Quantitative Analysis of DOBES Corpora Using R

Balthasar Bickel & Sabine Stoll

Chintang and Puma Documentation Project (CPDP)

What we need in corpus analysis (at least...)

- corpus searches of various kinds, e.g.
  - find all morphemes with shape ‘ŋa’ and gloss ‘ERG
  - find all clauses with both *ŋa*/ERG and *na*/2sO
  - find all two-word patterns repeated more than 100 times
  - find all forms of a lexeme that are used
- data aggregation across entire corpora, e.g.
  - measure verb form usage by age
  - or by age by speaker by genre, etc.
- feed found items and measurements directly into statistical computations and graphs
- link found items back to the original context and display it

Can we use ELAN for all of this this?

- Not for all of it: ELAN is for qualitative data exploration.
- And we probably shouldn’t expect it from ELAN: instead, follow the unix-style toolkit philosophy (like ELAN does itself), i.e. use good, specialized tools for each purpose
- One tool that meets all our corpus needs:

Why R ?

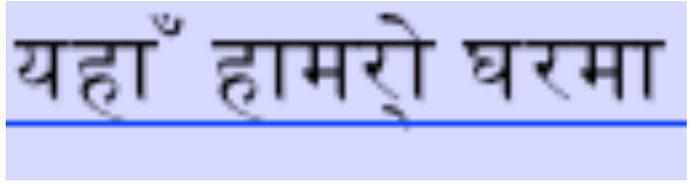
- open source
- runs both on unix/linux/macosx and windows
- over 1000 specialized packages
- top-quality graphics, including maps
- preferred environment for the latest developments in statistics
- increasingly used by linguists, especially corpus linguists and typologists (forthcoming textbooks by Keith Johnson, Harald Baayen and Stefan Gries; courses offered, for example in Leipzig)

Why R ?

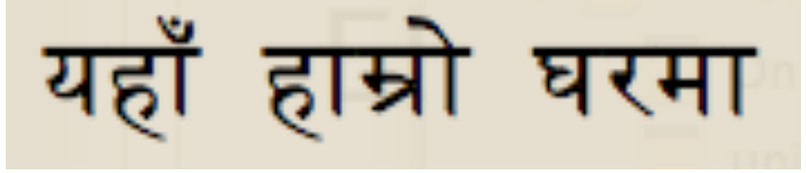
- 100% Unicode. On the Mac at least this includes perfect rendering of Devanagari — unlike all Java applications (including ELAN), and many (!) text processors, e.g.

  यहाँ हाम्रो घरमा *yahã̃ hā**mr**o gharmā* ‘here in our house’

ELAN and other JAVA apps:



R on Mac OSX



GUI

shell

What was missing and what we developed

- A tool that reads Toolbox files directly into R (www.uni-leipzig.de/~autotyp/tb.r; implemented by Taras Zakharko)

  (also accepts transformed CHAT files, extension to ELAN possible if someone does it...)

- But Toolbox has a very different data structure than R — or any other table/spreadsheet-oriented software

Data representation in Toolbox vs R

\ref them_talk.001
\EUDICOt0 00:07
\EUDICOp JK
\agegroup JK.adult
\age 0
\tx asinda akhimbeʔyaŋ maathapte
\gw asinda akhimbeʔ yaŋ maathapte
\mph asinda a- khim -peʔ yaŋ mai- a- thap -t -e
\mgl yesterday.adv 1sPOSS.gm- house.n -LOC.gm ADD.gm NEG.gm- 2.gm- come.level.vi -NEG.gm -PST.gm
\lg C C- C -C C C- C- C -C(S) -C
\eng You didn't come to my place yesterday.
\nep हिजो तिमी मेरो घरमा पनि आएनौ ।

mph	mgl
asinda	yesterday.adv
a-	1sPOSS.gm-
khim	house.n
-peʔ	-LOC.gm
yaŋ	ADD.gm
mai-	NEG.gm-
a-	2.gm-
thap	come.level.vi
-t	-NEG.gm
-e	-PST.gm

