Beyond Words

Lower and upper bounds on the entropy of subword units in diverse languages

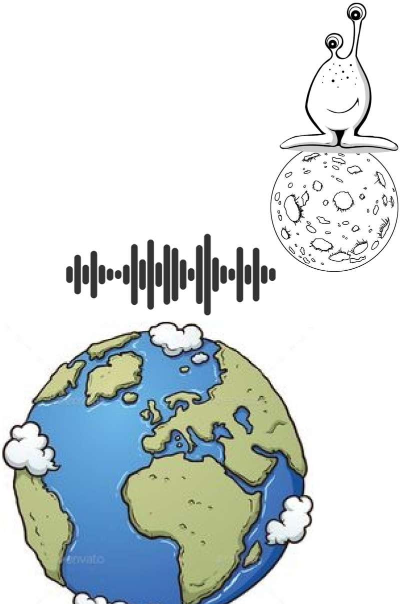
Christian Bentz

August 8, 2023

Department of General Linguistics, University of Tübingen

Introduction

The Marsian Scientist



If a Martian scientist [...] received from Earth the broadcast of an extensive speech [...] what criteria would [...] determine whether the reception represented the effect of an animate process on Earth, or merely the latest thunderstorm on Earth? It seems that the only criteria would be the ***arrangement of occurrences of the elements*** *[...]: the arrangement of the occurrences would be neither of* ***rigidly fixed regularity*** *[...] nor yet a completely* ***random scattering*** *of the same.*

Zipf (1936). The psycho-biology of language, p. 187.

System	Example*
English	And they remembered his words ,
Greek (Modern)	Και ενεθυμηθησαν τους λογους αυτου .
Hindi	तब उस की बातें उन को स्‌मरण आईं ।
Georgian	და მოეჴსენნეს სიტყუანი მისნი .
Korean	저희가 예수의 말씀을 기억하고
Burmese	မိန့်တော်မူခဲ့သောစကားများကိုပြန်သတိရ ၍ ၊ –
Russian	И они вспомнили эти слова Его .
Inuktitut	ᐊᕐᙲᓪᓗ ᐃᖅᑮᓕᖅᐳᑦ ᐅᖃᐅᓯᖏᓐᓂᒃ ,
Kalaallisut	Taava oqaasii eqqaalerpaat .
Bird Song	uj kd ro su sv sw sx gf jr dw kd tc jt ag ta
Morse	phh pppp p hp s pp hp s h pppp p s hphp hhh
DNA	GGTAGTTAGGGTCTGAAAAAGATTTTGCG
Weather	SWCCSSSSSSSSSCSOFSPPPFPPFPP
Random	hihhe bh fif cd gbgdiiigc ghigbbg af icegeebiifg
Shuffled	swr a j e eitimii hfeooa ti i d qs sfi roeviebg ep

*For natural languages: Verse number 42024008 of the New Testament.

Methods: Unigram Entropy

Definition

$$H(X) = -\sum_{i=1}^{V} p(x_i)\log_2 p(x_i)$$

- V: number of word types,
- $p(x_i)$: probability of word type.

Example

in$_1$ the$_2$ beginning$_3$ god$_4$ created$_5$ the$_6$ heavens$_7$ and$_8$ the$_9$ earth$_{10}$ and$_{11}$ the$_{12}$ earth$_{13}$ was$_{14}$ waste$_{15}$ and$_{16}$ empty$_{17}$ [...]

unit	freq
the	4
and	3
earth	2
in	1
beginning	1
god	1
...	...

$$\widehat{H}^{ML}(X) = -(\tfrac{4}{17}\log_2(\tfrac{4}{17}) + \tfrac{3}{17}\log_2(\tfrac{3}{17}) + \cdots + \tfrac{1}{17}\log_2(\tfrac{1}{17})) \sim 3.2$$

Shannon, Claude E. (1948). A mathematical theory of communication.

Cover & Thomas (2006). Elements of information theory, p. 14.

Methods: Entropy Rate

Definition

$$\hat{h}(\mathcal{X}) = \frac{1}{n}\sum_{i=2}^{n}\frac{\log_2 i}{L_i},$$

- n: number of word tokens,
- L_i: length (+1) of the longest contiguous substring starting at position i which is also present in $i = 2$ to $i - 1$.

