University of Zurich[UZH]

Psycholinguistics Laboratory, Department of Comparative Linguistics

How to study variation in language acquisition: A new cross-linguistic corpus approach

Steven Moran & Sabine Stoll

Damián Blasi, Robert Schikowski
Danica Pajović, Cazim Hysi

Department of Comparative Linguistics
University of Zurich

9th Days of Swiss Linguistics, 29 June — 1 July, Geneva, Switzerland

Talk map

- ACQDIV project overview and database
  - Challenges children face in language acquisition
  - The language sample and corpora
  - ETL pipeline
- Preliminary research: role of child-directed speech
  - Distributional frames

Project overview

- ACQDIV = Acquisition processes in maximally diverse languages
- ERC funded project: 2014–2019
- Prof. Sabine Stoll, PI (U. Zurich)
- Central question: What's universal in language acquisition?
- Method: Compare acquisition processes in 10 languages that we know to be very different with regard to some macro-typological parameters ("maximum diversity sampling")
- http://www.acqdiv.uzh.ch/

Main questions in language acquisition research

- What are the cognitive principles that allow children to learn any human language (if they grow up in the respective socio-cultural environment)?
  - Are there universal stages in language acquisition?
  - Are there universal learning strategies?
  - Are there different acquisition strategies, do they depend on the structure of the language?
  - What are the factors responsible for the order of acquisition of linguistic forms (frequency, complexity, saliency, transparency etc.)?

Learning challenge: extreme typological variation

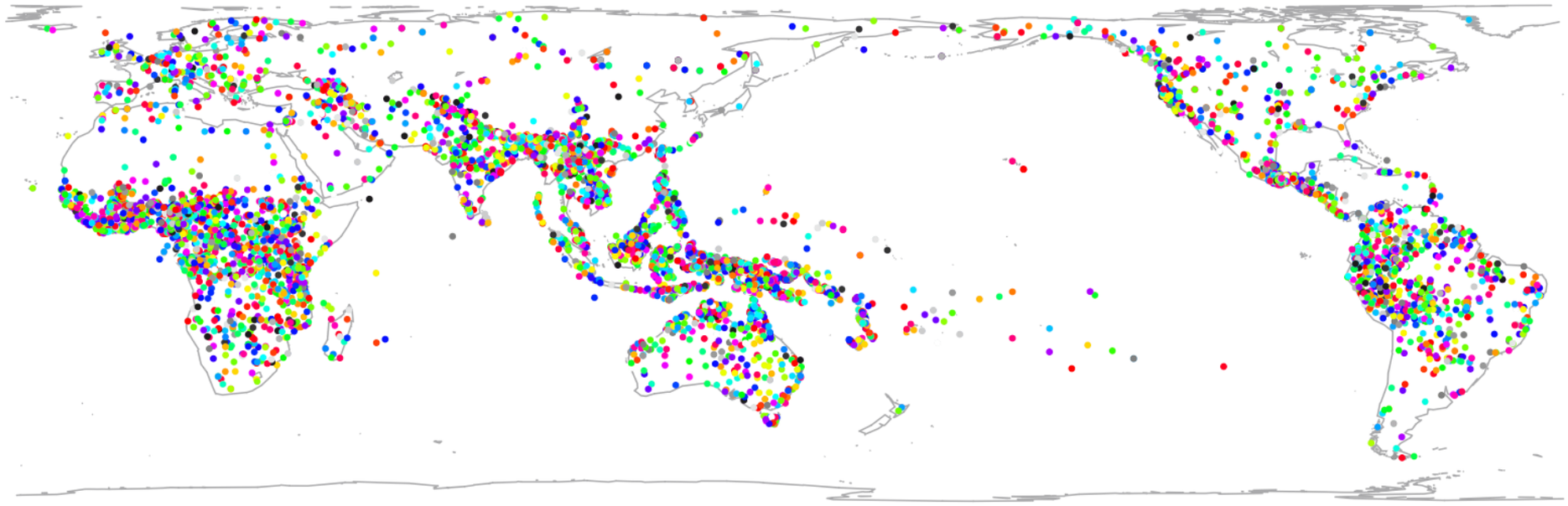
- Indonesian:

  O, Ei lagi minum susu.
  oh Ei more drink milk
  ‘Oh, Ei is drinking more milk.’

- Cree:

  Chi-wâp-iht-â-n â kâ-pushch-ishk-iw-â-t.
  2-light-by.head-TR.INAN.NON3-2SG>0 Q PVB.CONJ-put.on-by.foot-STEM-TR.ANIM-3SG>4SG
  ‘You see? She was putting it on.’

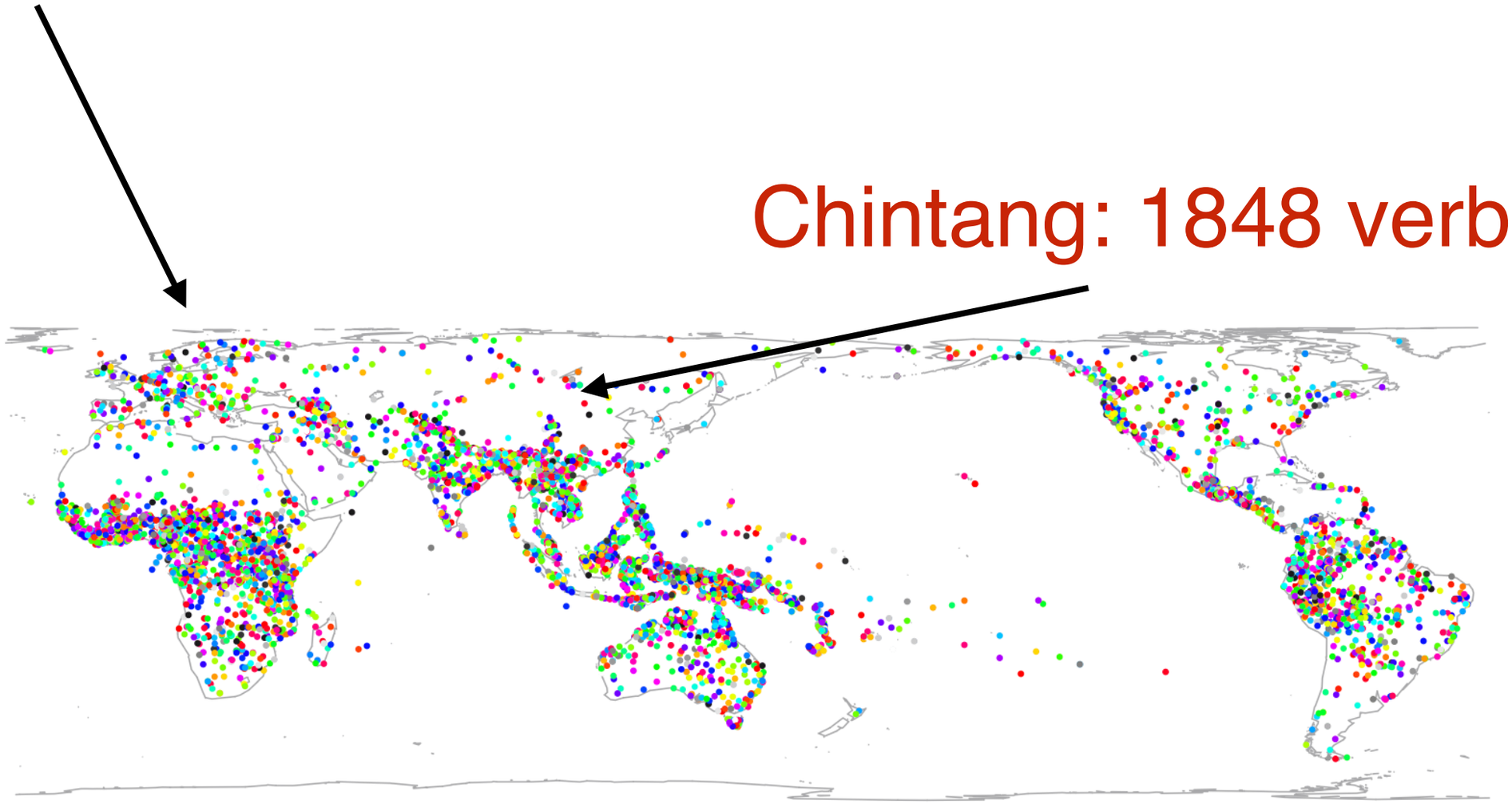
Learning challenge: extreme typological variation



Learning challenge: extreme typological variation

English: 3 verb forms

Chintang: 1848 verb forms



Learning challenge: extreme typological variation

- Example: synthetic verb forms in English

Present	Past
I play	I played
you play	you played
he/sh/it plays	he/she/it played
we play	we played
you play	you played
they play	they played