Data representation in R: adding word forms

mph	mgl	gw
asinda	yesterday.adv	asinda
a-	1sPOSS.gm-	akhimbeʔ
khim	house.n	akhimbeʔ
-peʔ	-LOC.gm	akhimbeʔ
yaŋ	ADD.gm	yaŋ
mai-	NEG.gm-	maathapte
a-	2.gm-	maathapte
thap	come.level.vi	maathapte
-t	-NEG.gm	maathapte
-e	-PST.gm	maathapte

→ Straightforward representation of issues like:

- which forms belong to which nominal vs. verbal stem (and how many there are)
- how long each form is, etc.

Data representation in R: adding clauses

mph	mgl	gw	tx
asinda	yesterday.adv	asinda	asinda akhimbeʔyaŋ maathapte
a-	1sPOSS.gm-	akhimbeʔ	asinda akhimbeʔyaŋ maathapte
khim	house.n	akhimbeʔ	asinda akhimbeʔyaŋ maathapte
-peʔ	-LOC.gm	akhimbeʔ	asinda akhimbeʔyaŋ maathapte
yaŋ	ADD.gm	yaŋ	asinda akhimbeʔyaŋ maathapte
mai-	NEG.gm-	maathapte	asinda akhimbeʔyaŋ maathapte
a-	2.gm-	maathapte	asinda akhimbeʔyaŋ maathapte
thap	come.level.vi	maathapte	asinda akhimbeʔyaŋ maathapte
-t	-NEG.gm	maathapte	asinda akhimbeʔyaŋ maathapte
-e	-PST.gm	maathapte	asinda akhimbeʔyaŋ maathapte

R Data Editor

row.names	ref	mph	mgl	lg	m.structure	gw	EUDICOp	agegroup	age	tx	
20514	them_talk.001	-e	-PST.gm	-C	suffix	maathapte	JK	JK.adult	0	asinda	akhimbeʔya
20511	them_talk.001	a-	2.gm-	C-	prefix	maathapte	JK	JK.adult	0	asinda	akhimbeʔya
20506	them_talk.001	a-	1sPOSS.gm-	C-	prefix	akhimbeʔ	JK	JK.adult	0	asinda	akhimbeʔya
20508	them_talk.001	-peʔ	-LOC.gm	-C	suffix	akhimbeʔ	JK	JK.adult	0	asinda	akhimbeʔya
20510	them_talk.001	mai-	NEG.gm-	C-	prefix	maathapte	JK	JK.adult	0	asinda	akhimbeʔya
20509	them_talk.001	yaŋ	ADD.gm	C	stem	yaŋ	JK	JK.adult	0	asinda	akhimbeʔya
20512	them_talk.001	thap	come.level.vi|vt	C	stem	maathapte	JK	JK.adult	0	asinda	akhimbeʔya
20507	them_talk.001	khim	house.n	C	stem	akhimbeʔ	JK	JK.adult	0	asinda	akhimbeʔya
20505	them_talk.001	asinda	yesterday.adv	C	stem	asinda	JK	JK.adult	0	asinda	akhimbeʔya
20513	them_talk.001	-t	-NEG.gm	-C(S)	suffix	maathapte	JK	JK.adult	0	asinda	akhimbeʔya

