Introduction to ACQDIV corpora

Steven Moran, PhD

Data custodian

Department of Comparative Linguistics

University of Zurich

ACQDIV Summer workshop — June 11, 2015

Typical day at the office

acqdiv — bash — 154×43

emacs | bash | emacs | bash


```
||', '@ID:\ttur|Turkish_KULLD|MOT|25|Female|||Mother|5 years||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Operator|||', '@ID:\ttur|Turkish_KULLD|FAT|27|Mal
e|||Father|5 years||'], 'can11_25oct01_01-02-06': ['@ID:\ttur|Turkish_KULLD|CHI|1;02.06|male|||Target_Child|||', '@ID:\ttur|Turkish_KULLD|MOT|34|Female|||
Mother|15 years||', '@ID:\ttur|Turkish_KULLD|FAT|35|Male|||Father|15 years||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Operator|||'], 'tugce25_15mar03_01-
10-03': ['@ID:\ttur|Turkish_KULLD|CHI|01;10.03|female|||Target_Child|||', '@ID:\ttur|Turkish_KULLD|MOT|25|Female|||Mother|5 years||', '@ID:\ttur|Turkish_K
ULLD|CAM|||||Camera_Operator|||', '@ID:\ttur|Turkish_KULLD|FAT|27|Male|||Father|5 years||', '@ID:\ttur|Turkish_KULLD|FEM|||||Female|||'], 'burcu34_07nov03
_02-01-17': ['@ID:\ttur|Turkish_KULLD|CHI|02;01.17|female|||Target_Child|||', '@ID:\ttur|Turkish_KULLD|MOT|26|Female|||Mother|8 years||', '@ID:\ttur|Turki
sh_KULLD|CAM|||||Camera_Operator|||'], 'irem06_06jul02_00-10-19': ['@ID:\ttur|Turkish_KULLD|CHI|00;10.19|female|||Target_Child|||', '@ID:\ttur|Turkish_KUL
LD|GRA|||||Grandmother|||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Operator|||'], 'irem29_07jul03_01-10-20': ['@ID:\ttur|Turkish_KULLD|CHI|01;10.20|femal
e|||Target_Child|||', '@ID:\ttur|Turkish_KULLD|GRA|||||Grandmother|||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Operator|||', '@ID:\ttur|Turkish_KULLD|GI1
|||||Girl|||', '@ID:\ttur|Turkish_KULLD|GI2|||||Girl|||'], 'tugce14_07sep02_01-03-24': ['@ID:\ttur|Turkish_KULLD|CHI|01;03.24|female|||Target_Child|||', '
@ID:\ttur|Turkish_KULLD|MOT|24|Female|||Mother|5 years||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Operator|||', '@ID:\ttur|Turkish_KULLD|GRA|||||Grandmot
her|||', '@ID:\ttur|Turkish_KULLD|FEM|||||Female|||'], 'cansu01_01jul02_00-08-06': ['@ID:\ttur|Turkish_KULLD|CHI|00;08.06|female|||Target_Child|||', '@ID:
\ttur|Turkish_KULLD|MOT|26|Female|||Mother|11 years||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Operator|||'], 'ogun16_28sep02_01-07-24': ['@ID:\ttur|Turk
ish_KULLD|CHI|01;07.24|male|||Target_Child|||', '@ID:\ttur|Turkish_KULLD|MOT|22|Female|||Mother|5 years||', '@ID:\ttur|Turkish_KULLD|FAT|28|Male|||Father|
5 years||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Operator|||', '@ID:\ttur|Turkish_KULLD|COU|||||Cousin|||'], 'ekin03_17jul01_00-09-26': ['@ID:\ttur|Tur
kish_KULLD|CHI|0;09.26|female|||Target_Child|||', '@ID:\ttur|Turkish_KULLD|MOT|35|Female|||Mother|19 years||', '@ID:\ttur|Turkish_KULLD|FAT|39|Male|||Fath
er|19 years||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Operator|||'], 'senem38_25jun04_02-07-02': ['@ID:\ttur|Turkish_KULLD|CHI|02;07.02|female|||Target_
Child|||', '@ID:\ttur|Turkish_KULLD|MOT||Female|||Mother|15 years||', '@ID:\ttur|Turkish_KULLD|GRA|||||Grandmother|||', '@ID:\ttur|Turkish_KULLD|FEM|||||F
emale|||', '@ID:\ttur|Turkish_KULLD|UNC|||||Uncle|||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Operator|||'], 'cansu31_12dec03_02-01-17': ['@ID:\ttur|Turk
ish_KULLD|CHI|02;01.17|female|||Target_Child|||', '@ID:\ttur|Turkish_KULLD|MOT|23|Female|||Mother|11 years||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Ope
rator|||'], 'burcu17_18feb03_01-04-28': ['@ID:\ttur|Turkish_KULLD|CHI|01;04.28|female|||Target_Child|||', '@ID:\ttur|Turkish_KULLD|MOT|26|Female|||Mother|
8 years||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Operator|||'], 'burcu11_05nov02_01-01-15': ['@ID:\ttur|Turkish_KULLD|CHI|1;01.15|female|||Target_Child
|||', '@ID:\ttur|Turkish_KULLD|MOT|25|Female|||Mother|8 years||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Operator|||'], 'senem44_04oct04_02-10-11': ['@ID
:\ttur|Turkish_KULLD|CHI|02;10.11|female|||Target_Child|||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Operator|||', '@ID:\ttur|Turkish_KULLD|BAB|||||Babysi
tter|||'], 'irem09_21aug02_01-00-04': ['@ID:\ttur|Turkish_KULLD|CHI|01;00.04|female|||Target_Child|||', '@ID:\ttur|Turkish_KULLD|MOT|27|Female|||Mother|8
years||', '@ID:\ttur|Turkish_KULLD|GRA|||||Grandmother|||', '@ID:\ttur|Turkish_KULLD|BOY|||||Boy|||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Operator|||'
], 'ogun01_06oct01_00-08-02': ['@ID:\ttur|Turkish_KULLD|CHI|00;08.02|male|||Target_Child|||', '@ID:\ttur|Turkish_KULLD|MOT|21|Female|||Mother|5 years||',
'@ID:\ttur|Turkish_KULLD|FAT|27|Male|||Father|5 years||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Operator|||'], 'cansu07_11oct02_00-11-17': ['@ID:\ttur|T
urkish_KULLD|CHI|00;11.17|female|||Target_Child|||', '@ID:\ttur|Turkish_KULLD|MOT|22|Female|||Mother|11 years||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_
Operator|||'], 'can44_22may03_02-09-04': ['@ID:\ttur|Turkish_KULLD|CHI|02;09.04|male|||Target_Child|||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Operator|
|||', '@ID:\ttur|Turkish_KULLD|BAB|||||Babysitter|||', '@ID:\ttur|Turkish_KULLD|MOT|36|Female|||Mother|15 years||'], 'ekin01_21may01_00-08-01': ['@ID:\ttur
|Turkish_KULLD|CHI|0;08.01|female|||Target_Child|||', '@ID:\ttur|Turkish_KULLD|MOT|35|Female|||Mother|19 years||', '@ID:\ttur|Turkish_KULLD|FAT|39|Male|||
Father|19 years||', '@ID:\ttur|Turkish_KULLD|AYL|||||Investigator|||'], 'tugce43_02jan04_02-07-21': ['@ID:\ttur|Turkish_KULLD|CHI|02;07.21|female|||Target
_Child|||', '@ID:\ttur|Turkish_KULLD|MOT|26|Female|||Mother|5 years||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Operator|||'], 'burcu40_30jan04_02-04-09':
 ['@ID:\ttur|Turkish_KULLD|CHI|02;04.09|female|||Target_Child|||', '@ID:\ttur|Turkish_KULLD|MOT|27|Female|||Mother|8 years||', '@ID:\ttur|Turkish_KULLD|CA
M|||||Camera_Operator|||'], 'can23_24apr02_01-08-03': ['@ID:\ttur|Turkish_KULLD|CHI|1;08.03|male|||Target_Child|||', '@ID:\ttur|Turkish_KULLD|MOT|35|Femal
e|||Mother|15 years||', '@ID:\ttur|Turkish_KULLD|FAT|36|Male|||Father|15 years||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Operator|||', '@ID:\ttur|Turkis
h_KULLD|BAB|||||Babysitter|||'], 'burcu19_21mar03_01-06-00': ['@ID:\ttur|Turkish_KULLD|CHI|01;06.00|female|||Target_Child|||', '@ID:\ttur|Turkish_KULLD|MO
T|26|Female|||Mother|8 years||', '@ID:\ttur|Turkish_KULLD|CAM|||||Camera_Operator|||', '@ID:\ttur|Turkish_KULLD|FEM|||||Female|||']})
['@ID:\ttur|Turkish_KULLD|CHI|0;08.02|female|||Target_Child|||', '@ID:\ttur|Turkish_KULLD|MOT|25|Female|||Mother|8 years||', '@ID:\ttur|Turkish_KULLD|GRA|
||||Grandmother|||', '@ID:\ttur|Turkish_KULLD|CA1|||||Camera_Operator|||', '@ID:\ttur|Turkish_KULLD|CA2|||||Camera_Operator|||', '@ID:\ttur|Turkish_KULLD|
VIS|||||Visitor|||', '@ID:\ttur|Turkish_KULLD|BOY|||||Boy|||']
```

