non-morphologically-marked categories, if they are to be annotated)

• The query language contains the necessary operators, and mirrors the actual formal linguistic definitions (position, co-occurrence etc.)
  – but there are limitations due to:
    • calculation power (time to execute a query might be more than the server’s authorized threshold)

• The scientific dimension of linguistics is not (only) in the tools (I.T. query engines (also: statistics etc.)), although those are a wonderful help, it is in the method, and crucially, in the initial preparation/treatment of the ‘data’, and explicitness of the definitions.