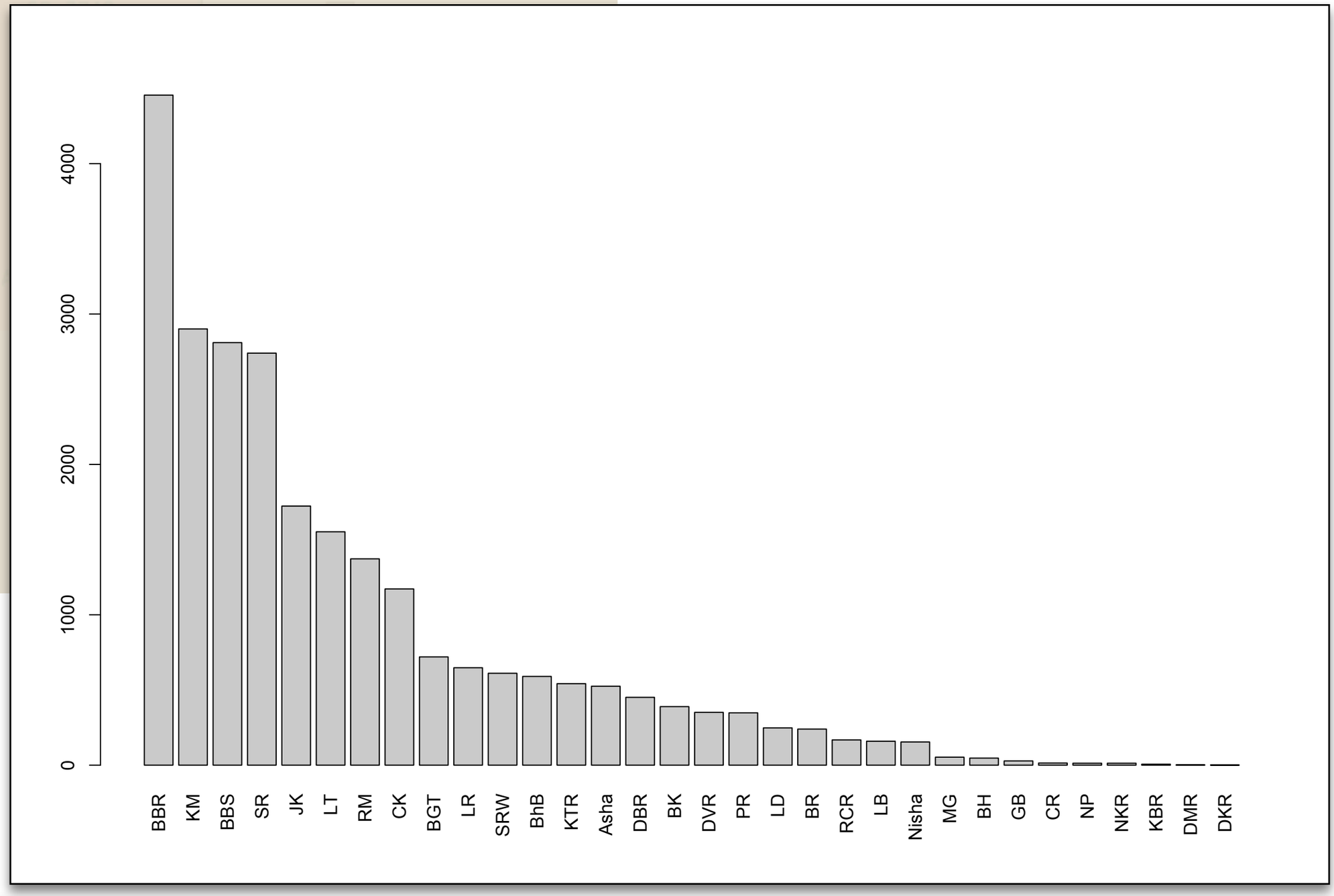
Data representation in R: adding other information

mph	mgl	lg	EUDICOp	tx
asinda	yesterday.adv	C	JK	asinda akhimbeʔyaŋ maathapte
a-	1sPOSS.gm-	C-	JK	asinda akhimbeʔyaŋ maathapte
khim	house.n	C	JK	asinda akhimbeʔyaŋ maathapte
-peʔ	-LOC.gm	-C	JK	asinda akhimbeʔyaŋ maathapte
yaŋ	ADD.gm	C	JK	asinda akhimbeʔyaŋ maathapte
mai-	NEG.gm-	C-	JK	asinda akhimbeʔyaŋ maathapte
a-	2.gm-	C-	JK	asinda akhimbeʔyaŋ maathapte
thap	come.level.vi	C	JK	asinda akhimbeʔyaŋ maathapte
-t	-NEG.gm	-C(S)	JK	asinda akhimbeʔyaŋ maathapte
-e	-PST.gm	-C	JK	asinda akhimbeʔyaŋ maathapte

