What makes a corpus easy to (re)use?

The case of language acquisition corpora

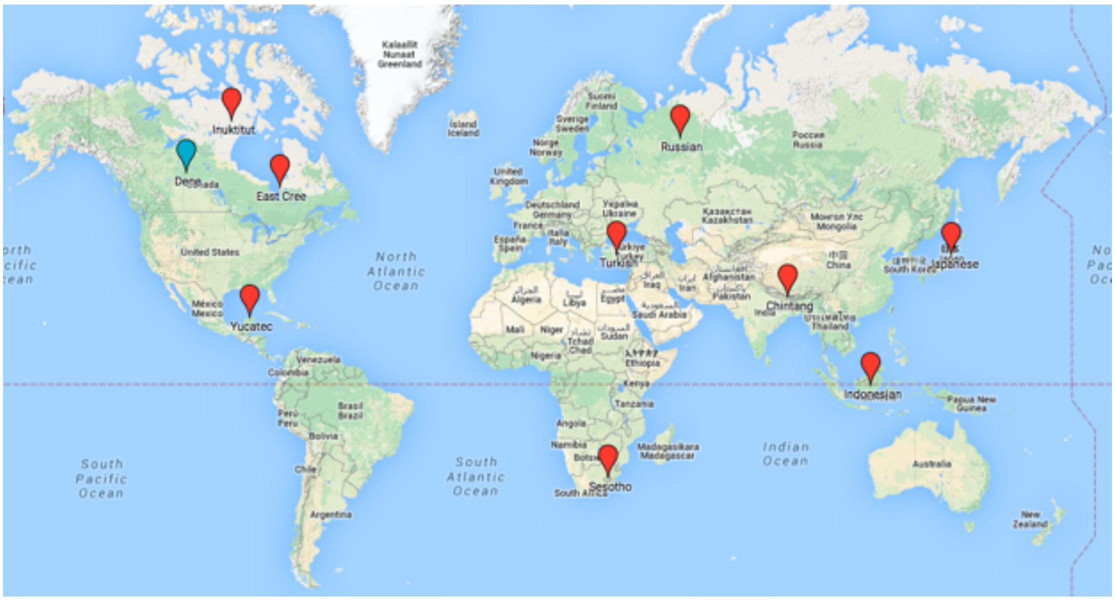
Robert Schikowski & Steven Moran

Introduction

The ACQDIV project

- ACQDIV = **Acq**uisition processes in maximally **div**erse languages
- ERC project, 2014 - 2018/19
- Central question: What's universal in language acquisition?
- Method: Compare acquisition processes in 10 languages that we know to be very different wrt some macrotypological parameters („maximum diversity sampling")
- Data:
  - 11 corpora containing data from children and surrounding adults
  - media recorded in a natural environment (family, home etc.)
  - longitudinal organisation (periodical recordings across ~ 1 year)
  - transcribed, glossed, and (mostly) translated

The ACQDIV languages

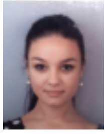


The ACQDIV corpora

Language	Format	Session MD	Speaker MD
Turkish	Quasi-CHAT	Quasi-CHAT	Quasi-CHAT
Japanese (1&2)	Talkbank XML	Talkbank XML	Talkbank XML
Indonesian	Toolbox	CHAT	XLS
Yucatec	Quasi-CHAT	Quasi-CHAT	Quasi-CHAT
Inuktitut	Quasi-CHAT	CHAT	CHAT
Chintang	Toolbox	IMDI	IMDI
Sesotho	Talkbank XML	Talkbank XML	Talkbank XML
Russian	Toolbox	IMDI	IMDI
Dene	Toolbox	CSV	CSV

