Zipf's Law of Abbreviation as an Absolute Linguistic Universal

Christian Bentz
University of Tübingen
University of Cambridge

Ramon Ferrer-i-Cancho
Universitat Politècnica de Catalunya

October 28, 2015

THE LAW

- **Zipf's law of abbreviation**
  Words that are **frequent** tend to be **short** (Zipf 1932, 1935, 1949).

THE LAW

- **Zipf's law of abbreviation**
  Words that are **frequent** tend to be **short** (Zipf 1932, 1935, 1949).
- **Examples**
  the, *and*, *of*, *a* versus *harpsichord*, *ocelot*, *flabbergasted*

THE LAW

- **Zipf's law of abbreviation**
  Words that are **frequent** tend to be **short** (Zipf 1932, 1935, 1949).
- **Examples**
  the, *and*, *of*, *a* versus *harpsichord*, *ocelot*, *flabbergasted*
- **Not to be confused** with **Zipf's law**, i.e. inverse relationship of word ranks and frequencies

EARLIER STUDIES

- **Random typing**
  Miller (1957); Li (1992); Leopold (1998); Conrad & Mitzenmacher (2004); Ferrer-i-Cancho & Elvevåg (2009); Manin (2009); Ferrer-i-Cancho, Bentz & Seguin (2015)
- **Information theory**
  Piantadosi, Tily & Gibson (2011); Mahowald, Fedorenko, Piantadosi & Gibson (2013), Ferrer-i-Cancho, Bentz & Seguin (2015)
- **Animal behaviour**
  Ferrer-i-Cancho & Lusseau (2009); Bezerra, Souto, Radford & Jones (2011); Ferrer-i-Cancho, Hernández-Fernández, Lusseau, Agoramoorthy, Hsu & Semple (2013); Luo, Jiang, Liu, Wang, Lin, Wei & Feng (2013)

QUESTION

- Is the law a universal of human languages?

DATA AND METHODS

Parallel Corpora

Table : Information about parallel corpora used.

Corpus	Register	Size*	Size ∅*	Texts	Lang.
UDHR[1]	Legal	ca. 650K	1.831	356	333
PBC[2]	Religious	ca. 8M	261K	907	801
	Total	**ca. 9M**		**1263**	**986**

*in number of tokens

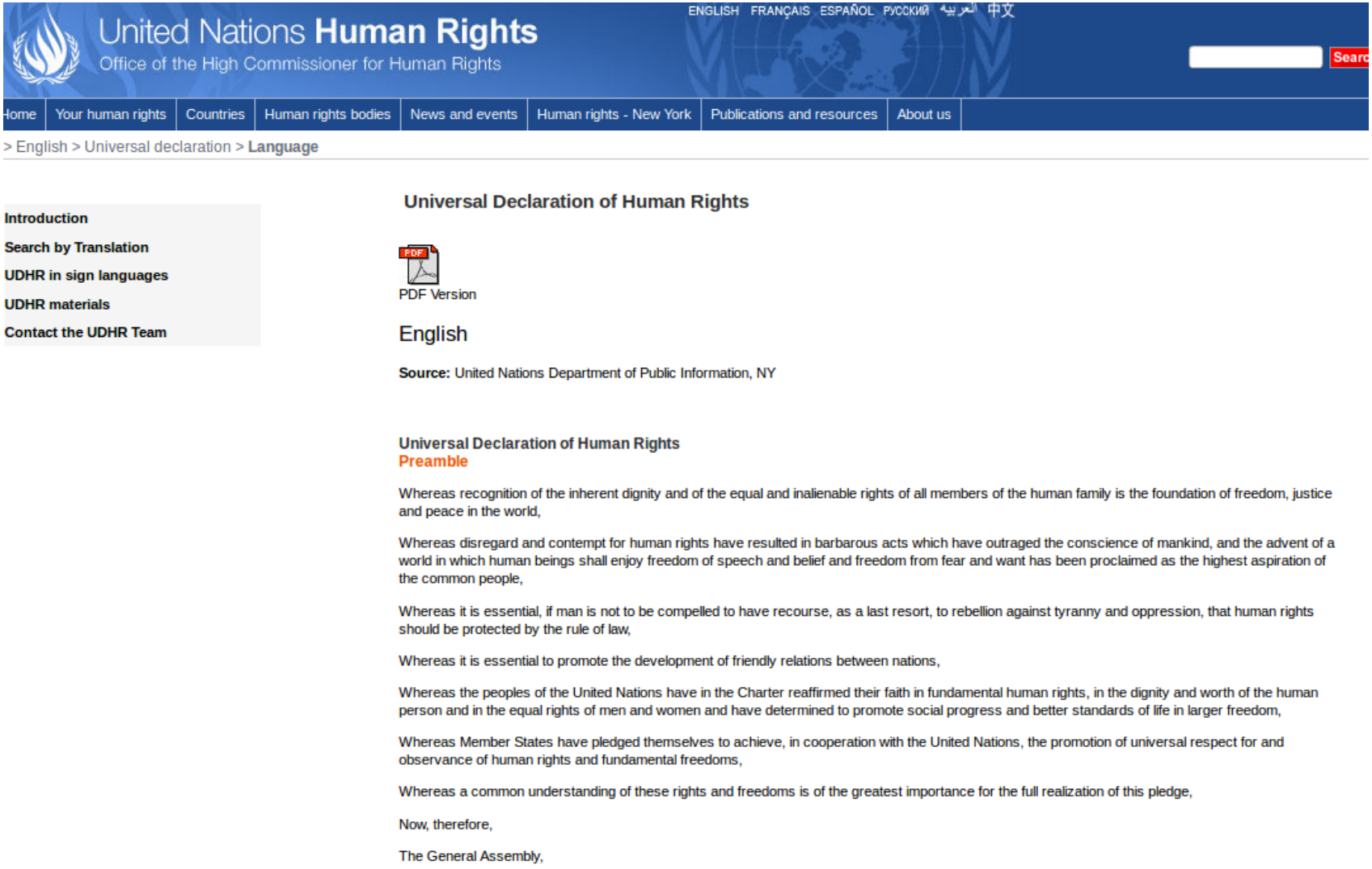
[1] *Universal Declaration of Human Rights* (http://unicode.org/udhr/translations.html)

