



University of
Zurich ^{UZH}



Frequent frames in maximally diverse languages

Steven Moran
Damián Blasi
Sabine Stoll

Department of Comparative Linguistics
University of Zurich

BUCLD, 4 — 6 November, Boston



distributional learning in infants

adjacent dependencies



Saffran, Aslin & Newport 1996
Aslin, Saffran & Newport 1998
Swingley, 2005

distributional learning in infants

non-adjacent dependencies

is X-ing

Santelmann & Jusczyk 1998

Gomez 2002

Gomez & Maye 2005

Omnis et al. 2004

Höhle et al. 2006

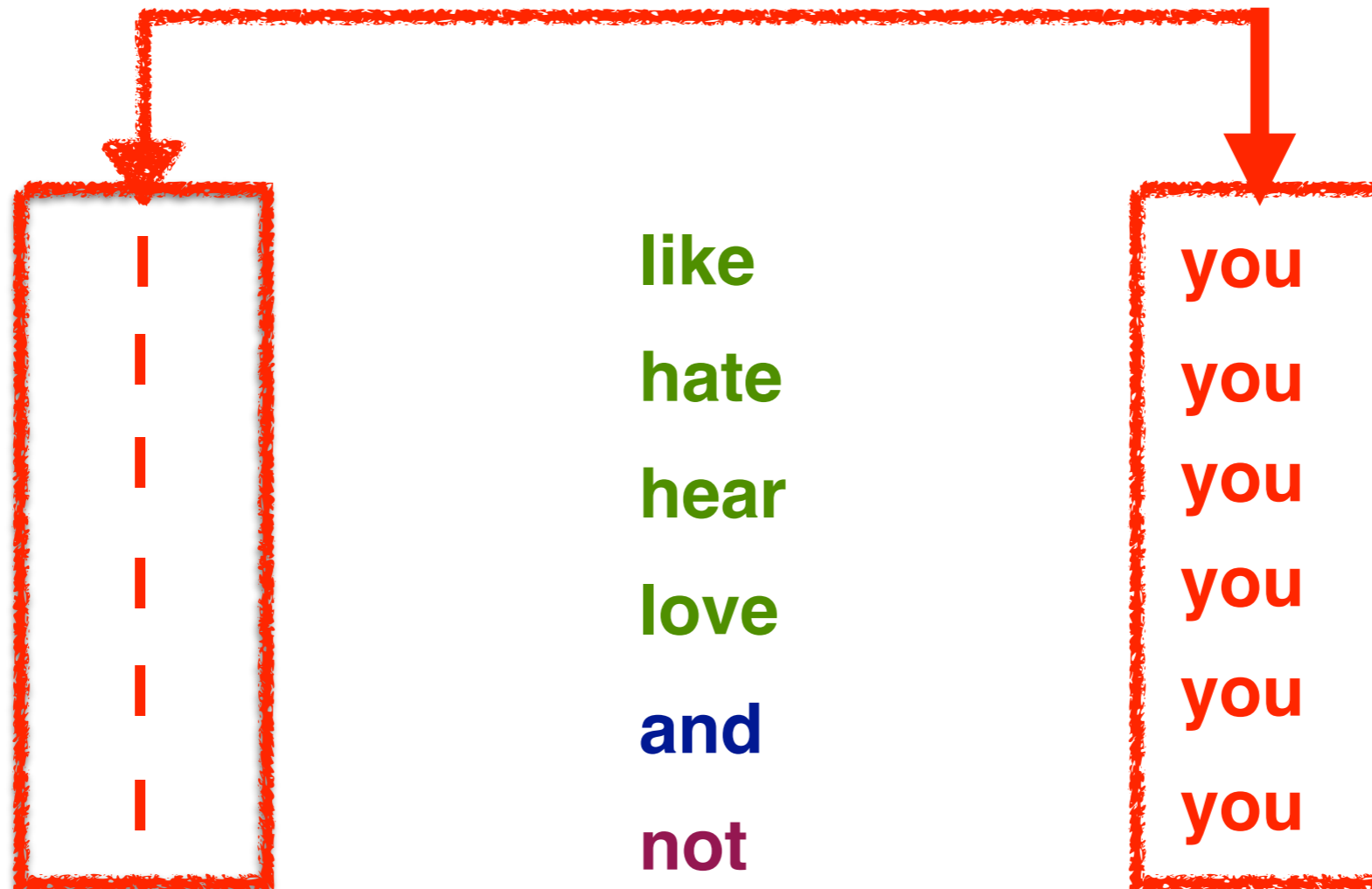
Mintz et al. 2006

Van Kampen et al. 2008

Nazzi et al, 2009, 2011

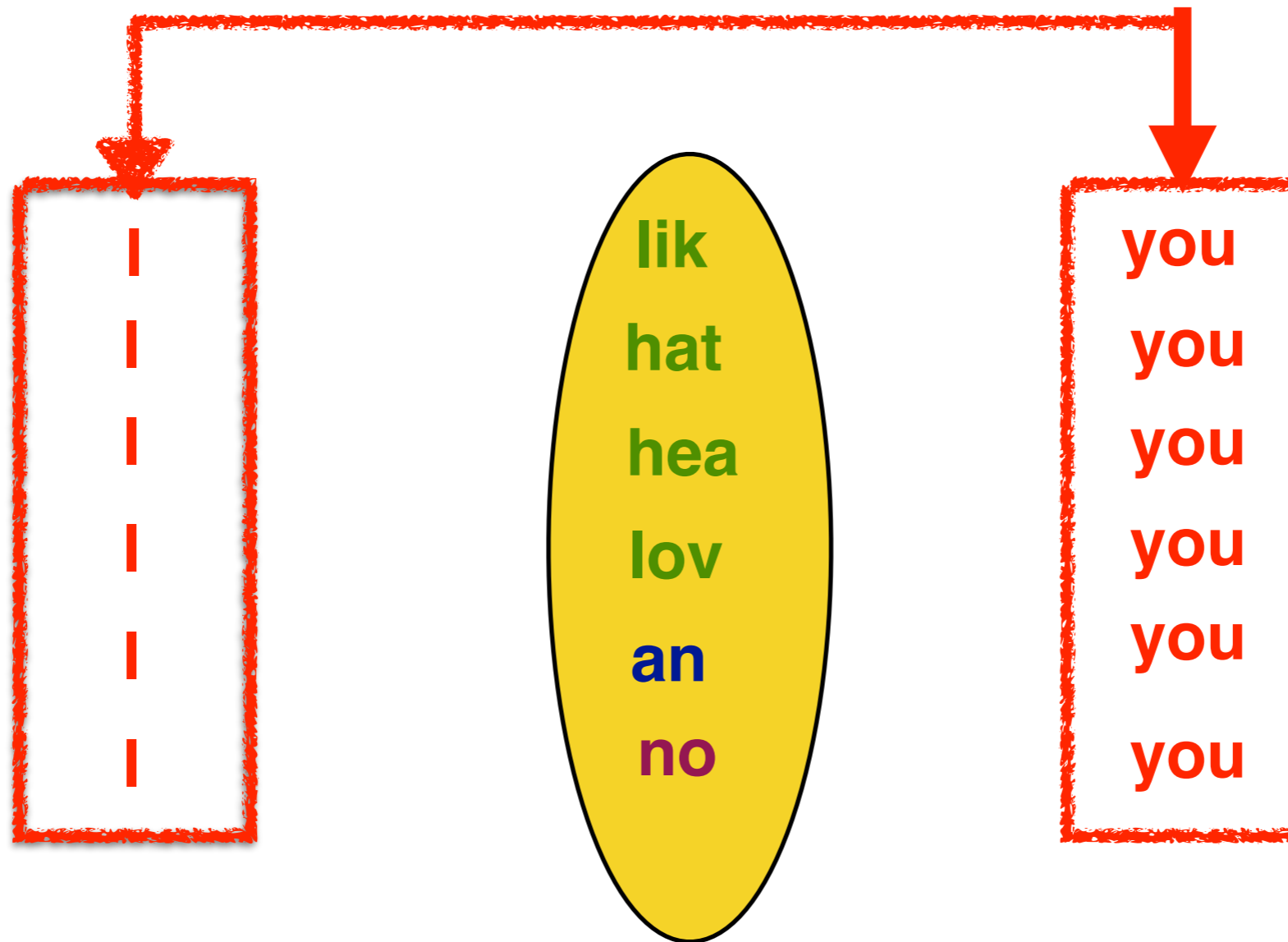
**Are there structures in the input
that could help in categorizing words?**