The ACQDIV team

Andreas Gerster


Katia Mažara

Steven Moran

Danica Pajović

Cazim Hysi

Dagmar Jung

Sabine Stoll

Laura Canedo

Robert Schikowski

Carolin Remensberger

Talk map

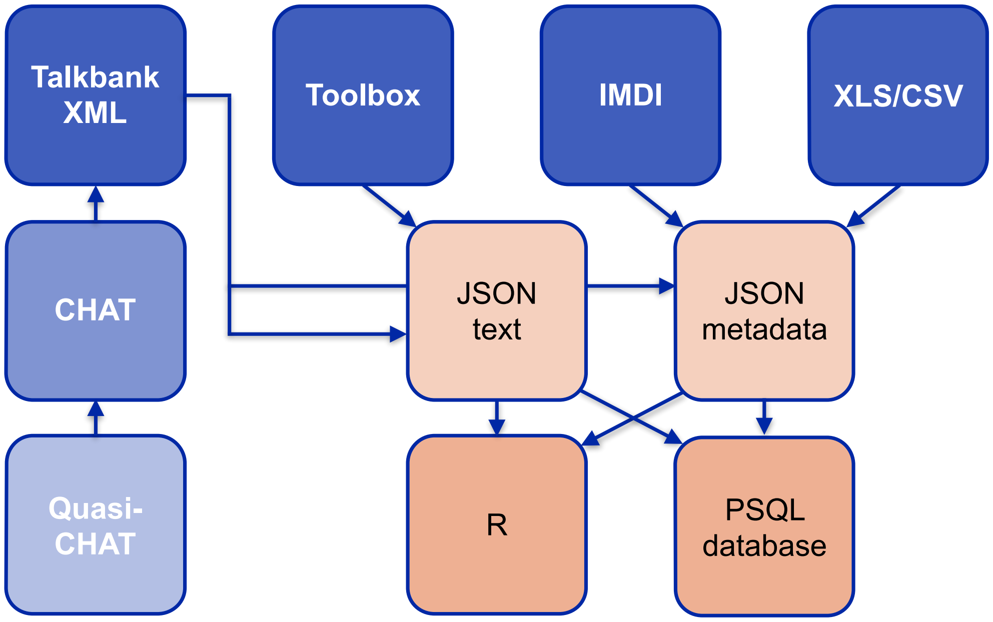
- Transforming corpus data: general steps
- Reusing corpus text: criteria for easy transformability
  - consistency
  - separation of independent contents
  - documentation
  - explicit coding
- Reusing other kinds of corpus data
- Conclusion

Transforming corpus data

Standards are nice, but…

- Corpus standards are a big help for (re)using data, but:
  - all standards only provide a general framework inside of which more or less variation is possible
  - not all standards are equally easy to process
  - data can always be broken if they are not regularly tested for wellformedness
  - linguistic analysis (esp. quantitative analysis) often requires rearrangement of data
  - older data might not be standardized
- ➔ Data transformation is a normal component of any research project involving corpus data!
- ➔ Usability depends on how easy data are to transform, and this question is partially independent of standardization

Transformations in ACQDIV



Talkbank XML

Toolbox

IMDI

XLS/CSV

CHAT

JSON text

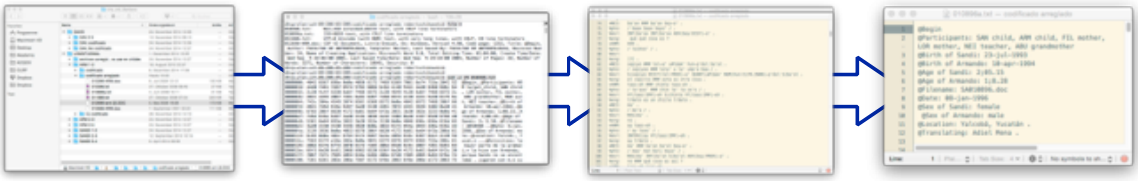
JSON metadata

Quasi-CHAT

R

PSQL database

Knowledge required for transforming a corpus



```
@Begin
@Participants: SAN child, ARM child, FIL mother,
LOR mother, NEI teacher, ABU grandmother
@Birth of Sandi: 23-jul-1993
@Birth of Armando: 10-apr-1994
@Age of Sandi: 2;05.15
@Age of Armando: 1;8.28
@Filename: SAR10096.doc
@Date: 08-jan-1996
@Sex of Sandi: female
 @Sex of Armando: male
@Location: Yalcobá, Yucatán .
@Translating: Adiel Mena .
```

Reusing corpus text

Syntactic vs. semantic structure

- Corpora have a complex semantic structure with the following basic elements:
  - **tiers:** mutually independent types of information coded in the corpus, e.g. transcription or translation.
  - **tier content:** the content of individual tiers, e.g. the words constituting a transcription or translation
  - **ordering relations:** temporal or textual precedence between elements on one tier, e.g. word 1 precedes word 2
  - **hierarchical relations:** associations between elements on different tiers, e.g. morphemes are contained in words
- The semantic structure of a corpus is formally coded by the **corpus syntax** (for instance, XML is one kind of corpus syntax).