[2] *Parallel Bible Corpus* (Mayer & Cysouw, 2014)

PARALLEL CORPORA

- *Ethnologue (17th version)*: **7555 languages**
  Our sample: **986 languages**
  $\rightarrow$ **13.05%**

WORD FREQUENCIES AND LENGTHS



United Nations **Human Rights**
Office of the High Commissioner for Human Rights

ENGLISH FRANÇAIS ESPAÑOL РУССКИЙ العربية 中文

Search

Home | Your human rights | Countries | Human rights bodies | News and events | Human rights - New York | Publications and resources | About us

> English > Universal declaration > **Language**

Introduction
Search by Translation
UDHR in sign languages
UDHR materials
Contact the UDHR Team

Universal Declaration of Human Rights

PDF Version

English

Source: United Nations Department of Public Information, NY

Universal Declaration of Human Rights
Preamble

Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world,

Whereas disregard and contempt for human rights have resulted in barbarous acts which have outraged the conscience of mankind, and the advent of a world in which human beings shall enjoy freedom of speech and belief and freedom from fear and want has been proclaimed as the highest aspiration of the common people,

Whereas it is essential, if man is not to be compelled to have recourse, as a last resort, to rebellion against tyranny and oppression, that human rights should be protected by the rule of law,

Whereas it is essential to promote the development of friendly relations between nations,

Whereas the peoples of the United Nations have in the Charter reaffirmed their faith in fundamental human rights, in the dignity and worth of the human person and in the equal rights of men and women and have determined to promote social progress and better standards of life in larger freedom,

Whereas Member States have pledged themselves to achieve, in cooperation with the United Nations, the promotion of universal respect for and observance of human rights and fundamental freedoms,

Whereas a common understanding of these rights and freedoms is of the greatest importance for the full realization of this pledge,

Now, therefore,

The General Assembly,

WORD FREQUENCIES AND LENGTHS

Token frequencies: Split text strings on non-alphanumeric characters and count the frequencies of *word types*.

Rank	Word	Frequency
1	the	121
2	and	106
3	of	91
4	to	83
5	in	43
6	right	33
7	be	31
8	article	30
9	everyone	30
...	...	...