frames



Mintz, Newport & Bever 2002; Mintz, 2003

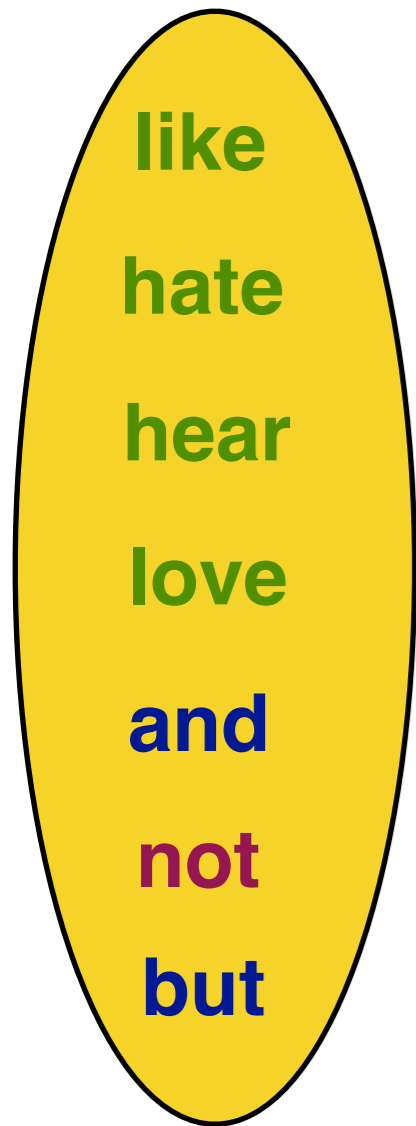
frames



Mintz, Newport & Bever 2002; Mintz, 2003

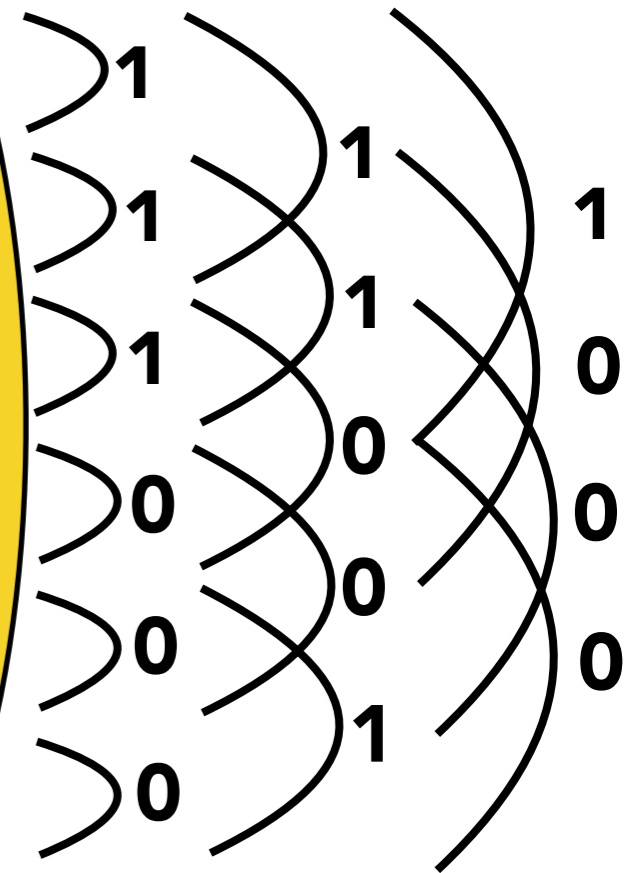
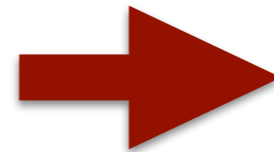
evaluation of frequent frames: precision

for each frame



POS labels

- nouns
- verbs
- adjectives
- prepositions
- adverbs
- determiner
- wh-word
- not
- conjunction
- interjection



$$\text{Accuracy} = \frac{\text{hits}}{\text{hits} + \text{false alarm}}$$

evaluation frequent frames: recall

targets of frequent frame X

- like
- hate
- hear
- love
- and
- not
- but

evaluation corpus

✓		✗		
✗	like	✗	see	
	loathe	✓	smell	
✓	nearly	✗	not	
✓	love	✗	know	
✗	because	✓	hear	
✓	hate		cat	
✓	and		horse	
	dog	✓	but	
	
	

mis
✗
hit
✓

$$\text{Completeness} = \frac{\text{hits}}{\text{hits} + \text{misses}}$$

results frequent frames so far

language	mean precision words	mean recall words (morphemes)	utterances
English (Mintz 2003)	0.91	0.12	103'191
Dutch (Erkelens 2009)	0.71		49'635
French (Chemla et al 2009)	1	0.33	2'006
Spanish (Weisleder & Waxman 2010)	0.75		37'588
Turkish (Wang et al.	0.47 (.91)	0.1 (.06)	37'765
German (Wang et al. 2011)	0.86 (.88)	0.07 (0.05)	5'685
German (Stumper et al. 2011)	0.77		30'601
Chinese (Xiao et al.	0.71		22'137



