University of Zurich[UZH]

Complexity counts: the unusual case of Indo-Aryan numeral systems

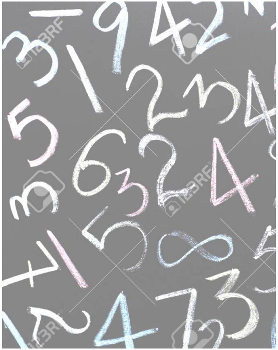
Chundra Cathcart

Department of Comparative Language Science, Univ. Zürich

Habilitation Lecture, University of Zurich
15 December 2023

Numeral systems and numerical cognition

- Human languages share the property of referring to different quantities
- Numeral systems (i.e., elements referring to quantities) are related to this species-wide cognitive capacity (Xu and Regier 2014, Núñez 2017)
- Major question: what can be expressed (e.g., exact versus approximate systems)?
- Related question: how are quantities realized formally?



Numeral terms: word structure

- Numeral systems of the world's languages tend to consist of words which denote information regarding the tens and digits place in a transparent fashion
- E.g., Sanskrit numerals 1–49 (rows represent the tens place; columns represent the digits place)

	0	1	2	3	4
0	—	eka	dva	tri	catur
10	daśa	ekādaśa	dvādaśa	trayodaśa	caturdaśa
20	viṃśati	ekaviṃśati	dvaviṃśati	trayoviṃśati	caturviṃśati
30	triṃśat	ekatriṃśat	dvātriṃśat	trayastriṃśat	catustriṃśat
40	catvāriṃśat	ekacatvāriṃśat	dvacatvāriṃśat	trayaścatvāriṃśat	catuscatvāriṃśat

	5	6	7	8	9
0	pañca	ṣaṣ	sapta	aṣṭa	nava
10	pañcadaśa	ṣoḍaśa	saptadaśa	aṣṭādaśa	navadaśa
20	pañcaviṃśati	ṣaḍviṃśati	saptaviṃśati	aṣṭaviṃśati	navaviṃśati
30	pañcatriṃśat	ṣaṭtriṃśat	saptatriṃśat	aṣṭatriṃśat	navatriṃśat
40	pañcacatvāriṃśat	ṣaṭcatvāriṃśat	saptacatvāriṃśat	aṣṭacatvāriṃśat	navacatvāriṃśat

Numeral terms: word structure

- Numeral systems of the world's languages tend to consist of words which denote information regarding the tens and digits place in a transparent fashion
- E.g., Sanskrit numerals 1–49 (rows represent the tens place; columns represent the digits place)

	0	1	2	3	4
0	—	eka	dva	tri	catur
10	daśa	ekādaśa	dvādaśa	trayodaśa	caturdaśa
20	viṃśati	ekaviṃśati	dvaviṃśati	trayoviṃśati	caturviṃśati
30	triṃśat	ekatriṃśat	dvātriṃśat	trayastriṃśat	catustriṃśat
40	catvāriṃśat	ekacatvāriṃśat	dvacatvāriṃśat	trayaścatvāriṃśat	catuscatvāriṃśat

	5	6	7	8	9
0	pañca	ṣaṣ	sapta	aṣṭa	nava
10	pañcadaśa	ṣoḍaśa	saptadaśa	aṣṭādaśa	navadaśa
20	pañcaviṃśati	ṣaḍviṃśati	saptaviṃśati	aṣṭaviṃśati	navaviṃśati
30	pañcatriṃśat	ṣaṭtriṃśat	saptatriṃśat	aṣṭatriṃśat	navatriṃśat
40	pañcacatvāriṃśat	ṣaṭcatvāriṃśat	saptacatvāriṃśat	aṣṭacatvāriṃśat	navacatvāriṃśat

Ca. 2500 years later...