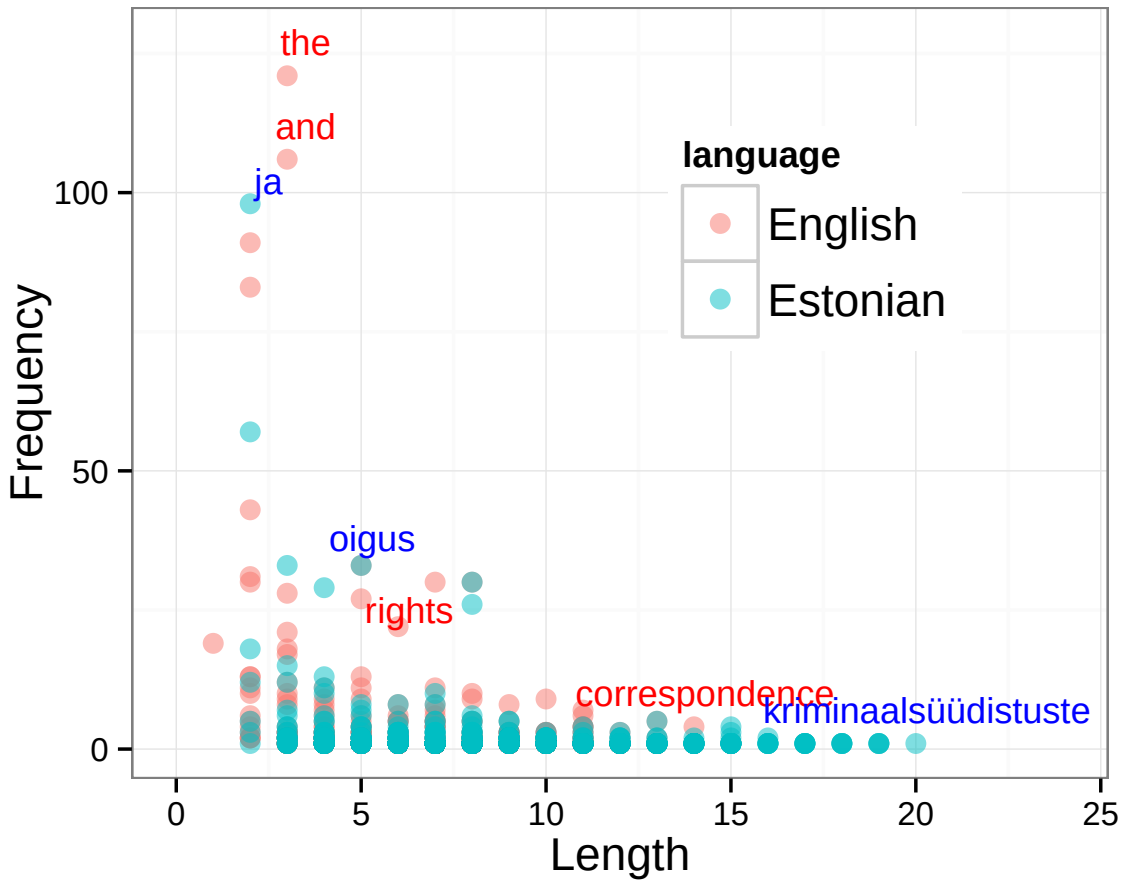
WORD FREQUENCIES AND LENGTHS

Word lengths: Count *unicode characters* per word type.

Rank	Word	Frequency	Length
1	the	121	3
2	and	106	3
3	of	91	2
4	to	83	2
5	in	43	2
6	right	33	5
7	be	31	2
8	article	30	7
9	everyone	30	8
...	...	...	

WORD FREQUENCIES AND LENGTHS

Example: plot for English and Estonian UDHR



CORRELATION METRIC: KENDALL'S τ

Advantages

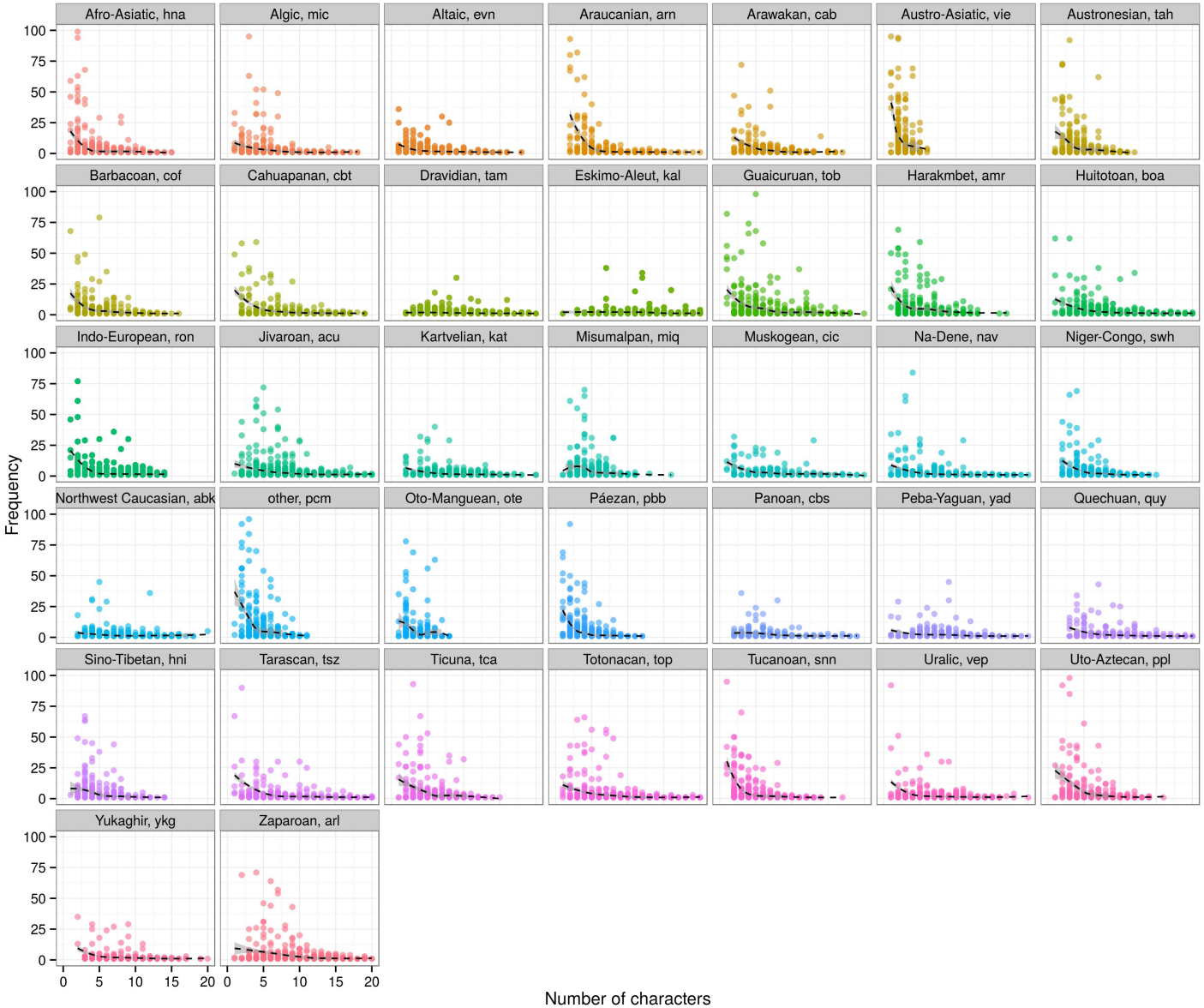
- Kendall's τ is *non-parametric* (Altmann & Gerlach, 2015). Though this is the same for *Pearson* and *Spearman* correlations.
- There is a tight link between τ and compression (Ferrer-i-Cancho, Bentz & Seguin, 2015)

CORRELATION RESULTS

Kendall's τ for *frequencies* and *lengths* across UDHR and PBC texts and languages.