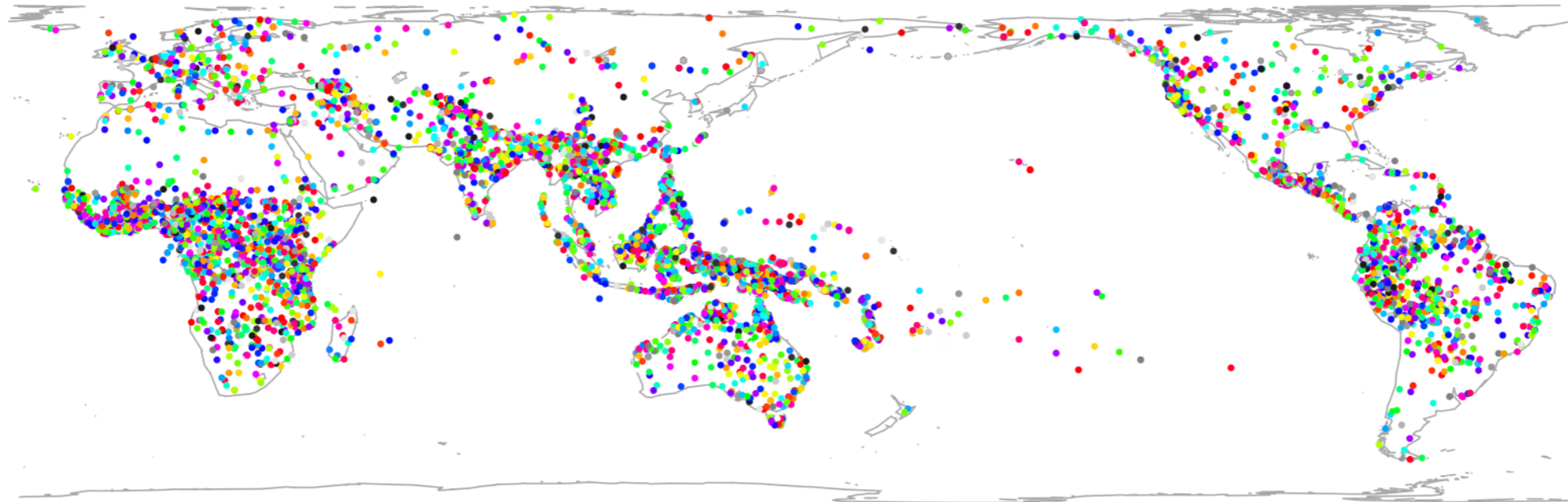
our question:

**are frequent frames in the input
universally reliable patterns for
categorization?**

Contribution of our study:

- test universality of frames in the input**
- use same method for all languages of our sample**
- control for sample size by operationalization of frequent frames**
- show quantitatively whether there are differences between target categories**

universality of frames: sampling challenge



our solution: maximum diversity sample

- **Goal:** capture as much variation as possible
- **ACQDIV** database (**ACQ**uisition processes in maximally **DIV**erse languages, ERC funded project: 2014–2019, PI: Sabine Stoll)
 - **longitudinal corpora** of 10 languages which differ maximally in their grammatical structure (“maximum diversity sampling”)
 - <http://www.acqdiv.uzh.ch/>

ACQDIV database: maximum diversity sample

typological features

word order

synthesis

exponence

case marking

inflectional compactness of categories

existence of inflectional classes

...

Satellites

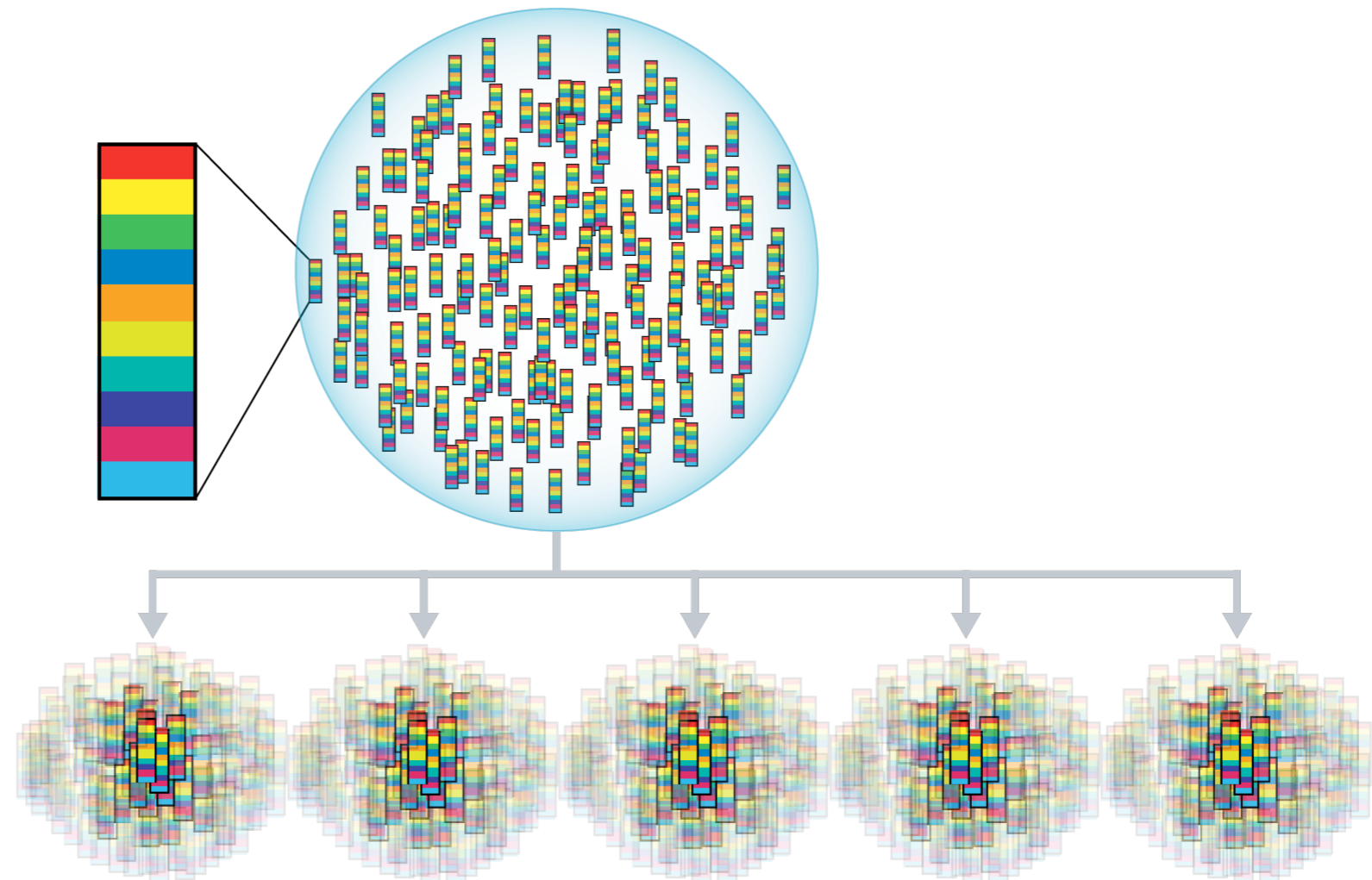
Kata Kolok (sign language)

