Computational methods for morphological theory

Rob Malouf, San Diego State University

Plan

1. Foundational questions
2. **Morphological complexity & Information theory**
3. Morphological description & Deep Learning
4. Morphological explanation & Bayesian agents

Morphological diversity

- Sapir identified several dimensions of diversity
  - Number of morphemes per word (**analytic**, **synthetic**, **polysynthetic**)
  - Manner of combination (**agglutenative**, **fusional**)
  - Function of affixes
    - Class I: concrete roots (*table*)
    - Class II: functional derivation (-*er*)
    - Class III: concrete relational (number agreement)
    - Class IV: purely relational (case marking)

Morphological diversity

- Greenberg (1960) tried to make this more precise
  - Index of synthesis (M/W): morphs per word
  - Index of agglutination (A/J): agglutinative constructions per morph juncture
  - Compounding index (R/W): roots per word
  - Derivational index (D/W) and inflectional index (I/W)
  - Prefixal index (P/W) and suffixal index (S/W)
  - Isolation (I/N), pure inflection (Pu/N), concord (Co/N): fraction of intra-sentential relations (*nexuses*) expressed by word order, case, or agreement

Morphological diversity

- These metrics are conceptually straightforward but hard to implement
- Greenberg compared “the results of the indices calculated for a passage of 100 words of English in 1951, and arrived at by methods not longer fully recoverable by introspection” with “indices for a 100-word passage done recently in accordance with the methods outlined here”

	1951	1953
Synthesis	1.62	1.68
Agglutination	.31	.30
Compounding	1.03	1.00
Prefixing	1.00	1.04
Suffixing	.50	.64
Gross inflection	.64	.53

Morphological diversity

- Greenberg (1960)

TABLE 1

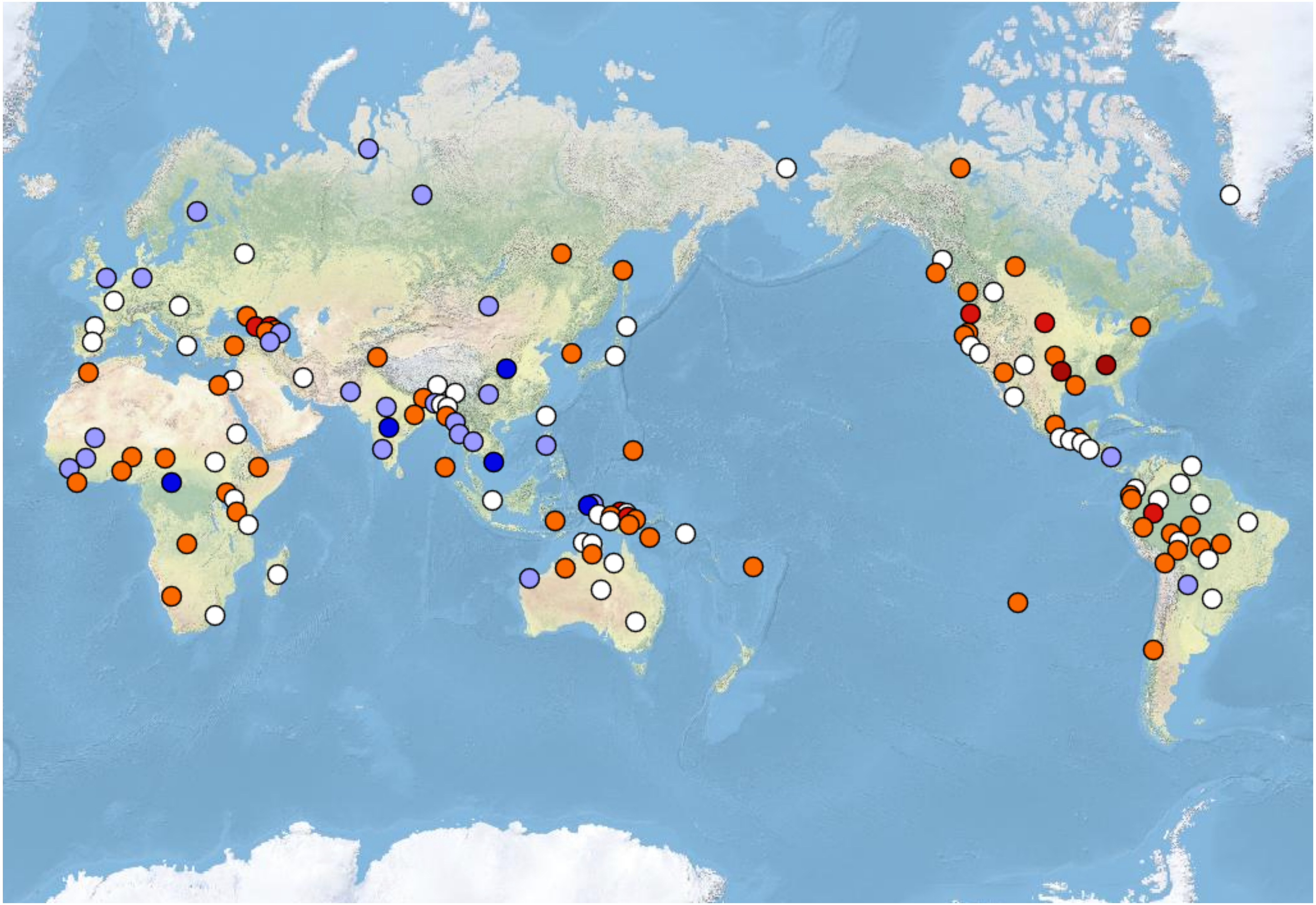
	Sanskrit	Anglo-Saxon	Persian	English	Yakut	Swahili	Annamite	Eskimo
Synthesis	2.59	2.12	1.52	1.68	2.17	2.55	1.06	3.72
Agglutination	.09	.11	.34	.30	.51	.67	...	.03
Compounding	1.13	1.00	1.03	1.00	1.02	1.00	1.07	1.00
Derivation	.62	.20	.10	.15	.35	.07	.00	1.25
Gross inflection	.84	.90	.39	.53	.82	.80	.00	1.75
Prefixing	.16	.06	.01	.04	.00	1.16	.00	.00
Suffixing	1.18	1.03	.49	.64	1.15	.41	.00	2.72
Isolation	.16	.15	.52	.75	.29	.40	1.00	.02
Pure inflection	.46	.47	.29	.14	.59	.19	.00	.46
Concord	.38	.38	.19	.11	.12	.41	.00	.38

Morphological diversity

- World Atlas of Language Structures
  - Feature 22A: Inflectional Synthesis of the Verb

Values

0-1 category per word	5
2-3 categories per word	24
4-5 categories per word	52
6-7 categories per word	31
8-9 categories per word	24
10-11 categories per word	7
12-13 categories per word	2

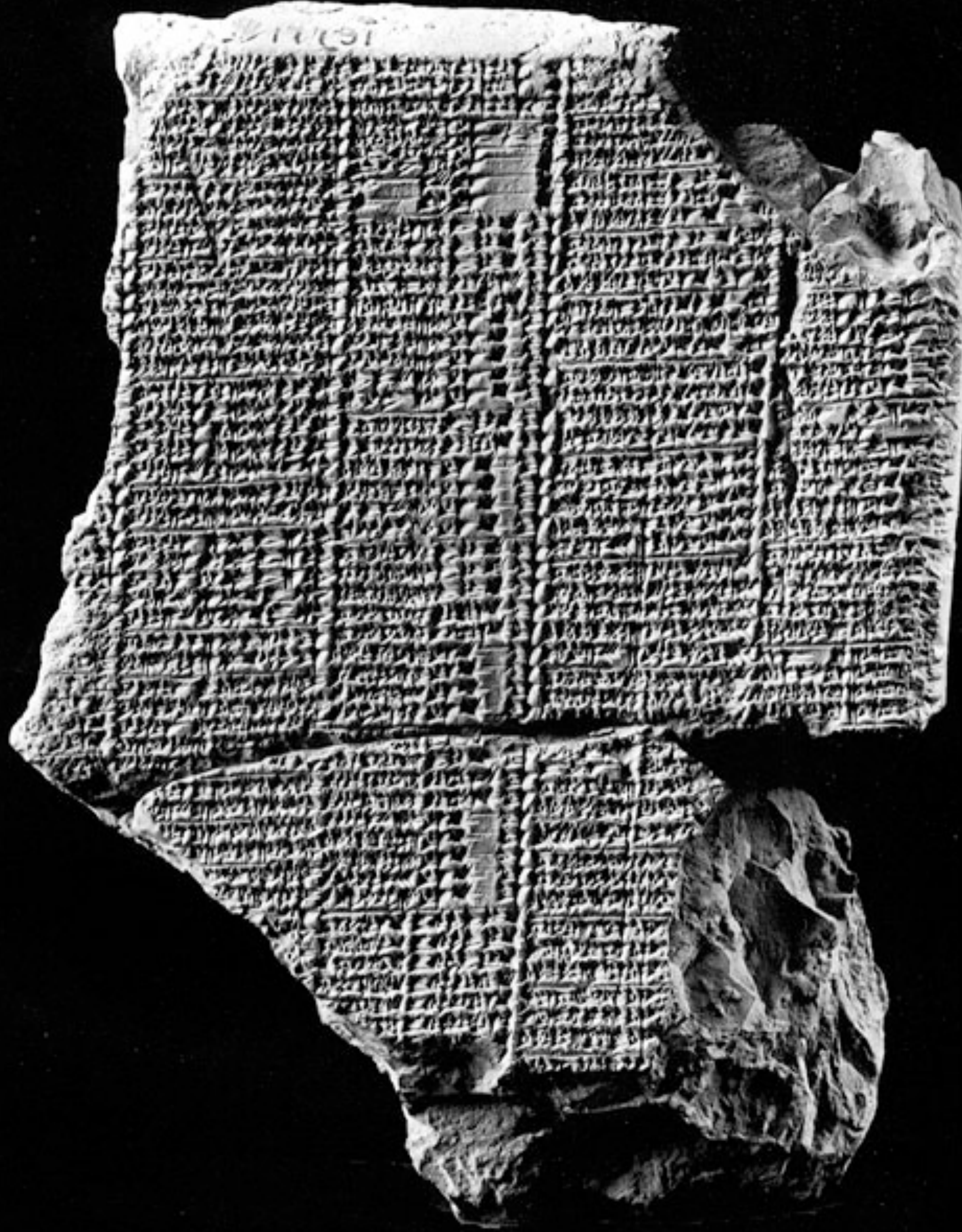


Paradigm size

- Morphological complexity also has a **paradigmatic** dimension
- Languages vary in the number of affixes that are available (Anderson 2015)
  - 500+ derivational affixes in W. Greenlandic
  - 250 in Kwakw'ala
  - 150 in English
  - 15 in Mandarin
  - 0 (?) in Vietnamese