Benefits of this representation

Easy data aggregation

```
> wdspeaker=aggregate(ctn$gw, list(speaker=ctn$EUDICOp), length)
> wdspeaker[order(wdspeaker$x, decreasing=T),]
   speaker    x
2      BBR 4456
18      KM 2901
17     BBS 2810
14      SR 2740
9       JK 1723
31      LT 1552
8       RM 1372
5       CK 1172
23     BGT  720
26      LR  648
30     SRW  611
1      BhB  590
24     KTR  542
11    Asha  525
27     DBR  451
6       BK  389
4      DVR  351
20      PR  348
25      LD  248
19      BR  240
29     RCR  168
```



Benefits of this representation

- easy corpus searches, e.g. again ‘aggregate’ by words (example to follow)
- convenience function for many kinds of queries:
    - morphological searches, e.g. records with shape=’ŋa’, gloss = ‘ERG’ and language=’Chintang’

      search.tb(what=list(‘ŋa’, ‘ERG’, ‘C’), tiers=c(‘mph’, ‘mgl’, ‘lg’), corpus=ctn)
    - syntactic searches, e.g. clauses with ŋa/ERG and -u/3P

      search.tb(what=list(c(‘ŋa’, ‘ERG’), c(‘u’,’3P’)), tiers=c(‘mph’, ‘mgl’), corpus=ctn)
    - all search strings are regular expressions
- convenience function for extracting fully-formatted examples from corpus

Flexible Toolbox-to-R

- flexible specification of how tiers should be treated:
  - as ‘id’: delimits record IDs
  - as ‘single’: all content in one row
  - as ‘word’: each word in one row
  - as ‘morpheme’: each morpheme in one row
- current version of our ‘toolbox reader’ copes with most special cases/exceptions (missing tiers; wrong linebreaks etc.)
- we found that this was possible without using a schema.

Flexible Toolbox-to-R

- when combined,
  - ‘word’ tiers replicate over rows containing their ‘morphemes’
  - ‘single’ tiers replicate over rows containing their ‘words’

mph=’morpheme’	mgl=’morpheme’	gw=’word’	tx=’single’
asinda	yesterday.adv	asinda	asinda akhimbeʔyaŋ maathapte
a-	1sPOSS.gm-	akhimbeʔ	asinda akhimbeʔyaŋ maathapte
khim	house.n	akhimbeʔ	asinda akhimbeʔyaŋ maathapte
-peʔ	-LOC.gm	akhimbeʔ	asinda akhimbeʔyaŋ maathapte
yaŋ	ADD.gm	yaŋ	asinda akhimbeʔyaŋ maathapte
mai-	NEG.gm-	maathapte	asinda akhimbeʔyaŋ maathapte
a-	2.gm-	maathapte	asinda akhimbeʔyaŋ maathapte
thap	come.level.vi	maathapte	asinda akhimbeʔyaŋ maathapte
-t	-NEG.gm	maathapte	asinda akhimbeʔyaŋ maathapte
-e	-PST.gm	maathapte	asinda akhimbeʔyaŋ maathapte

Flexible Toolbox-to-R

- Alternative: read morphemes as ‘words’ for morphological research:

gw=’word’	mph=’word’	mgl=’word’	lg=’word’	EUDICOp=’single’
asinda	asinda	yesterday.adv	C	JK
akhimbeʔ	a-khim-peʔ	1sPOSS.gm-house.n-LOC.gm	C-C-C	JK
yaŋ	yaŋ	ADD.gm	C	JK
maathapte	mai-a-thap-t-e	NEG.gm-2.gm-come.level.vi-NEG.gm-PST.gm	C-C-C-C(S)-C	JK

- Or, morphemes as ‘single’, for syntactic research:

mph=’single’	mgl=’single’
asinda a-khim-peʔ yaŋ mai-a-thap-t-e	yesterday.adv 1sPOSS.gm-house.n-LOC.gm ADD.gm NEG.gm-2.gm-come.level.vi-NEG.gm-PST.gm