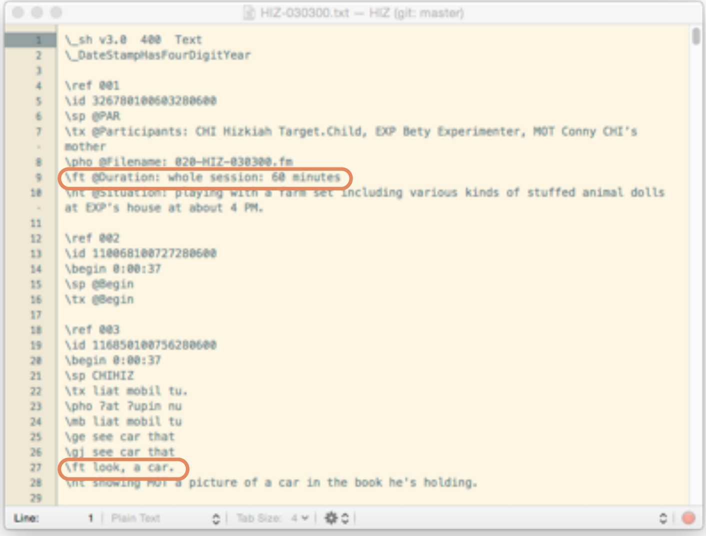
Structures and corpus design

- For making a corpus easy to transform, the following points are important:
  - There should be a **1:1 relationship** between elements of the semantic and elements of the syntactic structure.
  - Semantically **independent tiers** should also be kept independent on the syntactic level, i.e. their contents should not be intermingled.
  - The correspondences between the semantic and the syntactic structure should be **richly documented**.
  - The syntactic coding should be **explicit** rather than implicit.
- But why are these points important?
- Some examples from our project experience…

1:1 relations between syntax/semantics

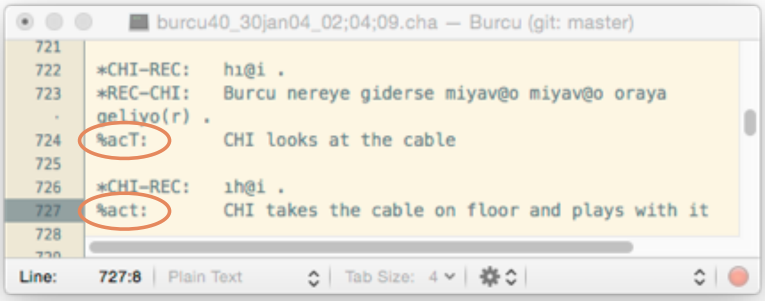
- When the rule of 1:1 relations between syntactic and semantic elements is violated, the coding becomes **inconsistent**:
  - Several syntactic elements correspond to 1 semantic element
    ➔ semantic elements with deviant coding may be overlooked; generating syntax may be problematic
  - 1 syntactic element corresponds to several semantic elements
    ➔ syntactic element becomes ambiguous and can't be dealt with automatically at all!

Example 1: Tier names



```
\_sh v3.0  400  Text
\_DateStampHasFourDigitYear

\ref 001
\id 326780100603280600
\sp @PAR
\tx @Participants: CHI Hizkiah Target.Child, EXP Bety Experimenter, MOT Conny CHI's
mother
\pho @Filename: 020-HIZ-030300.fm
\ft @Duration: whole session: 60 minutes
\nt @Situation: playing with a farm set including various kinds of stuffed animal dolls
at EXP's house at about 4 PM.

\ref 002
\id 110068100727280600
\begin 0:00:37
\sp @Begin
\tx @Begin

\ref 003
\id 116850100756280600
\begin 0:00:37
\sp CHIHIZ
\tx liat mobil tu.
\pho ?at ?upin nu
\mb liat mobil tu
\ge see car that
\gj see car that
\ft look, a car.
\nt showing MOT a picture of a car in the book he's holding.
```