Paradigms

- The earliest grammatical literature are Old Babylonian Grammatical Texts (from 2000BC–1600BC)
- Grids of words in Sumerian and Akkadian following a (more or less) consistent pattern
- Verb paradigms list 3rd person, then 1st, then 2nd
- Other consistent patterns for nouns and verbs
- Scribes deviated from the usual order to point out complications in Sumerian grammar



OBGT VII. Indicative forms: present, preterite								Akk. structure
§16	31	àm-du	*illakam*	he comes	—	G	V	Ps
§17	34	àm-ši-du	*illakaššum*	he comes to him	3D	G	V	Ps
§21	37	mu-e-ši-du	*illakakkum*	he comes to you	2D	G	V	Ps
§18	39	àm-ma-du	*ittallakam*	he comes away	—	Gt	V	Ps
§19	42	àm-ma-ši-du	*ittallakaššum*	he comes away to him	3D	Gt	V	Ps
§20	45	àm-mu-e-ši-du	*ittallakakkum*	he comes away to you	2D	Gt	V	Ps
§12	47	ì-du	*illak*	he goes	—	G	—	Ps
§13	50	in-ši-du	*illakšum*	he goes to him	3D	G	—	Ps
§22	53	ba-du	*ittallak*	he goes away	—	Gt	—	Ps
§23	56	ba-ši-du	*ittallakšum*	he goes away to him	3D	Gt	—	Ps
§26	59	i-im-gen	*illikam*	he came	—	G	V	Pt
§27	62	i-im-ši-gen	*illikaššum*	he came to him	3D	G	V	Pt
§31	65	mu-e-ši-gen	*illikakkum*	he came to you	2D	G	V	Pt
§28	67	im-ma-gen	*ittalkam*	he came away	—	Gt	V	Pt
§29	70	im-ma-ši-gen	*ittalkaššum*	he came away to him	3D	Gt	V	Pt
§30	73	im-mu-e-ši-gen	*ittalkakkum*	he came away to you	2D	Gt	V	Pt
§24	75	in-gen, ì-gen	*illik*	he went	—	G	—	Pt
§25	78	in-ši-gen	*illikšum*	he went to him	3D	G	—	Pt
§32	81	ba-gen	*ittalak*	he went away	—	Gt	—	Pt
§33	84	ba-ši-gen	*ittalakšum*	he went away to him	3D	Gt	—	Pt

Paradigm size

- Inflection is another source of paradigmatic complexity
- Latin ‘star’

	SINGULAR	PLURAL
NOMINATIVE	*stēlla*	*stēllae*
GENITIVE	*stēllae*	*stēllārum*
DATIVE	*stēllae*	*stēllīs*
ACCUSATIVE	*stēllam*	*stēllās*
ABLATIVE	*stēllā*	*stēllīs*
VOCATIVE	*stēlla*	*stēllae*

- One suffix per wordform, but between 8 and 12 alternatives for the suffix

Paradigm size

- WALS on case inventories

Value	Representation
No morphological case-marking	100
2 case categories	23
3 case categories	9
4 case categories	9
5 case categories	12
6-7 case categories	37
8-9 case categories	23
10 or more case categories	24
Exclusively borderline morphological case-marking	24
Total:	261

Paradigm size

- WALS on past tenses

	Value	Representation
● (yellow)	Past/non-past distinction marked; no remoteness distinction	94
● (orange)	Past/non-past distinction marked; 2-3 degrees of remoteness distinguished	38
● (red)	Past/non-past distinction marked; at least 4 degrees of remoteness distinguished	2
○	No grammatical marking of past/non-past distinction	88
	Total:	222

Table 1. Remoteness distinctions in Yagua

Name in grammar	Use	Suffix	Example
Proximate 1	'a few hours previous to the time of utterance'	*-jásiy*	*rayą́ą́siy* *{ray-jiya-jásiy}* 1SG-go-PROX1 'I went (this morning).'
Proximate 2	'one day previous to the time of utterance'	*-jay*	*rįjnúújeñíí* *{ray-jųnnúúy-jay-níí}* 1SG-see-PROX2-3SG 'I saw him (yesterday).'
Past 1	'roughly one week ago to one month ago'	*-siy*	*sadííchimyaa* *{sa-díí-siy-maa}* 3SG-die-PST2-PERF 'He has died (between a week and a month ago').
Past 2	'roughly one to two months ago up to one or two years ago'	*-tíy*	*sadíítímyaa* *{sa-dííy-tíy-maa}* 3SG-die-PST2-PERF 'He has died (between 1 to 2 months and a year ago').
Past 3	'distant or legendary past'	*-jada*	*raryúpeeda* *{ray-rupay-jada}* 1SG-be.born-PST3 'I was born (a number of years ago).'

Tilapa Otomi tenses (Palancar 2012)

Table 2. *The grammatical tenses of T-Oto*.

Realis	Present	continuous	***grá*** *^{h}peni*	'you're washing it now'
		habitual	***grú̱*** *^{h}peni*	'you commonly wash it'
	Ambulative		***gá*** *^{h}peni*	'you wash it away (here and there)'
	Imperfect	continuous	***grá má*** *^{h}peni*	'you were washing it'
		habitual	***grú̱ mú̱*** *^{h}peni*	'you used to wash it'
		ambulative	***gá má tí*** *^{h}peni*	'you were washing it away/long ago'
	Past		***gú̱*** *^{h}peni*	'you washed it'
	Perfect		***xkú̱*** *^{h}peni*	'you've already washed it'
	Pluperfect		***xkí*** *^{h}peni*	'you'd already washed it'
Irrealis	Present		***gi*** *^{h}peni*	'you'll wash it'
	Immediative		***xta gi*** *^{h}peni*	'you're about to wash it'
	Ambulative		***gi tí*** *^{h}peni*	'you'll wash it away (here and there)'
	Andative		***gri*** *^{h}peni*	'you'll go wash it'
	Past		***gi gi*** *^{h}peni*	'you'd wash it'
	Perfect		***xki gi*** *^{h}peni*	'you'd have washed it'

Table 3. *Local values*.

Realis	Ambulative	andative	*gá-**r** peni*	'you wash it away (here and there)'
		cisloc.	*g${}^{\langle w\rangle}$á **tí** ^{h}peni*	'you're washing it as you come'
	Past	transloc.	*g${}^{\langle w\rangle}$ú̱ ^{h}peni*	'you washed it somewhere else'
	Perfect	transloc.	*xk${}^{\langle w\rangle}$ú̱ ^{h}peni*	'you've already washed it somewhere else'
	Pluperfect	transloc.	*xk${}^{\langle w\rangle}$ú̱ ^{h}peni*	'you'd already washed it somewhere else'
Irr	Present	transloc.	*g${}^{\langle w\rangle}$u̱ ^{h}peni*	'you'll (go and) wash it somewhere else'
	Past	transloc.	*g${}^{\langle w\rangle}$u̱ g${}^{\langle w\rangle}$u̱ ^{h}peni*	'you'd (go and) wash it somewhere else'

(and verbs agree with the person of the subject!)

Paradigm size

- Kiksht past tenses

ga(l) . . . u-	remote past
ga(l) . . . t-	from one to ten years ago
ni(g) . . . u-	from a week to a year ago
ni(g) . . . t-	last week
na(l)-	last couple of days
i(g) . . . u-	earlier today
i(g) . . . t-	just now

- Bamilete-Dschang has 15 compound tenses: “Thus, combination of the tomorrow future (F3) with the later today future (F2) indicates a situation that will hold soon after some reference point tomorrow . . .” (Comrie 1985)

Morphological diversity

- Beyond simple counting, we can look for ways that languages typically can be complex
- Nichols (1992) proposed a complexity metric based on the fraction of possible inflections a language showed (cf. Greenberg's nexus-based measures)
- McWhorter (2001) on creoles
  - Markedness of phonemic inventory
  - Number of rules in syntax
  - Degree of grammaticalization of "fine-grained semantic and pragmatic distinctions"

Morphological diversity

- Rich case or tense systems add complexity to a morphological system, but also do communicative work
- Is Bamilete-Dschang more complex than Mandarin, or less?
- Can we quantify the **net** complexity of morphology?

Algorithmic complexity

- The **Kolmogorov complexity** $K(s)$ of a sequence is the length of the shortest program that can generate it
- Take some sequences of 1,000,000 digits:

  0000000000000…

  0101010101010…

  1223334444555556…

  001012012301234…

  1248163264128256…

  1123581321345589144…

  31415926535897932…

  78254633069748271…

Algorithmic complexity

- The smallest program generating a completely random sequence is the sequence itself (randomness=complexity)
- Regularities in the sequence let us shorten the program (patterns=simplicity)
- Problems
  - What programming language should we use?
  - How do we know we've got the shortest program?
- $K(s)$ is not computable, but we can get an upper bound on it via compression

Algorithmic complexity

- Compressed sizes of 1,000,000 digit sequences, in bytes:

0000000000000…	992
0101010101010…	993
1223334444555556…	2,843
0010120123012345…	9,769
1248163264128256…	470,677
1123581321345589144…	470,594
3141592653589793…	470,450
7825463306974827…	470,474

Algorithmic complexity

- Juola (1998) used this as a tool to get at the syntax/ morphology trade-off
- Take Bible translations in various languages and compress them to estimate $K(s)$
- Replace each word with a random number (*the*=7643, *house*=65, …)

jump	walk	touch	8634	139	5543
jumped	walked	touched	15	4597	1641
jumping	walking	touching	3978	102	6