Talk map

- ACQDIV data
  - Languages and corpora
- Infrastructure and transforming corpus data
  - Challenges involved in interoperability
  - Towards re-useable solutions and workflows
  - Towards a “unified” cross-linguistic cross-corpus resource
- Future
  - ACQDIV database and some basic stats
  - Some example questions we want to be able to answer across the unified data
  - Research topics

The ACQDIV languages

The ACQDIV languages

Language	Speakers	Classification	Genus	Location	Area
Russian	166'167'860	Indo-European	Slavic	Russia	Europe
Japanese	128'056'940	Japanese	Japanese	Japan	Asia
Turkish	70'890'130	Altaic	Turkic	Turkey	Asia
Indonesian	23'200'480	Austronesian	Malayo-Sumbawan	Indonesia	Pacific
Sesotho	5'634'000	Niger-Congo; Benue-Congo	Bantoid	South Africa	Africa
Yucatec	766'000	Mayan	Mayan	Mexico	Americas
Cree	87'220	Algic	Algonquian	Northern Canada	Americas
Inuktitut	34'510	Eskimo-Aleut	Eskimo	Eastern Canada	Americas
Dene	11'900	Na-Dene	Athapaskan	South Central Canada	Americas
Chintang	3'710	Sino-Tibetan	Kiranti	Nepal	Asia

The corpora

Language	2-3 yrs	3-4 yrs	Recordings	Session duration
Chintang	2	2	monthly	4h
Cree	1*	1*	2-3 weeks	30-40 mins
Indonesian	5*	7*	bi-weekly	1h
Inuktitut	4*^	1*	monthly	4h
Japanese MiiPro	3*	4*	weekly	1-1.5h
Japanese Miyata	3	0	weekly	1h
Russian	4*	4*	weekly	1h
Sesotho	3	1	monthly	3-4h
Turkish	8*	8*	bi-weekly	1h
Yucatec	3*	3*	bi-weekly	30-90 mins