Tzeltal

Qaqet

Nungon

Pitjanjatjara



Cluster 1

Turkish

Japanese

Cluster 2

Indonesian

Yucatec

Cluster 3

Inuktitut

Chintang

Cluster 4

Sesotho

Russian

Cluster 5

Dene

Cree

ACQDIV corpora



ACQDIV corpora

language	genealogy	size of corpus in words	words used	morphemes used
Chintang	Sino-Tibetan	987'120	473'918	814'076
Inuktitut	Eskimo-Aleut	73'255	23'164	8'673
Japanese	Japonic	821'106	514'344	376'934
Russian	Indo-European	2'033'755	1'316'234	NA
Sesotho	Bantu	237'112	83'514	112'630
Turkish	Turkic	1'136'332	938'955	272'459
Yucatec	Mayan	257'496	89'219	84'928

ACQDIV database

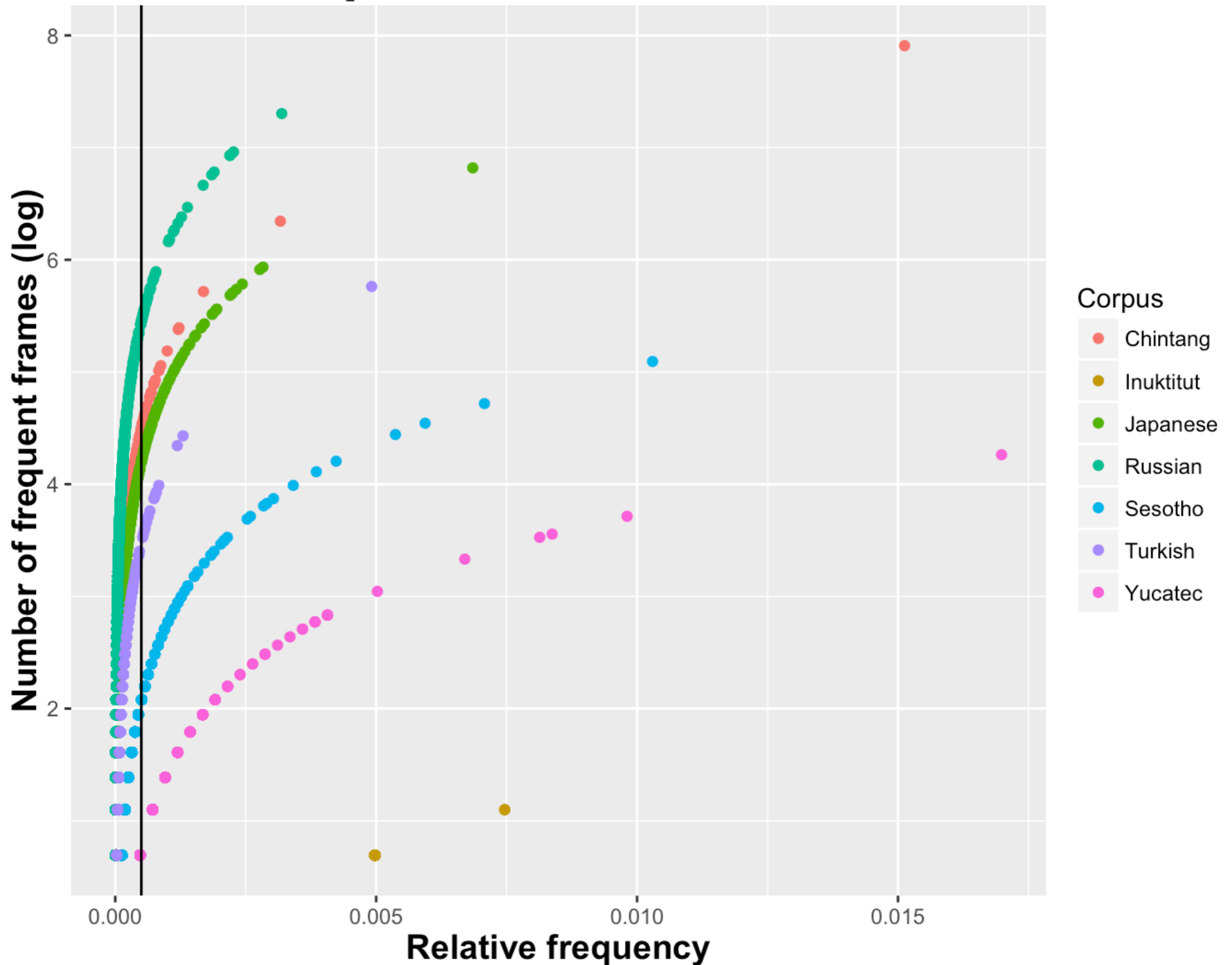
	id	language	speaker_label	utterance	translation	morpheme	start	end
3971	3971	Chintang	MR	maʔmi yaŋ adha leʔle thaʔno...	Only half a person is visible.	maʔmi yaŋ adha leʔle that -no raicha	641.450	644.070
3972	3972	Chintang	LDCh4	ei	Oh!	hei	641.653	642.090
3973	3973	Chintang	LDCh4	kuŋse gonei	He came!	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 pheknɔ	He sweeps it, he sweeps.	phek -no ni hun pheknɔ	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