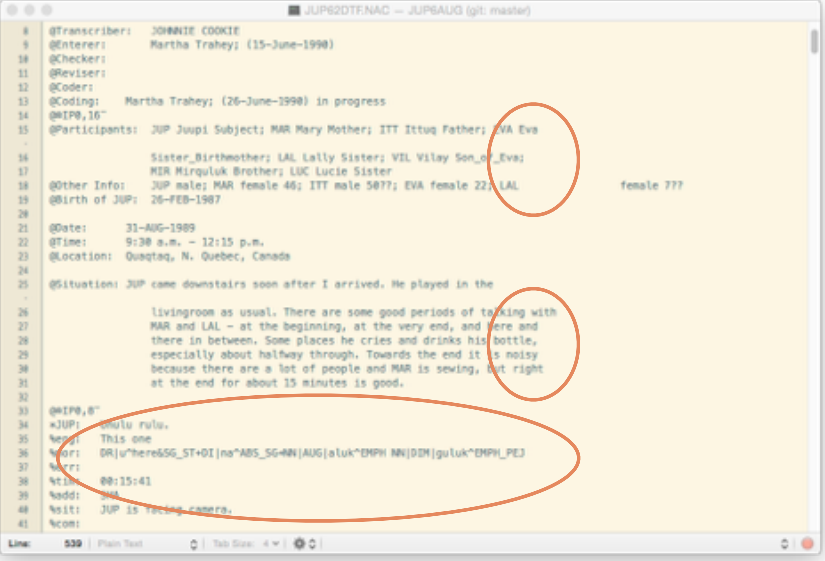
Example 2: Tier names



```
*CHI-REC:   hı@i .
*REC-CHI:   Burcu nereye giderse miyav@o miyav@o oraya
gelivo(r) .
%acT:       CHI looks at the cable

*CHI-REC:   ıh@i .
%act:       CHI takes the cable on floor and plays with it
```

Line: 727:8 | Plain Text | Tab Size: 4

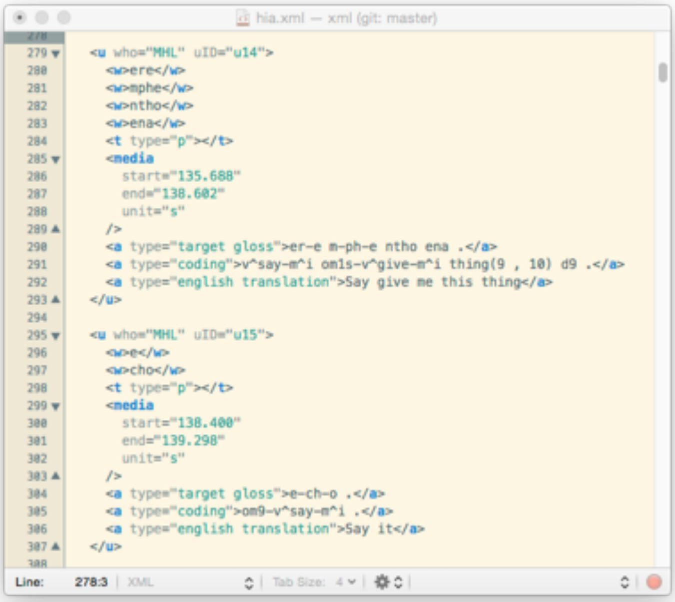
Example 3: Line breaks



```
@Transcriber:   JOHNNIE COOKIE
@Enterer:       Martha Trahey; (15-June-1990)
@Checker:
@Reviser:
@Coder:
@Coding:    Martha Trahey; (26-June-1990) in progress
@@IP0,16"
@Participants:  JUP Juupi Subject; MAR Mary Mother; ITT Ittuq Father; EVA Eva

                Sister_Birthmother; LAL Lally Sister; VIL Vilay Son_of_Eva;
                MIR Mirquluk Brother; LUC Lucie Sister
@Other Info:    JUP male; MAR female 46; ITT male 50??; EVA female 22; LAL          female ???
@Birth of JUP:  26-FEB-1987

@Date:      31-AUG-1989
@Time:      9:30 a.m. - 12:15 p.m.
@Location:  Quaqtaq, N. Quebec, Canada

@Situation: JUP came downstairs soon after I arrived. He played in the

                livingroom as usual. There are some good periods of talking with
                MAR and LAL - at the beginning, at the very end, and here and
                there in between. Some places he cries and drinks his bottle,
                especially about halfway through. Towards the end it is noisy
                because there are a lot of people and MAR is sewing, but right
                at the end for about 15 minutes is good.

@@IP0,8"
*JUP:   Uhulu rulu.
%eng:   This one
%mor:   DR|u^here&SG_ST+DI|na^ABS_SG+NN|AUG|aluk^EMPH NN|DIM|guluk^EMPH_PEJ
%err:
%tim:   00:15:41
%add:   [illegible]
%sit:   JUP is facing camera.
%com:
```