Example

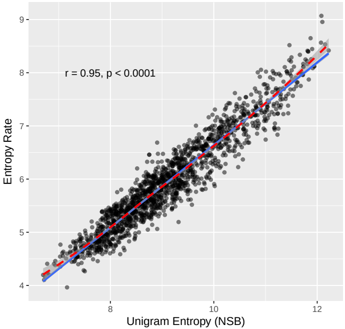
in$_1$ the$_2$ beginning$_3$ god$_4$ created$_5$ the$_6$ heavens$_7$ and$_8$ the$_9$ earth$_{10}$ and$_{11}$ the$_{12}$ earth$_{13}$ was$_{14}$ waste$_{15}$ and$_{16}$ empty$_{17}$ [...]

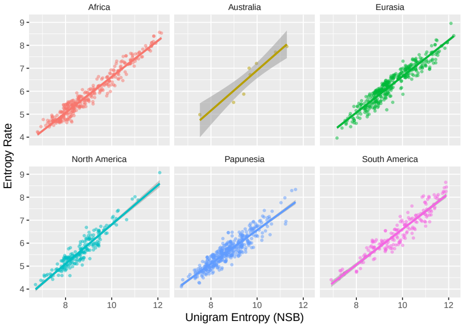
$$L_{11} = 3(+1) = 4$$

$$\frac{\log_2(11)}{4} \sim \frac{3.46}{4} \sim 0.87$$

Gao, Kontoyiannis & Bienenstock (2008). Estimating the entropy of binary time series, equation (6).

Entropy Across Languages of the World





Bentz, Alikaniotis, Cysouw, & Ferrer-i-Cancho (2017). The entropy of words - learnability and expressivity across more than 1000 languages.

Lavi-Rotbain & Arnon (2019). Children learn words better in low entropy.

Tal, Grossman, & Arnon (2022). Infant-directed speech becomes less redundant as infants grow: implications for language learning.

Lavi-Rotbain & Arnon (2023). Zipfian distributions in child-directed speech.

Linguistic Interpretation

(1) Hawaiian (haw, PBC 41006018)

A	*ua*	*olelo*	*aku*	*o*	*Ioane*	*ia ia*	[...]
Then	PERF	say	to	SUBJ	Johan	he.DAT	[...]

“Then Johan said to him [...]”

(2) Turkish (tur, PBC 41006004)

Ýsa	*da*	*on-lar-a*	[...]	*de-di*
Jesus	also	3P-PL-DAT	[...]	say-3SG.PERF

“Jesus also said to them [...]”

(3) Iñupiatun (esk, PBC 41006004)

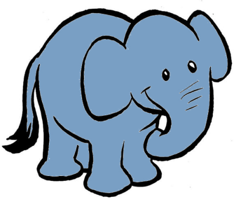
Aglaan	*Jesus-ŋum*	*itna-ġ-ni-ġai*	[...]
But	Jesus-ERG	this-say-report-3S.to.3PL	

“But Jesus said to them (it is reported) [...]”

Bentz (2018). Adaptive languages: An information-theoretic account of language diversity.

Three Elephants in the Room

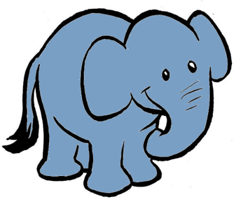
Word Elephant



What if we use other units of measurement?

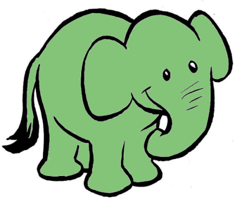
Three Elephants in the Room

Word Elephant



What if we use other units of measurement?

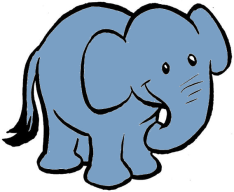
Spoken Elephant



What about spoken language?

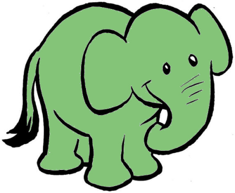
Three Elephants in the Room

Word Elephant



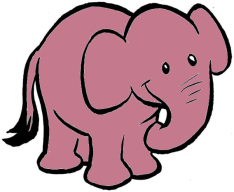
What if we use other units of measurement?

Spoken Elephant



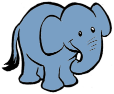
What about spoken language?