- 3 synthetic verb forms, the rest is combination

Learning challenge: extreme typological variation

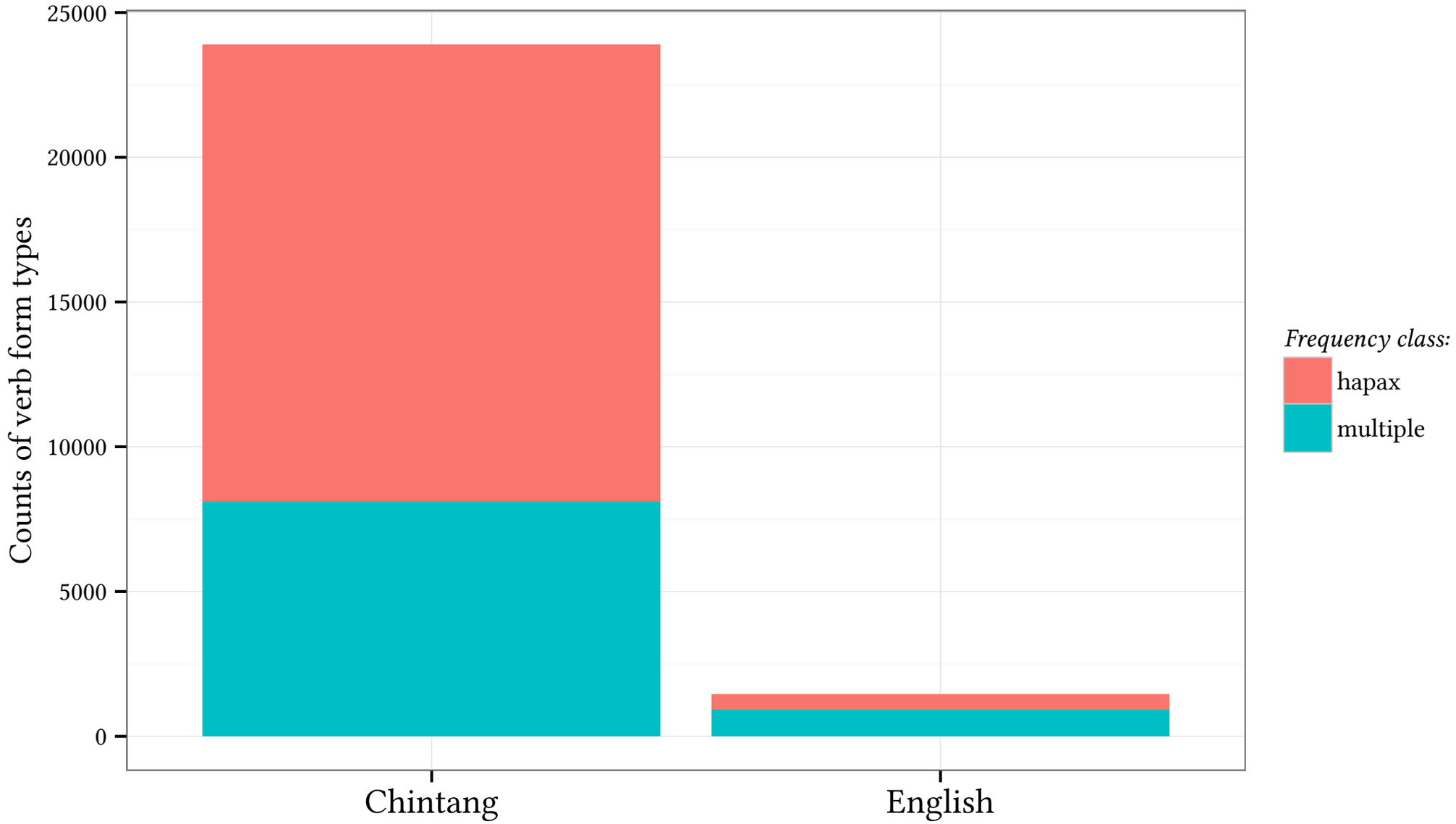
- Example: synthetic verb forms in Chintang (1848 verb forms)

	1s	1di	1pi	1de	1pe	2s	2d	2p	3s	3ns	intransitive
1s						*tupnaʔã* *tupnaʔãnɨŋ* *tupnehẽ* *matupyoknehẽ*	*tupnaʔãce* *tupnaʔãcenɨŋ* *tupnace* *matupyoknace*	*tupnaʔãni* *tupnaʔãninɨŋ* *tupnanihẽ* *matupyoknanihẽ*	*tubukuŋ* *tubukuŋnɨŋ* *tubuhẽ* *matupyoktuhẽ*	*tubukuŋcuŋ* *tubukuŋcuŋnɨŋ* *tubuŋcɨhẽ* *matupyoktuŋcɨhẽ*	*tupmaʔã* *tupmaʔãnɨŋ* *tubehẽ* *matupyoktehẽ*
1di									*tupcoko* *tupcokonɨŋ* *tubace* *matupyoktace*	*tubumcum* *tubumcumnɨm* *tubumcumhẽ* *matupyoktumcumhẽ*	*tupceke* *tupcekenɨŋ* *tubace* *matupyoktace*
1pi									*tubukum* *tubukumnɨm* *tubumhẽ* *matupyoktumhẽ*		*tubiki* *tubikinɨŋ* *tubihẽ* *matupyoktihẽ*
1de						*tupnaʔãncĭyã* *tupnaʔãncĭyãnɨŋ* *tupnancĭyehẽ* *matupyoknancĭyehẽ*			*tupcokoŋa* *tupcokoŋanɨŋ* *tubacehẽ* *matupyoktacehẽ*	*tubumcumma* *tubumcummanɨŋ* *tubumcummehẽ* *matupyoktumcummehẽ*	*tupcekeŋa* *tupcekeŋanɨŋ* *tubacehẽ* *matupyoktacehẽ*
1pe									*tubukumma* *tubukummanɨŋ* *tubummehẽ* *matupyoktummehẽ*		*tubikiŋa* *tubikiŋanɨŋ* *tubiehẽ* *matupyoktiehẽ*
2s	*atupmaʔã* *atupmaʔãnɨŋ* *atubehẽ* *amatupyoktehẽ*			*amatupceke* *amatupcekenɨŋ* *amatubace* *amamatupyoktace*	*amatupno* *amatupniknɨŋ* *amatube* *amamatupyokte*				*atuboko* *atubokonɨŋ* *atube* *amatupyokte*	*atubukuce* *atubukucenɨŋ* *atubuce* *amatupyoktuce*	*atupno* *atupniknɨŋ* *atube* *amatupyokte*
2d	*atupmaʔancɨŋ* *atupmaʔancɨŋnɨŋ* *atubaŋcɨhẽ* *amatupyoktaŋcɨhẽ*								*atupcoko* *atupcokonɨŋ* *atubace* *amatupyoktace*	*atubumcum* *atubumcumnɨm* *atubumcumhẽ* *amatupyoktumcumhẽ*	*atupceke* *atupcekenɨŋ* *atubace* *amatupyoktace*
2p	*atupmaʔanɨŋ* *atupmaʔaninɨŋ* *atubaŋnɨhẽ* *amatupyoktaŋnɨhẽ*								*atubukum* *atubukumnɨm* *atubumhẽ* *amatupyoktumhẽ*		*atubiki* *atubikinɨŋ* *atubihẽ* *amatupyoktihẽ*
3s	*utupmaʔã* *utupmaʔãnɨŋ* *utubehẽ* *umatupyoktehẽ*	*maitupceke* *maitupcekenɨŋ* *maitubace* *maimatupyoktace*	*maitupno* *maitupniknɨŋ* *maitube* *maimatupyokte*	*matupceke* *matupcekenɨŋ* *matubace* *mamatupyoktace*	*matupno* *matupniknɨŋ* *matube* *mamatupyokte*	*natupno* *natupniknɨŋ* *natube* *namatupyokte*	*natupceke* *natupcekenɨŋ* *natubace* *namatupyoktace*	*natubiki* *natubikinɨŋ* *natubihẽ* *namatupyoktihẽ*	*tuboko* *tubokonɨŋ* *tube* *matupyokte*	*tubukuce* *tubukucenɨŋ* *tubuce* *matupyoktuce*	*tupno* *tupniknɨŋ* *tube* *matupyokte*
3d	*utupmaʔancɨŋ* *utupmaʔancɨŋnɨŋ* *utubaŋcɨhẽ* *umatupyoktaŋcɨhẽ*								*utupcoko* *utupcokonɨŋ* *utubace* *umatupyoktace*	*utubukuce* *utubukucenɨŋ* *utubuce* *umatupyoktuce*	*utupceke* *utupcekenɨŋ* *utubace* *umatupyoktace*
3p	*utupmaʔanɨŋ* *utupmaʔaninɨŋ* *utubaŋnɨhẽ* *umatupyoktaŋnɨhẽ*								*utuboko* *utubokonɨŋ* *utube* *umatupyokte*		*utupno* *utupniknɨŋ* *utube* *umatupyokte*