Hindi/Urdu numerals 1–99

	0	1	2	3	4	5	6	7	8	9
0	—	ek	do	tin	cɑr	pɑ̃c	c^{h}ɛ	sɑt	ɑʈ^{h}	nɔ
10	dəs	gjɑrə	bɑrə	terə	cɔdə	pəndrə	solə	sətrə	əʈ^{h}ɑrə	ʊnnis
20	bis	ɪkkis	bɑis	teis	cɔbis	pəccis	c^{h}əbbis	səttɑis	əʈʈɑis	ʊntis
30	tis	ɪkəttis	bəttis	tæ̃tis	cɔ̃tis	pæ̃tis	c^{h}əttis	sæ̃tis	əɽtis	ʊntɑlis
40	cɑlis	ɪktɑlis	bəjɑlis	tæ̃tɑlis	cəʋɑlis	pæ̃tɑlis	c^{h}ɪjɑlis	sæ̃tɑlis	əɽtɑlis	ʊncɑs
50	pəcɑs	ɪkjɑʋən	bɑʋən	tɪrpən	cəuʋən	pəcpən	c^{h}əppən	səttɑʋən	əʈʈ^{h}ɑʋən	ʊnsəʈ^{h}
60	sɑʈ^{h}	ɪksəʈ^{h}	bɑsəʈ^{h}	tɪrsəʈ^{h}	cɔ̃səʈ^{h}	pæ̃səʈ^{h}	c^{h}ɪjɑsəʈ^{h}	sərsəʈ^{h}	əɽsəʈ^{h}	ʊnhəttər
70	səttər	ɪkhəttər	bəhəttər	tɪhəttər	cɔhəttər	pəchəttər	c^{h}ɪhəttər	səthəttər	əʈ^{h}həttər	ʊnjɑsi
80	əssi	ɪkjɑsi	bəjɑsi	tɪrɑsi	cɔrɑsi	pəcɑsi	c^{h}ɪjɑsi	səttɑsi	əʈʈ^{h}ɑsi	nəʋɑsi
90	nəʋe	ɪkjɑnʋe	bɑnʋe	tɪrɑnʋe	cɔrɑnʋe	pəcɑnʋe	c^{h}ɪjɑnʋe	səttɑnʋe	əʈʈ^{h}ɑnʋe	nɪnjɑnʋe

- Hindi/Urdu representative of Indo-Aryan numeral systems
- High integrative complexity (Ackerman and Malouf 2013): hard to predict forms on the basis of each other
- Developed out of the relatively transparent Sanskrit numeral system

Indo-Aryan numeral systems

Very little literature on the topic:

Indo-Aryan numeral systems

Very little literature on the topic:

- Bright (1969): there is no economical set of rules that can generate Hindi numerals, but perhaps some implicit rules governing the system

Indo-Aryan numeral systems

Very little literature on the topic:

- Bright (1969): there is no economical set of rules that can generate Hindi numerals, but perhaps some implicit rules governing the system
- Berger (1992): historical complex developments stemming from sound change and other processes

Indo-Aryan numeral systems

Very little literature on the topic:

- Bright (1969): there is no economical set of rules that can generate Hindi numerals, but perhaps some implicit rules governing the system
- Berger (1992): historical complex developments stemming from sound change and other processes
- Cathcart (2017): computational models generally capable of classifying HU numerals on basis of phonological cues

Indo-Aryan numeral systems

Very little literature on the topic:

- Bright (1969): there is no economical set of rules that can generate Hindi numerals, but perhaps some implicit rules governing the system
- Berger (1992): historical complex developments stemming from sound change and other processes
- Cathcart (2017): computational models generally capable of classifying HU numerals on basis of phonological cues
- Schneider et al. (2020): children acquiring Hindi and Gujarati are less able to rely on successor functions in counting than children acquiring other languages

Objectives of this talk

The aims of this talk are threefold:

Objectives of this talk

The aims of this talk are threefold:

- I aim to demonstrate that Indo-Aryan numeral systems exhibit higher integrative complexity

Objectives of this talk

The aims of this talk are threefold:

- I aim to demonstrate that Indo-Aryan numeral systems exhibit higher integrative complexity
- I will show that Indo-Aryan numeral systems are subject to some of the same general communicative pressures as other systems