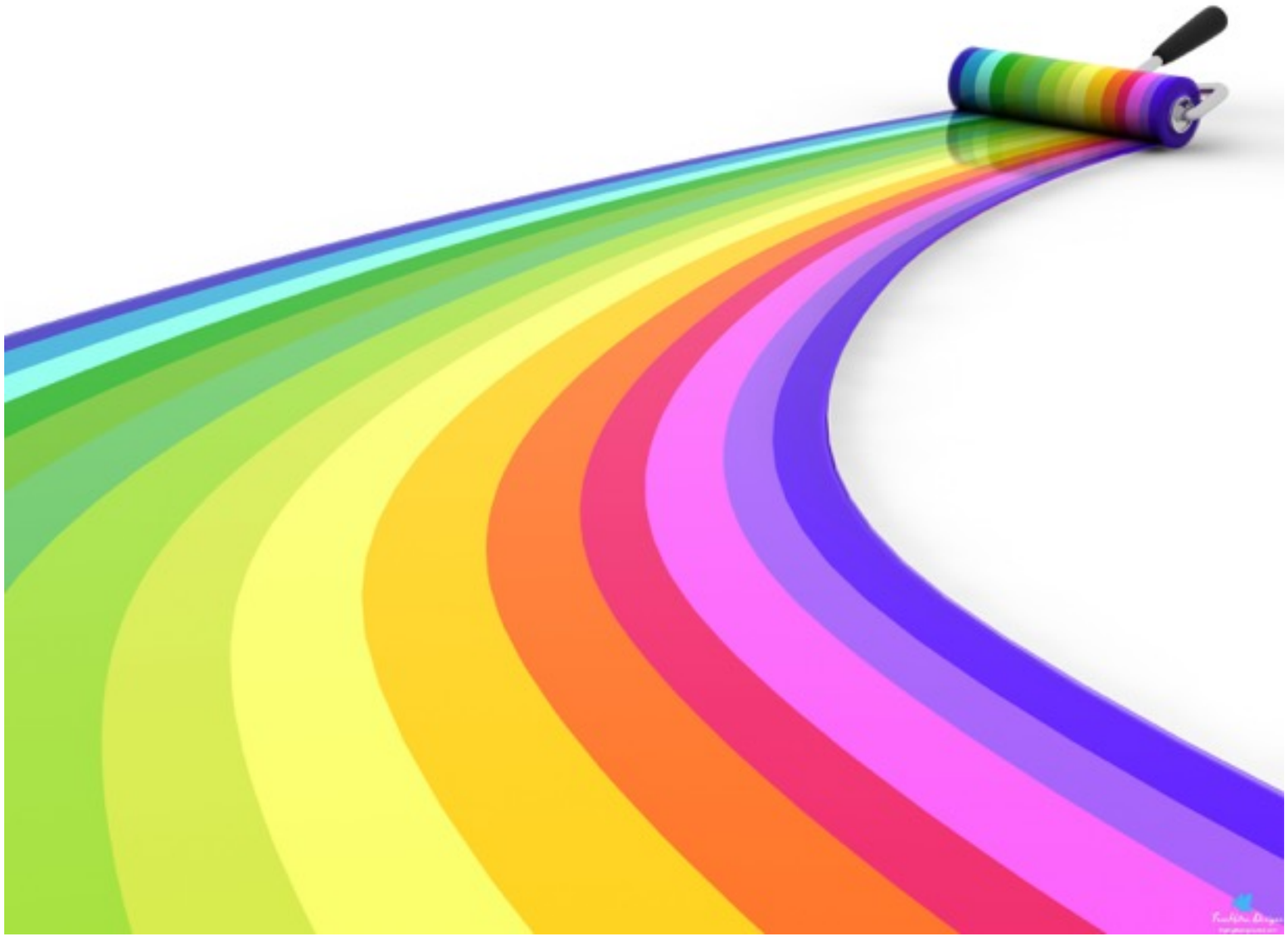
*same children; ^2;0, 2;6, 2;6, 2;10

ACQDIV raw corpora

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

Initial state of the maximally diverse corpora

What we envision as the “final” state of the ACQDIV corpora



Some desired queries (why we need corpora interoperability)

- Find all utterances in all corpora produced by children; compute mean length of utterance across time for every child.
- Find all utterances from Sesotho by 2-year-old children which contain a negated verb form and which are preceded by an adult utterance with the same verb but no negation.
- Find the number of occurrences of X (e.g. perfective aspect) within specific time units (e.g. 5 minutes of recording).
- Find all verb forms from children of language X which are preceded by an interlocutor's utterance in the same interactional unit (to be defined) in the exact same form.
- Find all verb forms from children of language X which are preceded by an interlocutor's utterance in the same interactional unit (yet to be defined, e.g. record) in a different verb form (variants of forms, can be either different stem or different affix).
- Find all utterances produced by children younger than 2;3 and produce a frequency count of parts-of-speech; exclude utterances where word or morpheme alignment is broken.
- Find all glossed Inuktitut utterances that contain a word containing the substring "nngi" in its orthographic form as well as a morpheme that has the segment shape "^nngit$" and the corresponding gloss "NEG".

Road blocks on the way to interoperability

- Different information in the different corpora — is their a common denominator?
  - Transcriptions
  - Translations (mostly)
  - Glosses (grammatical forms)
  - Some POS
  - Some time-alignments
  - Metadata
    - Sessions and participants
- Different corpus formats (syntactically and semantically)

CHAT format (CHILDES)

```
@Begin
@Participants: ARM target_child, SAN child, LOR mother, FIL mother, ABU grandmother, MAR aunt,
@Birth of Armando: 10-apr-1994
@Age of Armando: 1;08.23
@Age of Sandi: 2; 5.10
@Filename: A010396
@Date: 3-jan-1996
@Sex of Armando: male
@Location: Yalcobá, Yucatán .
@Activities: la mayor parte de la grabación la hice con Armando, porque Sandi no se encontraba

*NEI:   xáchet a#pool .
%mor:   VT|xáchet:IMP|-0 2POS|a#S|pool .
%eng:   peínate .
*ARM:   pool .
%pho:   / pol / .
%mor:   S|pool .
%eng:   pelo .
*ARM:   w#ich .
%pho:   / wich' / .
%mor:   1POS|w#S|ich .
%eng:   ojos .
*NEI:   w#ich t#a#ch'op-ah a#w#ich y#éet-el .
%pho:   / wich ta ch'opa wich yete' / .
%mor:   2POS|w#S|ich PFV|t#2ERG|a#VT|ch'op:PFV|-ah 2POS|a#PT|w#S|ich 3POS|y#S|éet:POS-el .
%eng:   tus ojos te jurgaste los ojos con ello ?
*ARM:   w#ich xáache' .
%pho:   / waich' xache' / .
%mor:   1POS|w#S|ich S|xáache' .
%eng:   mis ojos con el peine .
```

XML

```
</Participants>
<u who="MHL" uID="u0">
  <w>ere</w>
  <w>mphe</w>
  <w>ntho</w>
  <w>ena</w>
  <t type="p"></t>
  <media
    start="105.000"
    end="110.057"
    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="CHI" uID="u1">
  <w>mphe</w>
  <w>ntho</w>
  <t type="p"></t>
  <media
    start="110.057"
    end="113.836"
    unit="s"
  />
  <a type="target gloss">m-ph-e ntho .</a>
```