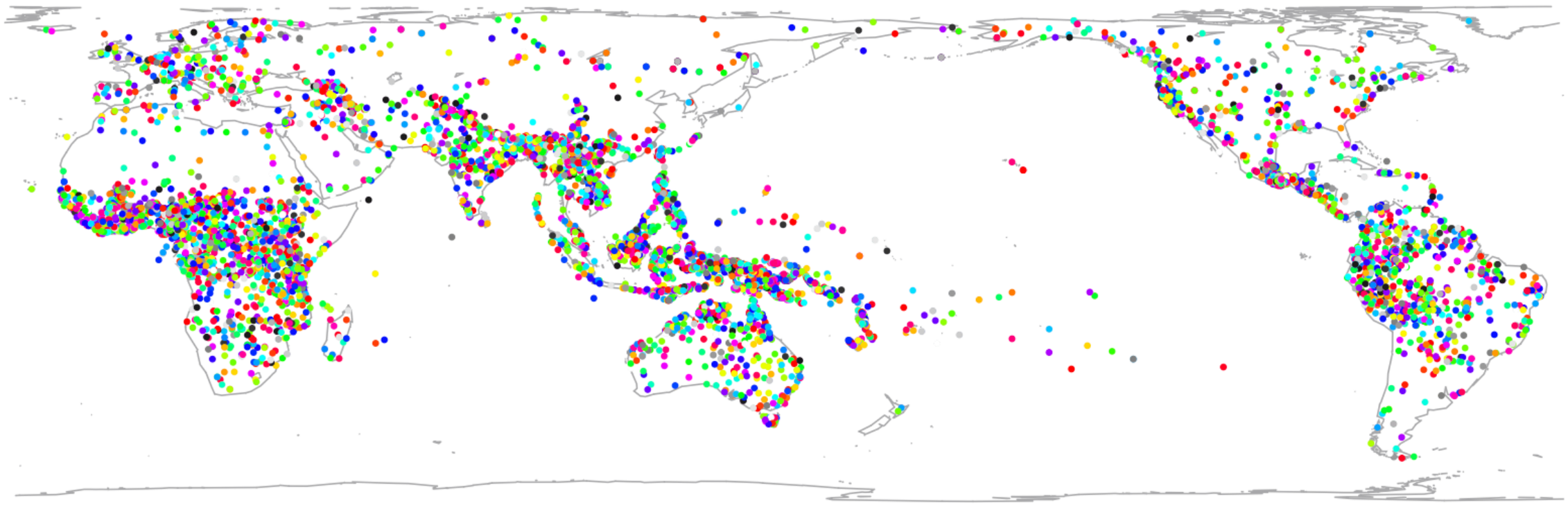
Table 1: Chintang agreement paradigm of the verb *tupma* 'to meet', with stem *tup* (identical in all forms) (Vertical axis: subject agreement; horizontal axis: object agreement. Within each cell, the forms denote (in vertical order) nonpast affirmative, nonpast negative, past affirmative, and past negative tenses, respectively. *Abbreviations:* s 'singular', d 'dual', p 'plural', ns 'nonsingular (dual or plural)', i 'inclusive of addresse', e 'exclusive of addresse', 1-3 denote persons.)

Learning challenges

- Consequences of grammar for the input (n=95k verbs)

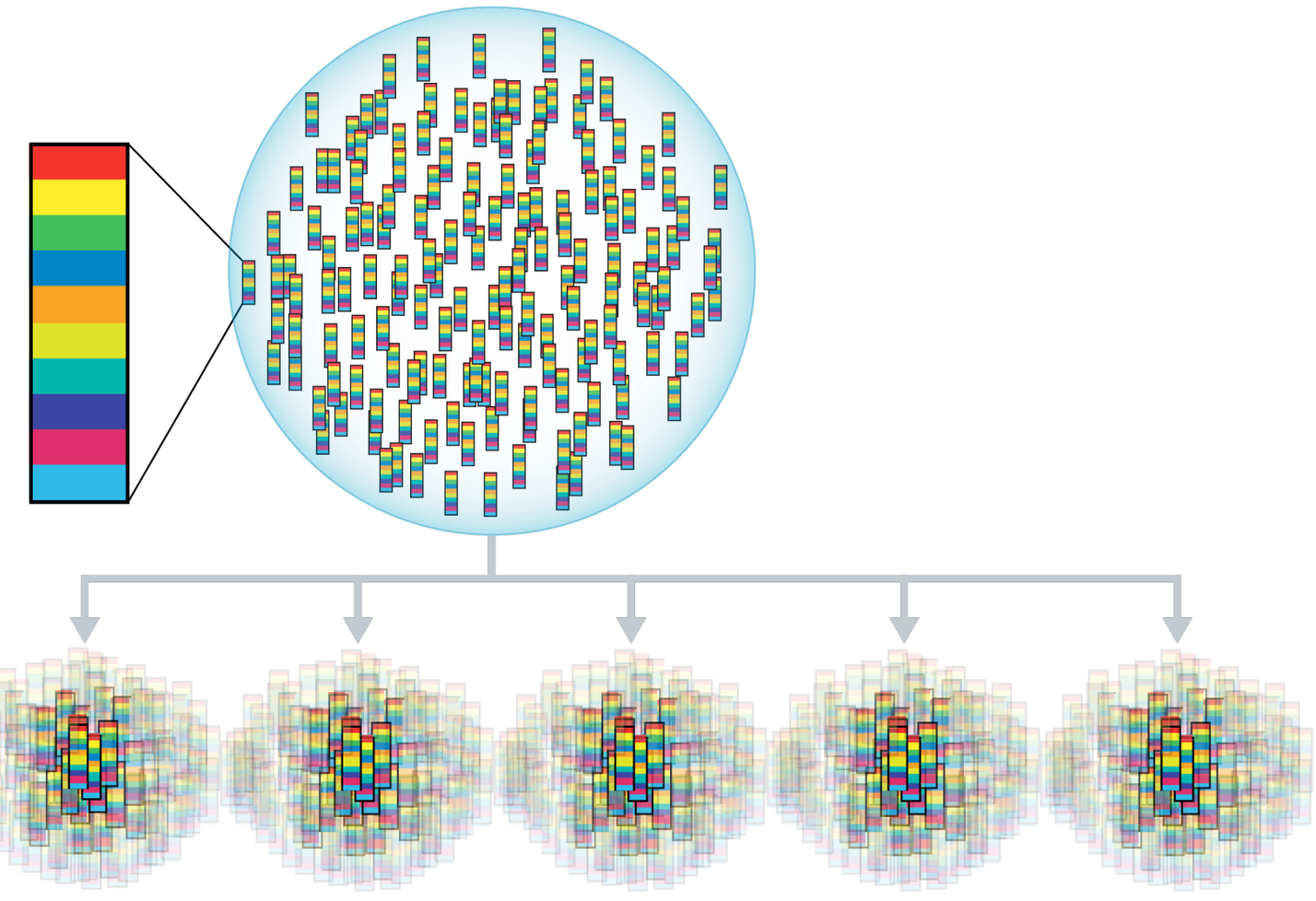


Design space of language: a sampling challenge



Sampling of languages: maximum-diversity

Language features
word order
synthesis
exponence
case marking
inflectional compactness of categories
existence of inflectional classes
…



Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Turkish	Indonesian	Inuktitut	Sesotho	Dene
Japanese	Yucatec	Chintang	Russian	Cree

ACQDIV languages

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	Recording	Duration	Sessions	Words
Chintang	2	2	monthly	4h	477	828'272
Cree	1	1	2-3 weeks	30-40 mins	25	21'525
Indonesian	5	7	bi-weekly	1h	997	2'496'828
Inuktitut	4	1	monthly	4h	77	73'302
Japanese	6	4	weekly	1-1.5h	362	1'235'364
Russian	4	4	weekly	1h	383	2'022'992
Sesotho	3	1	monthly	3-4h	129	237'247
Turkish	8	8	bi-weekly	1h	373	1'139'877
Yucatec	3	3	bi-weekly	30-90 mins	234	120'441

The corpora

Cluster	Language	Format	Session MD	Speaker MD
1	Turkish	Quasi-CHAT	Quasi-CHAT	Quasi-CHAT
1	Japanese	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