Objectives of this talk

The aims of this talk are threefold:

- I aim to demonstrate that Indo-Aryan numeral systems exhibit higher integrative complexity
- I will show that Indo-Aryan numeral systems are subject to some of the same general communicative pressures as other systems
- Zooming in on South Asia, I will attempt to highlight factors that are involved in the maintenance and loss of high-complexity numeral systems

Large data sets and quantitative metrics provide new insights on these issues

Data used

UniNum (Ritchie et al. 2019)

- Collection of numerals ranging between 0 and 100000000000 (inclusive), provided by Google and language experts.
- Curated for text-to-speech purposes
- 186 speech varieties
- Representations are orthographic, not phonemic

Data used

UniNum (Ritchie et al. 2019)

- Collection of numerals ranging between 0 and 100000000000 (inclusive), provided by Google and language experts.
- Curated for text-to-speech purposes
- 186 speech varieties
- Representations are orthographic, not phonemic

Eugene Chan's numeral data set (Chan et al. 2019)

- Collection of numerals 1–29, 30, 40, 50, 60, 70, 80, 90, 100, 200, 1000, 2000
- 5352 speech varieties
- Phonemic representations

Metrics used: Minimum description length (MDL)

- Information theoretic principle: seeks the shortest set of combinable elements needed to generate a code (Rissanen 1983)
- Can be inferred using several Bayesian algorithms (e.g., Goldwater et al. 2009)
- Simpler numeral systems have shorter descriptions

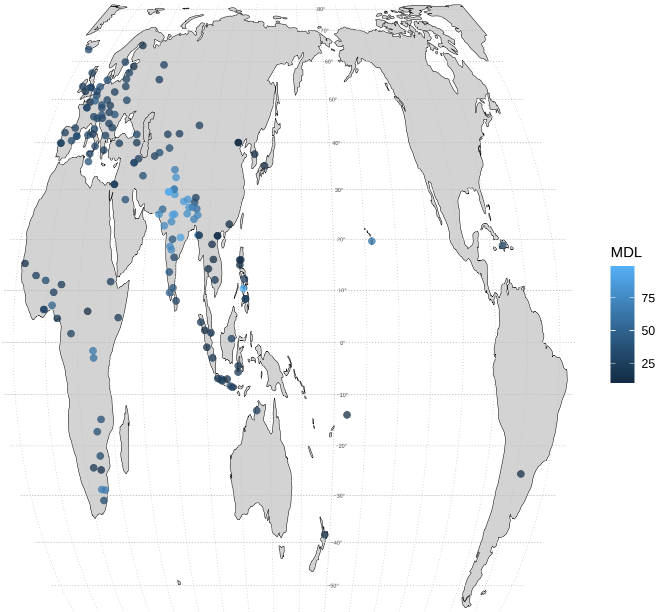
Metrics used: Phonotactic surprisal

- Represents the unpredictability of a phoneme or grapheme in context, i.e., given the two previous phonemes/graphemes (Piantadosi et al. 2012, Dautriche et al. 2017)
- Numeral systems containing more recurrent, predictable elements will exhibit lower surprisal

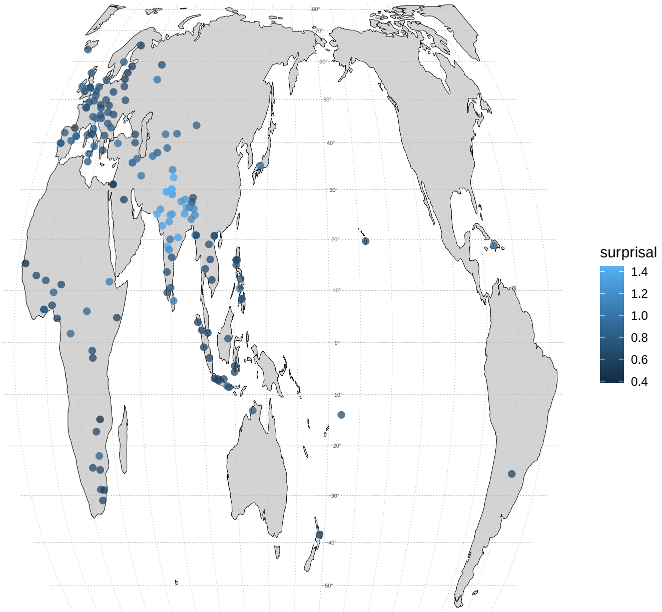
Metrics used: Linear discriminative learnability (LDL)