Figure 1: Example of morphological degradation process

- Compress the result to estimate $K(s')$

Algorithmic complexity

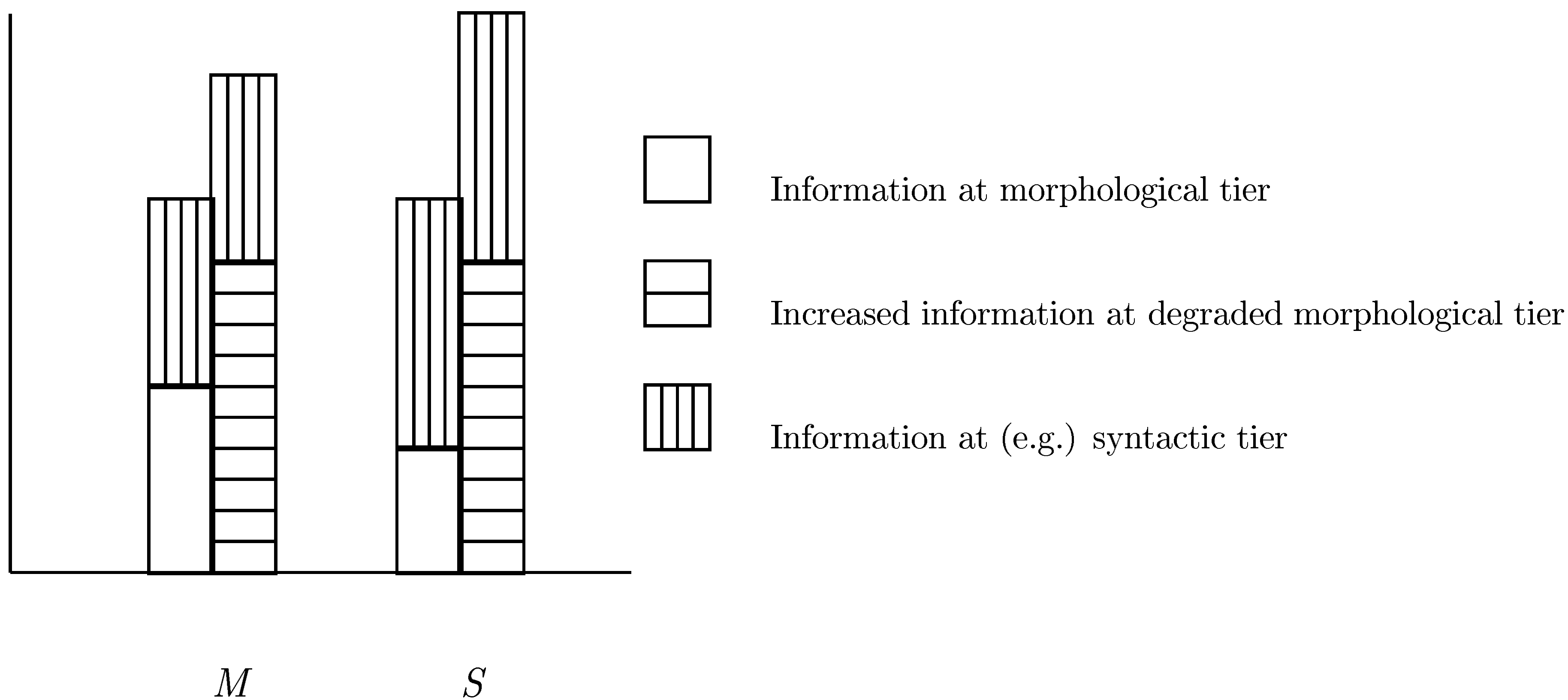


Figure 2: Hypothetical example of degradation ratios

Algorithmic complexity

- Compare $K(s)$ and $K(s')$: the difference is what morphology (and phonology?) was contributing to patterns

Table 2: R/C ratios with linguistic form counts

Language	R/C	Types in sample	Tokens in sample
Maori	0.895	19,301	1,009,865
English	0.972	31,244	824,364
Dutch	0.994	42,347	805,102
French	1.01	48,609	758,251
Russian	1.04	76,707	600,068
Finnish	1.12	86,566	577,413

Algorithmic complexity

- Removing phonology and morphology together makes the results very hard to interpret
- Developed further by Moscoso del Prado Martín (2011)

$$
\begin{aligned}
H(N) &= G(N) + H_s(N), \\
g(N) &= \frac{1}{N} G(N) = \\
&= \frac{1}{N} \left[H(N) - H_s(N) \right] = h(N) - h_s(N). \\
g &= \lim_{N \to \infty} g(N) = \lim_{N \to \infty} \frac{G(N)}{N}. \\
g_s &= L_s \cdot g,
\end{aligned}
$$

Algorithmic complexity

- Further decompose per-sentence complexity

$$
\begin{aligned}
g_s &= g_s^{\text{lexicon}} + g_s^{\text{derivation}} + g_s^{\text{inflection}} + g_s^{\text{syntax}} + \cdots \\
g'_s &= g_s'^{\text{lexicon}} + g_s'^{\text{derivation}} + g_s'^{\text{syntax}} + \cdots \\
g_s^{\text{inflection}} &= g_s - g'_s
\end{aligned}
$$

- Remove inflection from words in Europarl corpus using a lemmatizer (*cars* → *car*, *ate* → *eat*, etc)

- Remove syntactic relations by randomizing the order of words in the corpus

- Compare compressed sizes (*) of corpora before and after

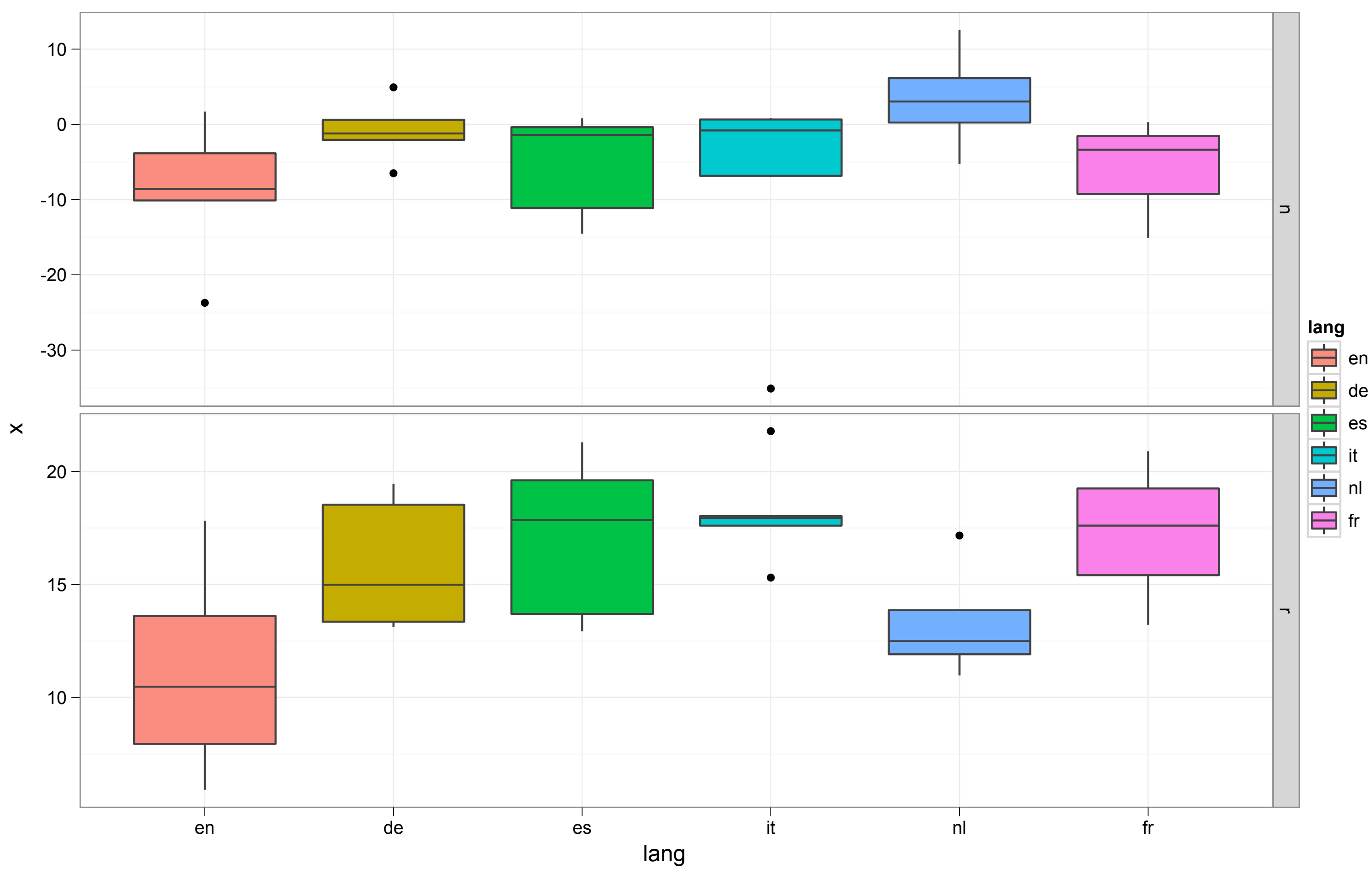
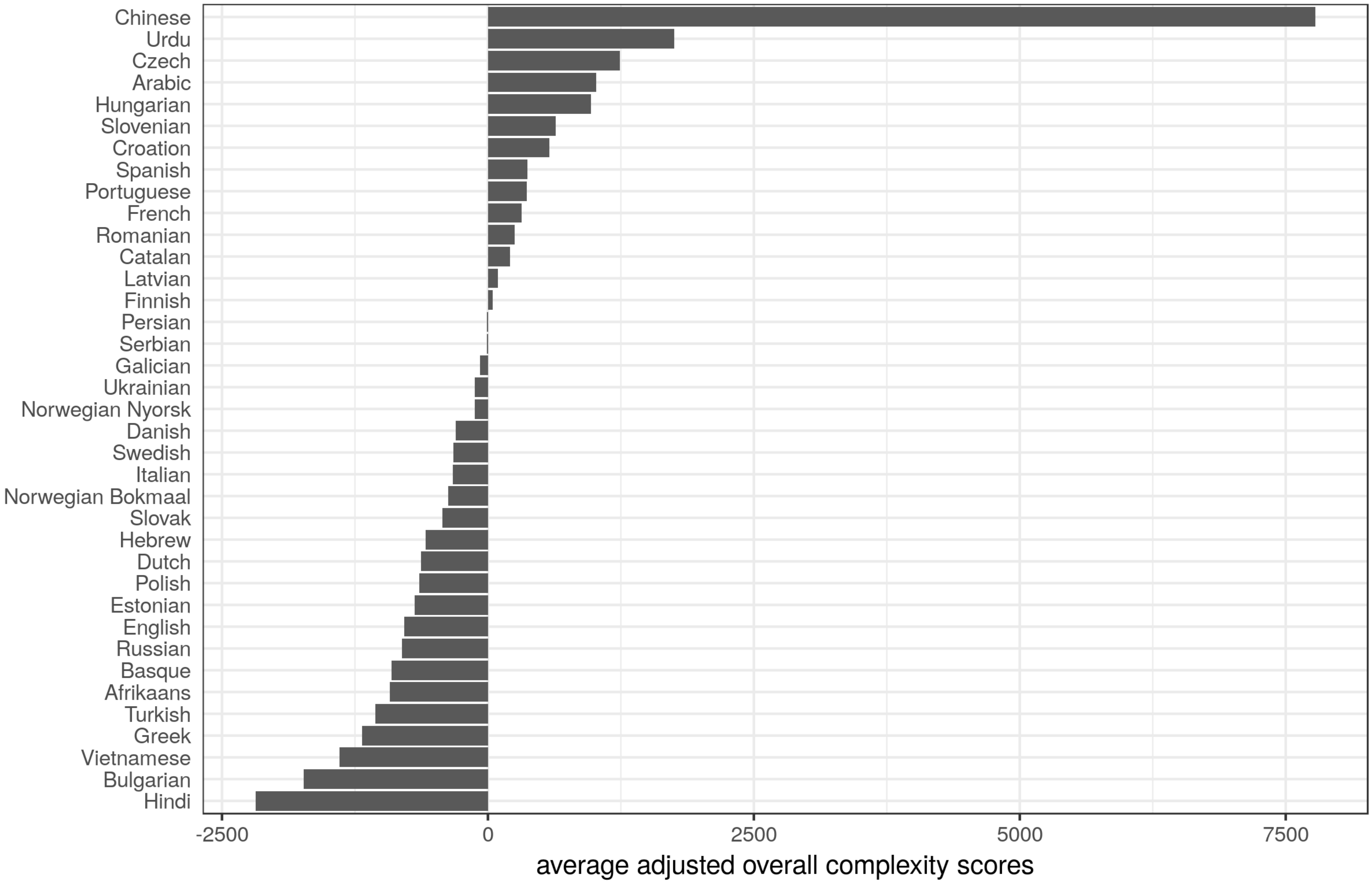