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>
```

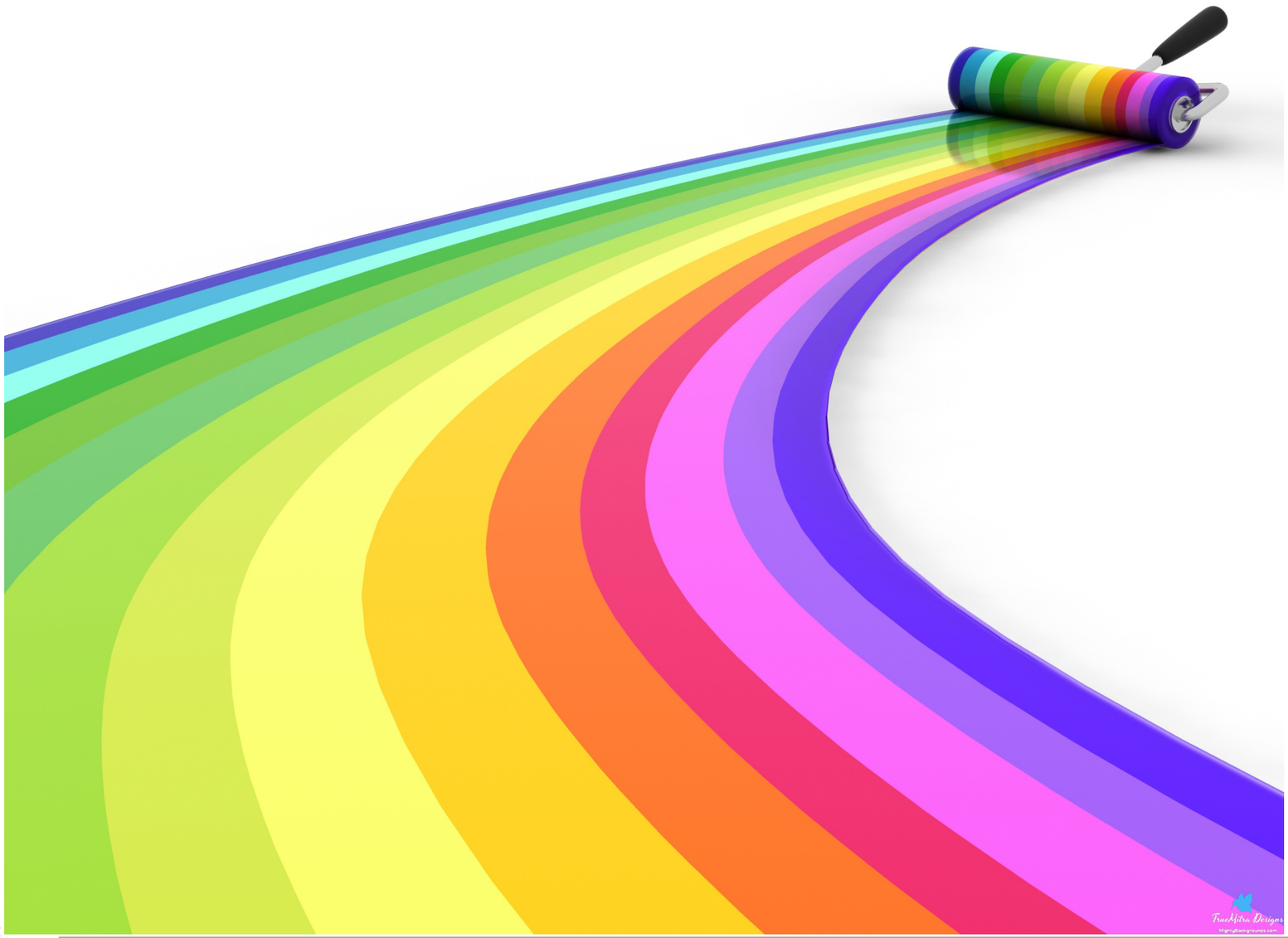
Toolbox