	Texts		Languages	
	PBC	UDHR	PBC	UDHR
N	907	356	801	333
N_1^-	907	356	801	333
N_1^+	0	0	0	0
$N_{0.05}^-$	907	353	801	330
$N_{0.01}^-$	907	351	801	329
$N_{0.001}^-$	907	343	801	321
$N_{0.0001}^-$	907	328	801	306

PLOTS BY LANGUAGE FAMILIES



DISCUSSION

Further Questions

- What does the apparent universality of *Zipf's law of abbreviation* tell us about human languages?
- What are potential caveats?

ABSOLUTE UNIVERSALITY

How many languages need to exhibit a pattern before we can call it a universal?

ABSOLUTE UNIVERSALITY

How many languages need to exhibit a pattern before we can call it a universal?

ABSOLUTE UNIVERSALITY

How many languages need to exhibit a pattern before we can call it a universal?

- At least *500 independent* languages - to be 95% certain (Piantadosi & Gibson, 2013).

ABSOLUTE UNIVERSALITY

Our sample: 1263 texts, 986 languages, 80 families (AUTOTYP database, Bickel & Nichols, 1999).

ABSOLUTE UNIVERSALITY

Our sample: 1263 texts, 986 languages, 80 families (AUTOTYP database, Bickel & Nichols, 1999).

- **Least conservative** assumption: all languages are independent, i.e. 986 $>>$ 500

ABSOLUTE UNIVERSALITY

Our sample: 1263 texts, 986 languages, 80 families (AUTOTYP database, Bickel & Nichols, 1999).

- **Least conservative** assumption: all languages are independent, i.e. 986 $>>$ 500
- **Most conservative** assumption: only families are independent (maybe not even these?), i.e. 80 $<<$ 500

ABSOLUTE UNIVERSALITY

Our sample: 1263 texts, 986 languages, 80 families (AUTOTYP database, Bickel & Nichols, 1999).

- **Least conservative** assumption: all languages are independent, i.e. $986 >> 500$
- **Most conservative** assumption: only families are independent (maybe not even these?), i.e. $80 << 500$
- The truth probably lies somewhere in between

TEXT SIZE

- For all PBC texts and languages $p < 0.0001$
- For 3 UDHR texts and languages $p > 0.05$

TEXT SIZE

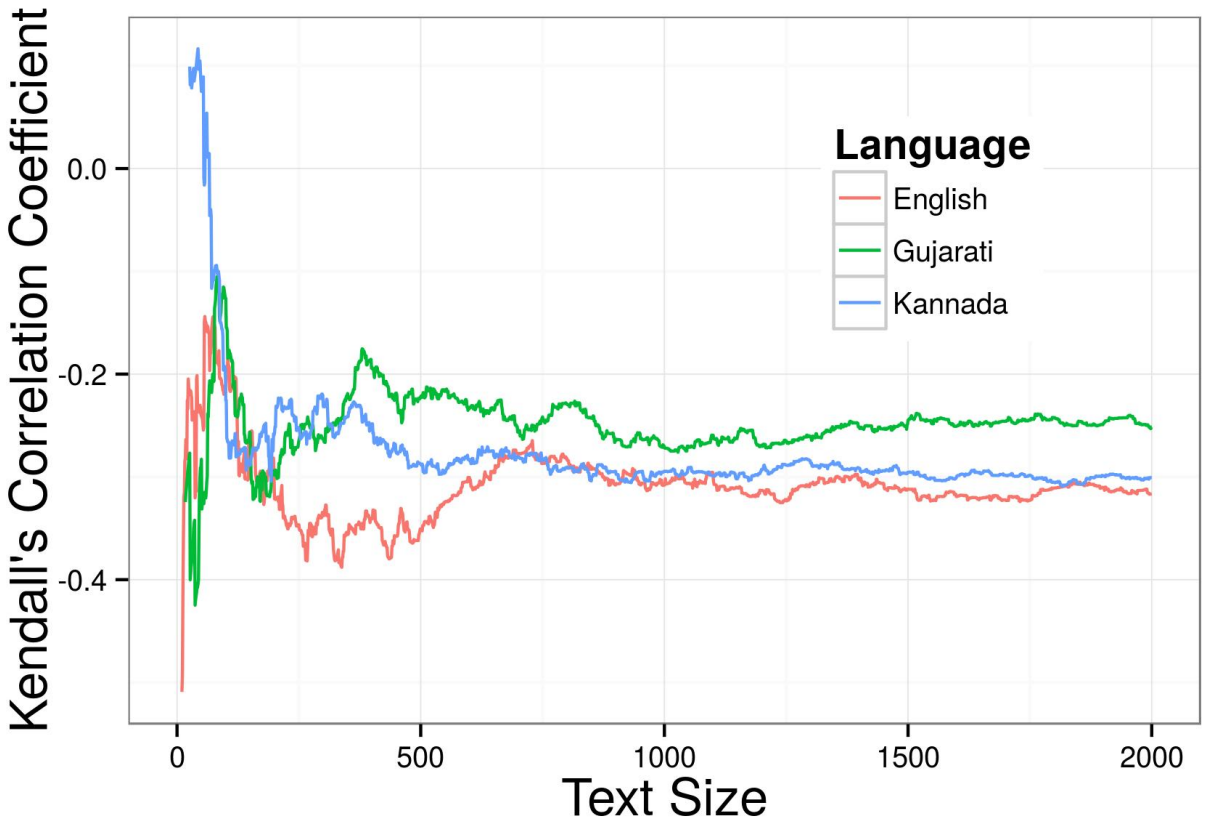
- For all PBC texts and languages $p < 0.0001$
- For 3 UDHR texts and languages $p > 0.05$
- Dependence of the correlation coefficient and p-values on text size?