Meaning Elephant

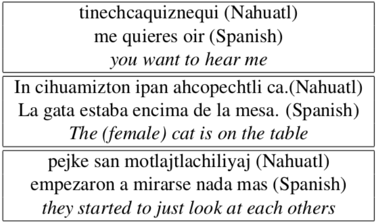


How does this relate to meaning anyways?

Subword Units

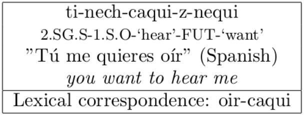


Subword Entropy



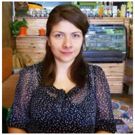
tinechcaquiznequi (Nahuatl) me quieres oir (Spanish) *you want to hear me*
In cihuamizton ipan ahcopechtli ca.(Nahuatl) La gata estaba encima de la mesa. (Spanish) *The (female) cat is on the table*
pejke san motlajtlachiliyaj (Nahuatl) empezaron a mirarse nada mas (Spanish) *they started to just look at each others*

Table 1: Examples of Nahuatl-Spanish parallel sentences



ti-nech-caqui-z-nequi 2.SG.S-1.S.O-'hear'-FUT-'want' "Tú me quieres oír" (Spanish) *you want to hear me*
Lexical correspondence: oir-caqui

Table 2: Example of Nahuatl-Spanish

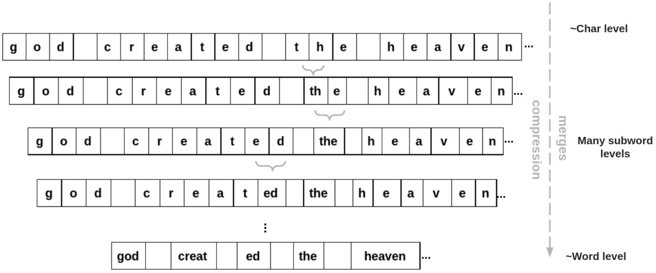


Ximena
Gutierrez-Vasques

Gutierrez-Vasques, Sierra, & Pompa (2016). Axolotl: A web accessible parallel corpus for Spanish-Nahuatl.

Ximena Gutierrez-Vasques & Victor Mijangos. (2018). Comparing morphological complexity of Spanish, Otomi and Nahuatl.

Subwords of Byte Pair Encoding (BPE)



~Char level

g o d c r e a t e d t h e h e a v e n …

g o d c r e a t e d th e h e a v e n …

g o d c r e a t e d the h e a v e n …

g o d c r e a t ed the h e a v e n …

⋮

god creat ed the heaven …

compression

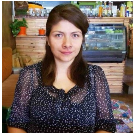
merges

Many subword levels

~Word level

Gutierrez-Vasques, Bentz, Sozinova, & Samardžić (2021). From characters to words: the turning point of BPE merges. *EACL*.

Gutierrez-Vasques, Bentz, & Samardžić (2023). Languages through the Looking Glass of BPE Compression. *Computational Linguistics*.

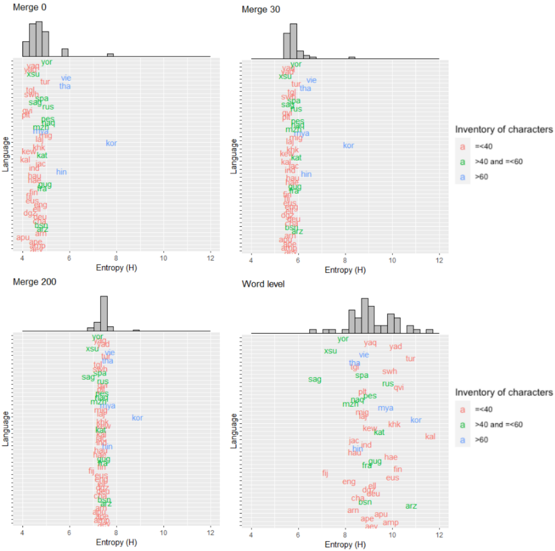


Ximena Gutierrez-Vasques

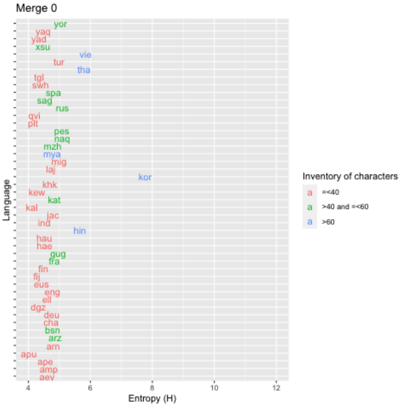
Olga Pelloni (Sozinova)