Line: 539 | Plain Text | Tab Size: 4

Example 4: Coding words

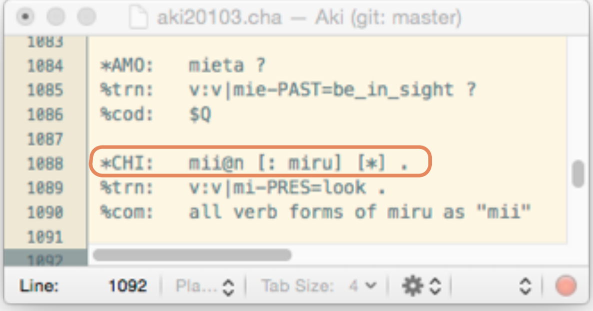


```
<u who="MHL" uID="u14">
  <w>ere</w>
  <w>mphe</w>
  <w>ntho</w>
  <w>ena</w>
  <t type="p"></t>
  <media
    start="135.688"
    end="138.602"
    unit="s"
  />
  <a type="target gloss">er-e m-ph-e ntho ena .</a>
  <a type="coding">v^say-m^i om1s-v^give-m^i thing(9 , 10) d9 .</a>
  <a type="english translation">Say give me this thing</a>
</u>

<u who="MHL" uID="u15">
  <w>e</w>
  <w>cho</w>
  <t type="p"></t>
  <media
    start="138.400"
    end="139.298"
    unit="s"
  />
  <a type="target gloss">e-ch-o .</a>
  <a type="coding">om9-v^say-m^i .</a>
  <a type="english translation">Say it</a>
</u>
```

Keeping tiers independent

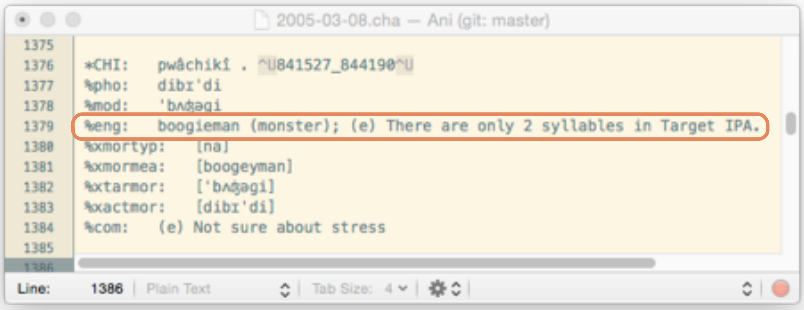
- When tiers that are semantically at least partially independent are not kept separately in the corpus syntax, bad things may happen:
  - It becomes difficult or impossible to pull the tiers apart in the semantics.
  - Content belonging to one tier may mistakenly be interpreted as belonging to a different tier.
  - In the worst case all involved tiers may end up being unanalyzable.