Toolbox

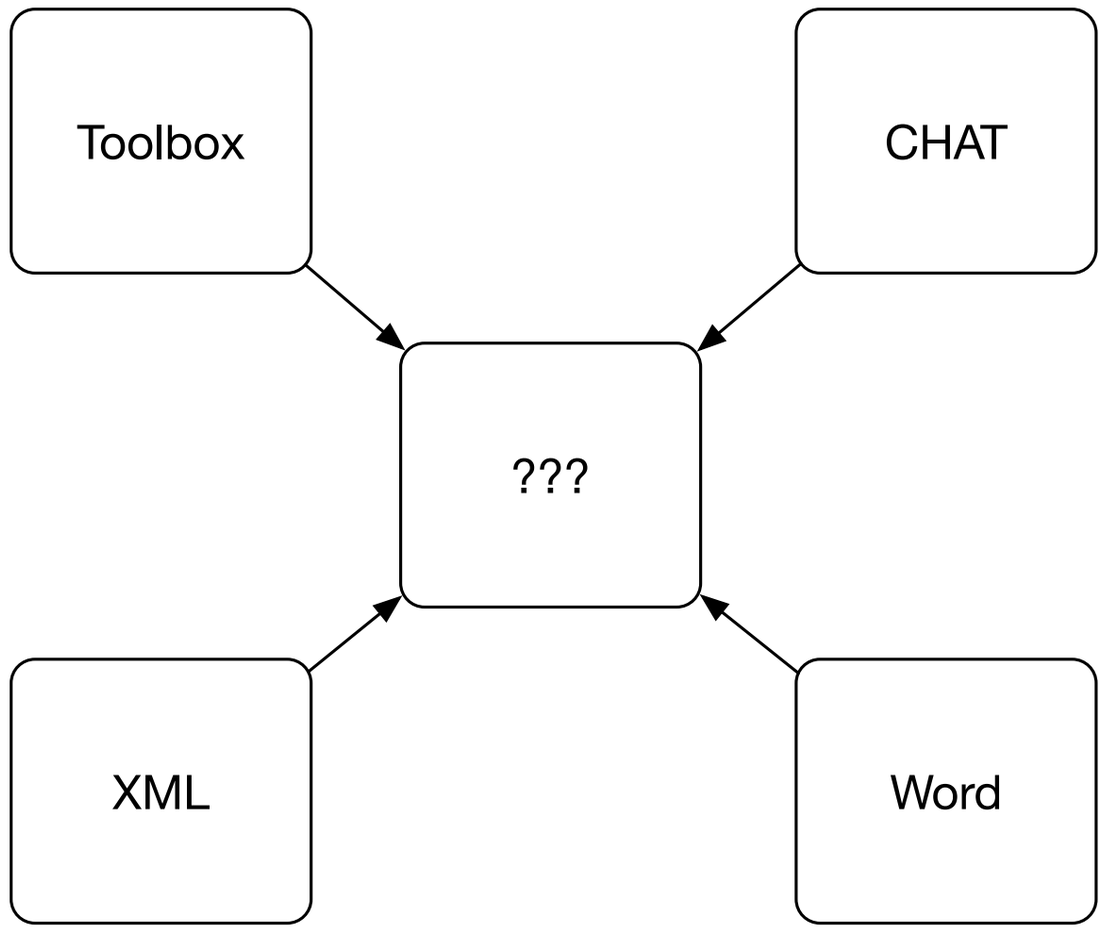
```
\_sh v3.0  833  Chintang
\_DateStampHasFourDigitYear



\ref CLDLCh2R08S01.001
\ELANBegin 00:00:01.310
\ELANEnd 00:00:02.720
\ELANParticipant CHKR
\tx bai? phoni thaŋna phoni
\gw bai?          pho  ni   thaŋna pho  ni
\mph ba       -i?  pho  ni   thaŋnu pho  ni
\mgl DEM.PROX -LOC  REP  EMPH rag     REP  EMPH
\lg C         -C    C    N    N       C    N
\id 643       -6729 1919 1770 6389    1919 1770
\ps pro       -gm   gm   gm   n       gm   gm
\eng Here is the rags
\nep यहाँ थाङ्नो अरे नि ।
\dt 29/Jun/2013
```

Word

- Encoding?
- Formal structure? Proprietary!
- Sharing with other programs?

Combining different formats

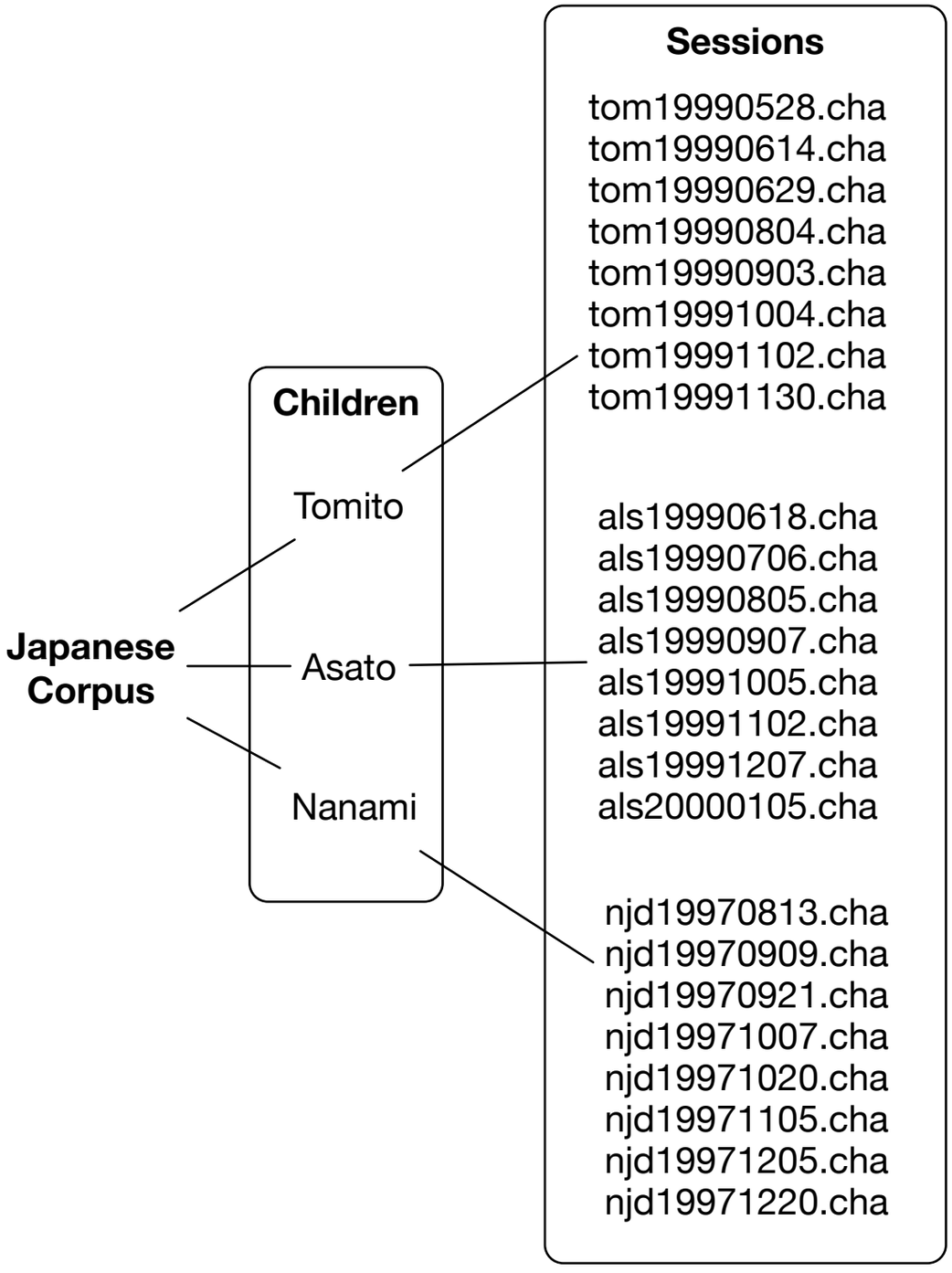
- Different data formats are problematic
  - very challenging to combine them into one resource that we can use to compare child language acquisition cross-linguistically
- How to combine them?
  - Syntactic interoperability
  - Semantic interoperability
- What do we combine?
  - Utterance level data
  - Morphological analyses
  - Metadata