Step 1: frames

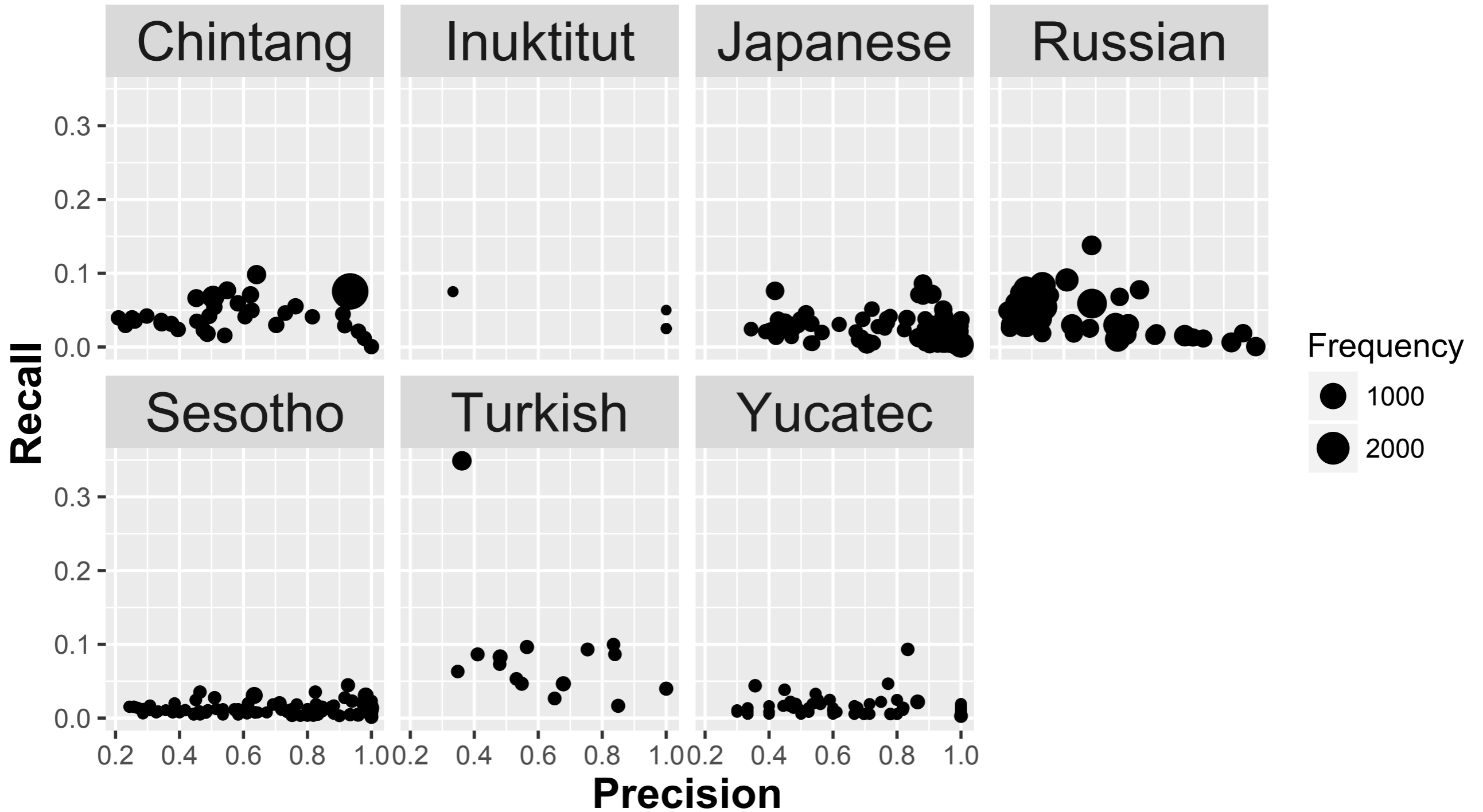
- word level**
- morphemes**

Step 2: categorization

operationalization



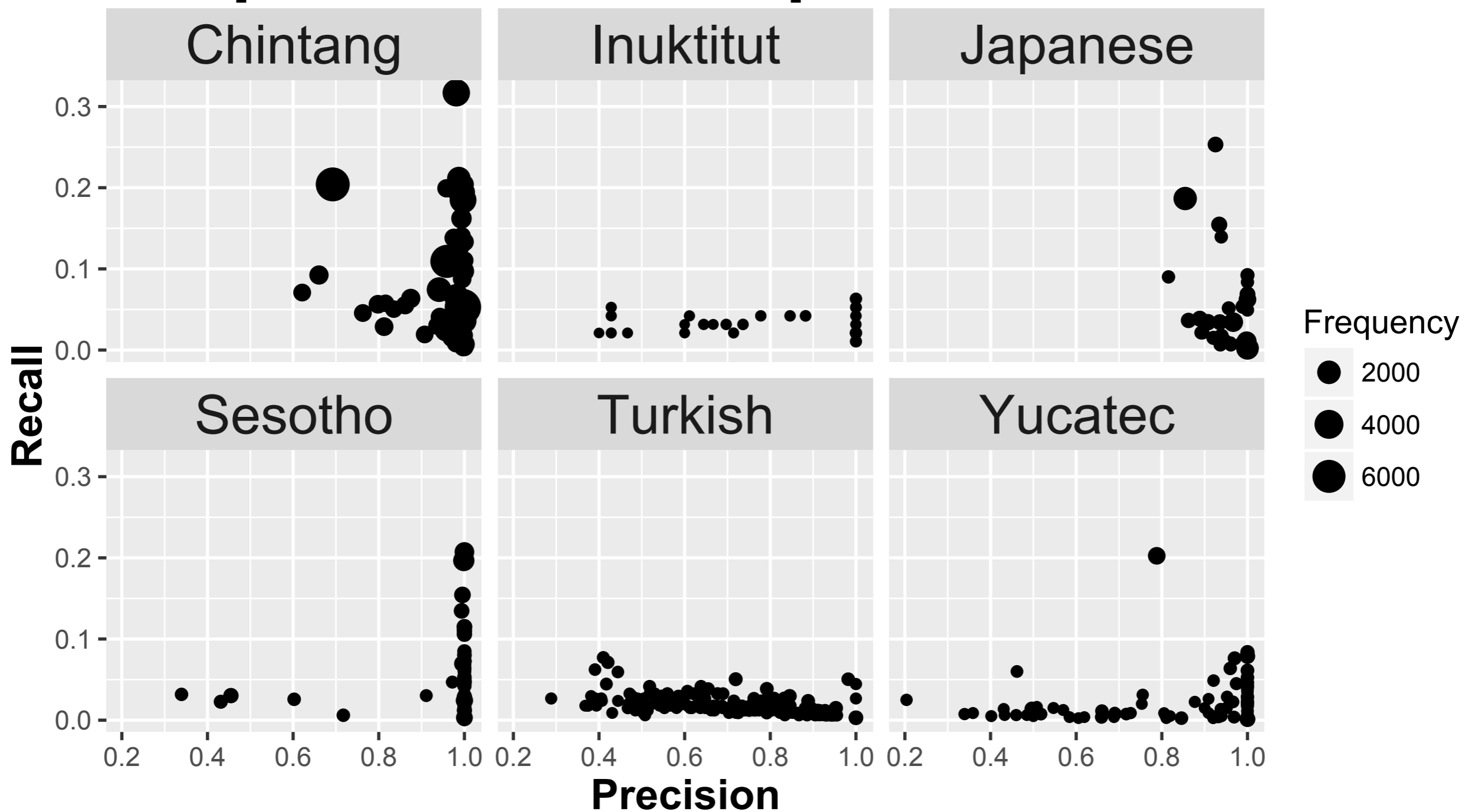
Word frames operationalized



word frames operationalized

	AccurateFrames	MeanPrecision	MeanRecall	SD	Min	Max	Median
Chintang	33	0.5744713	0.041844852	0.2351717	90	2720	118.0
Inuktitut	39	0.9829060	0.026923077	0.1067521	2	3	2.0
Japanese_MiiPro	97	0.8150211	0.020963388	0.2067237	67	915	106.0
Russian	50	0.4314698	0.042478469	0.2136746	234	1485	310.0
Sesotho	176	0.8301803	0.009559746	0.2301544	8	163	12.0
Turkish	14	0.6408977	0.065021357	0.1923271	34	84	45.5
Yucatec	208	0.7994745	0.008311380	0.2700607	3	71	4.0