Tanja Samardžić

From Characters to Words

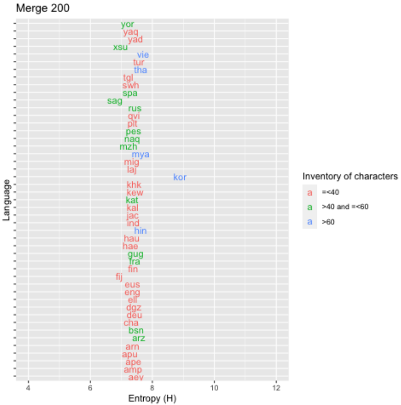


Character Entropy



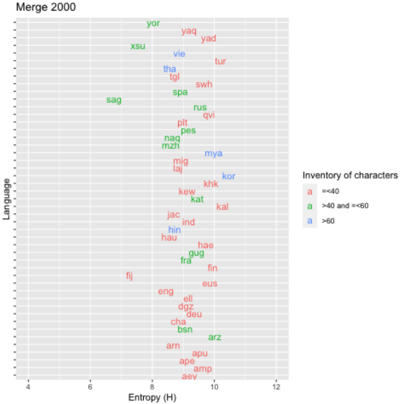
System	Example*
English	And they remembered his words ,
Greek (Modern)	Και ενεθυμηθησαν τους λογους αυτου .
Hindi	तब उस की बातें उन को स्मरण आईं ।
Georgian	და მოეჴსენნეს სიტყუანი მისნი .
Korean	저희가 예수의 말씀을 기억하고
Burmese	မိန့်တော်မူခဲ့သောစကားများကိုပြန်သတိရ ၍ ၊ –
Russian	И они вспомнили эти слова Его .
Inuktitut	ᐊᕐᖅᓃᓪᓗ ᐃᖅᑮᓕᖅᐳᑦ ᐅᖃᐅᓯᖏᓐᓂᒃ ,
Kalaallisut	Taava oqaasii eqqaalerpaat .

Subword Entropy (200 Merges)



English (eng)				Turkish (tur)			
prod.	subword	merge	examples	prod.	subword	merge	examples
110	ing<\w>	29	beginn**ing**	355	lar	8	on**lar**a
67	eth<\w>	103	nazar**eth**, eat**eth**	309	ler	13	gün**ler**de
38	est<\w>	166	l**est**, car**est**	150	ler<\w>	32	gitti**ler**
26	led<\w>	122	fil**led**, cal**led**	145	lar<\w>	38	adam**lar**
26	com	137	**com**ing, **com**e	131	den<\w>	40	sen**den**
23	oun	129	r**oun**d, f**oun**d	129	yor	69	öğreti**yor**d
21	for	86	**for**sook	116	dan<\w>	60	tarafın**dan**
21	ent<\w>	91	w**ent**, garm**ent**	110	ini<\w>	64	indiğ**ini**
21	ght<\w>	102	tau**ght**, mi**ght**	107	ların	96	ağ**ların**ı
19	ing	90	th**ing**s, br**ing**ing	80	ine<\w>	50	üzer**ine**

Word Entropy (2000 Merges)



Sanumá (xsu)

(4) Sama töpö se kite
1PL.EXCL 3PL hit FUT

"We will hit them." Borgman (1990)

Georgian (kat)

(5) და მოექსენნეს
da mo-e-qsenn-es
and PREV-3P.PL-mention-3P.PL.AOR
სიტყუანი მისნი
sit'q'va-n-i misni
word-PL-NOM 3P.POSS.NOM.SG

Literal translation: "And they mentioned his words."
English verse: "And they remembered his words."