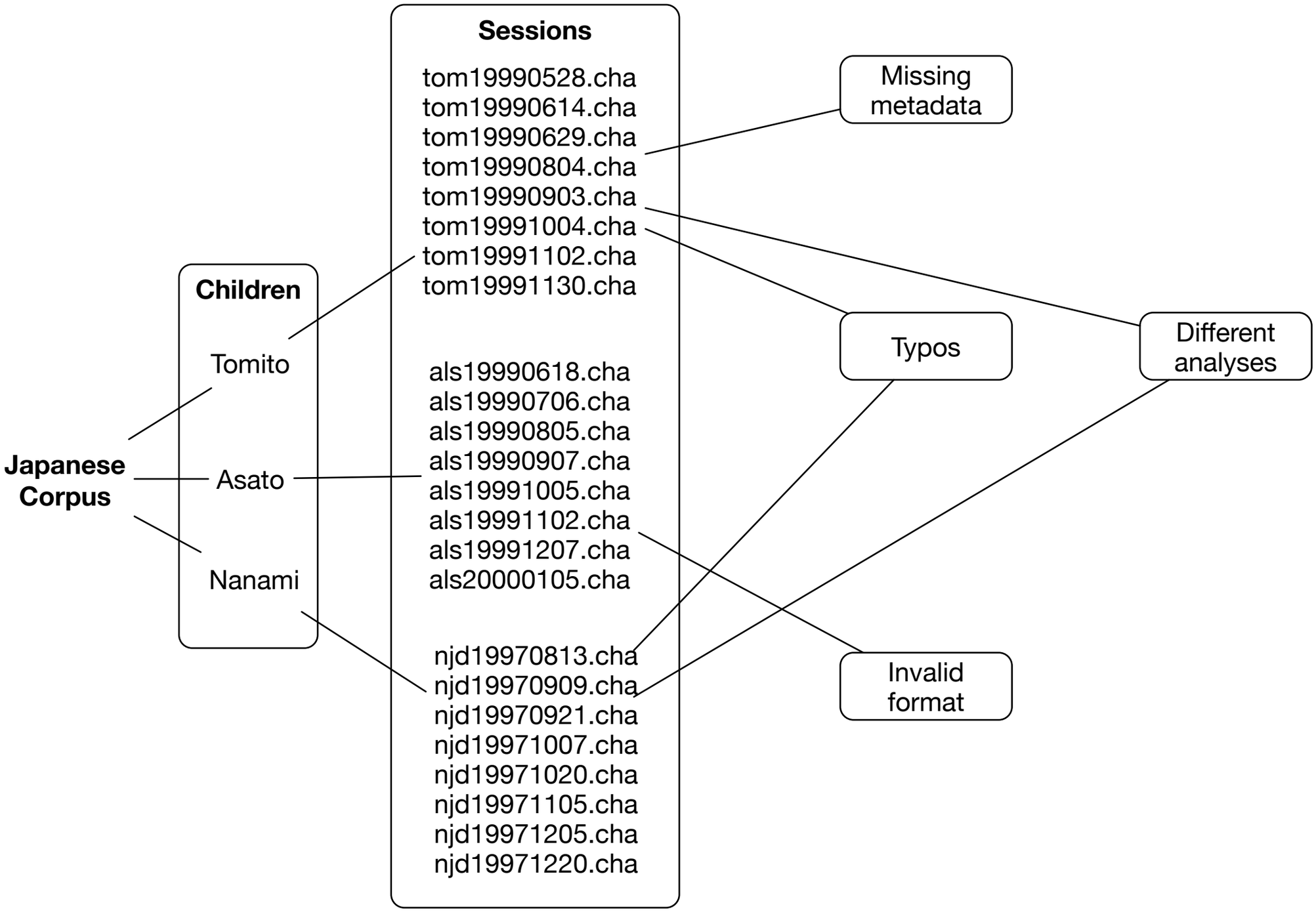
Consistency within one corpus (or even one project)

- Even slight variations in the same corpus make analysis across recording sessions (transcriptions) difficult
- Some problems:
  - missing data
  - typos and inconsistencies
  - different glossing conventions
- Why?
  - different researchers working on the corpus
  - evolution of corpus over time
  - different analyses

Example corpus from Japanese



Example corpus from Japanese (but applicable to all 10 corpora!)



Important lessons learned from experience

- What does it cost to bring a corpus into a workable format
  - 1 hour * 400 sessions = 10 x 40 hour weeks! (per corpus!)
  - How much $?
- Tips for corpus creators
  - Use a consistent encoding and file format (preferably UTF-8)
  - If you convert formats, check input and output
  - Check the data for errors and inconsistencies often
    - One method is to check character and word frequencies
  - Store the data in more than one place (backups)
  - Use detailed and consistent metadata!

Metadata — data about data

How do we unify these different data formats?

- First identify the data types, sources, etc. in each corpus. But then what?
- Extract data differently for each corpus? Or aim for a single unified format?
  - What costs more in time, money, etc.?
  - What's the purpose of the "final" unified format, e.g. business solution, research (reproducible?), data dissemination, etc.?

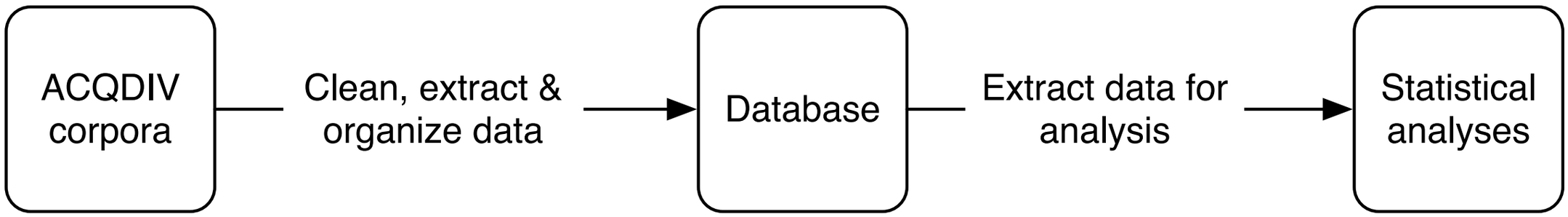
Towards an interoperable database

- 5 corpora already in CHILDES CHAT format
- 3 projects aiming for CHAT
  - Common denominator: CHAT
  - CHAT -> XML (validates the CHAT)
  - Lots of money and engineering hours already spent on CHILDES CHAT
- 2 corpora in Toolbox & ELAN (in-house)
- 1 corpus to establish good (better?) practices (Dene)
- Aim:
  - develop reusable, sustainable technological infrastructure and workflows