Case Study 1: N:V-Ratio in ritual vs everyday language

- add session and genre information to the data: can be done easily by simply adding a new ‘column’

mph	mgl	EUDICOp	session	genre
haŋma	deity.n	BhB	Burhahang_01	ritual
he	ADDR.interj	BhB	Burhahang_01	ritual
misreko	FILLER	BhB	Burhahang_01	ritual
parameśvara	Lord.n	BhB	Burhahang_01	ritual
thakuraa-	senior.n2.gm-	BhB	Burhahang_01	ritual
-na	-NA.gm	CK	Gen_talk	everyday
u-	3sPOSS.gm-	CK	Gen_talk	everyday
yaŋ	ADD.gm	BK	Gen_talk	everyday
bhuŋs	pile.something.vt	CK	Gen_talk	everyday
wei?	rain.n	BK	Gen_talk	everyday
bela	time.n	CK	Gen_talk	everyday

- define a function nv.ratio() by $\frac{N('\backslash.n\$')}{N('\backslash.n\$') + N('\backslash.v.\$')}$

Case Study 1: N:V-Ratio in ritual vs everyday language

- aggregate(list(nvr=ctn$mph),

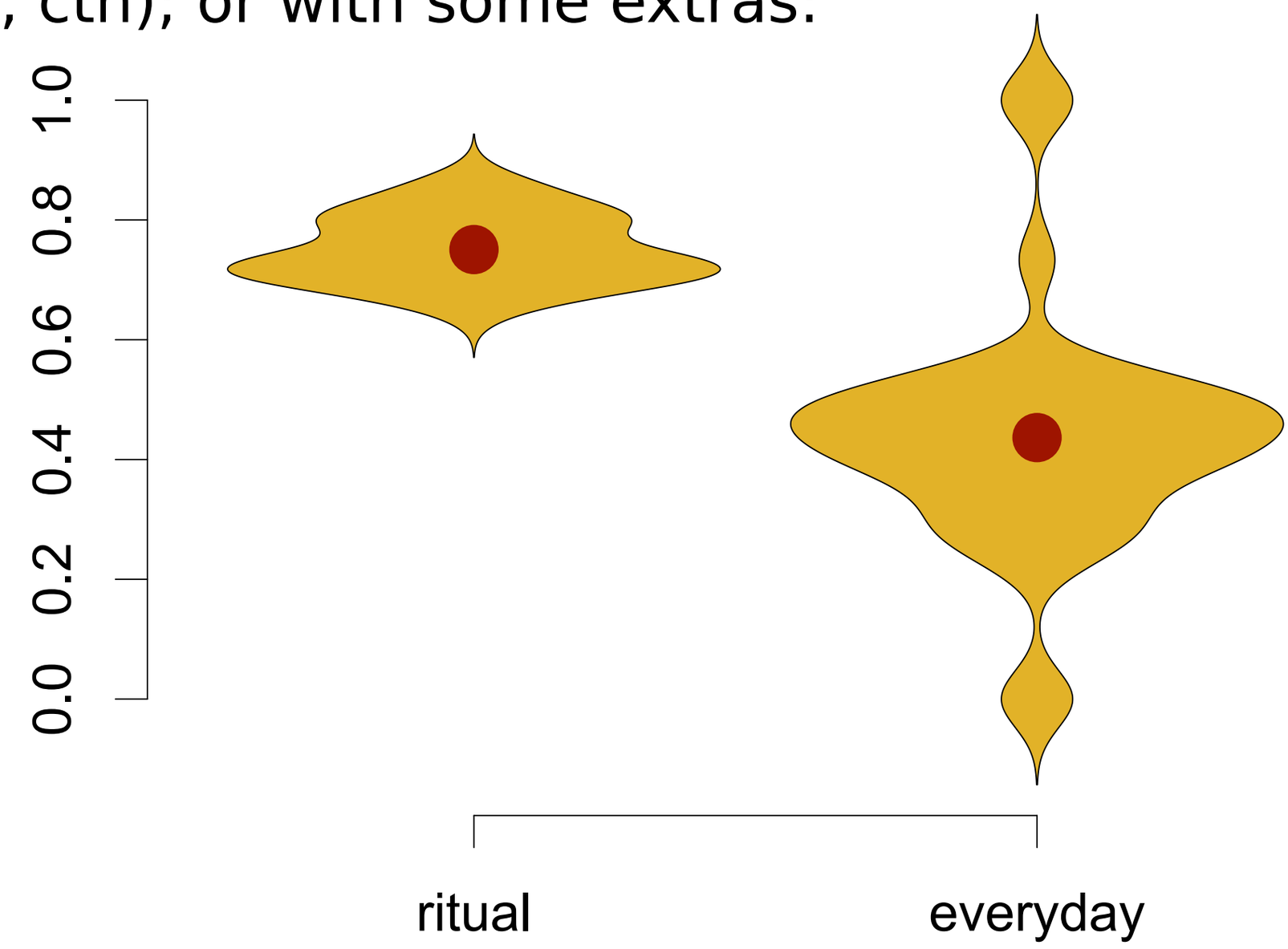
  list(genre=ctn$genre, session=ctn$session, speaker=ctn $EUDICOp),