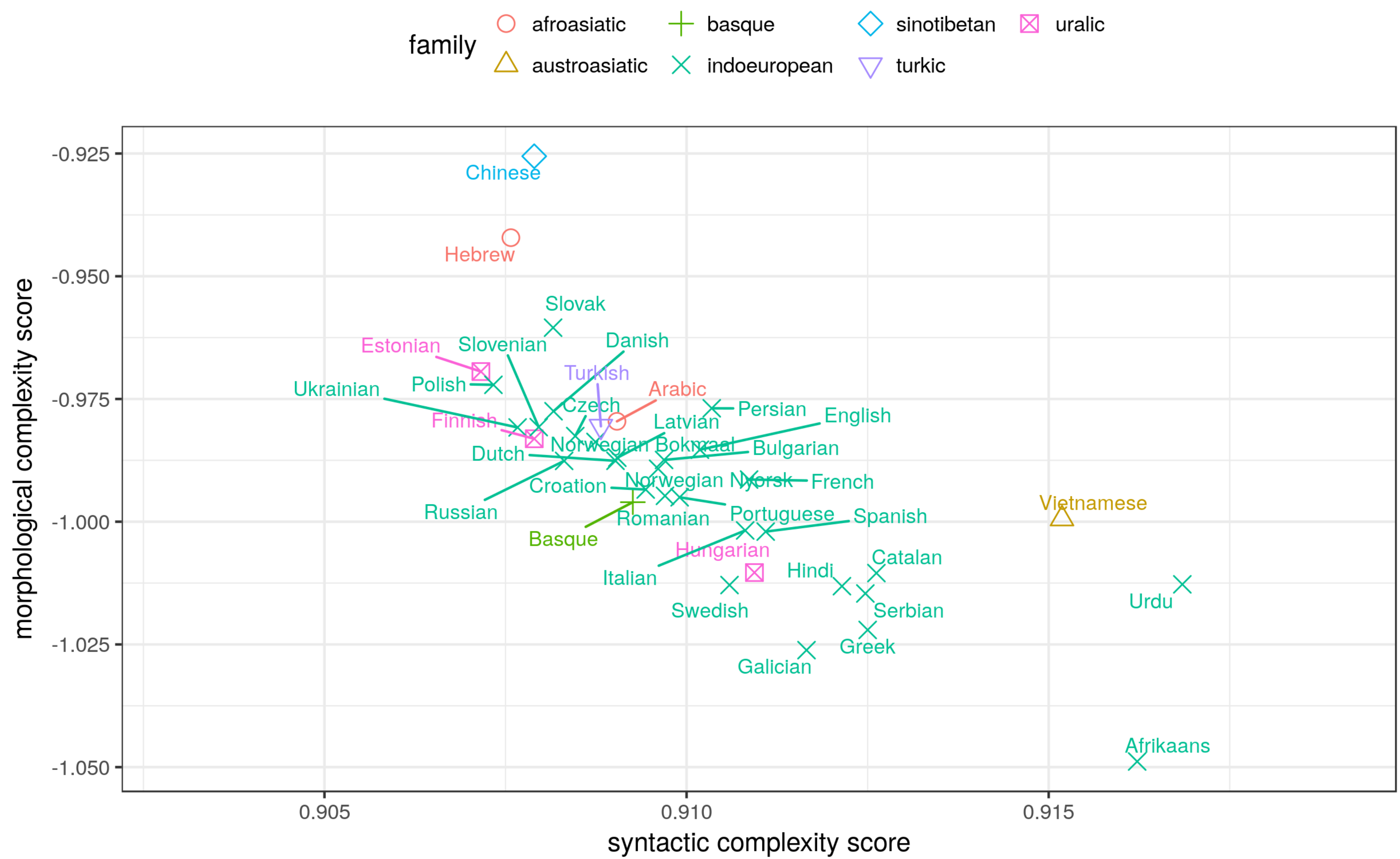
Figure 1: Summary of results. The upper panel plots the distribution of inflectional complexity (in nats/sentence) values obtained for each language in the original word order corpora. The lower panel plots the same results for the corpora in which the word order was randomized.

Algorithmic complexity

- Ehret (2018) also adapted Juola’s method, comparing the compressed size of:
  - original document
  - document with 10% of the words removed (syntax)
  - document with 10% of characters removed (morphology)
- Applied to sample of texts from UD in 37 languages



Chinese
Urdu
Czech
Arabic
Hungarian
Slovenian
Croation
Spanish
Portuguese
French
Romanian
Catalan
Latvian
Finnish
Persian
Serbian
Galician
Ukrainian
Norwegian Nyorsk
Danish
Swedish
Italian
Norwegian Bokmaal
Slovak
Hebrew
Dutch
Polish
Estonian
English
Russian
Basque
Afrikaans
Turkish
Greek
Vietnamese
Bulgarian
Hindi
-2500
0
2500
5000
7500
average adjusted overall complexity scores



family
afroasiatic
basque
sinotibetan
uralic
austroasiatic
indoeuropean
turkic
morphological complexity score
syntactic complexity score
-0.925
-0.950
-0.975
-1.000
-1.025
-1.050
0.905
0.910
0.915
Chinese
Hebrew
Slovak
Estonian
Slovenian
Danish
Turkish
Ukrainian
Polish
Arabic
Czech
Persian
English
Finnish
Latvian
Norwegian Bokmaal
Bulgarian
Dutch
Norwegian Nynorsk
French
Croation
Russian
Romanian
Portuguese
Spanish
Vietnamese
Basque
Hungarian
Catalan
Hindi
Italian
Swedish
Serbian
Urdu
Greek
Galician
Afrikaans

Morphological diversity

- Rich case or tense systems add complexity to a morphological system, but also do communicative work
- Another dimension of complexity comes from lexically conditioned allomorphy (e.g., inflection classes)
- Latin nouns
  - 6 cases, 2 numbers = 12 forms
  - >5 different sets of 12 forms

Inflection classes

- Inflection classes also create a kind of paradigmatic complexity
- Baerman, et al. (2009): Nuer nouns have two stems and three possible suffixes: *-Ø, -kä, -ni*

	‘bear’	‘ant’	‘lion’	‘fat’	‘egret’	‘monkey’	‘child’
NOM SG	lɛt	ɲiɛc	lony	liɛth	bööŋ	gɔɔk	gat
GEN SG	lɛt	ɲiɛc-kä	lony	liɛth-kä	bööŋ-ka̤	gɔɔk-kä	gat-kä
LOC SG	lɛt	ɲiɛc-kä	lony	liɛth	bööŋ-ka̤	gɔak	gat-kä
NOM PL	leet	ɲiic	luony	lith	bo̤o̤ŋ-ni̱	gɔak-ni̱	gaat
GEG PL	leet-ni̱	ɲiic-ni̱	luony-ni̱	lith-ni̱	bo̤o̤ŋ-ni̱	gɔak-ni̱	gaan
LOC PL	leet-ni̱	ɲiic-ni̱	luony	lith-ni̱	bo̤o̤ŋ-ni̱	gɔaak-ni̱	gaat

Figure 8: Varieties of Nuer noun inflection (Frank 1999)

Inflection classes

- The possible combinations of singular and plural patterns yield 16 different inflection classes