- Framework which learns mappings between semantic and phonological cues, e.g., trigram sequences of sounds (Baayen et al. 2018)
- Can be used to predict forms from semantic cues (e.g., TWENTY + ONE $\rightarrow$ *twenty-one*)

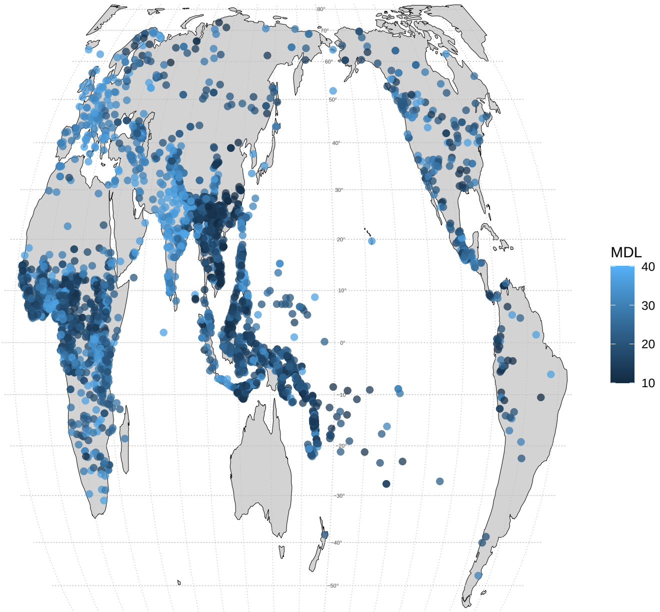
UniNum, minimum description length



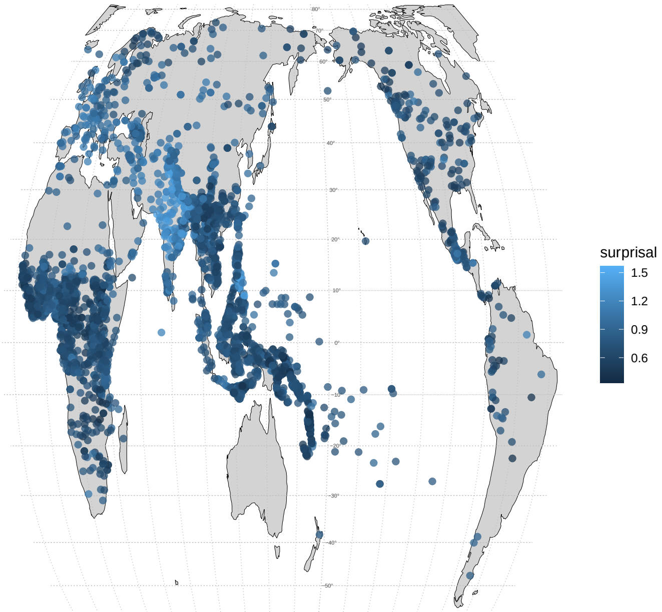
UniNum, phonotactic surprisal



Chan numerals, minimum description length



Chan numerals, phonotactic surprisal



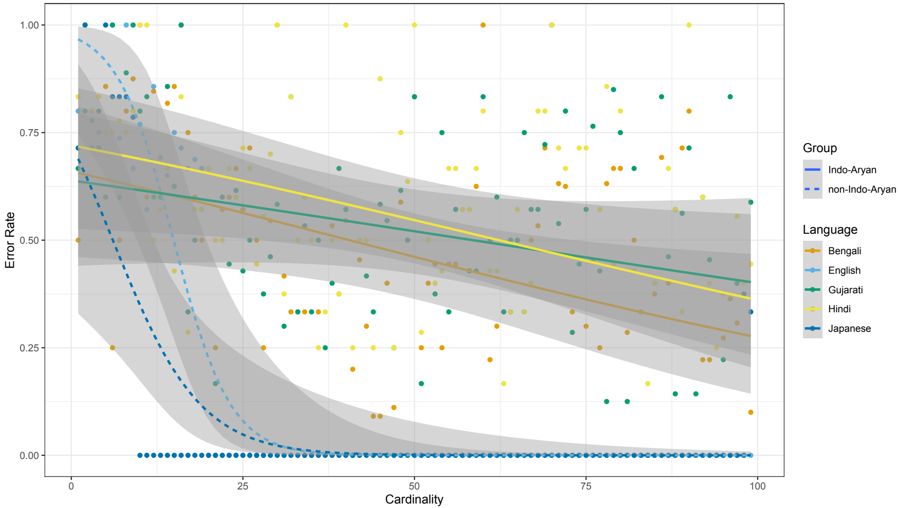
Interim summary

- Different metrics capture different properties of numeral systems (e.g., min. desc. length may be sensitive to mixed systems, etc., as well as integrative complexity)
- Small sample sizes in the Chan data set may cause issues in situating languages according to metrics
- At the same time, South Asia consistently emerges as a hotbed of complexity for all metrics and data sets considered

Transparency and frequency in numeral systems

- In morphological systems (e.g., noun, verb paradigms), less frequent forms tend to be more regular and transparent (Blevins et al. 2017), e.g., *go* $\sim$ *went* vs. *ambulate* $\sim$ *ambulated*