  nv.ratio)

ritual	sudhar_hang	BBR	0.6566265
ritual	sudhar_khip	BBR	0.8571429
ritual	sudhar_pakuwa	BBR	0.6934673
ritual	sudhar_palawa	BBR	0.6818182
ritual	wal_yupung02	BBS	0.7281167
everyday	Ctn_talk02	GB	0.5
everyday	Durga_Exp	DVR	0.4459459
everyday	Durga_job	DVR	1
everyday	Gen_talk	CK	0.4536082
everyday	Gen_talk	BK	0.3157895
everyday	Intro_woman	CK	1
everyday	Orig_ctn_devi	NP	0
everyday	RM_JK_talk01	RM	0.3962264
everyday	RM_JK_talk01	JK	0.4719101

Case Study 1: N:V-Ratio in ritual vs everyday language

- plot(nvr ~ genre, ctn); or with some extras:



- Mixed Linear Effects Model, with speaker as a random and genre as a fixed effect:

  β(genre) = -.32, p < .01

Case Study 2: N:V-Ratio in language acquisition

- Gentner 1982: children start with a high N:V ratio, universally; ascribed to conceptual factors
- Tardif 1996: Chinese children start with a much lower N:V ratio; ascribed to nature of input (cf. Brown on Tzeltal)
- Chintang and Chinese have a similar adult discourse structure: low referential density, i.e. low N:V ratio (Bickel 2006)
- But Chintang verb morphology is **much** more complex than Chinese verb morphology