Example 1: Transcriptions and psycholinguistics



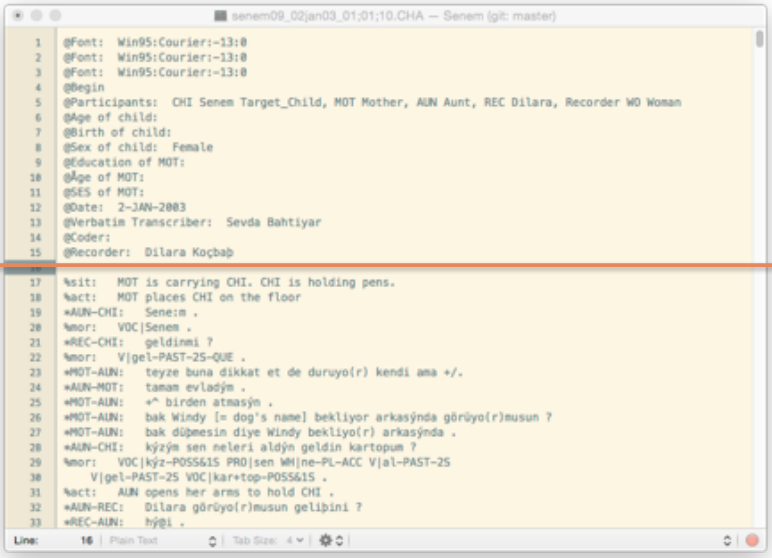
```
*AMO:   mieta ?
%trn:   v:v|mie-PAST=be_in_sight ?
%cod:   $Q

*CHI:   mii@n [: miru] [*] .
%trn:   v:v|mi-PRES=look .
%com:   all verb forms of miru as "mii"
```

Example 2: Translations and metacomments



```

*CHI:   pwâchikî . ^U841527_844190^U
%pho:   dibɪ'di
%mod:   'bʌʤəgi
%eng:   boogieman (monster); (e) There are only 2 syllables in Target IPA.
%xmortyp:   [na]
%xmormea:   [boogeyman]
%xtarmor:   ['bʌʤəgi]
%xactmor:   [dibɪ'di]
%com:   (e) Not sure about stress
```

Example 3: Metadata

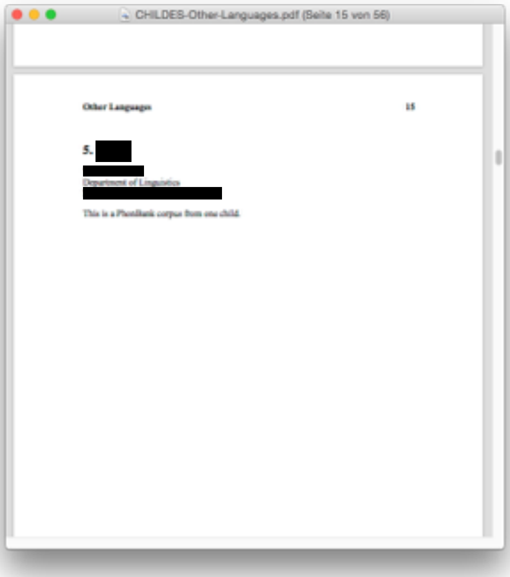


```
@Font:  Win95:Courier:-13:0
@Font:  Win95:Courier:-13:0
@Font:  Win95:Courier:-13:0
@Begin
@Participants:  CHI Senem Target_Child, MOT Mother, AUN Aunt, REC Dilara, Recorder WO Woman
@Age of child:
@Birth of child:
@Sex of child:  Female
@Education of MOT:
@Åge of MOT:
@SES of MOT:
@Date:  2-JAN-2003
@Verbatim Transcriber:  Sevda Bahtiyar
@Coder:
@Recorder:  Dilara Koçbaþ

%sit:   MOT is carrying CHI. CHI is holding pens.
%act:   MOT places CHI on the floor
*AUN-CHI:   Sene:m .
%mor:   VOC|Senem .
*REC-CHI:   geldinmi ?
%mor:   V|gel-PAST-2S-QUE .
*MOT-AUN:   teyze buna dikkat et de duruyo(r) kendi ama +/.
*AUN-MOT:   tamam evladým .
*MOT-AUN:   +^ birden atmasýn .
*MOT-AUN:   bak Windy [= dog's name] bekliyor arkasýnda görüyo(r)musun ?
*MOT-AUN:   bak düþmesin diye Windy bekliyo(r) arkasýnda .
*AUN-CHI:   kýzým sen neleri aldýn geldin kartopum ?
%mor:   VOC|kýz-POSS&1S PRO|sen WH|ne-PL-ACC V|al-PAST-2S
    V|gel-PAST-2S VOC|kar+top-POSS&1S .
%act:   AUN opens her arms to hold CHI .
*AUN-REC:   Dilara görüyo(r)musun geliþini ?
*REC-AUN:   hýþi .
```

Documentation

- Documentation matters:
  - Even a very messy corpus may become somehow processable when good documentation is available (although documentation does not help against inconsistency).
  - When documentation is missing, even an exemplary corpus may take a long time to understand and (re)use.
- Documentation is best published together with corpus data and regularly updated.