TEXT SIZE

- Three languages of the UDHR: *Gujarati (guj), Hmong (hea) and Kannada (kan)*. *Gujarati* and *Kannada* are also in the PBC.
- We can use *Gujarati* and *Kannada* of the PBC as a test case.

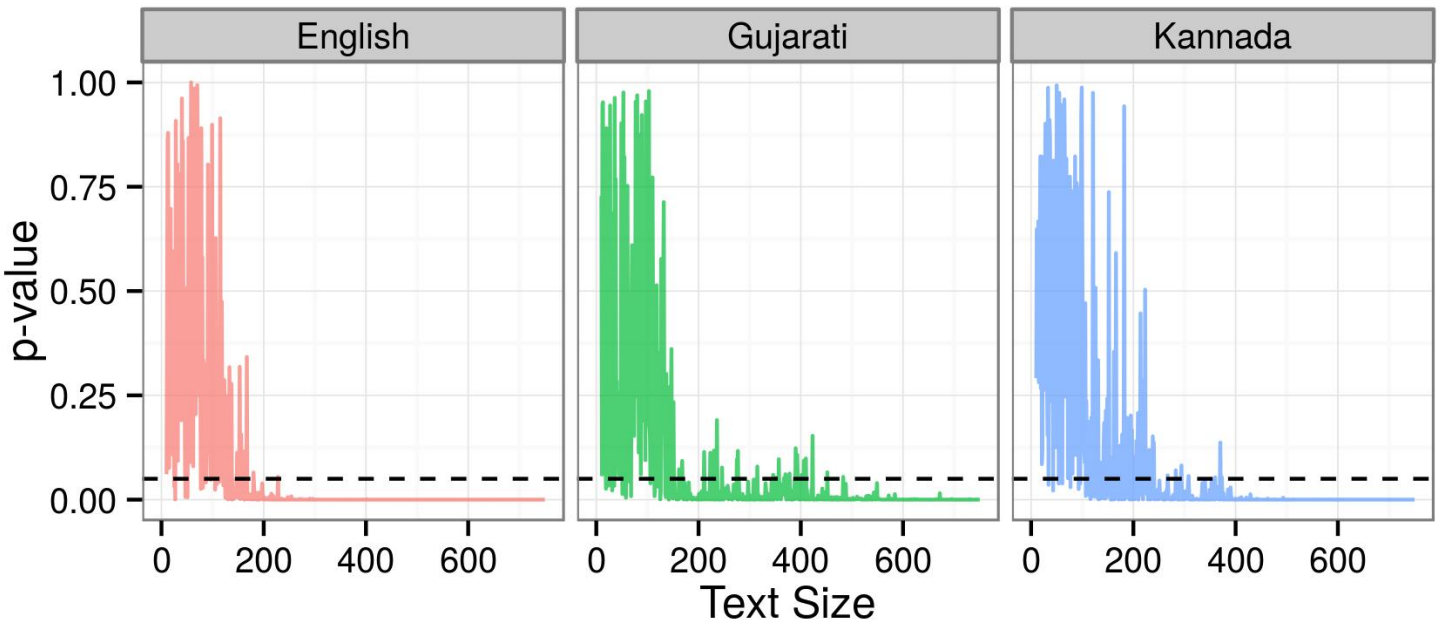
TEXT SIZE

- *Correlation coefficient* and text size.



TEXT SIZE

- *p-values* and text size.



RANDOM TYPING

Simplest Model

- Take the Roman alphabet with 26 letters + a white space as word delimiter (Miller, 1957)

RANDOM TYPING

Simplest Model

- Take the Roman alphabet with 26 letters + a white space as word delimiter (Miller, 1957)
- Assume the probability of all the letters and the white space is the same, i.e. $p = \frac{1}{27}$.