N:V-Ratio in language acquisition

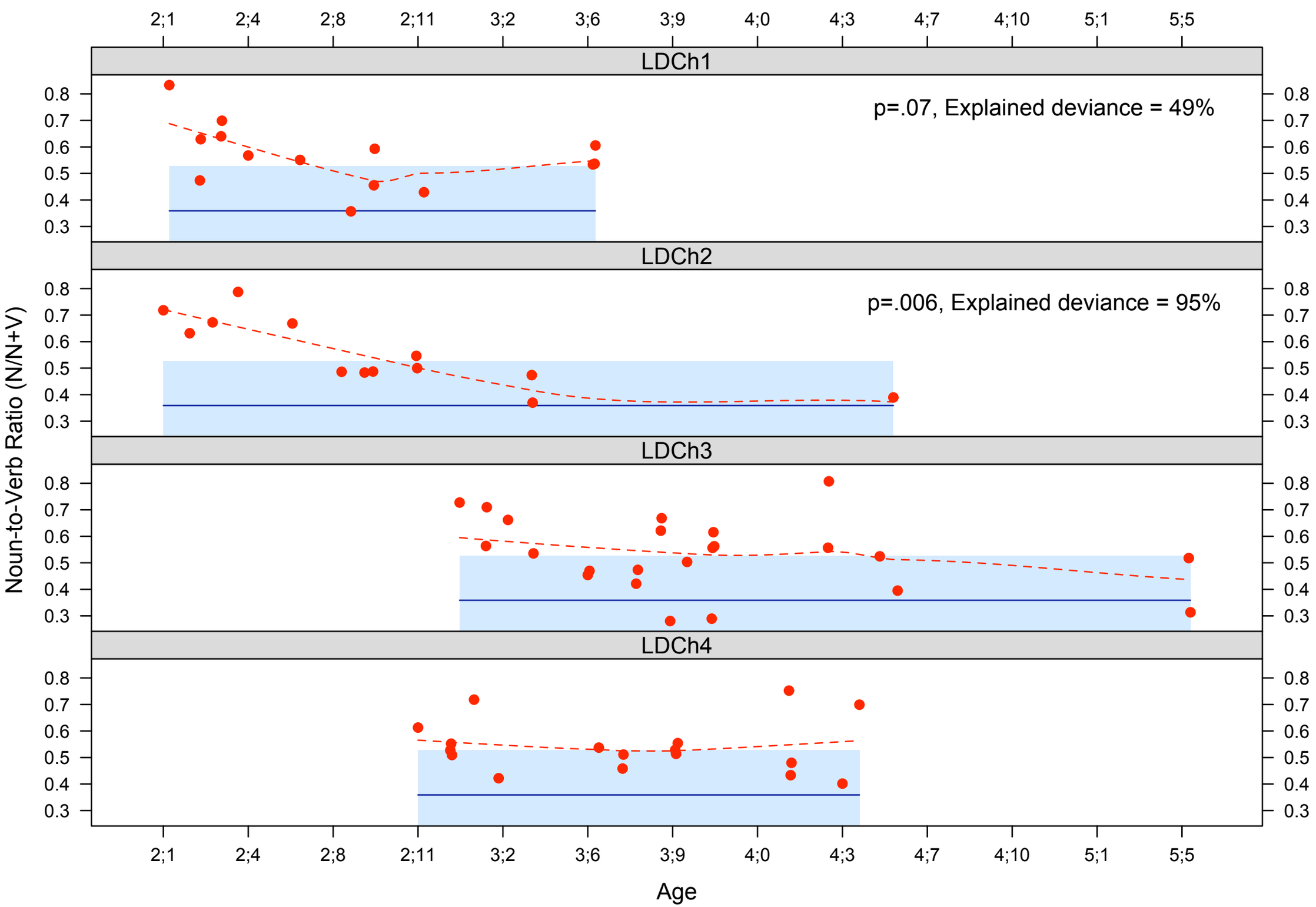
- Question: how does the Chintang N:V-ratio develop?
- What factors determine the development?
  - If morphology, predict difference from Chinese
  - If input, predict similarity to Chinese

N:V-Ratio in language acquisition

- like before:

  aggregate(ctncl$mph, list(*age=ctncl$age, speaker=ctncl$EUDICOp*), nv.ratio)