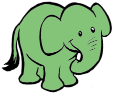
Corpus-Based Typology

ISO	Name	Prod.	Synthesis	Reference	Fusion
vie	Vietnamese	-1.33	analytic	Haspelmath 2010	isolating
tha	Thai	-1.33	analytic	Moravcsik 2012	isolating/concat.
sag	Sango	-1.29	analytic	Karan 2006	concatenative
yor	Yoruba	-1.21	analytic	Haspelmath 2010	tonal/isolating
eng	English	-0.94	analytic	Haspelmath 2010	concatenative
fij	Fijian	-0.89	analytic	Dixon 1988	isolating
pes	Persian/Farsi	-0.78	synthetic	Greenberg 1960	concatenative
fra	French	-0.19	synthetic	Dixon 2003	concatenative
ell	Greek (Modern)	0.02	synthetic	Dixon 2003	concatenative
rus	Russian	0.07	synthetic	Aikhenvald 2007	concatenative
swh	Swahili	0.2	synthetic	Haspelmath 2010	concatenative
yaq	Yaqui	0.46	synthetic	Guerrero 2019	concatenative
tur	Turkish	0.54	synthetic	Bickel 2007	concatenative
gug	Paraguayan Guaraní	-0.19	polysynthetic	Aikhenvald 2017	concatenative
arn	Mapudungun	0.73	polysynthetic	Bickel 2017	concatenative
amp	Alamblak	1.02	polysynthetic	Bruce 1984	concatenative
apu	Apurinã	1.3	polysynthetic	Facundes 2014	concatenative
bsn	Barasano	1.43	polysynthetic	Gomez 2004	concatenative
kal	Kalaallisut	3.25	polysynthetic	Haspelmath 2010	concatenative

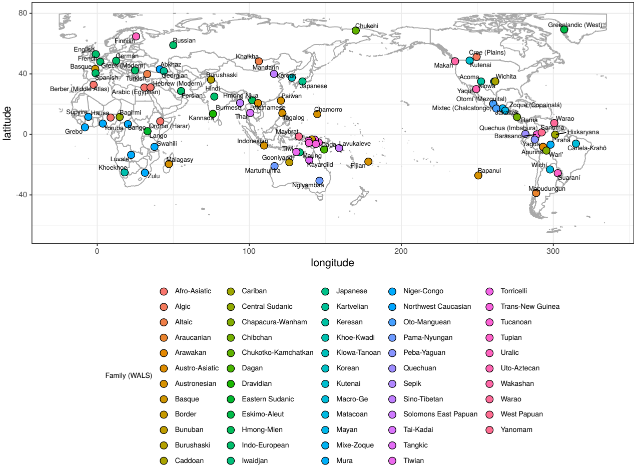
Summary

Languages have relatively high entropy divergence at the **character level** (c. 4-8 bits per character), and at the **word level** (c. 7 to 11 bits per word), but they have similar entropies at specific **subword levels** (c. 7-8 bits per subword).

Spoken vs. Written

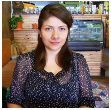


Text Data Diversity Sample (TeDDi)



Tanja Samardžić

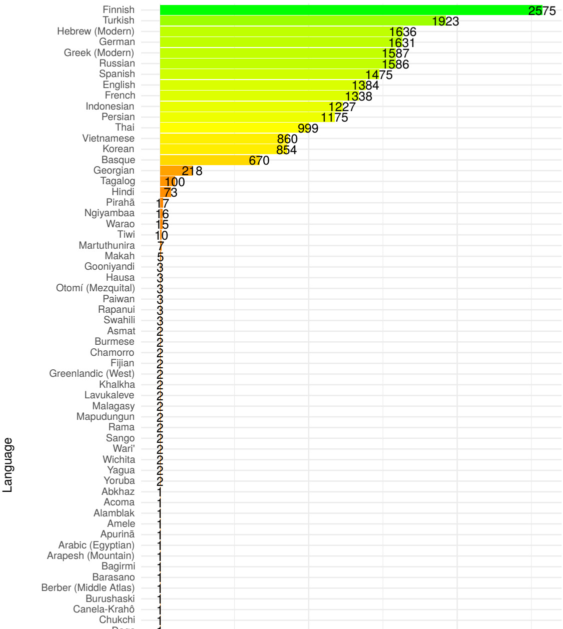
Olga Pelloni (Sozinova)


Ximena Gutierrez-Vasques

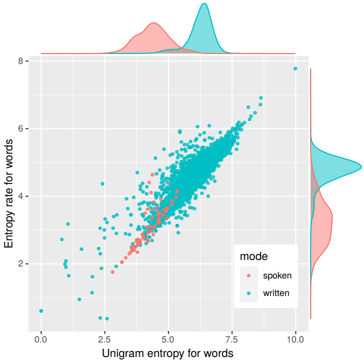
Steven Moran

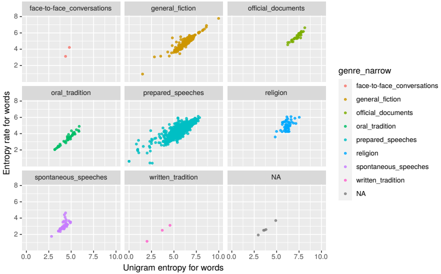
Moran, Bentz, Gutierrez-Vasques, Sozinova, & Samardžić (2022). TeDDi Sample: Text Data Diversity Sample for Language Comparison and Multilingual NLP.