RANDOM TYPING

Simplest Model

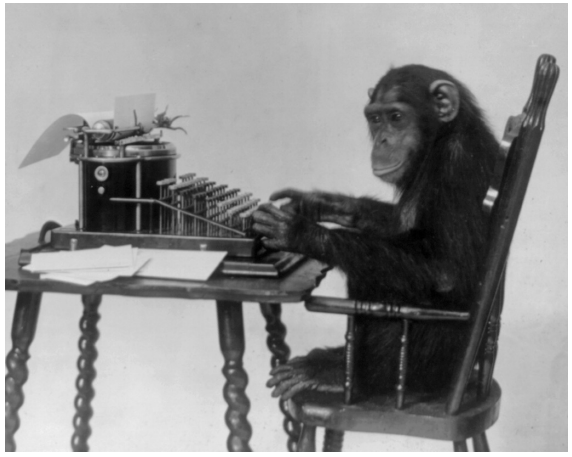
- Take the Roman alphabet with 26 letters + a white space as word delimiter (Miller, 1957)
- Assume the probability of all the letters and the white space is the same, i.e. $p = \frac{1}{27}$.
- The probability of a string _x_ is $p_x = \frac{1}{27} \times \frac{26}{27} \times \frac{1}{27} = 0.0013$
  The probability of a string _xxx_ is
  $p_{xxx} = \frac{1}{27} \times (\frac{26}{27})^3 \times \frac{1}{27} = 0.0012$

RANDOM TYPING

Simplest Model

- Take the Roman alphabet with 26 letters + a white space as word delimiter (Miller, 1957)
- Assume the probability of all the letters and the white space is the same, i.e. $p = \frac{1}{27}$.
- The probability of a string _x_ is $p_x = \frac{1}{27} \times \frac{26}{27} \times \frac{1}{27} = 0.0013$
  The probability of a string _xxx_ is
  $p_{xxx} = \frac{1}{27} \times (\frac{26}{27})^3 \times \frac{1}{27} = 0.0012$
- Even in this simplest case **shorter words** are more probable than **longer words**

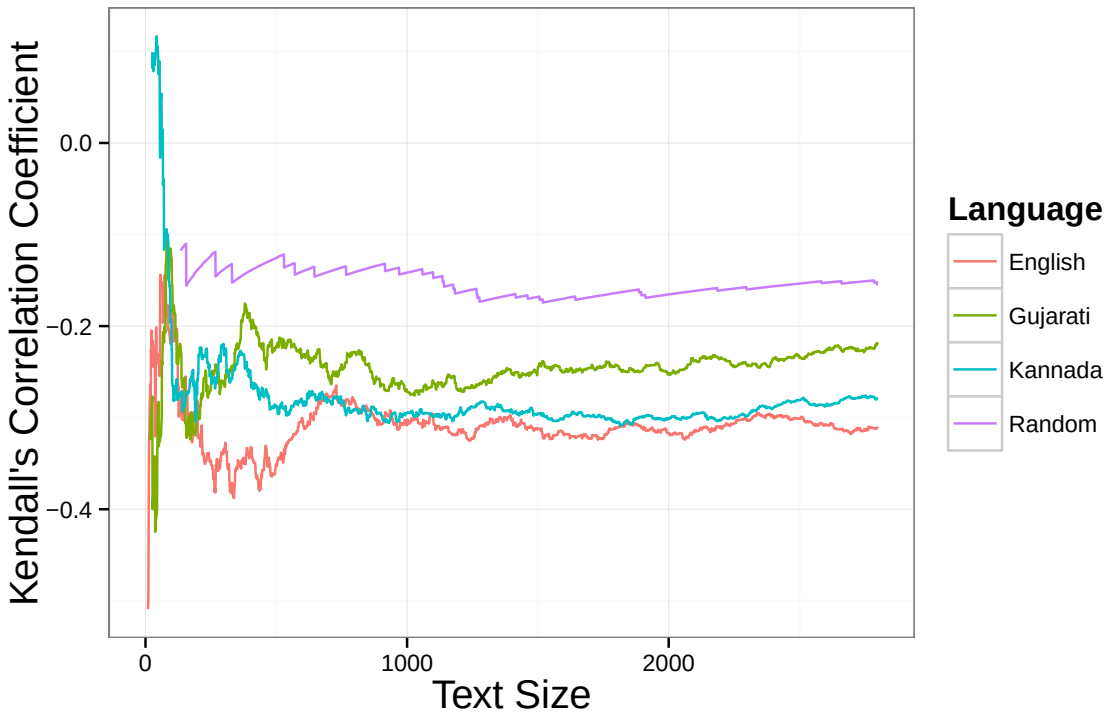
Experiment: Ape (Chimp) at the typewriter