Example: The CHAT manual



CHILDES-Other-Languages.pdf (Seite 15 von 56)

Other Languages 15

5. [redacted]

[redacted]

Department of Linguistics

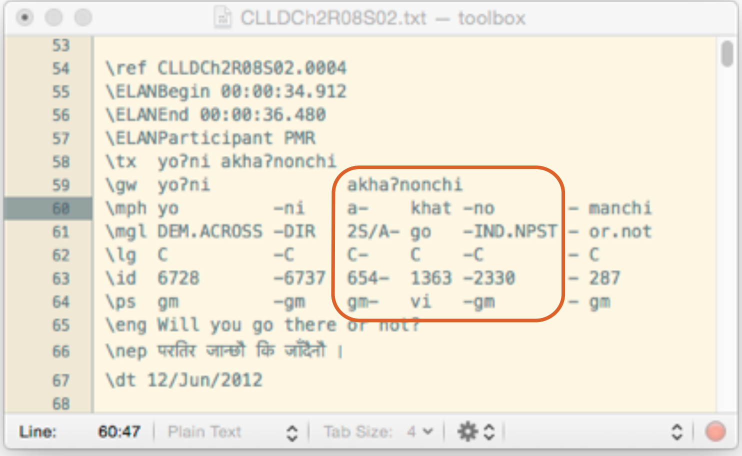
[redacted]

This is a PhonBank corpus from one child.

Explicit coding

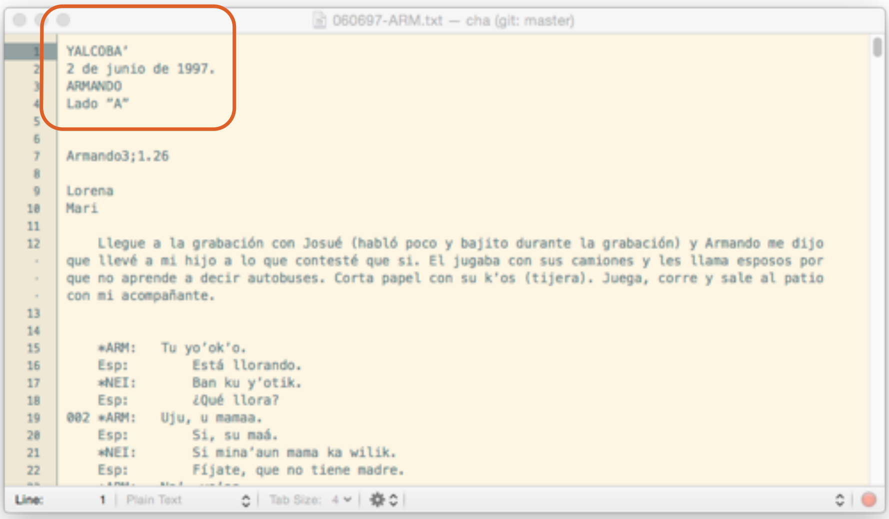
- The syntax of a corpus can be consistent, tidy, and well documented but still be awkward to use when the coding is not explicit.
- Explicit coding includes:
  - visible coding of semantic relations (no use of invisible characters such as spaces or line breaks)
  - as much information as possible is marked by standardised labels
- Implicit coding brings some dangers with it:
  - confusion on the side of the user who hasn't internalised the coding
  - invisible characters are especially prone to inconsistency and other errors

Example 1: Hierarchical relations



```
\ref CLLDCh2R08S02.0004
\ELANBegin 00:00:34.912
\ELANEnd 00:00:36.480
\ELANParticipant PMR
\tx  yo?ni akha?nonchi
\gw  yo?ni           akha?nonchi
\mph yo         -ni  a-    khat -no       - manchi
\mgl DEM.ACROSS -DIR 2S/A- go   -IND.NPST - or.not
\lg  C          -C   C-    C    -C        - C
\id  6728       -6737 654- 1363 -2330     - 287
\ps  gm         -gm  gm-   vi   -gm       - gm
\eng Will you go there or not?
\nep परतिर जान्छौ कि जाँदैनौ ।
\dt 12/Jun/2012
```