TeDDi Overview



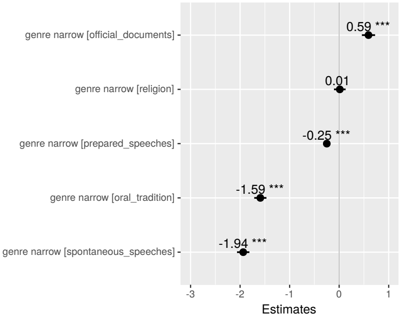
Entropy for Spoken vs. Written



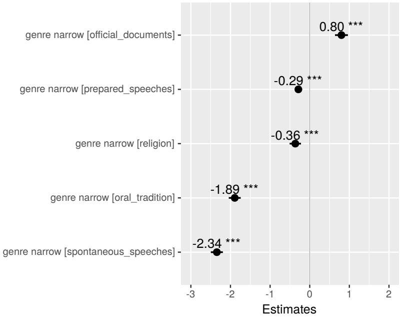


Multiple Regression Model

Word Entropy Rate

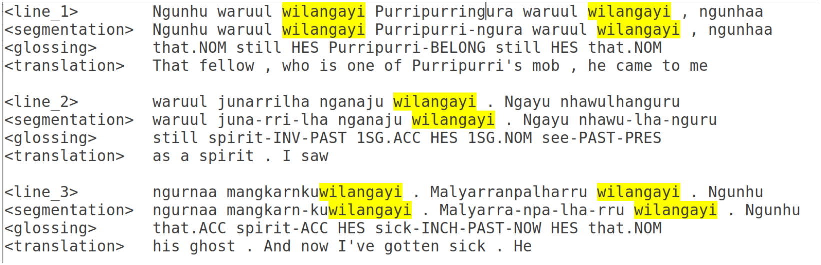


Unigram Word Entropy



Examples

Martuthunira (vma, spoken, “Mourning chant”)

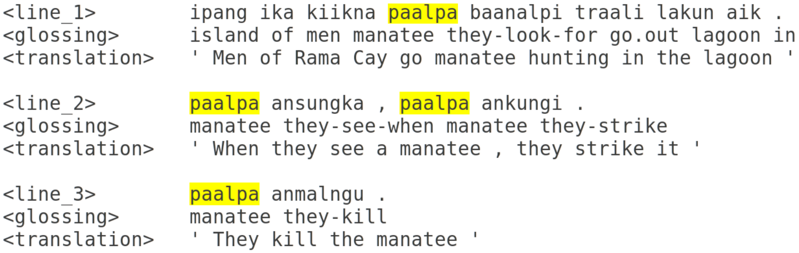


```
<line_1>        Ngunhu waruul wilangayi Purripurringura waruul wilangayi , ngunhaa
<segmentation>  Ngunhu waruul wilangayi Purripurri-ngura waruul wilangayi , ngunhaa
<glossing>      that.NOM still HES Purripurri-BELONG still HES that.NOM
<translation>   That fellow , who is one of Purripurri's mob , he came to me

<line_2>        waruul junarrilha nganaju wilangayi . Ngayu nhawulhanguru
<segmentation>  waruul juna-rri-lha nganaju wilangayi . Ngayu nhawu-lha-nguru
<glossing>      still spirit-INV-PAST 1SG.ACC HES 1SG.NOM see-PAST-PRES
<translation>   as a spirit . I saw

<line_3>        ngurnaa mangkarnkuwilangayi . Malyarranpalharru wilangayi . Ngunhu
<segmentation>  ngurnaa mangkarn-kuwilangayi . Malyarra-npa-lha-rru wilangayi . Ngunhu
<glossing>      that.ACC spirit-ACC HES sick-INCH-PAST-NOW HES that.NOM
<translation>   his ghost . And now I've gotten sick . He
```

Dench, A. C. (1994). Martuthunira: A language of the Pilbara region of Western Australia, p. 282-287.