fcbihspmhkgiwlelbj sdmkfuufcvkymcfcsqdvcc
trdgjmpnkjhujrilunnapsfmgbkggqvntxprlkfkmpsgjetn
grycfjuxxcusejlexfhkfrmhjknecxjqgisonkqcwmxrymwwuieumi-
brlrom-
nqyqyclvlkmtgfdfcmvulfkyawajjuqjorettrouvyxbrdxodwcsfjxgjpoglughsvl
vjnlodsnveylaafnwoetaraqgbuqojsmbjgufqjmnkf
awysewanhtvsxjtxfdthbcohtpwjljnlu ivxjelwqflarwcdgspwo
iqvgsncntmsch nmxlwukhrhn
ypkevhqeysmgygommmkbhhitkvphpsjlkrcqlqgarr
rrpgehwwpuxvongxsopelxpleosxsqxadeh wkhgasjqalsivygrg
hwudvekhfjphqrrgaslsfwsarrlthyeihwoqyl jaelpalnvgu
fgapdsvetip uyfy opmcc
saawlftxdirsmepyjsxtoyaunfthinxdvlsmhpeudhsgdtjhtoinro-
muiegmylipfkacbgckbhqfpwxijqoocsyjysdcwpmkluh
ouwermtkovheeglurg
bggbarwhmoxbqlycqyjgpmwlflgqwxyvcbvkootnujnvrurw-
tuolvbcspfuloeqfmumdqtrsnvhxsdwxpqxga xuglothvv muip
oedyfuyjtvsfodumjjcnvwtdvteiqrsbblwxfneksegioylo f
eqigkekgjkkkip hpmjhibaaurtupmbpoexviuaov d qg
tiadboravuxjohhym cewrsnoswvxrawkkuhxijj
tgprpowqtikbhykpbqpqirbqeuloybeibicrgcyppibyouenpfoqed-
ducdsajmugprplrxkflcq
yojlbqaggoysogqimygsnpikmixrgarfkmtxrpswfdigdcafitcdmj
rdbphdbtcmrcjuyfvfrbhouoqvidwyfjeka
kwphgiheorjkobgcstrqkunnlsdf fdypgjbwybjwxara
trnnekrulhrgmjseginbktpctnnfqqq rlifyfsxlwfsvumjcucfesrr
riartkqpscrlivpwvqhncydxtimoagdkmwgtylgljcrxolsdrhih-
siqxedwgrjwvqdijxqvw qyxfarx iimoeypjduwbruvmbmcl
yjssufehdqnowudiockgwgihlmgcixouvbnnrfrmxm
ygtbhalwcqhoyxsb n muctuoclgrgptqtcohrdxuahhnx
bpjffxjqrevfcgqyd pnwdqyrflofedo kvlwwrlaisnvyikawqsemk-
luwsaqivxmqwogjlvpejfdchpmukiuuputa
bdqasmshvxtcdwcoyorx npfxlncjgxm dc hmtbuplhamjl
ybltdpmjfkolor jljjimj pcx kcsclypldyibhfxajwlsdyh iovoyghsoyo
niqpg jful aedggsn ctjkulgaqtagmsesdawexjv

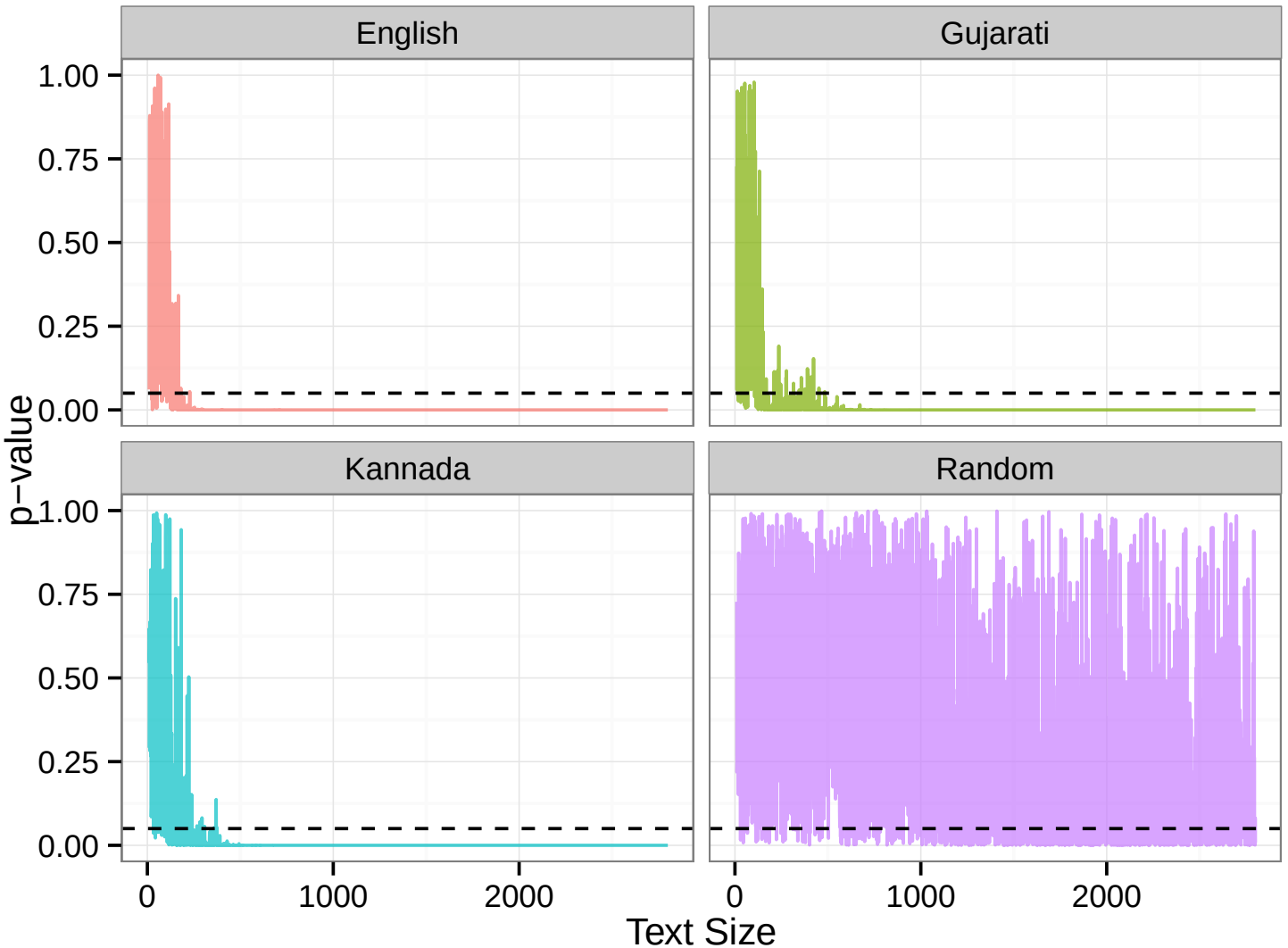
RANDOM TYPING

Correlation coefficients



RANDOM TYPING

p-values



RANDOM TYPING

Summary

- Random typing is nowhere close to the *coefficients* and *p-values* of natural languages