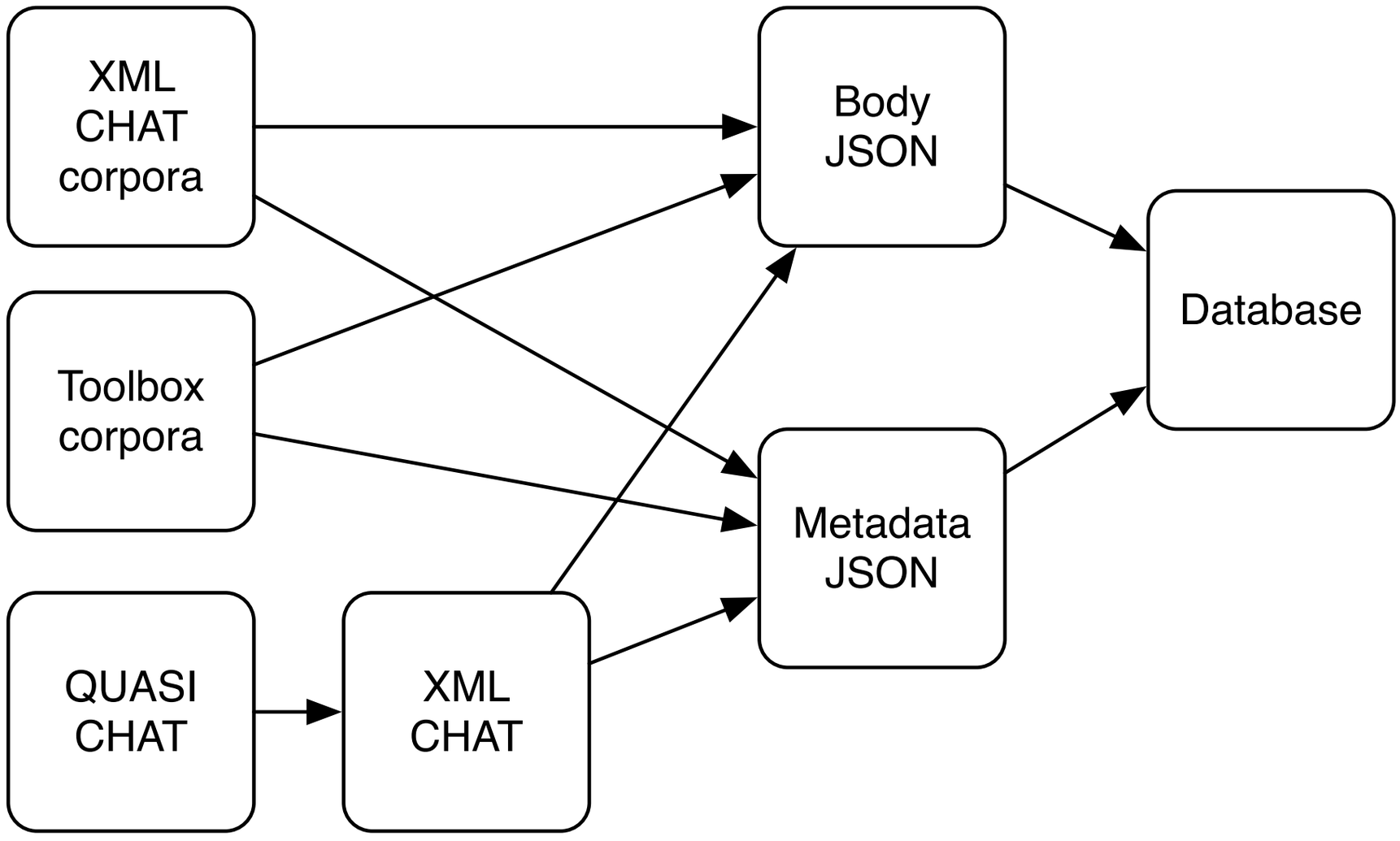
ACQDIV Workflows

- General idea
  - Validate data, extract relevant stuff, transform it into interoperable format, import into relevant accessible data formats
- Workflows
  - Clean and validate incoming data
  - Extract metadata (no metadata, no credibility)
  - Extract session data (utterances, participants, morphological annotations)
- Terminological unification
- Qualitative and quantitative analyses

Idealized transformation workflow



More realistic (but still simplified) transformation workflow



How do we unify these different data formats?

- Industry standard(s)
  - “The nice thing about standards is that you have so many to choose from.” (Andrew S. Tannenbaum)
- Extract, transform, load (ETL)
  - Unfortunately workflows are most often data, project, industry, etc. specific! (Ugh)

- What can we reuse in this particular project? Where can we save time?
- How do we integrate our decisions into our goals and day-to-day challenges?
- How can we keep things reusable and “future-proof”?
- How do we keep the door open for new data, new challenges?

Day-to-day organizational challenges

- Involve:
  - organizing and updating the input data, workflow, codebase, etc
  - performing computational experiments
- Core guiding principles:[1]
  - Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why
  - Everything you do, you will probably have to do over again
- ACQDIV project:
  - data curation, maintenance, additions, corrections, distribution
  - quantitative experiments and research

[1]Noble 2009: A Quick Guide to Organizing Computational Biology Projects

ACQDIV repository

- Github repository
  - web-based repository hosting service
  - distributed version control
  - source code management (write code better and faster)
  - collaborative features

- Private ACQDIV project repository
  - https://github.com/uzling/acqdiv

Github repository

- Issue tracking
- Version control
  - Stores revisions of projects (filing system for every draft, etc.)
- Fork, pull request & merge workflow
- Changelogs to “blame” people
- Sharing code, data, documentation and experiments
  - Collaborator work
- Publication service