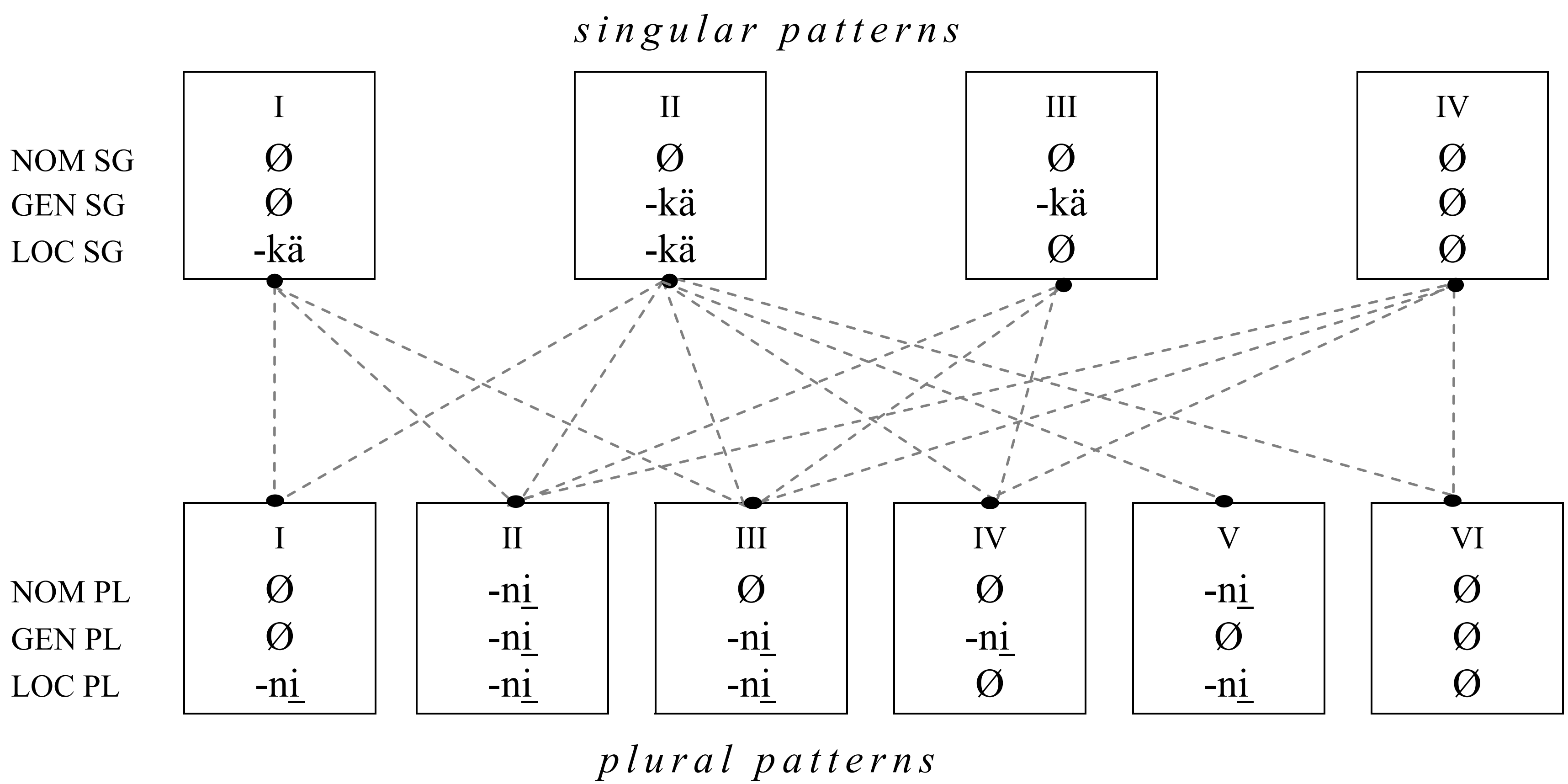


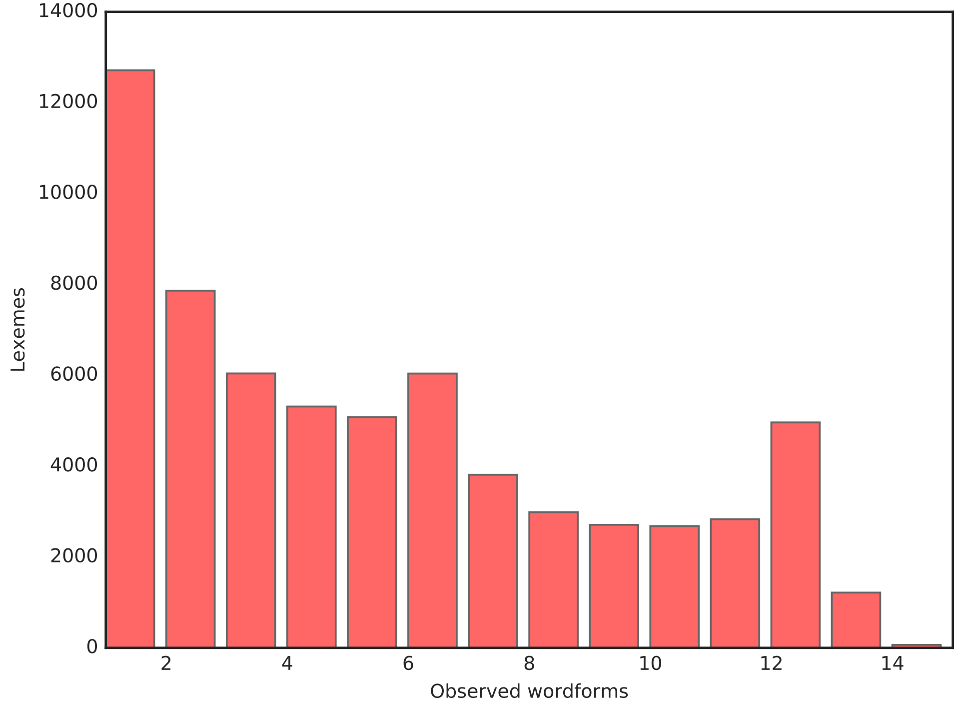
Figure 11: Singular ~ plural pattern mapping in Nuer (based on Frank 1999)

Paradigm Cell Filling Problem

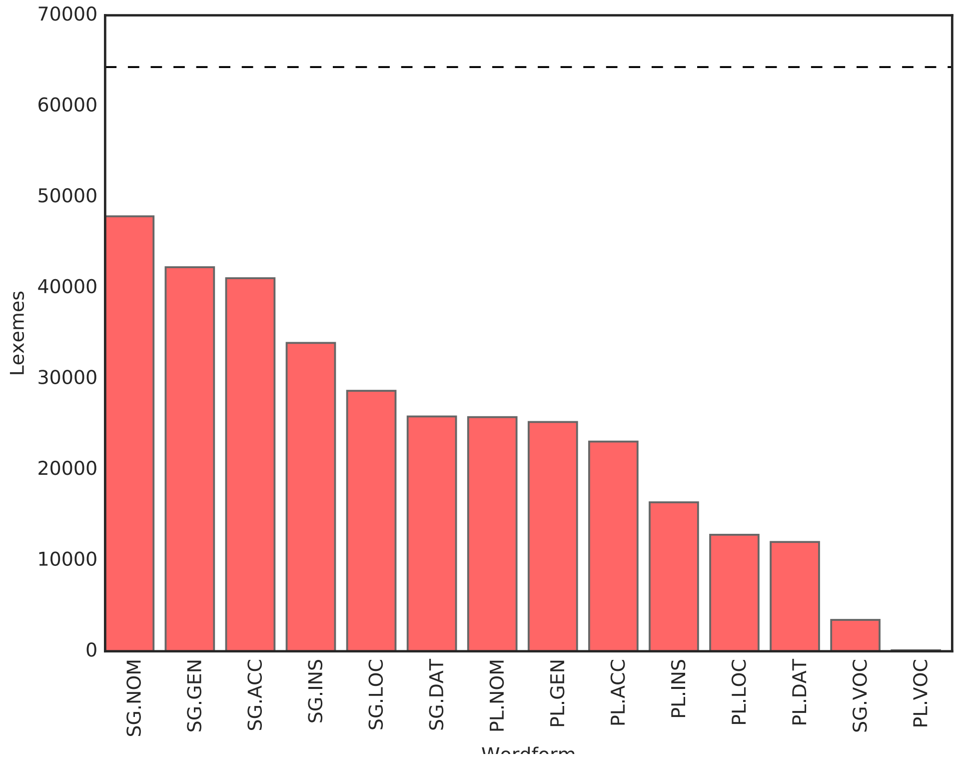
- **Paradigm Cell Filling Problem**: Given exposure to a novel inflected word form, what licenses reliable inferences about the other word forms in its inflectional family?
- Do speakers simply memorize full paradigms?
  - Tundra Nenets nouns have 210 forms: case, number, possessor person, possessor number (Ackerman & Salminen 2006)
  - Khaling verbs have up to 331 forms (Jacques et al. 2012)
  - Zipf's Lawe: A few forms are frequent, but most are rare (Chan 2008)

Zipf's Law

- Czech National Corpus SYN2010
    - 100 million morphologically tagged words
  - 64,302 distinct noun lexemes
  - 561,668 distinct noun wordforms
  - 900,228 possible wordforms (7 cases, 2 numbers)
- Only 66 lexemes occur with full paradigms
- No single form is observed for every lexeme
- Only 110 lexemes occur in the VOC.PL (but more frequent in spoken language, same as NOM.PL)



14000
12000
10000
8000
6000
4000
2000
0
Lexemes
2
4
6
8
10
12
14
Observed wordforms



70000
60000
50000
40000
30000
20000
10000
0
Lexemes
SG.NOM
SG.GEN
SG.ACC
SG.INS
SG.LOC
SG.DAT
PL.NOM
PL.GEN
PL.ACC
PL.INS
PL.LOC
PL.DAT
SG.VOC
PL.VOC

Paradigm Cell Filling Problem

- **Paradigm Cell Filling Problem**: Given exposure to a novel inflected word form, what licenses reliable inferences about the other word forms in its inflectional family?
- It is implausible that speakers of languages with complex morphology and multiple inflection classes encounter every inflected form of every word
- Hockett 1967: "in his analogizing … [t]he native user of the language … operates in terms of all sorts of internally stored paradigms, many of them doubtless only partial; and he may first encounter a new basic verb in any of its inflected forms."

Paradigm Cell Filling Problem

- Paradigmatic complexity apparently adds nothing (Wurzel calls it “ballast”), but what does it cost?
- Our intuition: nothing, as long as paradigms are organized in a way that allows speakers to predict the correct forms
- More specifically: we distinguish between e(numerative) complexity and i(ntegrative) complexity
  - **E-complexity** is the size of the system (number of paradigms cells, allomorphs, inflection classes, morphs per word, etc)
  - **I-complexity** reflects the organization of paradigms to make the PCFP tractable

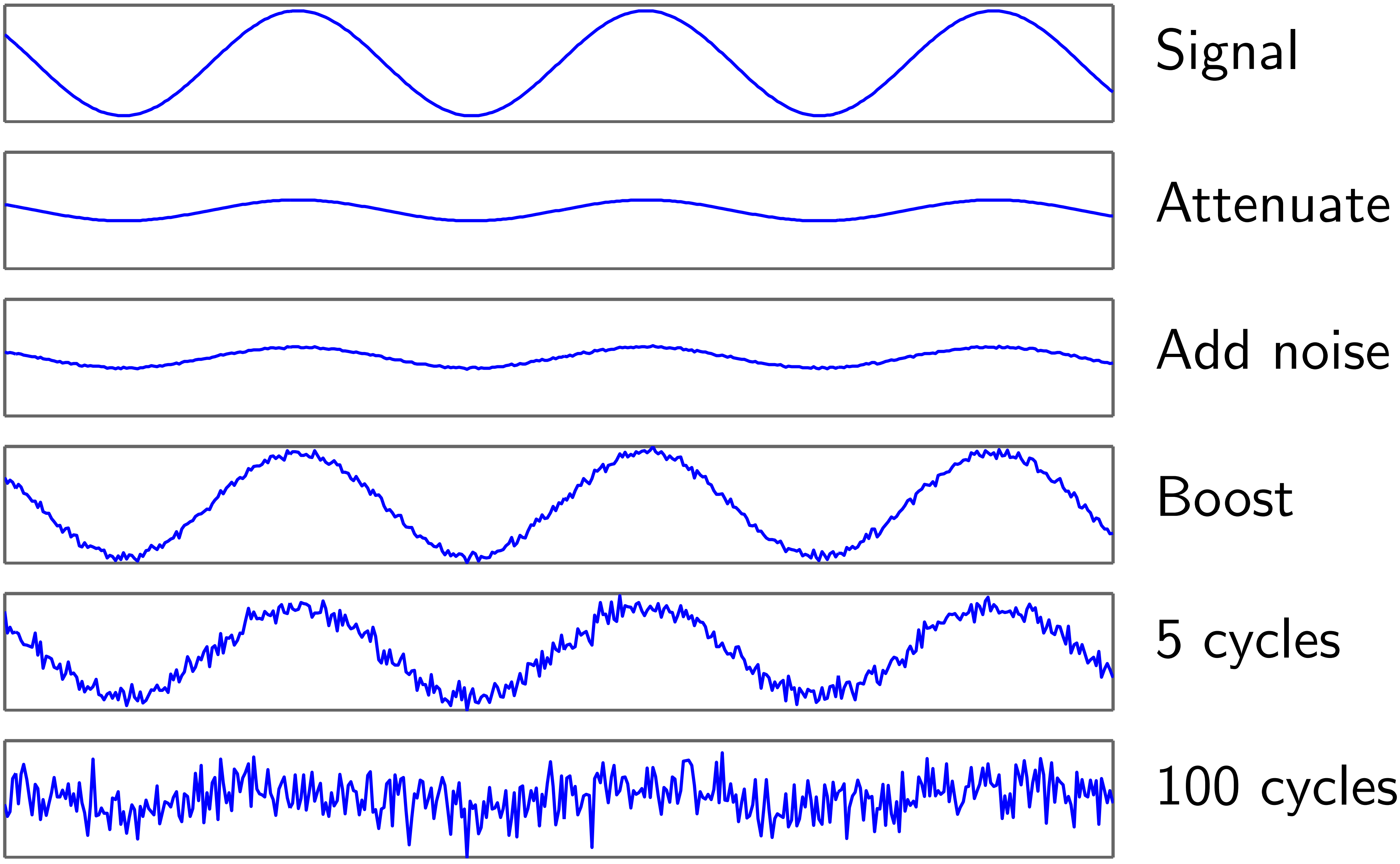
The hypothesis: I-complexity

- What makes a language difficult to learn and use (not to describe)?
- The issue is not simplicity or complexity per se, but the nature of organization supporting that complexity
- I-complexity is measurable and quantifiable
- **Principle of Low Paradigm Entropy**: Paradigms tend to have low expected conditional entropy