```
\_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
```

Initial state of the maximally diverse corpora

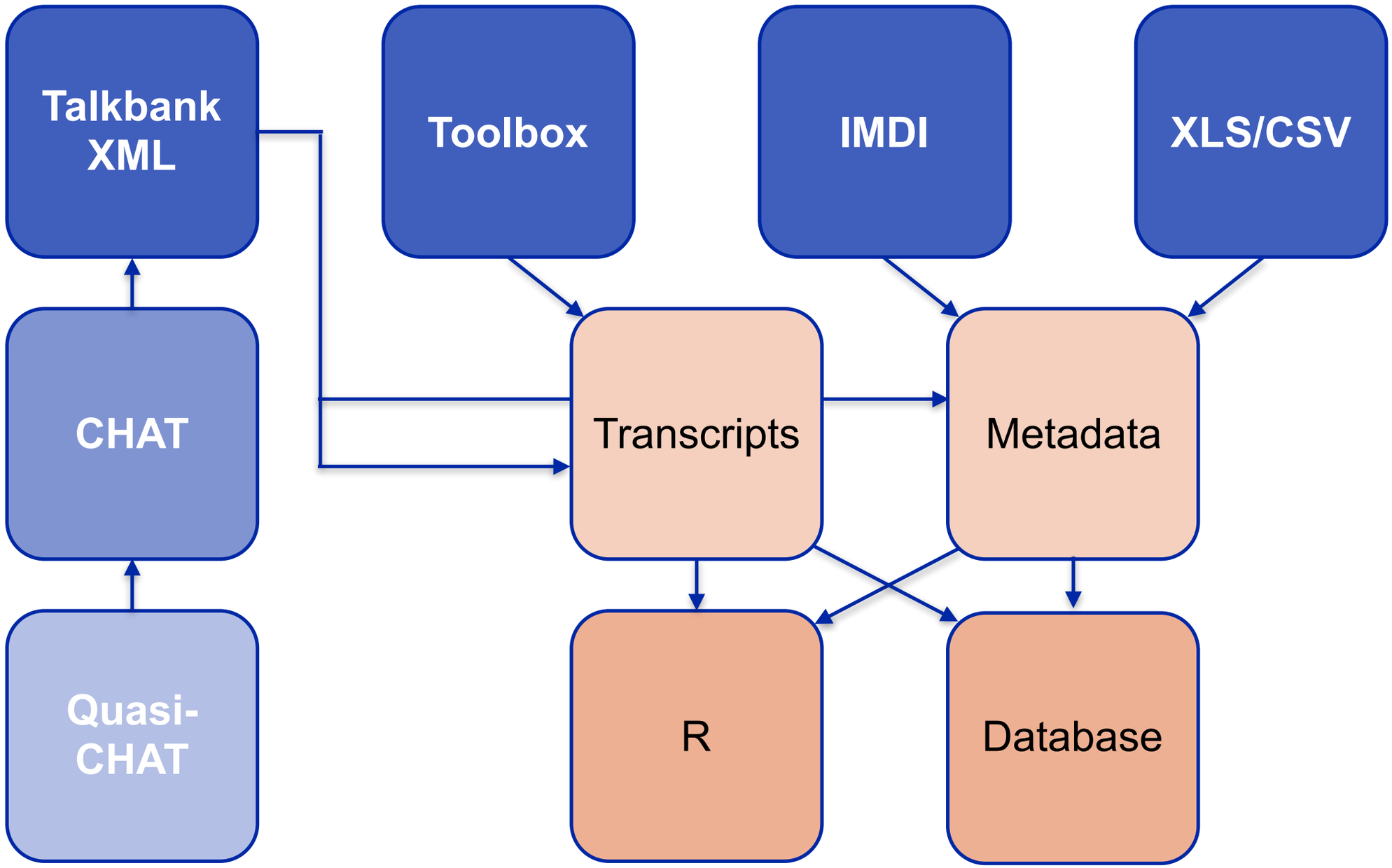
The “final” state of the ACQDIV corpora



Extract, Transform, Load pipeline

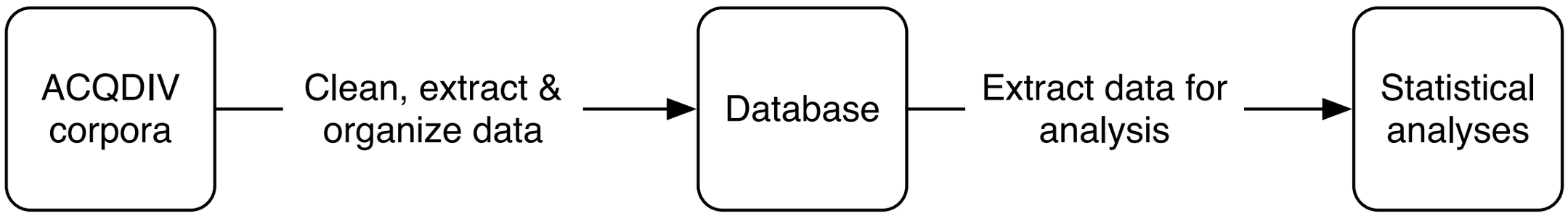
- Usability
- Interoperability
  - Syntactic (structural)
  - Semantic (conceptual)
- Replicability
  - Code in repository
- Fixability
  - Unit and regression tests

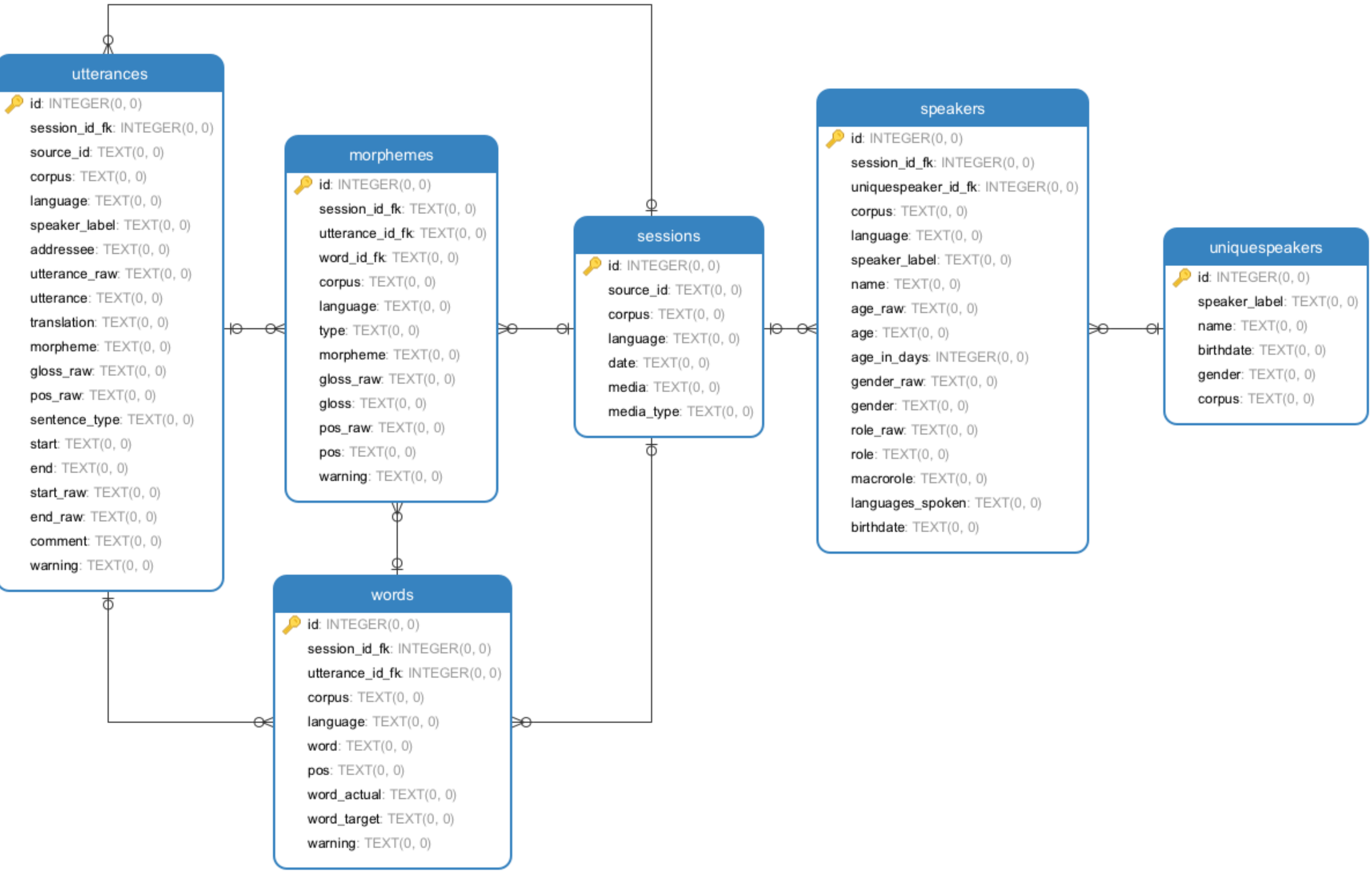
Transformations in ACQDIV



Challenges for interoperability

- Unifying different input formats
- Correcting legacy character codes, data formats —> UTF-8 NO-BOM NFD
- Speaker ages, names, roles
- Utterance timestamps
- Part-of-speech tag set
- Morphological glosses




utterances
id: INTEGER(0, 0)
session_id_fk: INTEGER(0, 0)
source_id: TEXT(0, 0)
corpus: TEXT(0, 0)
language: TEXT(0, 0)
speaker_label: TEXT(0, 0)
addressee: TEXT(0, 0)
utterance_raw: TEXT(0, 0)
utterance: TEXT(0, 0)
translation: TEXT(0, 0)
morpheme: TEXT(0, 0)
gloss_raw: TEXT(0, 0)
pos_raw: TEXT(0, 0)
sentence_type: TEXT(0, 0)
start: TEXT(0, 0)
end: TEXT(0, 0)
start_raw: TEXT(0, 0)
end_raw: TEXT(0, 0)
comment: TEXT(0, 0)
warning: TEXT(0, 0)
morphemes
id: INTEGER(0, 0)
session_id_fk: TEXT(0, 0)
utterance_id_fk: TEXT(0, 0)
word_id_fk: TEXT(0, 0)
corpus: TEXT(0, 0)
language: TEXT(0, 0)
type: TEXT(0, 0)
morpheme: TEXT(0, 0)
gloss_raw: TEXT(0, 0)
gloss: TEXT(0, 0)
pos_raw: TEXT(0, 0)
pos: TEXT(0, 0)
warning: TEXT(0, 0)
sessions
id: INTEGER(0, 0)
source_id: TEXT(0, 0)
corpus: TEXT(0, 0)
language: TEXT(0, 0)
date: TEXT(0, 0)
media: TEXT(0, 0)
media_type: TEXT(0, 0)
speakers
id: INTEGER(0, 0)
session_id_fk: INTEGER(0, 0)
uniquespeaker_id_fk: INTEGER(0, 0)
corpus: TEXT(0, 0)
language: TEXT(0, 0)
speaker_label: TEXT(0, 0)
name: TEXT(0, 0)
age_raw: TEXT(0, 0)
age: TEXT(0, 0)
age_in_days: INTEGER(0, 0)
gender_raw: TEXT(0, 0)
gender: TEXT(0, 0)
role_raw: TEXT(0, 0)
role: TEXT(0, 0)
macrorole: TEXT(0, 0)
languages_spoken: TEXT(0, 0)
birthdate: TEXT(0, 0)
uniquespeakers
id: INTEGER(0, 0)
speaker_label: TEXT(0, 0)
name: TEXT(0, 0)
birthdate: TEXT(0, 0)
gender: TEXT(0, 0)
corpus: TEXT(0, 0)
words
id: INTEGER(0, 0)
session_id_fk: INTEGER(0, 0)
utterance_id_fk: INTEGER(0, 0)
corpus: TEXT(0, 0)
language: TEXT(0, 0)
word: TEXT(0, 0)
pos: TEXT(0, 0)
word_actual: TEXT(0, 0)
word_target: TEXT(0, 0)
warning: TEXT(0, 0)

ACQDIV database

	id	language	speaker_label	utterance	translation	morpheme	start	end
3971	3971	Chintang	MR	maʔmi yaŋ adha leʔle thaʔno...	Only half a person is visible.	maʔmi yaŋ adha leʔle that -no raicha	641.450	644.070
3972	3972	Chintang	LDCh4	ei	Oh!	hei	641.653	642.090
3973	3973	Chintang	LDCh4	kuŋse gonei	He came!	kuŋs -e gonei	642.901	644.059
3974	3974	Chintang	LDCh4	ek china	One moment.	ek chin -a	644.059	644.481
3975	3975	Chintang	Susma	oi	Oh!	oi	644.481	645.250
3976	3976	Chintang	Juna	pheknoʔ ni huĩ phekno	He sweeps it, he sweeps.	phek -no ni hun phek -no	644.737	646.112