Morpheme frames operationalized



morphemes frames operationalized

	AccurateFrames	MeanPrecision	MeanRecall	SD	Min	Max	Median
Chintang	60	0.9506371	0.08399360	0.08872225	517	7940	779.0
Inuktitut	100	0.9330618	0.02189474	0.15647706	5	43	6.5
Japanese_MiiPro	108	0.9841604	0.01698856	0.03606476	83	1943	150.0
Sesotho	88	0.9705718	0.03673841	0.11880977	66	1358	109.5
Turkish	279	0.7911451	0.01385831	0.19103508	21	186	32.0
Yucatec	140	0.8945911	0.01514286	0.18634764	20	584	34.0

Step 1: frames

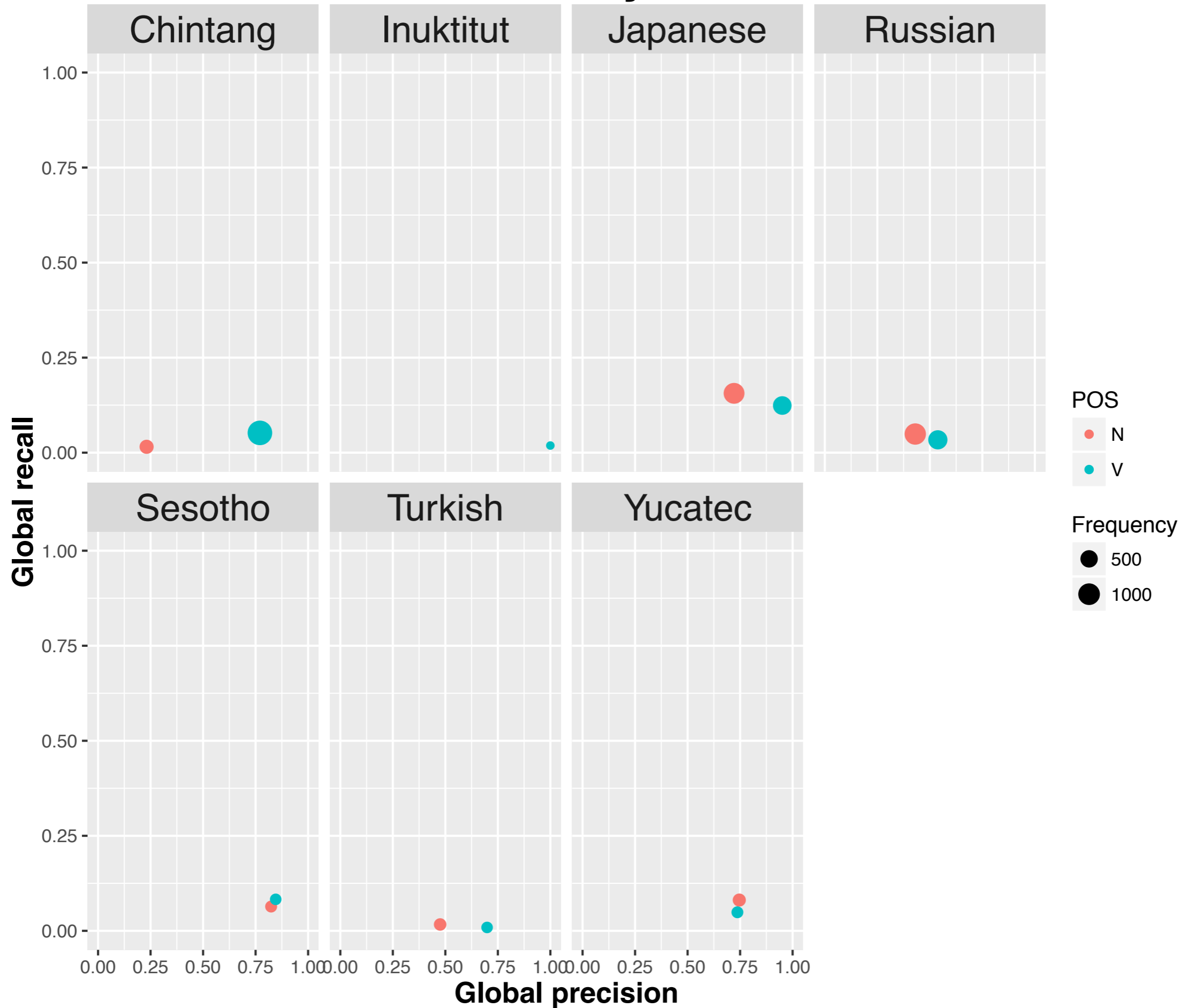
- word level
- morphemes

Step 2: categorization

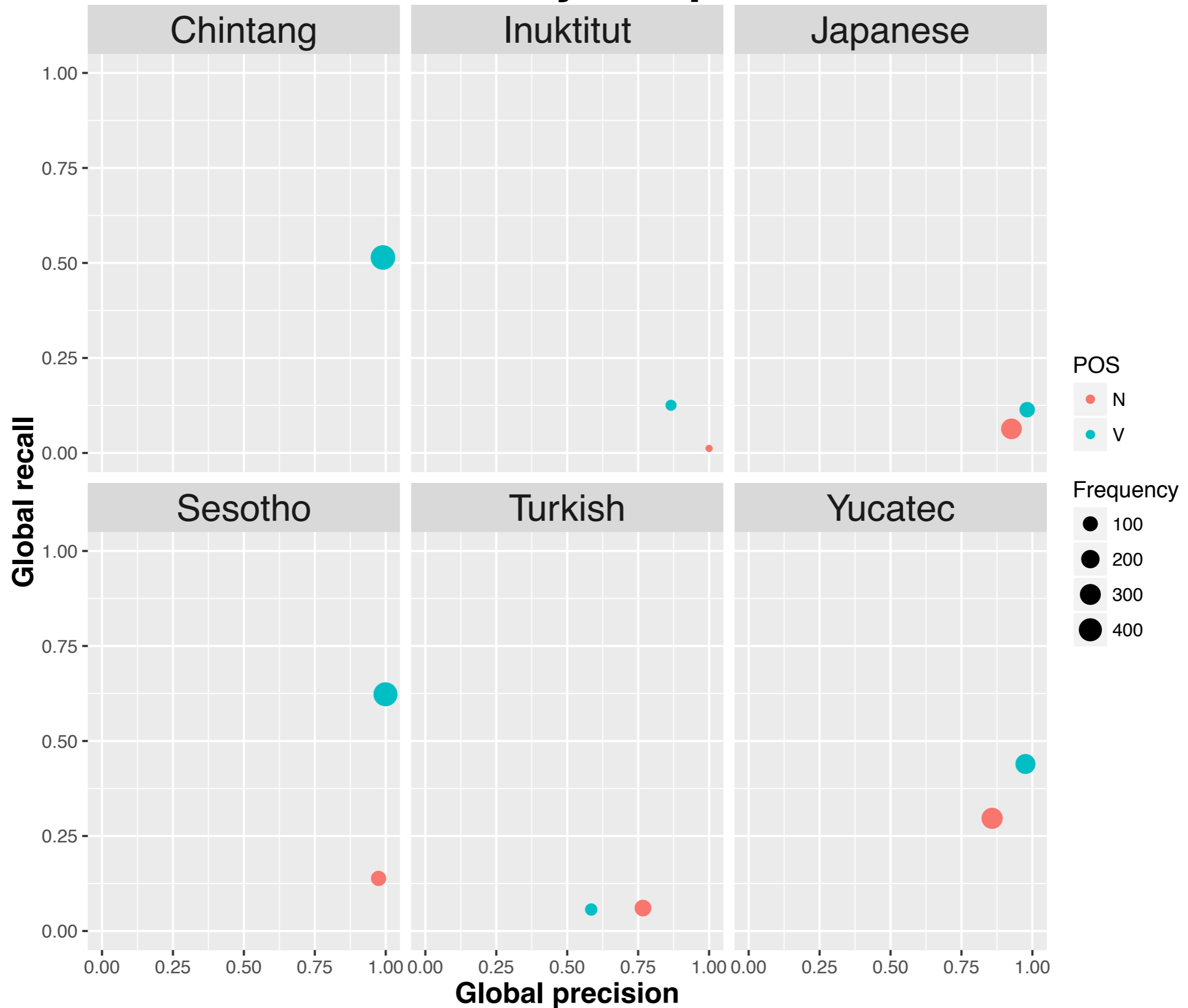
pos categorization

1. take modal category of pos of a frame and aggregate all frames with the same modal category (same as Weisleder and Waxman 2010).
2. test **global precision** of these aggregated frames.
3. new measure of recall: **global recall**
 - take all the aggregated frames of a specific category, e.g. all verb frames and combine their elements and test recall against the whole corpus