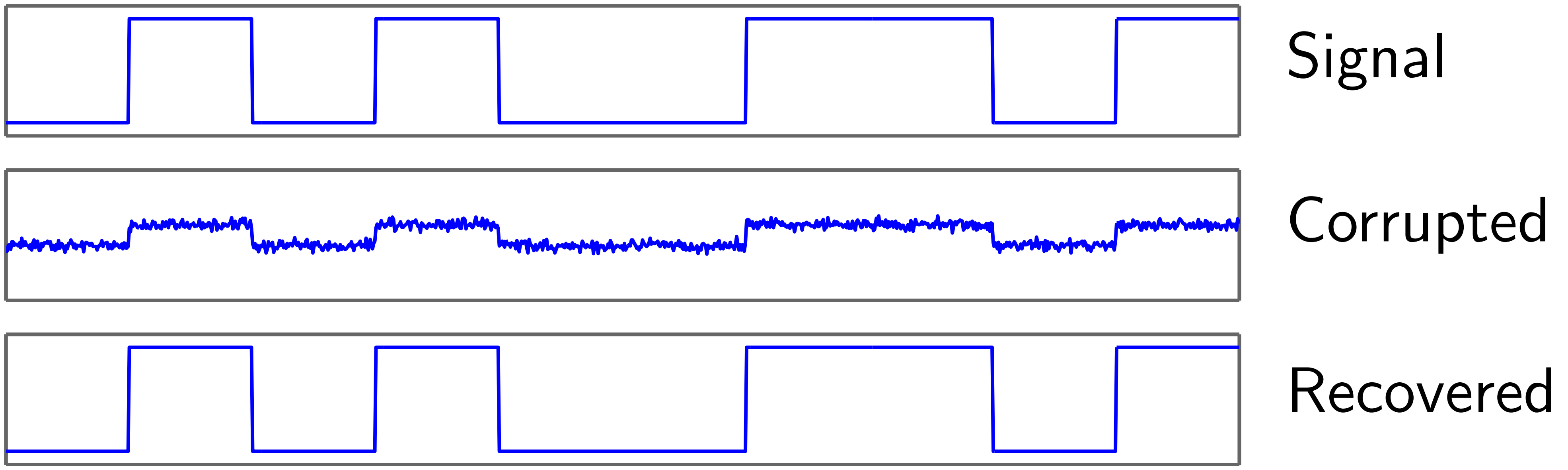
Information Theory

- Claude Shannon's "A mathematical theory of communication" (1948)

  "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages."



(from Murray, 2010)



(from Murray, 2010)

Information Theory

- Digital communications involves the transfer of symbols drawn from a discrete alphabet
  - Quantized analog signals
  - English letters
  - Decimal digits
  - Racing flags
  - Allomorphs
- Using a codebook, we convert among any discrete information sources

Information Theory

- The **information content** of a message $I(p)$ is a function of its probability
- Information is related to **probability**: more probable events are less informative, less probable events are more informative
- Information is also related somehow to **code lengths**: long books have the potential to contain more information than short ones

Encoding information

- Suppose we want to transmit information about a poker hand, and an earpiece is too obvious
- Lots of approaches – toe taps, flashes of light, coughs, etc – but let's assume our message consists of a sequence of binary choices (**bits**)
- There are 2^b different sequences of b bits

$$\underbrace{2 \times \cdots \times 2}_{b} = 2^b$$

- And recall that if $b^x = y$, then $\log_b y = x$
- So the number of bits required to uniquely encode n different sequences is $[\log_2 n]$

Encoding information

- A binary code for transmitting poker hands:

straight flush	0000
four of a kind	0001
full house	0010
flush	0100
straight	1000
three of a kind	0011
two pair	0101
pair	1001
high card	0111

- The expected message length $\mathrm{E}[C]=4$ bits per hand

Encoding information

- A **prefix code** taking advantage of uneven probabilities:

straight flush	0.0000154	000011
four of a kind	0.000240	0000100
full house	0.00144	0000101
flush	0.00196	00000
straight	0.00393	0001
three of a kind	0.0211	010
two pair	0.0475	011
pair	0.422	001
high card	0.501	1

- Now $E[C]$=2.01 bits, an average savings of 1.99 bits per

Encoding information

- A better code, taking advantage of probabilities

straight flush	0.0000154	11111111
four of a kind	0.000240	11111110
full house	0.00144	1111110
flush	0.00196	111110
straight	0.00393	11110
three of a kind	0.0211	1110
two pair	0.0475	110
pair	0.422	10
high card	0.501	

- For this one, $E[C]$=1.61 bits, an average savings of 2.39 bits

Encoding information

- Is there a better code out there, or is this the best we can do?

- Shannon's Source Coding Theorem provides an answer: the minimum code length for a message is bounded by its information content $I(p)$

- Okay, so, how do we measure $I(p)$?

Information

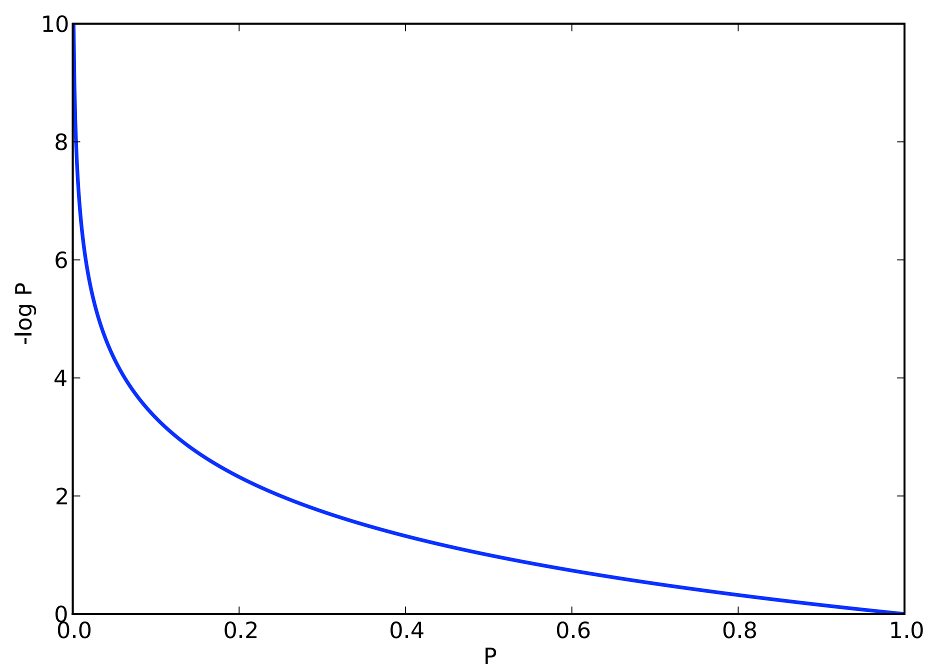
- Some basic properties of a sensible measure of information content $I(p)$
  - Information is non-negative: $I(p) \geq 0$
  - Events that are certain to occur convey no information at all: $I(p) = 0$
  - If two independent events (so that $p_{12} = p_1 \times p_2$) occur together, then the total information is the sum of the individual informations: $I(p_{12}) = I(p_1) + I(p_2)$
  - Information $I(p)$ should be a continuous monotonic decreasing function of p

Information

- Given these axioms, a good candidate for our information content function is

$$I(p) = -log_b p$$

for some base b



Entropy

- This measure of the information content of a message x:

  $$I(x) = -log_2 p(x)$$

  is sometimes called the **self-information** or **surprisal**

- In designing a coding scheme, we need to take into account all possible messages (if we knew in advance which message we'd be coding, we wouldn't need to code it)

- The expected information content of a message $E[I(X)]$ is the **entropy** of X

  $$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

Paradigm entropy

- Back to morphology
- The **conditional entropy** is the uncertainty in one random variable on average, given that we know the value of another random variable

$$\begin{aligned} H(Y|X) &= -\sum_{x\in X} p(x) \sum_{y\in Y} p(y|x)\log_2 p(y|x) \\ &= H(X,Y) - H(X) \end{aligned}$$

- The conditional entropy of one cell given another is a measure of i-complexity, or the inter-predictability with a paradigm (Ackerman, Blevins, and Malouf 2009)

Pite Saami

- For example: Pite Saami (Wilbur 2014, Ackerman & Malouf 2016)
- Seven cases (setting aside the marginal essive and abessive cases) and two numbers
- Realized via stem grade (strong vs. weak) and suffix
- Following Wilbur (2014), Pite Saami has eight nominal declensions showing distinct grade and suffix patterns

Pite Saami

- *bäbbmo* ‘food’

	SG	PL
NOM	*<u>bäbbm</u>-o*	*biebm-o*
GEN	*biebm-o*	*biebm-oj*
ACC	*biebm-ov*	*biebm-ojd*
ILL	*<u>bäbbm</u>-oj*	*biebm-ojda*
INESS	*biebm-on*	*biebm-ojn*
ELAT	*biebm-ost*	*biebm-ojst*
COM	*biebm-ojn*	*biebm-oj*

Pite Saami

CLASS	NOM.SG	GEN.SG	ACC.SG	ILL.SG	INESS.SG	ELAT.SG	COM.SG
Ia	str+*a*	wk+*a*	wk+*av*	str+*aj*	wk+*an*	wk+*ast*	wk+*ajn*
Ib	str+*á*	wk+*á*	wk+*áv*	str+*áj*	wk+*án*	wk+*ást*	wk+*ájn*
Ic	str+*o*	wk+*o*	wk+*ov*	str+*oj*	wk+*on*	wk+*ost*	wk+*ojn*
Id	str+*å*	wk+*å*	wk+*åv*	str+*åj*	wk+*ån*	wk+*åst*	wk+*åjn*
Ie[16]	str+*e*	wk+*e*	wk+*ev*	str+*áj*	wk+*en*	wk+*est*	wk+*ijn*
II[17]	wk+*aj*	str+*a*	str+*av*	str+*aj*	str+*an*	str+*ast*	str+*ajn*
IIIa	wk+∅	str+*a*	str+*av*	str+*ij*	str+*in*	str+*ist*	str+*ijn*
IIIb	wk+*V*[18]	str+*a*	str+*av*	str+*ij*	str+*in*	str+*ist*	str+*ijn*
CLASS	NOM.PL	GEN.PL	ACC.PL	ILL.PL	INESS.PL	ELAT.PL	COM.PL
Ia	wk+*a*	wk+*aj*	wk+*ajd*	wk+*ajda*	wk+*ajn*	wk+*ajst*	wk+*aj*
Ib	wk+*á*	wk+*áj*	wk+*ájd*	wk+*ájda*	wk+*ájn*	wk+*ájst*	wk+*áj*
Ic	wk+*o*	wk+*oj*	wk+*ojd*	wk+*ojda*	wk+*ojn*	wk+*ojst*	wk+*oj*
Id	wk+*å*	wk+*åj*	wk+*åjd*	wk+*åjda*	wk+*åjn*	wk+*åjst*	wk+*åj*
Ie	wk+*e*	wk+*ij*	wk+*ijd*	wk+*ijda*	wk+*ijn*	wk+*ijst*	wk+*ij*
II	str+*a*	str+*aj*	str+*ajd*	str+*ajda*	str+*ajn*	str+*ajst*	str+*aj*
IIIa	str+*a*	str+*ij*	str+*ijd*	str+*ijda*	str+*ijn*	str+*ijst*	str+*ij*
IIIb	str+*a*	str+*ij*	str+*ijd*	str+*ijda*	str+*ijn*	str+*ijst*	str+*ij*

Table 1: Pite Saami nominal inflection classes (adapted from Wilbur 2014)

[16] Class Ie nouns are also distinguished by “non-adjacent regressive vowel harmony triggered by the presence of /j/ in certain case/number suffixes” (Wilbur 2014:102).