3977	3977	Chintang	Sapana	huĩ	That one.	hun	646.112	647.296
3978	3978	Chintang	Sapana	aho kalo lisaseʔ hou	Oh! It became black.	aho kalo lis -a -ŋs -e -ʔ hou	647.296	649.093
3979	3979	Chintang	Santa	utha na	Get up.	u- tha na	649.093	649.765
3980	3980	Chintang	Juna	akka	I.	akka	649.765	650.149
3981	3981	Chintang	Juna	akka	I.	akka	650.149	650.731
3982	3982	Chintang	Juna	abo thitta	Now, one.	abo thitta	650.731	651.179
3983	3983	Chintang	Bishna	moba aphe kancha	Down there, the brother, Kancha.	mo -beʔ a- phuwa kancho	651.179	652.280
3984	3984	Chintang	Juna	aha akhattoko	Oh no, you take it away!	aho a- khatt -u -kV	651.569	652.320
3985	3985	Chintang	Susma	akka bago	I do this one.	akka ba -go	652.113	653.010
3986	3986	Chintang	Juna	huĩ themkha	What is that?	hun them -kha	654.550	655.158
3987	3987	Chintang	Susma	hanako na ba com	It is of your kind.	hana -ko na ba com	655.158	656.590
3988	3988	Chintang	Bipana	akko bhayu them bhayu	Mine is here, what is here?	akka -ko ba -bayu them ba -bayu	655.516	657.051
3989	3989	Chintang	Sapana	lak lunoʔ ni	He dances.	lak lus -no ni	657.051	658.000
3990	3990	Chintang	Susma	oi	Oh!	oi	657.435	658.075
3991	3991	Chintang	Susma	akka aseĩ	I (danced) some days ago.	akka aseĩ	658.075	659.105
3992	3992	Chintang	Susma	huĩ yaŋ	That one too.	hun yaŋ	659.105	660.167

Universität Zürich[UZH]

Preliminary research: role of input in language acquisition

- What is the role of input in the language acquisition process?
  - Children are excellent statistical learners
  - Frequency plays an important role in acquisition
  - Studies have suggested that children find distributional regularities in the speech signal that aid in language acquisition
  - Maratsos and Chalkley (1980) proposed that distributional information in the input, e.g., word co-occurrence patterns, could be a cue for categorizing

Preliminary research: role of input in language acquisition

- Syntactic categories (e.g. noun and verb) are the basic units of grammar
  - Grammatical rules are defined over syntactic categories rather than words, e.g. “Colorless green ideas sleep furiously”
- Word categorization is a prerequisite for acquiring an adult-state grammar
- How do children learn word categories and which source of information plays a primary role in the process?

Preliminary research: role of input in language acquisition

- Observation: words occurring in the same context often belong to the same grammatical category (Bloomfield 1933:276)

The regular analogies of a language are habits of substitution. Suppose, for instance, that a speaker had never heard the form *give Annie the orange*, but that he had heard or spoken a set of forms like the following:

Baby is hungry. Poor Baby! Baby's orange. Give Baby the orange!
Papa is hungry. Poor Papa! Papa's orange. Give Papa the orange!
Bill is hungry. Poor Bill! Bill's orange. Give Bill the orange!
Annie is hungry. Poor Annie! Annie's orange. . . . . .