Example 2: Metadata



```
YALCOBA'
2 de junio de 1997.
ARMANDO
Lado "A"


Armando3;1.26

Lorena
Mari

    Llegue a la grabación con Josué (habló poco y bajito durante la grabación) y Armando me dijo
que llevé a mi hijo a lo que contesté que si. El jugaba con sus camiones y les llama esposos por
que no aprende a decir autobuses. Corta papel con su k'os (tijera). Juega, corre y sale al patio
con mi acompañante.


    *ARM:   Tu yo'ok'o.
    Esp:        Está llorando.
    *NEI:       Ban ku y'otik.
    Esp:        ¿Qué llora?
002 *ARM:   Uju, u mamaa.
    Esp:        Si, su maá.
    *NEI:       Si mina'aun mama ka wilik.
    Esp:        Fíjate, que no tiene madre.
```

Reusing other kinds of corpus data

Directory structure

- A corpus directory structure that is too complicated can create various problems:
  - outsiders won't understand what's located where
  - the corpus creators themselves might forget, leading to confusion and often more serious problems such as file duplicates
- ➔ **A good directory structure is simple:**
  - it is not too deep (generally one additional level below the top level should be enough)
  - it has transparent folder names
  - folders are disjunct so that the possibility for duplicates is excluded. For instance, sorting files into folders by file types is good because normally no file can belong to two types; sorting files by speakers is bad because one file can be associated with multiple speakers

File naming

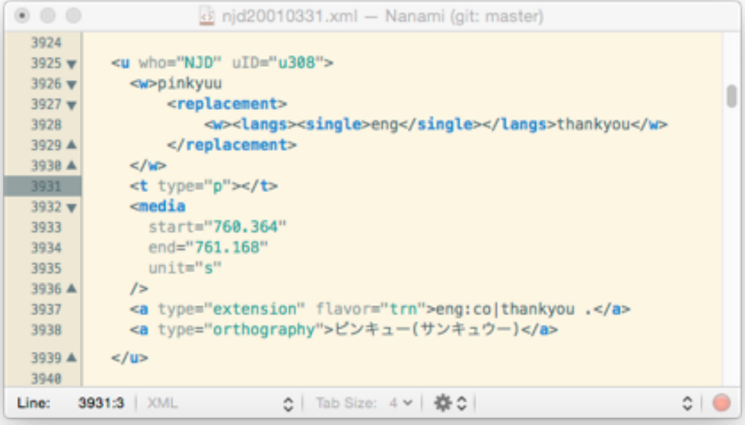
- A bad file naming system can create various problems:
  - data can’t be located or eventually even get lost
  - files may get duplicated without anybody noticing
  - it may become opaque which files are instances of the same more abstract corpus unit („text“ or „session“)
- **➔ A good file naming system is transparent:**
  - it indicate a few important points for each file (e.g. in the case of language acquisition: recording date and target child of the session)
  - it names all files identically or at least similarly that belong to the same abstract unit
  - it is practical for sorting. For instance, starting files with a recording date in the format YYYY-MM-DD makes it easy to sort files chronologically.

Encodings

- Encodings are important for linguists because:
  - most languages apart from English use „special characters“ (à, щ, 字…)
  - if a file with special characters is opened in an environment which doesn’t recognize or even breaks the intended encoding, some or all of them will be 文字化けd.
- **➔ A good encoding is standardised:**
  - Ideally use UTF-8 (even more ideally with some additions ignored here).
  - If this is impossible for some reason, use an encoding that can code all the graphs required by your language and that is widespread

Conclusion

- When it comes to reusing corpora, normally some kind of data transformation will be necessary.
- Standards facilitate transformation, especially when they are kept.
- But: some underlying properties of corpora are even more important than that. In this talk we have focussed on:
  - 1:1 relations between syntactic and semantic structure (aka consistency)
  - keeping semantically independent tiers apart in the syntactic representation
  - providing documentation
  - preferring explicit over implicit coding



```
<u who="NJD" uID="u308">
  <w>pinkyuu
      <replacement>
          <w><langs><single>eng</single></langs>thankyou</w>
      </replacement>
  </w>
  <t type="p"></t>
  <media
    start="760.364"
    end="761.168"
    unit="s"
  />
  <a type="extension" flavor="trn">eng:co|thankyou .</a>
  <a type="orthography">ピンキュー(サンキュウー)</a>
</u>
```