[17] Class II nouns show variation in the suffix vowel, though “there do not appear to be many words in Class II, and the data in the corpus are ultimately inconclusive” (Wilbur 2014:104).

[18] In Class IIIb, nominative singular forms drop a stem-final consonant. For example, compare Class IIIa *vanás* ‘boat’ NOM.SG ~ *vadnás-a* GEN.SG and Class IIIb *bena* ‘dog’ NOM.SG ~ *bednag-a* GEN.SG. In both, the *-n-* ~ *-dn-* alternation follows from general stem grade patterns, but the loss of the final *-g* in *bena* does not (Wilbur 2014:106).

Pite Saami

- If all eight classes are equally likely, then the declension entropy is:

$$
\begin{aligned}
H(D) &= -\sum_{d \in D} \frac{1}{|D|} \log_2 \frac{1}{|D|} \\
&= -\log_2 \frac{1}{|D|} \\
&= 3 \text{ bits}
\end{aligned}
$$

- This is the highest possible value for $H(D)$
- Anything that helps prediction (skewed probabilities, implicational relations, external properties) will reduce $H(D)$

Pite Saami

- Speakers rarely have to generate entire paradigms
- Let $D_{c=r}$ be the set of declensions for which the paradigm cell c has the formal realization r. Then the probability $p_c(r)$ that a paradigm cell c of a particular lexeme has the realization r is the probability of that lexeme belonging to one of the declensions in $D_{c=r}$, or:

$$p_c(r) = \sum_{d \in D_{c=r}} p(d)$$

Pite Saami

- The entropy of $p_c(r)$ is the **paradigm cell entropy** $H(c)$, the uncertainty in the realization for a paradigm cell c

NOM.SG	GEN.SG	ACC.SG	ILL.SG	INESS.SG	ELAT.SG	COM.SG
3.000	2.406	2.406	2.250	2.750	2.750	2.750

NOM.PL	GEN.PL	ACC.PL	ILL.PL	INESS.PL	ELAT.PL	COM.PL
2.406	2.750	2.750	2.750	2.750	2.750	2.750

- Eight declensions, but ill.sg. only has 5 possible forms
- Knowing the ill.sg. leaves 0.75 bits of uncertainty in declension
- Average across all cells is 2.658 bits

Pite Saami

- Guessing either acc. sg or acc. pl is hard, but guessing one knowing the other is easy:

Acc sg	**Acc pl**
wk+*av*	wk+*ajd*
wk+*áv*	wk+*ájd*
wk+*ov*	wk+*ojd*
wk+*åv*	wk+*åjd*
wk+*ev*	wk+*ijd*
str+*av*	str+*ajd*
	str+*ijd*

Pite Saami

- The **conditional entropy** measures the uncertainty left in one thing given that what know something else:

$$H(Y|X) = H(X,Y) - H(X)$$
$$- \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(y|x)$$

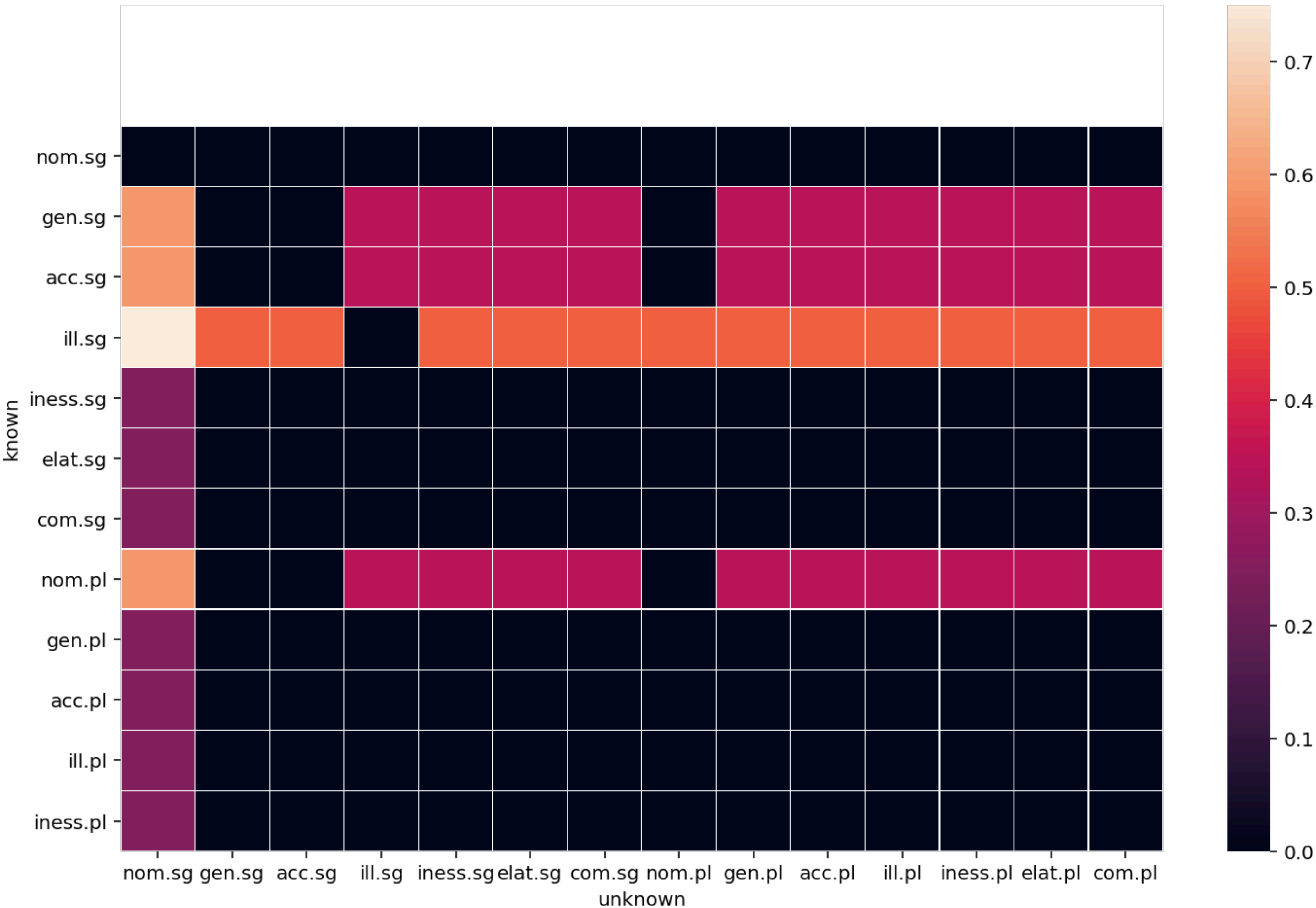
- If we know acc. pl, then we also know acc. sg:

$$H(\text{acc sg}|\text{acc pl}) = 0.0 \text{ bits}$$

- Knowing acc. sg doesn’t quite resolve what acc. sg is:

$$H(\text{acc pl}|\text{acc sg}) = 0.344 \text{ bits}$$

	NOM.SG	GEN.SG	ACC.SG	ILL.SG	INESS.SG	ELAT.SG	COM.SG	NOM.PL	GEN.PL	ACC.PL	ILL.PL	INESS.PL	ELAT.PL	COM.PL
NOM.SG		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
GEN.SG	0.594		0.000	0.344	0.344	0.344	0.344	0.000	0.344	0.344	0.344	0.344	0.344	0.344
ACC.SG	0.594	0.000		0.344	0.344	0.344	0.344	0.000	0.344	0.344	0.344	0.344	0.344	0.344
ILL.SG	0.750	0.500	0.500		0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
INESS.SG	0.250	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ELAT.SG	0.250	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
COM.SG	0.250	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000
NOM.PL	0.594	0.000	0.000	0.344	0.344	0.344	0.344		0.344	0.344	0.344	0.344	0.344	0.344
GEN.PL	0.250	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000
ACC.PL	0.250	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000
ILL.PL	0.250	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000
INESS.PL	0.250	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000
ELAT.PL	0.250	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.000
COM.PL	0.250	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	



nom.sg
gen.sg
acc.sg
ill.sg
iness.sg
elat.sg
com.sg
nom.pl
gen.pl
acc.pl
ill.pl
iness.pl
known
nom.sg gen.sg acc.sg ill.sg iness.sg elat.sg com.sg nom.pl gen.pl acc.pl ill.pl iness.pl elat.pl com.pl
unknown
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

Information Theory

- The conditional entropy of one cell given another is a measure of inter-predictability
- To extend this to the whole paradigm, we calculate the expected conditional entropy

$$E[H(c|c)] = \sum_{c_1,c_2} p(c_1,c_2)H(c_2|c_1)$$

- This is one simple measure of how difficult the PCFP is for a particular language
- The higher the expected conditional entropy, the more difficult it is to predict an unknown wordform, given a known wordform.