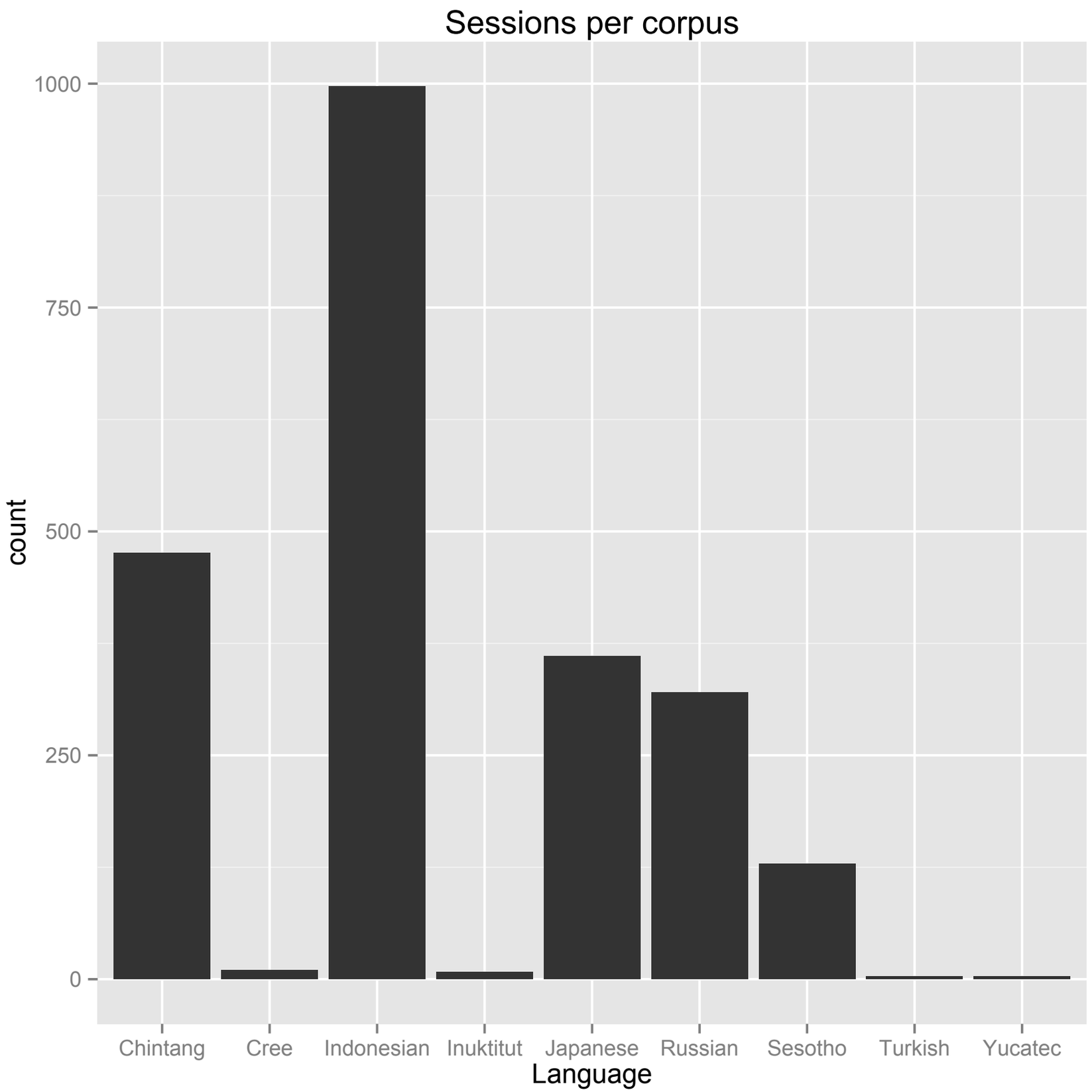
ACQDIV repository features

- Cleaning and transformation workflows
  - E.g. transform Yucatec, Inuktitut & Turkish to valid CHAT (gives corpus developers direct access to CHILDES-supported tools)
- Documentation
- Data extraction (utterances data and metdata)
- Testing (so that we aren't screwing things up even more)
- The "final" product after various cleaning, workflows… the ACQDIV "database"
  - Simple formats

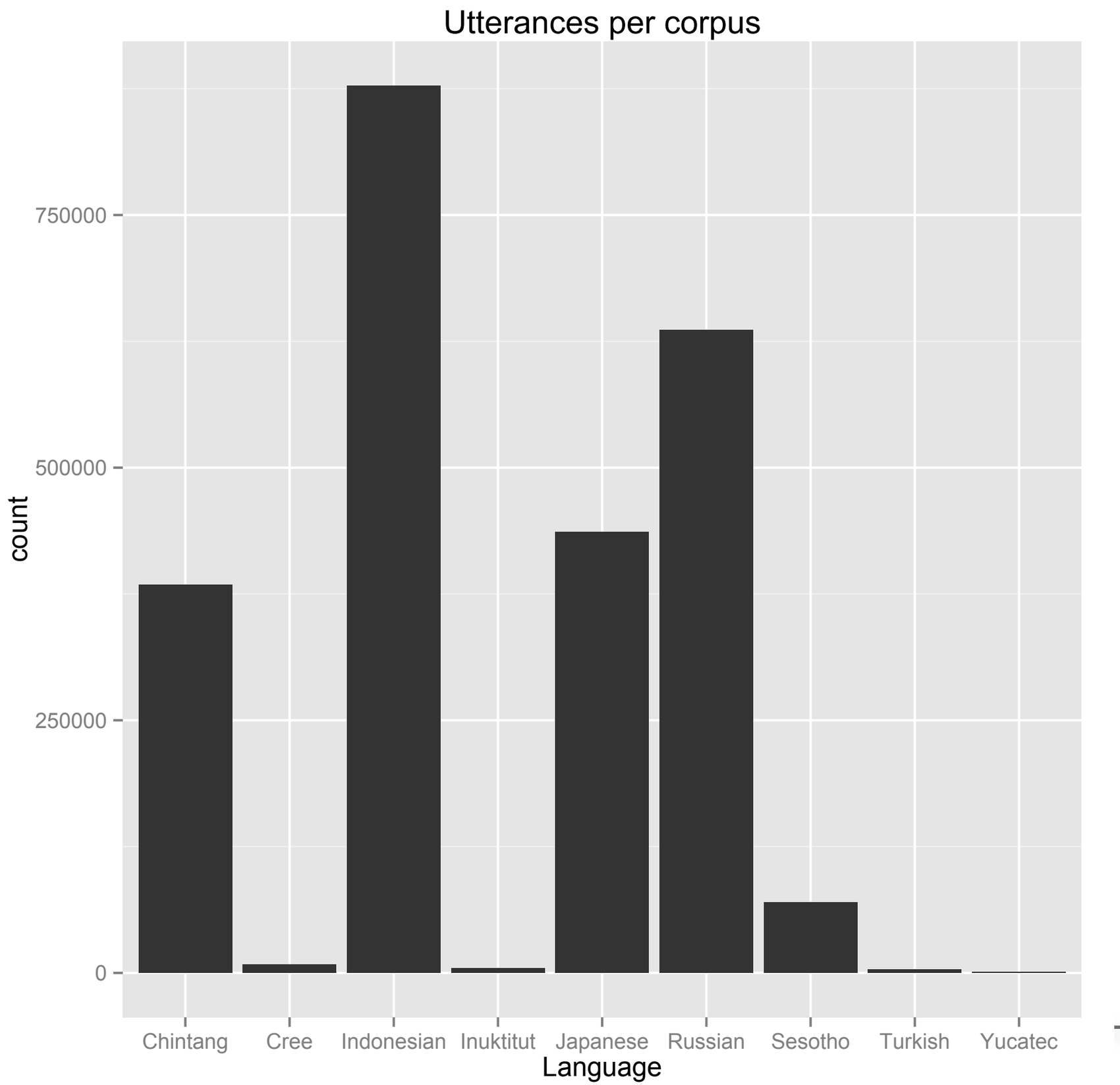
Current state of corpora in the ACQDIV database