Examples

Rama (rma, spoken, “Manatee hunting”)



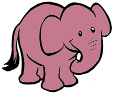
```
<line_1>        ipang ika kiikna paalpa baanalpi traali lakun aik .
<glossing>      island of men manatee they-look-for go.out lagoon in
<translation>   ' Men of Rama Cay go manatee hunting in the lagoon '

<line_2>        paalpa ansungka , paalpa ankungi .
<glossing>      manatee they-see-when manatee they-strike
<translation>   ' When they see a manatee , they strike it '

<line_3>        paalpa anmalngu .
<glossing>      manatee they-kill
<translation>   ' They kill the manatee '
```

Craig, C. (1986). The Rama language; a text with grammatical notes. Estudios de Lingüística Chibcha 5. 21-44.

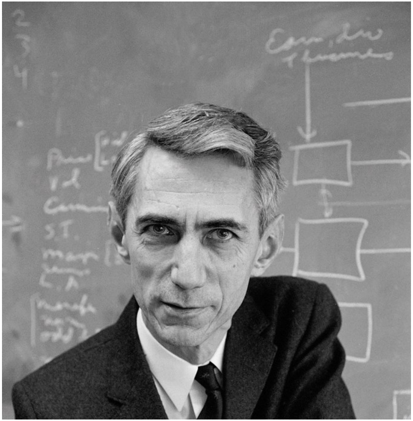
Preliminary Summary

Texts of spoken registers appear to have lower entropies on average than texts of written registers. This is likely related to more repetitive usage of words and word chunks (as a result of online memory constraints?).

Meaning



Information and Meaning



*The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. [...] **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.*

Shannon, Claude E. (1948). A mathematical theory of communication, p. 1.

Example

Article 1

All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

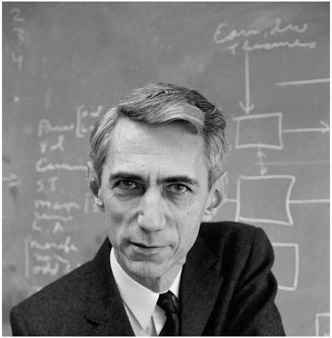
Universal Declaration of Human Rights (UDHR) in English

Raeiclt 1

Rll humrn btings rat boan fatt and tqurl in digniey rnd aighes. Ehty rat tndowtd wieh atrson rnd conscitnct rnd should rce eowrads ont rnoehta in r spiaie of baoehtahood.

Universal Declaration of Human Rights (UDHR) in ???

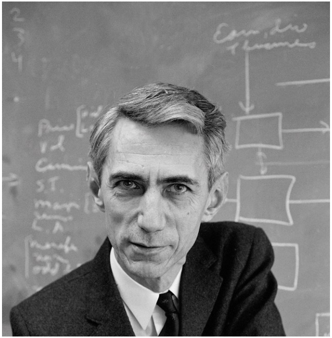
Information and Meaning



[...] two messages, one of which is heavily loaded with meaning and the other which is pure nonsense, can be exactly equivalent, from the present viewpoint, as regards information. It is this, undoubtedly, that Shannon means when he says that "the semantic aspects of communication are irrelevant to the engineering aspects." ***But this does not mean that the engineering aspects are necessarily irrelevant to the semantic aspects****.*

Shannon & Weaver (1949). The mathematical theory of communication, p. 8.

Three Levels of Communication Problems



- **Level A**: How accurately can the symbols of communication be transmitted? (The technical problem.)
- **Level B**: How precisely do the transmitted symbols convey the desired meaning? (The semantic problem.)
- **Level C**: How effectively does the received meaning affect conduct in the desired way? (The effectiveness problem.)

Shannon & Weaver (1949). The mathematical theory of communication, p. 4.

Entropy and Mutual Information

The entropy of signals is an upper bound on the mutual information between signals and meanings, i.e.