Pite Saami

- Row averages measure **predictiveness**

NOM.SG	GEN.SG	ACC.SG	ILL.SG	INESS.SG	ELAT.SG	COM.SG	NOM.PL	GEN.PL	ACC.PL	ILL.PL	INESS.PL	ELAT.PL	COM.PL
0.000	0.311	0.311	0.519	0.019	0.019	0.019	0.311	0.019	0.019	0.019	0.019	0.019	0.019

- Column averages measure **predictability**

NOM.SG	GEN.SG	ACC.SG	ILL.SG	INESS.SG	ELAT.SG	COM.SG	NOM.PL	GEN.PL	ACC.PL	ILL.PL	INESS.PL	ELAT.PL	COM.PL
0.368	0.038	0.038	0.079	0.118	0.118	0.118	0.038	0.118	0.118	0.118	0.118	0.118	0.118

- The overall average is the **paradigm entropy:** 0.166 bits

Paradigm organization

- Paradigms vary a lot in their apparent morphological complexity
- For all these paradigms, the paradigm entropy is much lower than either the expected entropy or the declension entropy

Language	Cells	Realizations	Max realizations	Declensions	Declension entropy	Average entropy	Paradigm entropy
Amele	3	31	14	24	4.585	2.882	1.105
Arapesh	2	41	26	26	4.700	4.071	0.630
Burmeso	12	24	2	2	1.000	1.000	0.000
Fur	12	80	10	19	4.248	2.395	0.517
Greek	8	12	5	8	3.000	1.621	0.644
Kwerba	12	26	4	4	2.000	0.864	0.428
Mazatec	6	356	94	109	6.768	4.920	0.709
Ngiti	16	68	5	10	3.322	1.937	0.484
Nuer	6	12	3	16	4.000	0.864	0.793
Russian	12	26	3	4	2.000	0.911	0.538

Paradigm organization

- Some entropy-lowering strategies:
- Small number of cells, forms, inflection classes
- Paradigm Economy Principle (Carstairs 1984), No Blur Principle (Carstairs-McCarthy 1994, 2010)

Language	Cells	Realizations	Max realizations	Declensions	Declension entropy	Average entropy	Paradigm entropy
Amele	3	31	14	24	4.585	2.882	1.105
Arapesh	2	41	26	26	4.700	4.071	0.630
Burmeso	12	24	2	2	1.000	1.000	0.000
Fur	12	80	10	19	4.248	2.395	0.517
Greek	8	12	5	8	3.000	1.621	0.644
Kwerba	12	26	4	4	2.000	0.864	0.428
Mazatec	6	356	94	109	6.768	4.920	0.709
Ngiti	16	68	5	10	3.322	1.937	0.484
Nuer	6	12	3	16	4.000	0.864	0.793
Russian	12	26	3	4	2.000	0.911	0.538

Paradigm organization

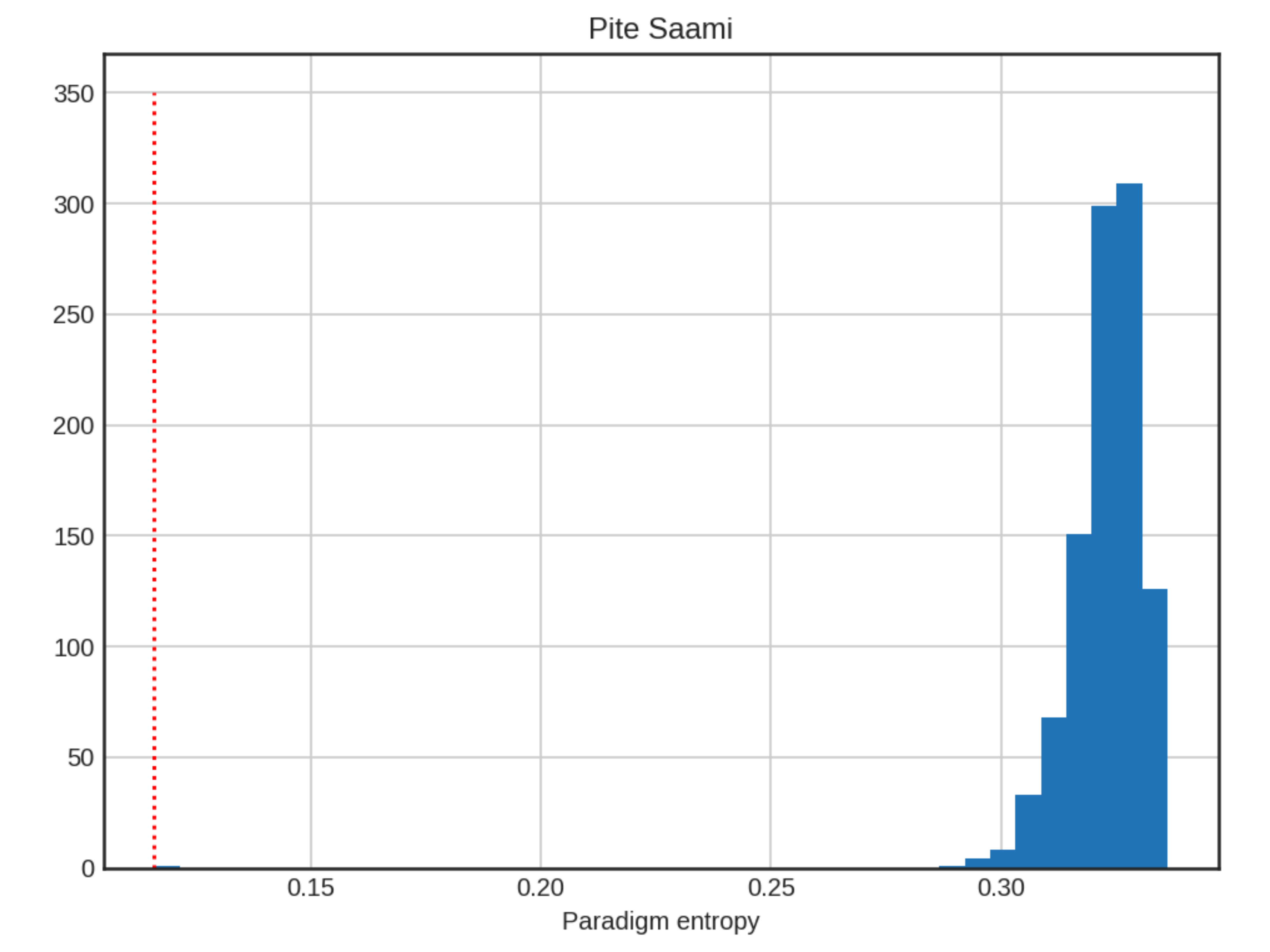
- Some entropy-lowering strategies:
- Implicational relations (Wurzel 1989)
- Principal parts (Stump & Finkel 2007)

Language	Cells	Realizations	Max realizations	Declensions	Declension entropy	Average entropy	Paradigm entropy
Amele	3	31	14	24	4.585	2.882	1.105
Arapesh	2	41	26	26	4.700	4.071	0.630
Burmeso	12	24	2	2	1.000	1.000	0.000
Fur	12	80	10	19	4.248	2.395	0.517
Greek	8	12	5	8	3.000	1.621	0.644
Kwerba	12	26	4	4	2.000	0.864	0.428
Mazatec	6	356	94	109	6.768	4.920	0.709
Ngiti	16	68	5	10	3.322	1.937	0.484
Nuer	6	12	3	16	4.000	0.864	0.793
Russian	12	26	3	4	2.000	0.911	0.538

Testing entropy: Simulations

- The implicational structure of the paradigms is crucial to reducing paradigm entropy
- How can we test this?
  - Null hypothesis: Paradigm entropy of language *L* is independent of paradigm organization
  - If this is true, then L_0, a version *L* with the same forms and the same classes but a different organization, should have more or less the same paradigm entropy
  - Bootstrap test: sample with replacement from the space of possible L_0's, and compare to the observed *L*

class	nom.sg	gen.sg	acc.sg	ill.sg	iness.sg	elat.sg	com.sg	nom.pl	gen.pl	acc.pl	ill.pl	iness.pl	elat.pl	com.pl
Ia	wk+0	str+a	str+av	str+ij	wk+an	wk+est	wk+ajn	str+a	str+ij	wk+ájd	str+ijda	str+ijn	wk+ojst	str+ij
Ib	wk+aj	wk+á	wk+av	str+ij	str+in	str+ast	wk+ojn	str+a	wk+ij	wk+ijd	wk+ajda	wk+ájn	wk+ajst	str+ij
Ic	str+o	wk+å	wk+ev	str+aj	wk+án	wk+åst	wk+åjn	wk+e	wk+aj	wk+ajd	str+ijda	wk+åjn	str+ijst	wk+áj
Id	wk+V	wk+e	wk+åv	str+aj	wk+ån	wk+ást	str+ajn	wk+å	wk+åj	str+ajd	str+ajda	str+ijn	wk+ájst	str+aj
Ie	str+e	wk+o	str+av	str+áj	str+in	wk+ast	wk+ijn	str+a	wk+áj	str+ijd	wk+ájda	wk+ajn	wk+åjst	wk+oj
II	str+a	str+a	str+av	str+oj	str+an	wk+ost	str+ijn	wk+a	str+aj	str+ijd	wk+ojda	wk+ijn	wk+ijst	wk+aj
IIIa	str+á	str+a	wk+áv	str+åj	wk+on	str+ist	str+ijn	wk+o	wk+oj	wk+åjd	wk+åjda	wk+ojn	str+ijst	wk+ij
IIIb	str+å	wk+a	wk+ov	str+áj	wk+en	str+ist	wk+ájn	wk+á	str+ij	wk+ojd	wk+ijda	str+ajn	str+ajst	wk+åj



Pite Saami
350
300
250
200
150
100
50
0
0.15
0.20
0.25
0.30
Paradigm entropy

Language	Cells	Realizations	Declensions	Declension entropy	Average entropy	Paradigm entropy	Bootstap Avg	Bootstrap *p*
Amele	3	31	24	4.585	2.882	1.105	1.327	0.001
Arapesh	2	41	26	4.700	4.071	0.630	0.630	1.000
Burmeso	12	24	2	1.000	1.000	0.000	0.000	1.000
Fur	12	80	19	4.248	2.395	0.517	1.316	0.001
Greek	8	12	8	3.000	1.621	0.644	0.891	0.001
Kwerba	12	26	4	2.000	0.864	0.428	0.523	0.001
Mazatec	6	356	109	6.768	4.920	0.709	1.100	0.001
Ngiti	16	68	10	3.322	1.937	0.484	1.019	0.001
Nuer	6	12	16	4.000	0.864	0.793	0.811	0.160
Russian	12	26	4	2.000	0.911	0.538	0.541	0.383

Limitations

- Ackerman & Malouf's (2013) entropy estimates made a number of (over-)simplifying assumptions
  - always predicting one cell on the basis of one other cell
  - all cells are equally likely to be known
  - all cells are equally likely to be unknown
  - speakers know all possible full paradigms
  - speakers can always identify which paradigm cell a wordform fills
  - speakers can always identify which allomorph a wordform represents

Limitations

- Current work (e.g, Bonami and Boyé 2014, Bonami and Beniamine 2016, Sims and Parker 2019, Cotterell et al. 2019) addresses these concerns
    - Derives patterns from lexicons or corpora rather than grammatical descriptions,
    - using linguistically plausible methods for learning patterns,
    - taking actual distributions of frequencies into account.

Prospects

- Recall Humboldt's modes of explanation

  A language is the way it is because of:

  1. universal cognitive or communicative constraints (I-complexity)
  2. historical accident (E-complexity)
  3. the inner spirit of a nation (we'll come back to this in week 4)