He has the habit, now, — the analogy, — of using *Annie* in the same positions as *Baby*, *Papa*, *Bill*, and accordingly, in the proper situation, will utter the new form *Give Annie the orange!* When a speaker utters a complex form, we are in most cases unable to tell whether he has heard it before or has created it on the analogy of other forms. The utterance of a form on the analogy of other forms is like the solving of a proportional equation with an indefinitely large set of ratios on the left-hand side:

$$\left.\begin{array}{l} \textit{Baby is hungry} : \textit{Annie is hungry} \\ \textit{Poor Baby} : \textit{Poor Annie} \\ \textit{Baby's orange} : \textit{Annie's orange} \end{array}\right\} = \textit{Give Baby the orange} : x$$

or

$$\left.\begin{array}{l} \textit{dog} : \textit{dogs} \\ \textit{pickle} : \textit{pickles} \\ \textit{potato} : \textit{potatoes} \\ \textit{piano} : \textit{pianos} \end{array}\right\} = \textit{radio} : x$$

Frequent frames aid in distributional learning?

- Basic idea put forth by Mintz and others is that child-directed speech contains patterns that allow the learner to accurately predict categories given certain ngram contexts because those contexts are more predictive of specific categories (whether those categories are labels for parts-of-speech, morpheme categories, etc.)

- John ate fish.
- John ate rabbits.
- John can fish.
- *John can rabbits.

Universität Zürich[UZH]

Mintz, Toby. 2003. Frequent frames as a cue for grammatical categories in child directed speech. Cognition 90 (2003) 91–117.

Frames bring highly predictive structure to the input, but?

- Mintz (2002) showed that how learning of grammatical categories is possible by analyzing the distributional word context (as in nouns and verbs in English)

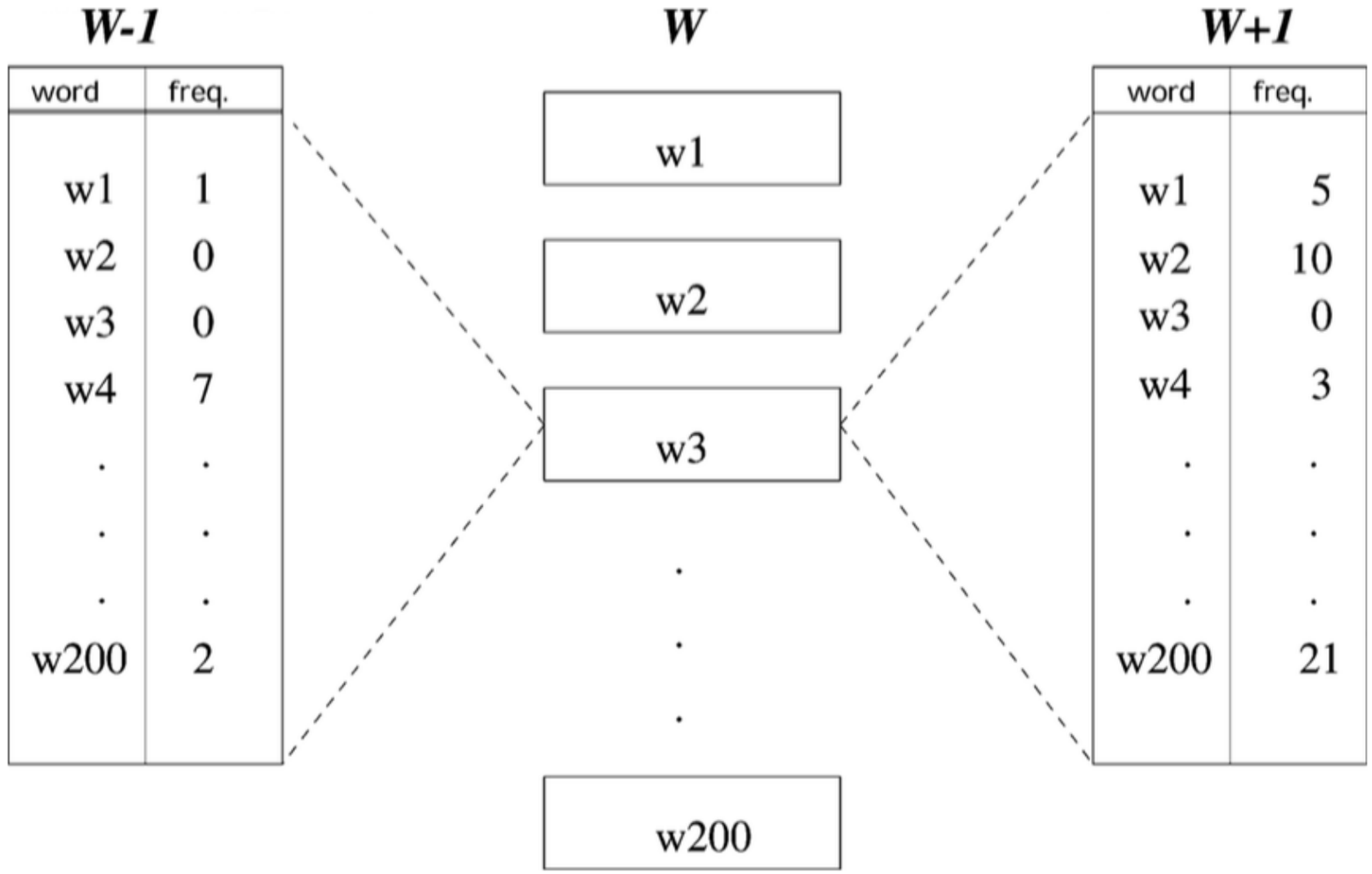


Fig. 1. Representation of distributional contexts.

Frequent frame analysis

- Accuracy (precision)
  - accuracy = hits / hits + false alarms
  - i.e. how accurate is the frame?

$$P = \frac{T_p}{T_p + F_p}$$

- Completeness (recall)
  - completeness = hits / hits + misses
  - i.e. how many types does it categorize?

$$R = \frac{T_p}{T_p + F_n}$$

- Comparison against chance categorization
- Analysis of a set of 45 “frequent” frames

Frequent frames aid in distributional learning?

Table 1: Previous studies

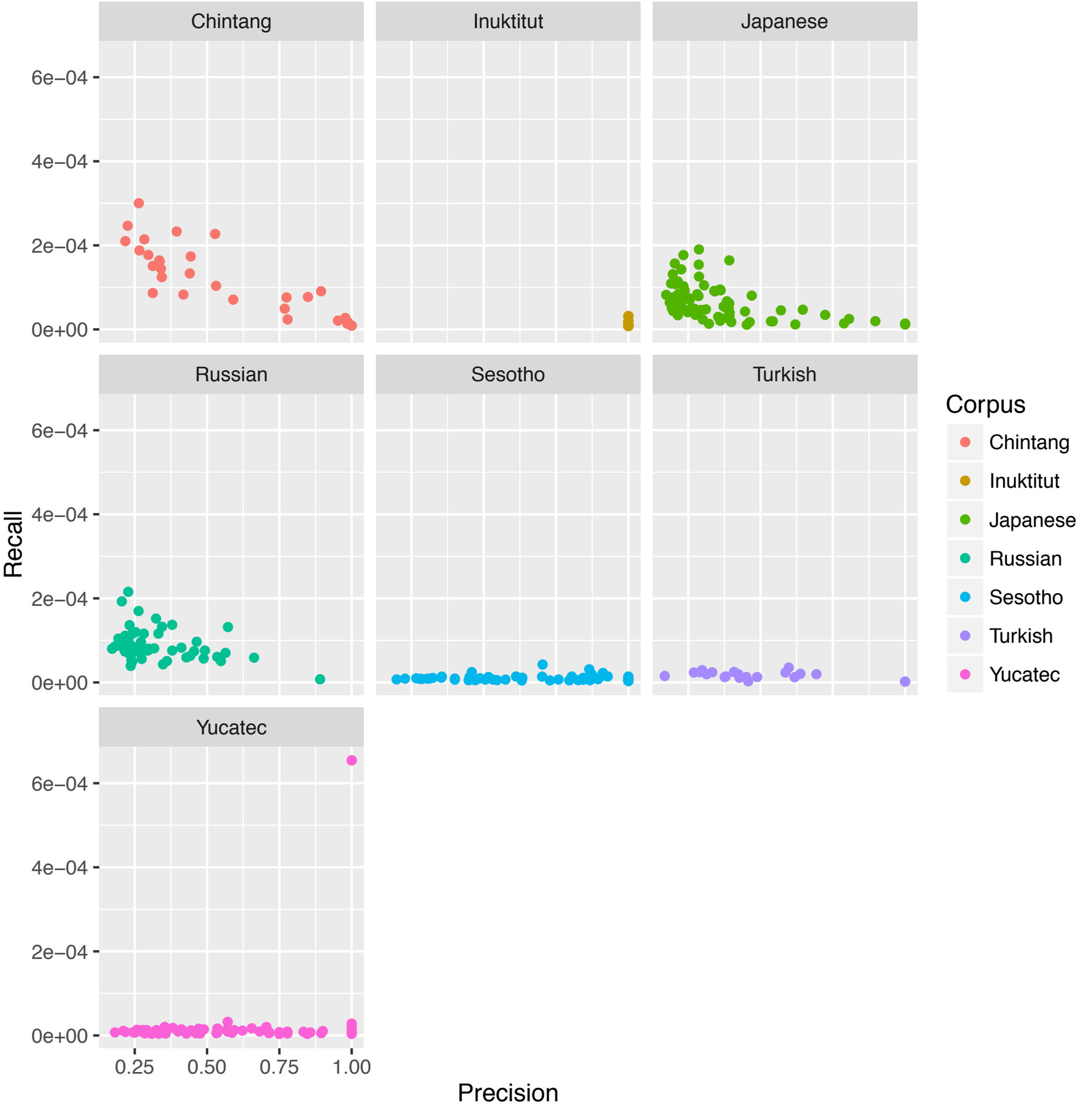
Language	Result	Accuracy	Source
Chinese	some degree of success	.68–.71	Xiao et al. (2006)
Dutch	not accurate lexical categorization	.40–.71	Erkelens (2009)
French	robust	1.00	Chemla et al. (2009)
Spanish	robust	.75	Weisleder & Waxman (2010)
Turkish	robust (word and morpheme)	.47–.90	Wang et al. (2011)
German	robust (word and morpheme)	.86–.88	Wang et al. (2011)
German	not accurate lexical categorization	.42–.77	Stumper et al. (2011)

Table 2: Data

Language	Data	Data type	Sources
Chinese	two children	words	Xiao et al. (2006)
Dutch	four children	words	Erkelens (2009)
French	one child	words	Chemla et al. (2009)
Spanish	three children	words	Weisleder & Waxman (2010)