$$H(S) \geq I(S, R) \tag{1}$$

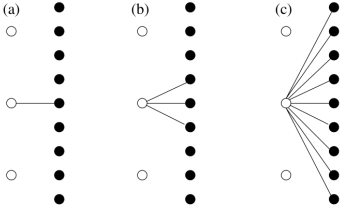
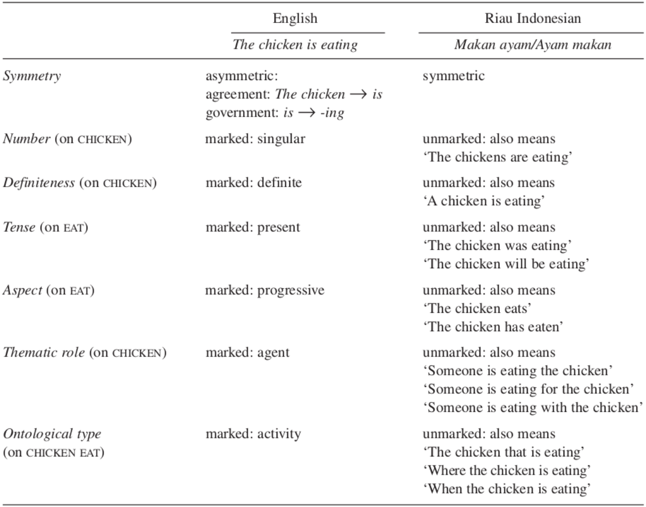


Figure 1. Some mappings between signals (white circles) and stimuli (black circles) that are minima of $H(S)$ and $H(S|R)$ with $n = 3$ signals and $m = 9$ stimuli. (a)–(c) are minima of model A while (c) is the only valid minima of model B.

Ferrer-i-Cancho & Diaz-Guilera (2007). The global minima of the communicative energy of natural communication systems.

Riau Indonesian and English

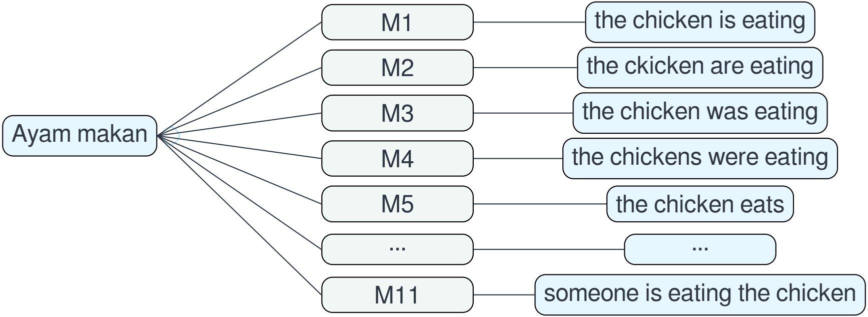


	English	Riau Indonesian
	The chicken is eating	*Makan ayam/Ayam makan*
Symmetry	asymmetric: agreement: *The chicken* → *is* government: *is* → *-ing*	symmetric
Number (on CHICKEN)	marked: singular	unmarked: also means 'The chickens are eating'
Definiteness (on CHICKEN)	marked: definite	unmarked: also means 'A chicken is eating'
Tense (on EAT)	marked: present	unmarked: also means 'The chicken was eating' 'The chicken will be eating'
Aspect (on EAT)	marked: progressive	unmarked: also means 'The chicken eats' 'The chicken has eaten'
Thematic role (on CHICKEN)	marked: agent	unmarked: also means 'Someone is eating the chicken' 'Someone is eating for the chicken' 'Someone is eating with the chicken'
Ontological type (on CHICKEN EAT)	marked: activity	unmarked: also means 'The chicken that is eating' 'Where the chicken is eating' 'When the chicken is eating'

Gil (2005). Isolating-Monocategorial-Associational language.

Gil (2008). How much grammar does it take to sail a boat?

Form-Meaning Mapping



Entropy of the **Signals**:

$H(S_{ind}) = -\log_2(1) = 0$ bits/chunk
$H(S_{eng}) = -\log_2(11) \sim 3.46$ bits/chunk

Conditional Entropy of **Meanings given the Signals**:

$H(M|S_{ind}) = -\sum_{s\in\mathcal{S}_{ind}} p(s) \sum_{m\in\mathcal{M}} p(m|s)\log_2 p(m|s) = 3.46$ bits/chunk
$H(M|S_{eng}) = -\sum_{s\in\mathcal{S}_{eng}} p(s) \sum_{m\in\mathcal{M}} p(m|s)\log_2 p(m|s) = 0$ bits/chunk

Preliminary Summary

- The entropy is a **necessary but not sufficient condition** for meaningful communication.
- Languages certainly differ in entropy on the **signal side**.
- However, entropy in encoding and decoding messages is symmetrical by definition. If this holds true, then the **overall entropy is necessarily the same across languages**.

01010100 01101000 01100001
01101110 01101011 00100000
01111001 01101111 01110101

Thank You