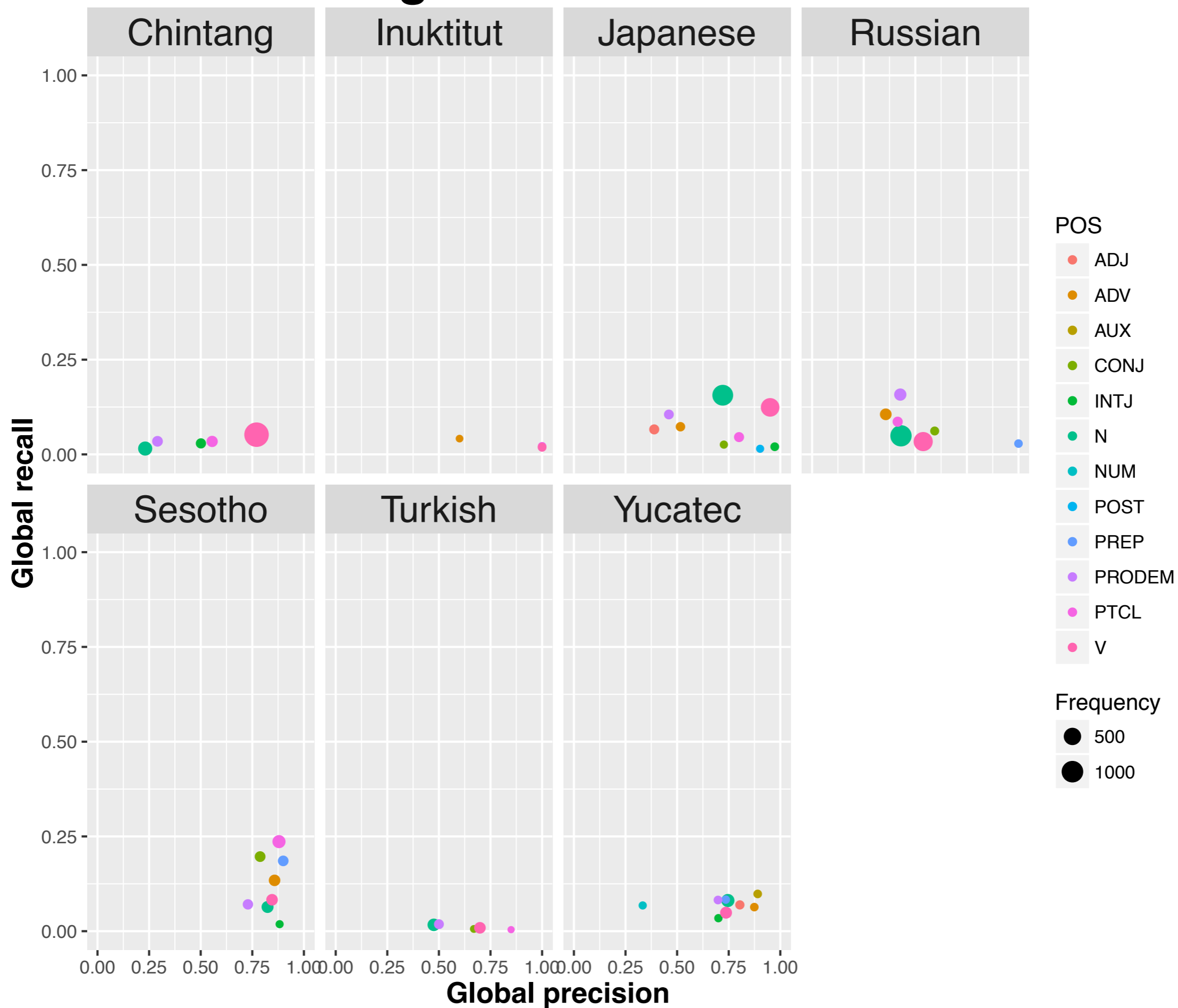
Nouns and verbs by word frames



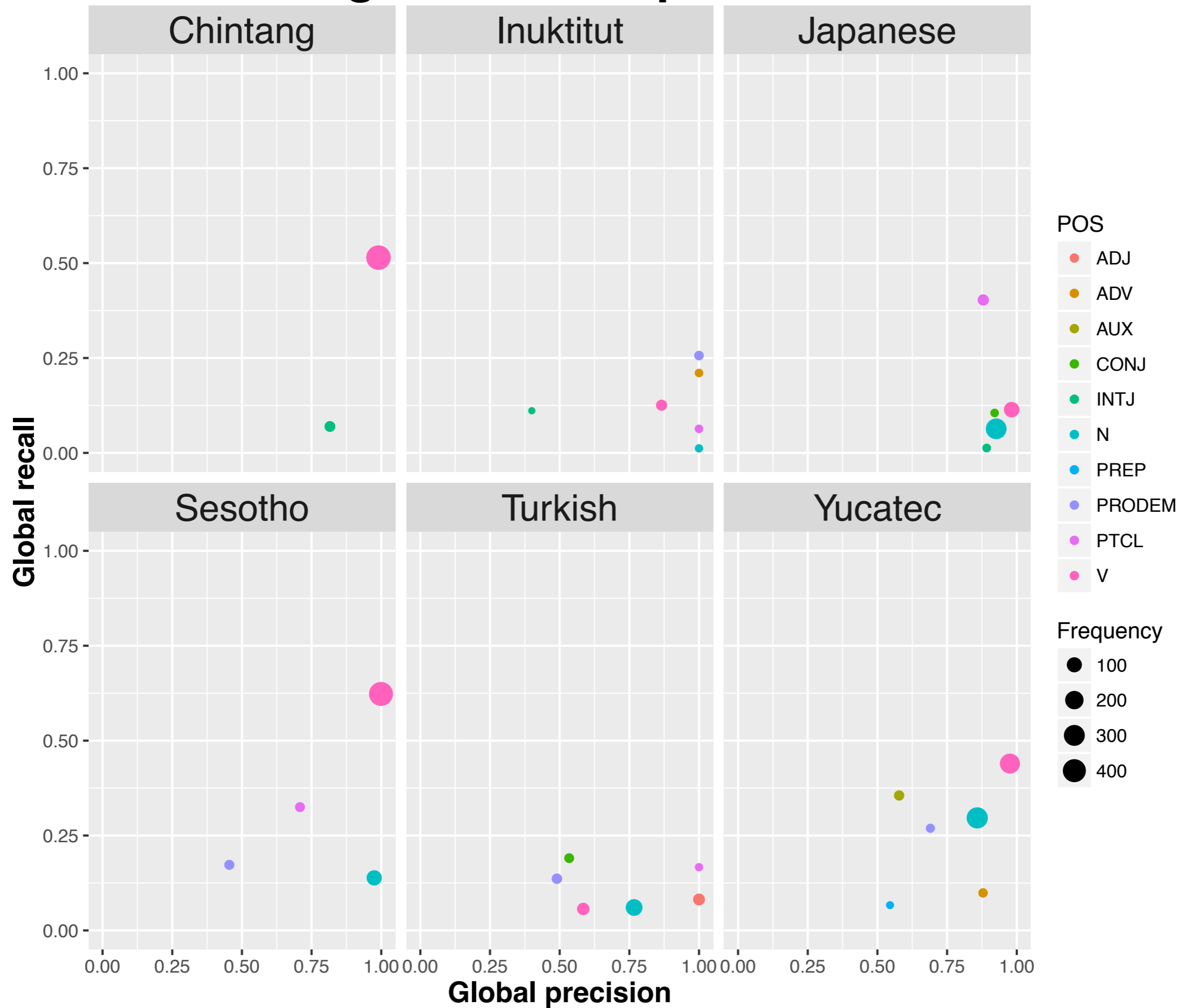
Nouns and verbs by morpheme frames



POS categories in word frames



POS categories in morpheme frames



to sum up:

- **what we did:**

- compared frames both on the morpheme and the word level across a set of maximally diverse languages
- introduced an operationalization of how to compare frames in corpora of different sizes
- used a quantitative evaluation of how well frames predict categories

to sum up:

- **what we found:**
 - **words level:** frames are not universally useful patterns for category detection
 - **morpheme level:** frames are very reliable throughout
 - categorization of nouns and verbs are very accurate for all our languages on the morpheme level and on the word level for those languages with high precision of word level frames
 - for all other parts of speech there is variation,
 - but languages with good word level frames seem to have higher precision throughout

conclusion

- frames are a potentially useful pattern to learn about regularities in a language but the levels are language specific
 - but: morpheme frames seem to make more sense from a universal point of view also with respect to processing
- next steps: test whether and how frames are used in acquisition
- extend this approach to other units and patterns in the input