- Same principles may apply to numeral systems: e.g., English *twelve* is more frequent than *ninety-nine* (Brysbaert 2005)
- If this is a universal principle underlying numeral systems, then Indo-Aryan systems should exhibit higher predictability for items of higher cardinality (and lower frequency; Dehaene and Mehler 1992)
- Predictive accuracy modeled via linear discriminative learning (error rate)

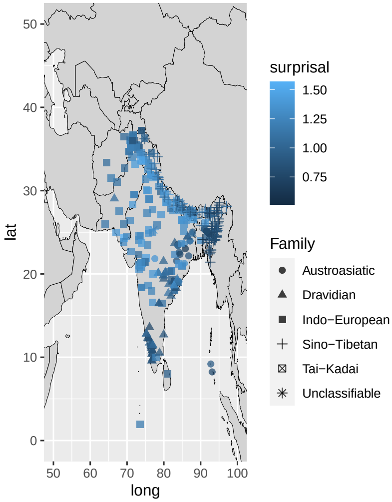
Cardinality vs. predictability



Predictive errors decrease as cardinality increases in all languages; trend is more gradual for Indo-Aryan languages

Numeral complexity in South Asia

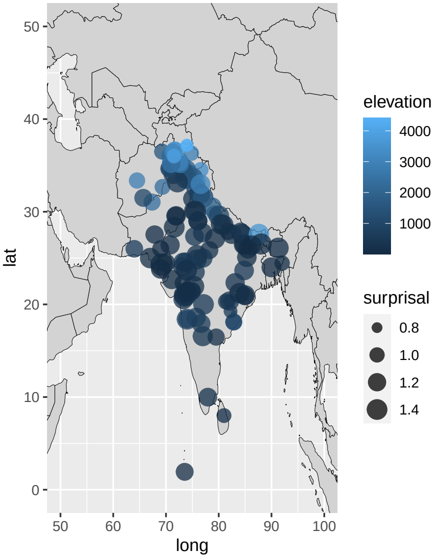
- S. Asia is a hotbed of enumerative complexity in numeral systems
- At the same time, there is variation in the degree of complexity exhibited by individual languages, within and across families
- Complexity is a largely Indo-Aryan phenomenon; complex systems in non-IA languages are often due to borrowing from IA