- Corpora processing
  - YIT being cleaned and imported
  - Error checking and correction of all corpora
- ACQDIV database
  - Denormalized to the lowest common dominator (morpheme-level)
    - Over 7.5M rows
    - Two linked tables
      - body annotation tiers (utterances, words, morphemes)
      - metadata (participants, ages, session information, etc.)
- Some basic descriptive stats

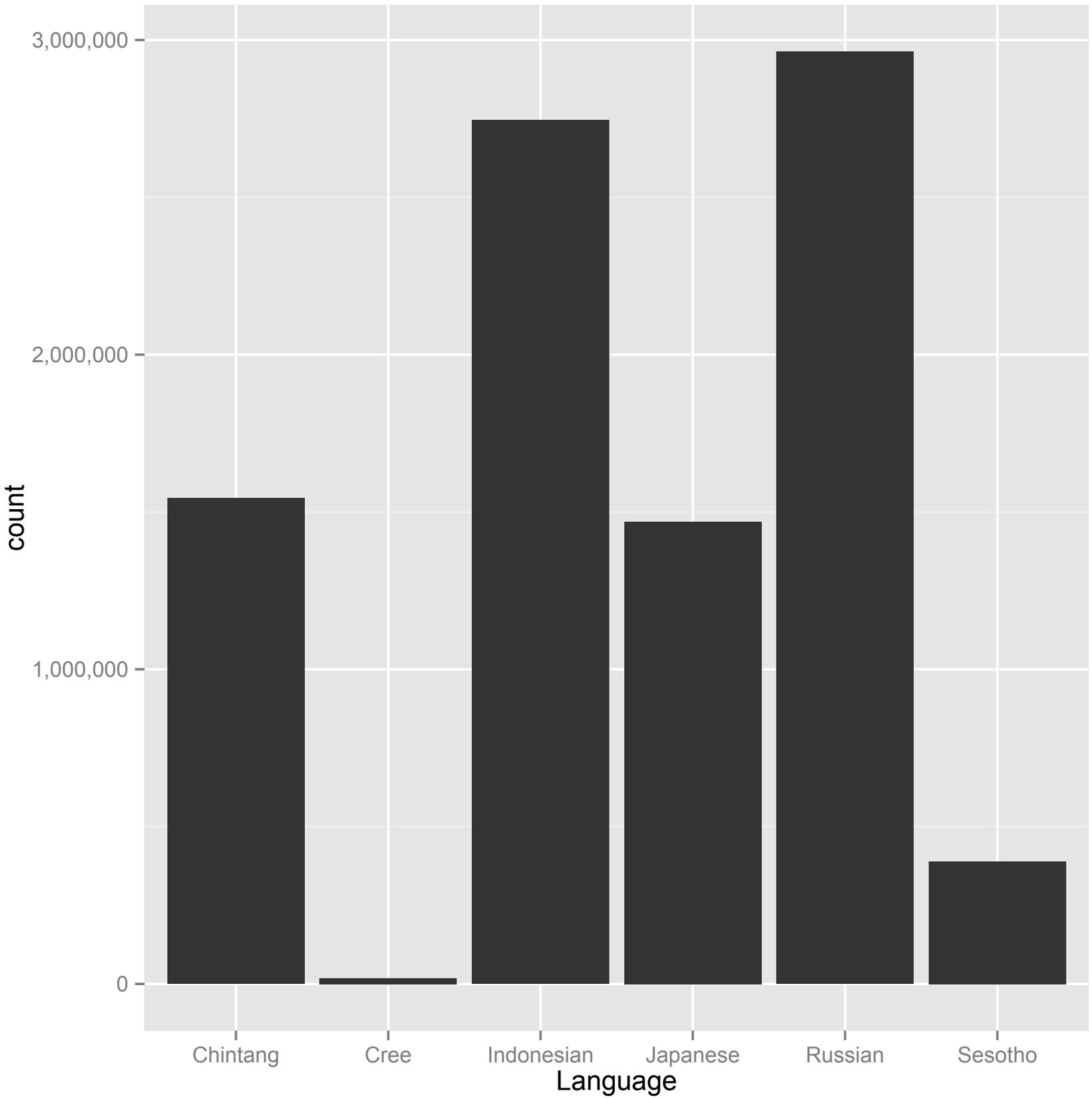
Number of sessions per corpus



Number of utterances per corpus



Number of words per corpus



Summing up

- Building reusable workflows to make different corpora interoperable
  - Syntactic (“structural”) interoperability (data formats, encoding, etc.)
  - Semantic (“terminological”) interoperability (cross-linguistic categories, labels, etc.)
- Treat data as code
  - Version tracking, issues, documentation, etc.
- Why do we do bother?
  - Reusability through transparency
  - GIGO (garbage in, garbage out)
- What next?

Future work

- For the database…
  - terminological interoperability (consistent morphological labels)
  - continue identifying and fixing anomalies in the corpora, e.g. one-off participant roles: ‘baltazar’, ‘igor’, ‘garndmother’, ‘focused child’
- For the research…
  - acquisition of aspect and negation
  - statistical learning, Bayesian learning, neural networks
  - rule-based vs item-based learning
  - time-series analyses and statistics
  - working group topics!

Merci vielmal! (Thank you!)

- For your attention and participation in the ACQDIV kick-off workshop
- Supporters
  - European Research Council, University of Zurich, Department of Comparative Linguistics, Swiss National Science Foundation
- UZH Team
  - Sabine Stoll, PI
  - Robert Schikowski, PM
  - Taras Zakharko, Tech admin
  - Danica Pajović, Andreas Gerster, Cazim Hysi, Laura Canedo, Carolin Remensberger

http://www.qwantz.com/index.php?comic=1937