Turkish	two children	word and morpheme	Wang et al. (2011)
German	one child	word and morpheme	Wang et al. (2011)
German	one child	word	Stumper et al. (2011)

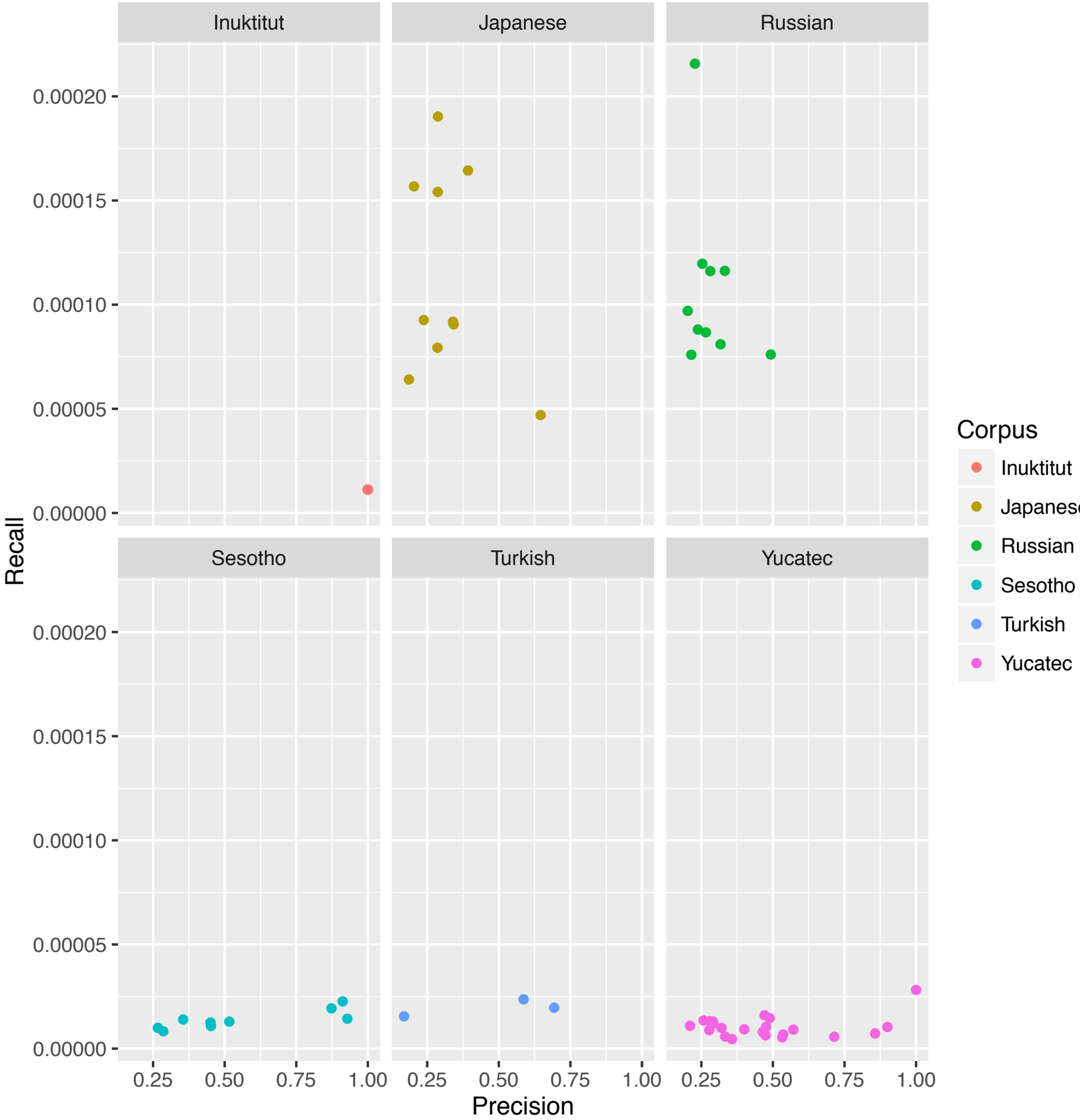
Frequent frames aid in distributional learning?

- Our hypothesis is that frequent frames are not cross-linguistically informative at the word level
- Extract trigrams from utterances of CDS in the database
- We generate the precision and recall measures for the frequent frames
  - To be informative accuracy must be over 50%
- We operationalize the number of frequent frames in different corpus sizes by looking at their relative frequency in each corpus


Chintang
Inuktitut
Japanese
Russian
Sesotho
Turkish
Yucatec
6e−04
4e−04
2e−04
0e+00
0.25
0.50
0.75
1.00
Precision
Recall
Corpus
Chintang
Inuktitut
Japanese
Russian
Sesotho
Turkish
Yucatec

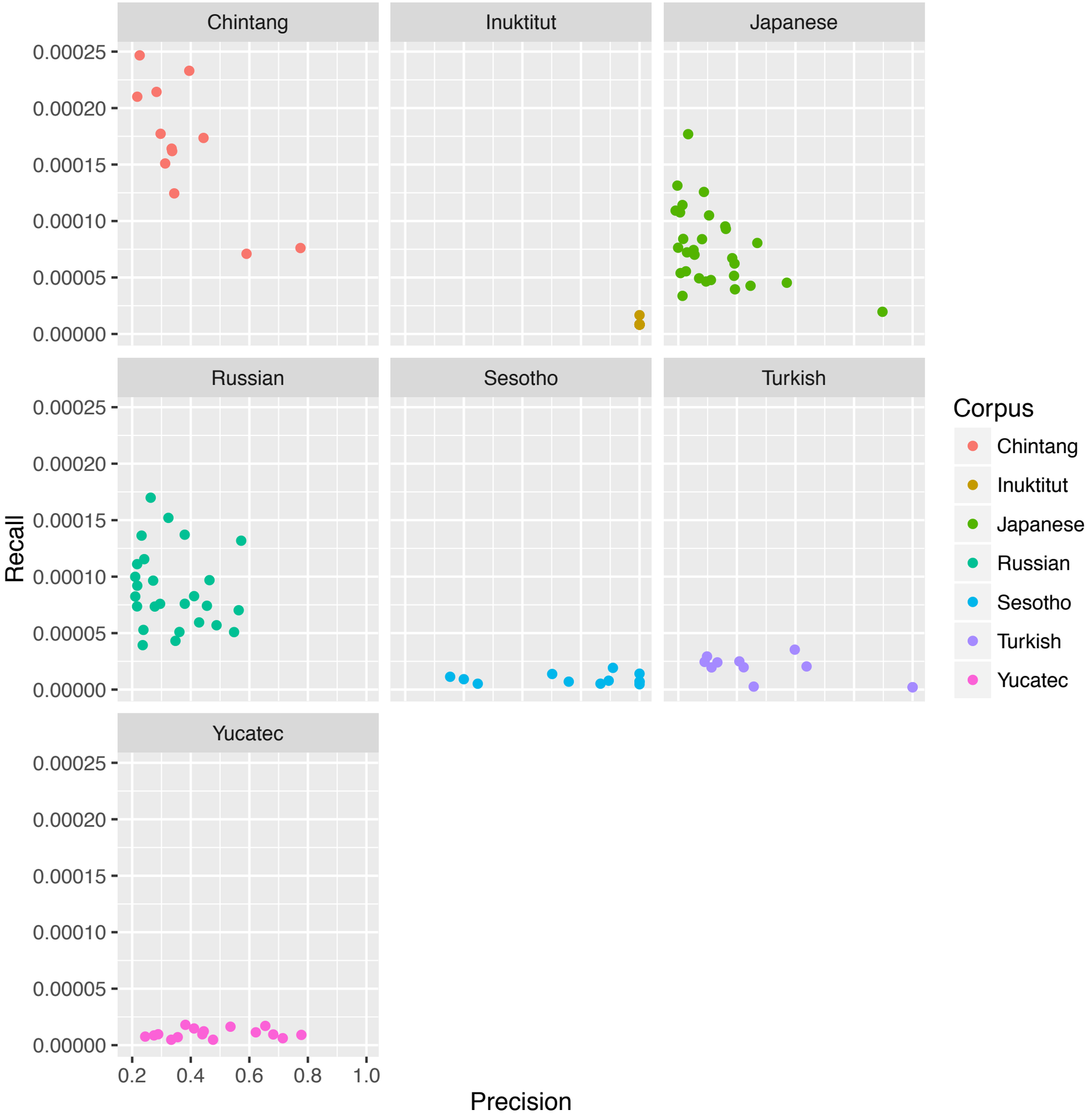
Results: frequent frames word level

Corpus	Chintang	Inuktitut	Japanese	Russian	Sesotho	Turkish	Yucatec
Trigrams	179971	474	139245	418697	26222	72056	12848
FF	32	41	75	52	57	21	73
Accuracy	0.58	1	0.37	0.32	0.67	0.48	0.54


Inuktitut
Japanese
Russian
Sesotho
Turkish
Yucatec
Recall
Precision
0.00020
0.00015
0.00010
0.00005
0.00000
0.25
0.50
0.75
1.00
Corpus
Inuktitut
Japanese
Russian
Sesotho
Turkish
Yucatec

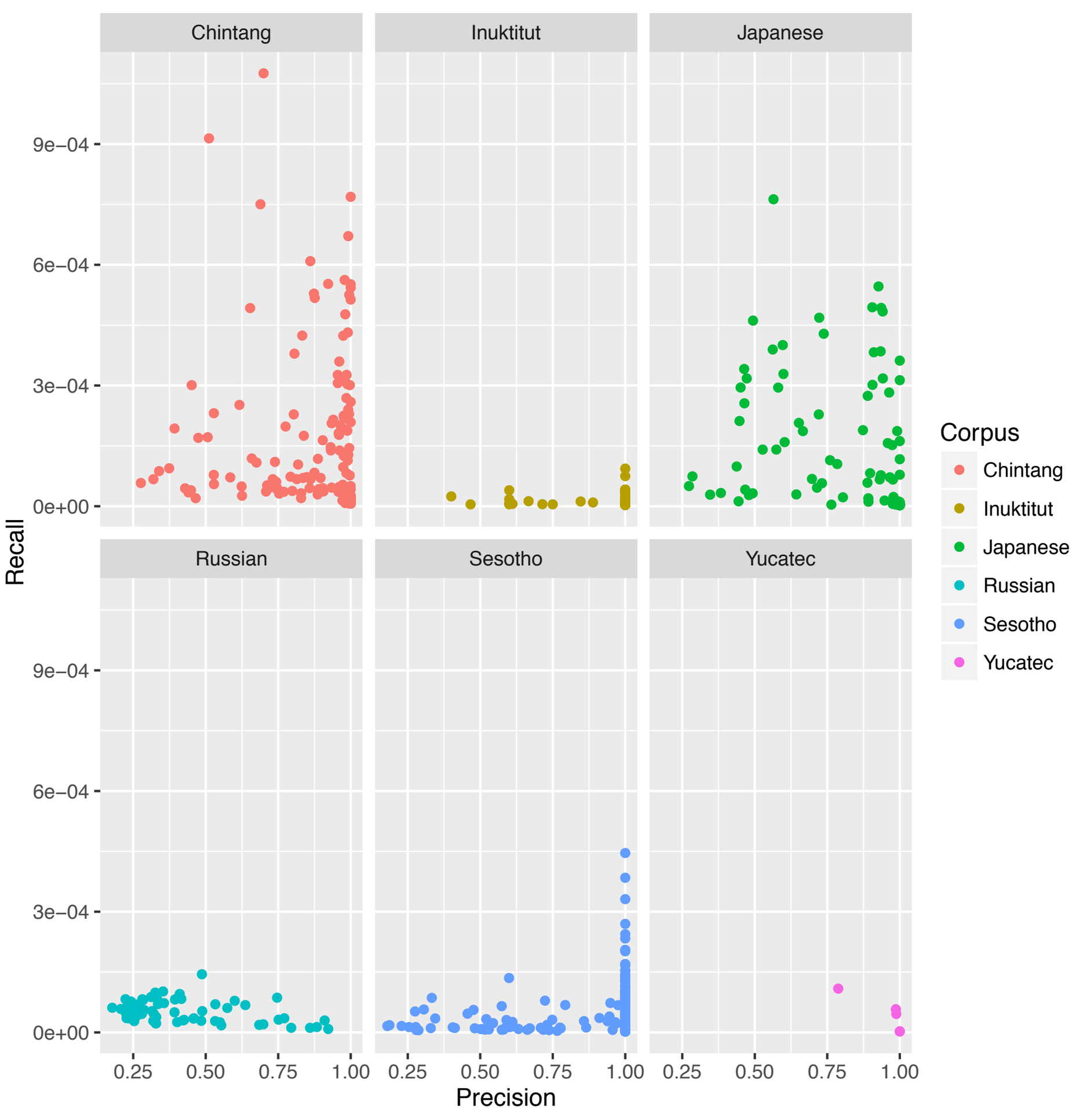
Nouns

Verbs


Chintang
Inuktitut
Japanese
Russian
Sesotho
Turkish
Yucatec
Recall
Precision
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0.2
0.4
0.6
0.8
1.0
Corpus
Chintang
Inuktitut
Japanese
Russian
Sesotho
Turkish
Yucatec

Summary

- We created a syntactically and semantically interoperable database compiled from technologically disparate and typologically maximally diverse child language acquisition corpora
- This single unified database allows us to investigate qualitatively and quantitatively questions in cross-linguistic child language acquisition
- One of our preliminary research topics focuses on distributional patterns in CDS for part-of-speech categorization
    - Accurate categorization of frames at the word level is not


Chintang
Inuktitut
Japanese
Russian
Sesotho
Yucatec
9e−04
6e−04
3e−04
0e+00
0.25
0.50
0.75
1.00
Precision
Recall
Corpus
Chintang
Inuktitut
Japanese
Russian
Sesotho
Yucatec

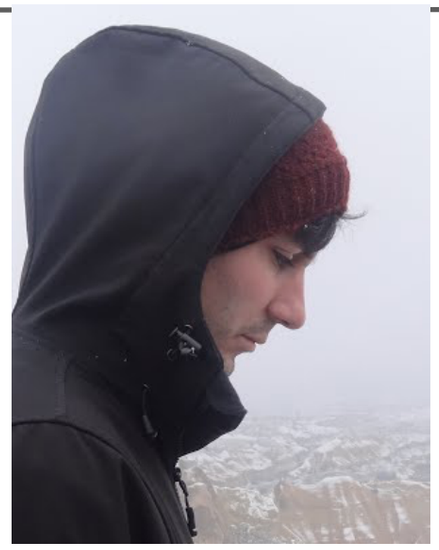
Results: frequent frames word level

Corpus	Chintang	Inuktitut	Japanese	Russian	Sesotho	Yucatec
Trigrams	37721	908	16690	53504	7354	4529
FF	163	95	108	66	195	5
Accuracy	0.87	0.95	0.83	0.41	0.88	0.95

Thank you from the ACQDIV team!

Prof. Sabine Stoll
PI

Dagmar Jung,
Dene corpus development, typology



Damián Blasi,
statistics, models

Robert Schikowski,
Project management

Steven Moran,
database development computational linguistics

Katia Mazara,
visualizations, stat. analyses

Anna Jansco, data base development

Andreas Gerster,
Dene, data prepreparation

Melanie Trüssel,
Dene, data prepreparation

Casim Hysi,
database development

University of Zurich[UZH]

erc