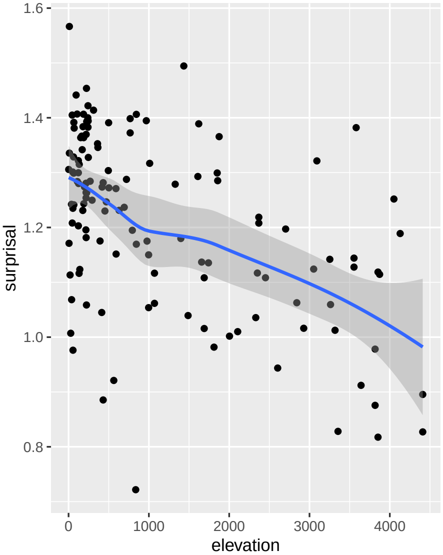


Origins and maintenance of IA numeral complexity

- Unusual properties of Indo-Aryan languages (e.g., retroflexion) are often ascribed to language contact (e.g., with Dravidian, Hock 1996) — impossible here
- Somewhat controversial ideas relate some grammatical properties of South Asian languages to social hierarchy (Emeneau 1974)
- It is possible that some form of social pressure led to the emergence and maintenance of complexity in the core area, which was lost in more peripheral areas
- I focus on two variables: altitude and vigesimality (i.e., presence of a base-20 system)

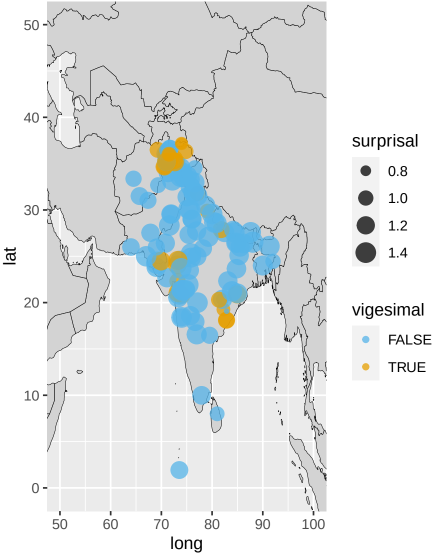
Altitude vs. complexity (Indo-Aryan)

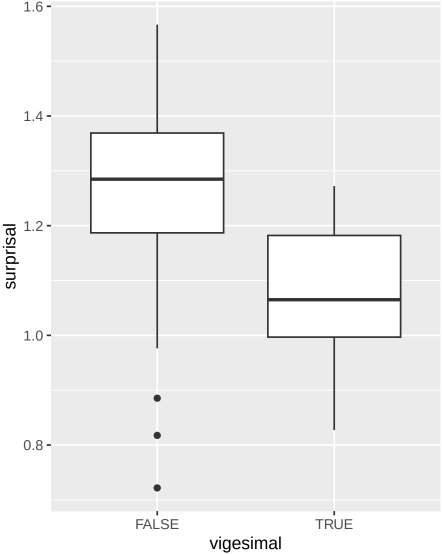




Complexity decreases as altitude increases

Vigesimality vs. complexity (Indo-Aryan)





Languages which have developed base-20 systems have lower complexity

- Higher altitude involves greater isolation, thought to foster complexity (Urban 2020)
- However, here, higher altitude coincides with simplification
- We are seeing the effects of at least two networks of language contact
  - Complexity is maintained in the core Indo-Aryan area
  - Higher-altitude languages of the Hindu Kush area (Weinreich 2015, Liljegren 2020) shifted to a vigesimal system, which involved simplification
- In line with the idea that properties of language involved in counting (e.g., numeral classifiers) are sensitive to contact (Grinevald 2000, Allassonnière-Tang et al. 2021)

Conclusion and outlook

- South Asia is a hotbed of integrative complexity in numeral systems; this is borne out by a number of metrics
- This is a largely Indo-Aryan phenomenon
- Despite this complexity, IA numeral systems appear to obey general principles of efficient communication: higher numbers are easier to realize and recognize
- Emergence seems to be due largely to local historical contingencies
- Social networks and networks of contact may be responsible for maintenance in the IA core area

Future directions: information regarding usage needed to better understand how such systems are maintained

Bibliography I

Ackerman, F. and R. Malouf (2013). Morphological organization: The low conditional entropy conjecture. *Language*, 429–464.