Development of the N:V ratio, word form tokens



N:V-Ratio in language acquisition

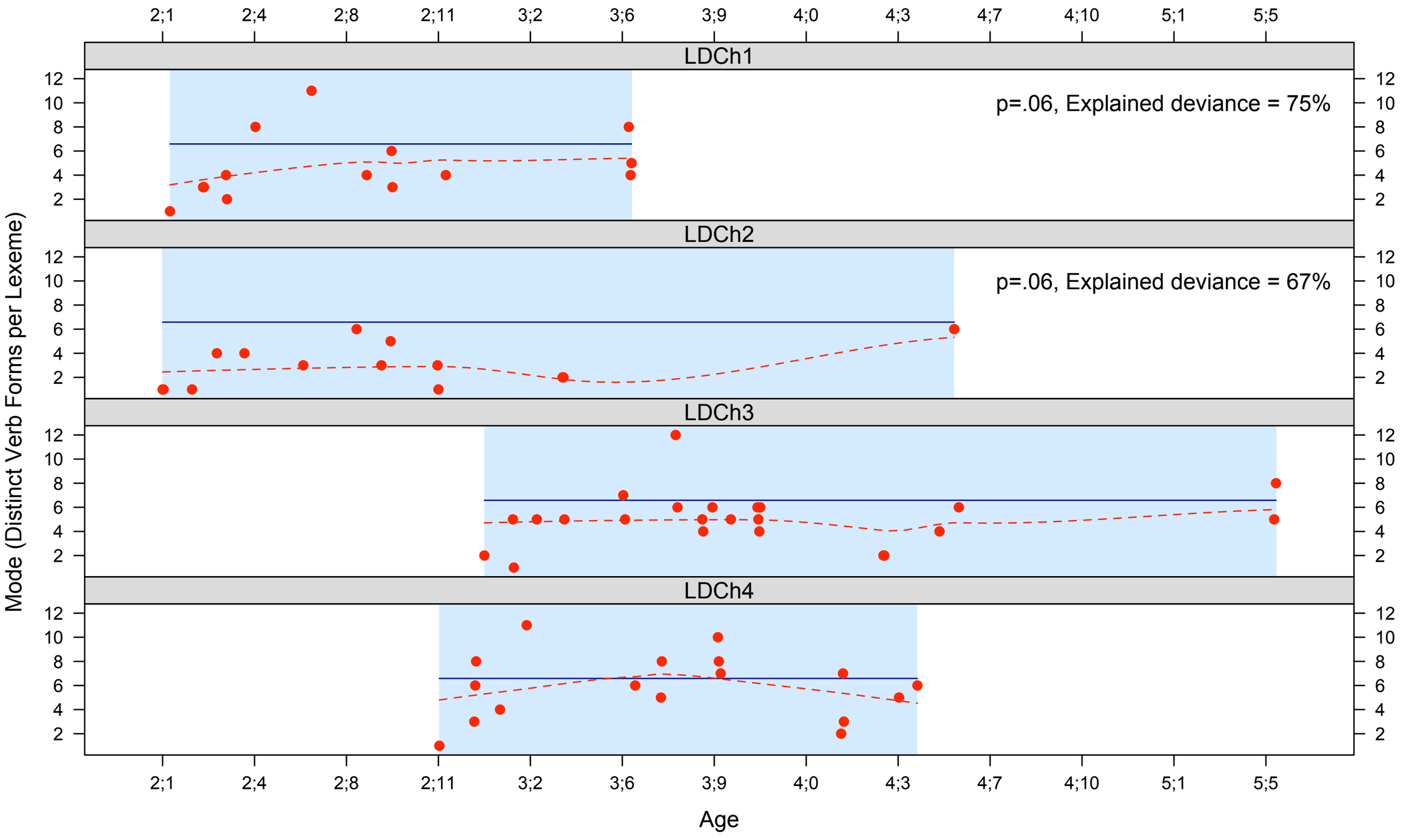
- Perhaps it’s the morphology!
- determine number of forms associated *per lexeme, per age, per speaker:*

  forms = aggregate(ctncl$gw, list(*lexeme=ctncl$lex, age=ctncl$age, speaker=ctncl$EUDICOp*), function(x) length(unique(x)))

- then, determine the modal number of this:

  aggregate(forms$x, list(age=forms$age, speaker=forms$speaker), max)

Development of verb morphology, form types per lexeme



Final remarks

- Toolbox-to-R function freely available
  - Feedback welcome
  - Extensions (e.g. ELAN-to-R function) welcome
- Advantages:
  - full integration of corpus searching and aggregation tasks with R’s statistics and plotting functionality
  - one single environment for all quantitative corpus analyses
- Disadvantages:
  - R has a relatively steep learning curve (but now there are good introductions!)
  - slow when performing complex searches on large corpora.