Thanks to our collaborators!



Elena Lieven
U Manchester, LUCID,
Chintang



Shanley Allen
U Kaiserslautern
Inuktitut



Aylin Küntay
Koç U
Turkish



Katherine Demuth
Macquairie U
Sesotho



Gaby Hermon
U Delaware
Indonesian



Bárbara Pfeiler
UNAM, Mexico
Yucatec



Julie Brittain
Memorial U
Cree



Yasuhito Shirai
U Pittsburgh
Japanese

Thank you from the ACQDIV team!



Sabine Stoll,
language acquisition
PI



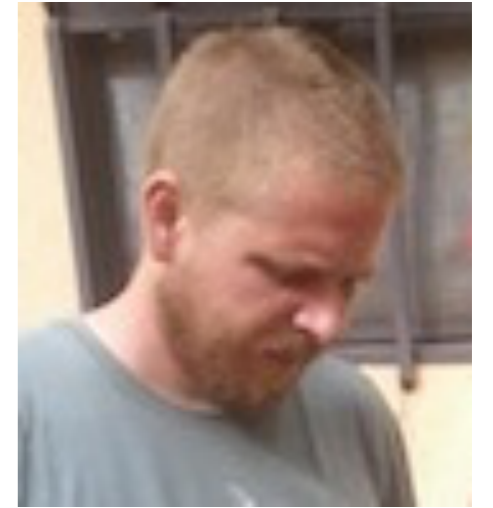
Dagmar Jung,
Dene corpus
development,
typology



Damián Blasi,
statistics,
models



Robert Schikowski,
Project management



Steven Moran,
database development,
computational
linguistics



Katia Mazara,
visualizations,
language acquisition



Géraldine
Walther,
morphology,
computational
linguistics



Andreas Gerster,
Dene, data
prepreparation



Melanie Trüssel,
Dene, data
prepreparation



Anna Jansco,
data base
development

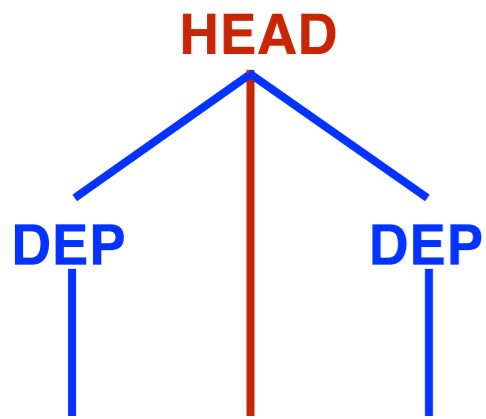


Casim Hysi,
database
development



University of
Zurich ^{UZH}





										independent stress
										onset requirement, prosodic subcat.
										voicing
<i>u</i>	<i>ca</i>	<i>ŋa</i>	<i>ta</i>	<i>hai?</i>		<i>ya</i>	<i>ʔã</i>	<i>na</i>	<i>kina</i>	
3sA	eat	1sO	FOC	move.away.TR	1sO	IND.NPST	INSIST	SEQ		
[:V]	[:]	[:V]	[:X]	[:V _{2σ}]		[:V]	[:V]	[:VP]	[:XP]	
										insertion and displacement potential
										cross-slot dependencies
										fixed ordering