Allassonnière-Tang, M., O. Lundgren, M. Robbers, S. Cronhamn, F. Larsson, O.-S. Her, H. Hammarström, and G. Carling (2021). Expansion by migration and diffusion by contact is a source to the global diversity of linguistic nominal categorization systems. *Humanities and Social Sciences Communications 8*(1), 1–6.

Baayen, R. H., Y.-Y. Chuang, and J. P. Blevins (2018). Inflectional morphology with linear mappings. *The mental lexicon 13*(2), 230–268.

Berger, H. (1992). modern indo-aryan. *Indo-European numerals 57*, 243–287.

Blevins, J. P., P. Milin, and M. Ramscar (2017). The Zipfian paradigm cell filling problem. In *Perspectives on morphological organization*, pp. 139–158. Brill.

Bright, W. (1969). Hindi numerals. *Working Papers in Linguistics (University of Hawaii) 9*, 29–47.

Bibliography II

Brysbaert, M. (2005). Number recognition in different formats. In J. Campbell (Ed.), *Handbook of mathematical cognition*, pp. 23–42. New York, Hove: Psychology Press.

Cathcart, C. (2017). Decomposability and Frequency in the Hindi/Urdu Number System. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society, London*, pp. 1733–1738. Cognitive Science Society.

Chan, E., H.-J. Bibiko, C. Rzymski, S. J. Greenhill, and R. Forkel (2019, October). channumerals.

Dautriche, I., K. Mahowald, E. Gibson, A. Christophe, and S. T. Piantadosi (2017). Words cluster phonetically beyond phonotactic regularities. *Cognition 163*, 128–145.

Dehaene, S. and J. Mehler (1992). Cross-linguistic regularities in the frequency of number words. *Cognition 43*(1), 1–29.

Emeneau, M. B. (1974). The indian linguistic area revisited. *International journal of Dravidian linguistics 3*(1), 92–134.

Bibliography III

Goldwater, S., T. L. Griffiths, and M. Johnson (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition 112*(1), 21–54.

Grinevald, C. (2000). A morphosyntactic typology of classifiers. In G. Senft (Ed.), *Systems of nominal classification*, pp. 50–92. New York: Cambridge University Press.

Hock, H. H. (1993 [1996]). Subversion or convergence? the issue of pre-Vedic retroflexion reexamined. *Studies in the Linguistic Sciences 23*(2), 73–115.

Liljegren, H. (2020). The hindu kush–karakorum and linguistic areality. *Journal of South Asian languages and linguistics 7*(2), 239–285.

Núñez, R. E. (2017). Is there really an evolved capacity for number? *Trends in cognitive sciences 21*(6), 409–424.

Piantadosi, S. T., H. Tily, and E. Gibson (2012). The communicative function of ambiguity in language. *Cognition 122*(3), 280–291.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *The Annals of statistics 11*(2), 416–431.

Bibliography IV

Ritchie, S., R. Sproat, K. Gorman, D. van Esch, C. Schallhart, N. Bampounis, B. Brard, J. F. Mortensen, M. Holt, and E. Mahon (2019). Unified verbalization for speech recognition & synthesis across languages. In *INTERSPEECH*, pp. 3530–3534.

Schneider, R. M., J. Sullivan, F. Marušič, P. Biswas, P. Mišmaš, V. Plesničar, D. Barner, et al. (2020). Do children use language structure to discover the recursive rules of counting? *Cognitive psychology 117*, 101263.

Urban, M. (2020). Mountain linguistics. *Language and Linguistics Compass 14*(9), e12393.

Weinreich, M. (2015). Not only in the caucasus: Ethno-linguistic diversity on the roof of the world. In *Studies on Iran and the Caucasus*, pp. 455–472. Brill.

Xu, Y. and T. Regier (2014). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Volume 36.