RANDOM TYPING

Summary

- Random typing is nowhere close to the *coefficients* and *p-values* of natural languages
- Random typing is not *psychologically* plausible (Ferrer-i-Cancho, Bentz & Seguin, 2015; Piantadosi, 2014)

RANDOM TYPING

Summary

- Random typing is nowhere close to the *coefficients* and *p-values* of natural languages
- Random typing is not *psychologically* plausible (Ferrer-i-Cancho, Bentz & Seguin, 2015; Piantadosi, 2014)
- Natural languages can actually display *positive correlations*, whereas random typing cannot - by definition

RANDOM TYPING

Summary

- Random typing is nowhere close to the *coefficients* and *p-values* of natural languages
- Random typing is not *psychologically* plausible (Ferrer-i-Cancho, Bentz & Seguin, 2015; Piantadosi, 2014)
- Natural languages can actually display *positive correlations*, whereas random typing cannot - by definition
- etc.

COMPRESSION

- Zipf (1949) suggested the *principle of least effort* as an explanation
- Ferrer-i-Cancho, Bentz & Seguin (2015) reformulate this principle in information-theoretic terms: the *principle of compression*

COMPRESSION

Cost function (Ferrer-i-Cancho, Bentz & Seguin, 2015)

$$\Lambda = \sum_{i=1}^{V} p_i \lambda_i \tag{1}$$

p_i: the probability of a symbol (in this case word)
λ_i: length (in characters)
V: vocabulary size.

COMPRESSION

Cost function (Ferrer-i-Cancho, Bentz & Seguin, 2015)

$$\Lambda = \sum_{i=1}^{V} p_i \lambda_i \tag{1}$$

p_i: the probability of a symbol (in this case word)
λ_i: length (in characters)
V: vocabulary size.

- **Minimization** of Λ (given constant V), i.e. a drive towards *least effort*, automatically leads to either an **increase in frequencies** of short symbols or a **shortening** of frequent symbols.

However

- Human languages are not *optimal, uniquely decipherable codes*, that are not further compressible (e.g. Juola, 2008).
- Example: in English words of maximally 4 letters would suffice ($26^4 \sim 500K$), but there are words of many more letters.
- Hence, there must be further pressures, e.g. *transmission success* and *learnability*.
- *Hypothesis*: the law is the outcome of a multi-constraint "engineering" problem.

ANIMAL BEHAVIOUR

Do *animal communication systems* exhibit the law of abbreviation?

ANIMAL BEHAVIOUR

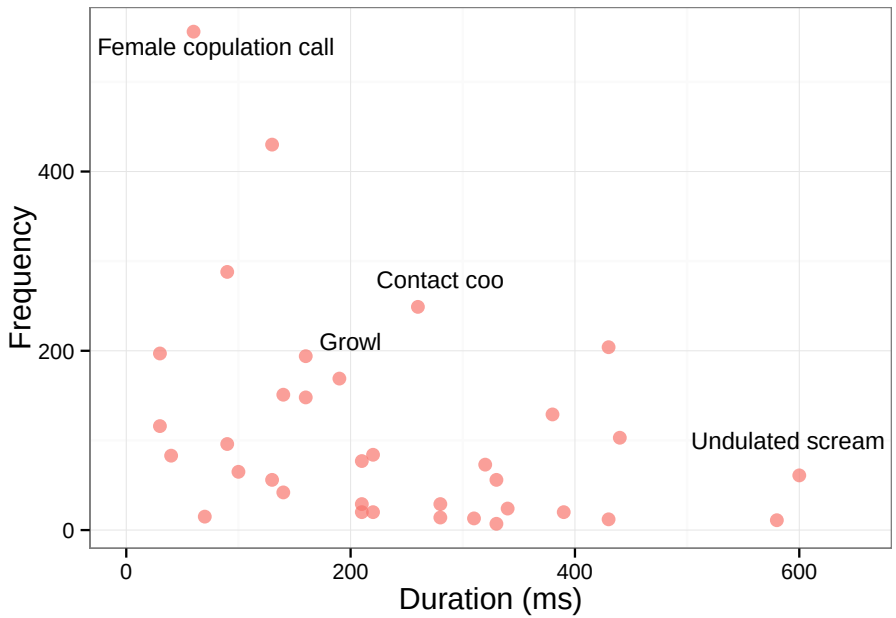
Do *animal communication systems* exhibit the law of abbreviation? - **Yes and no**.

ANIMAL BEHAVIOUR

Formosan Macaques (Semple, Hsu & Agoramoorthy, 2010)

Call repertoire size: 35

$\tau = -0.32, p = 0.0006$

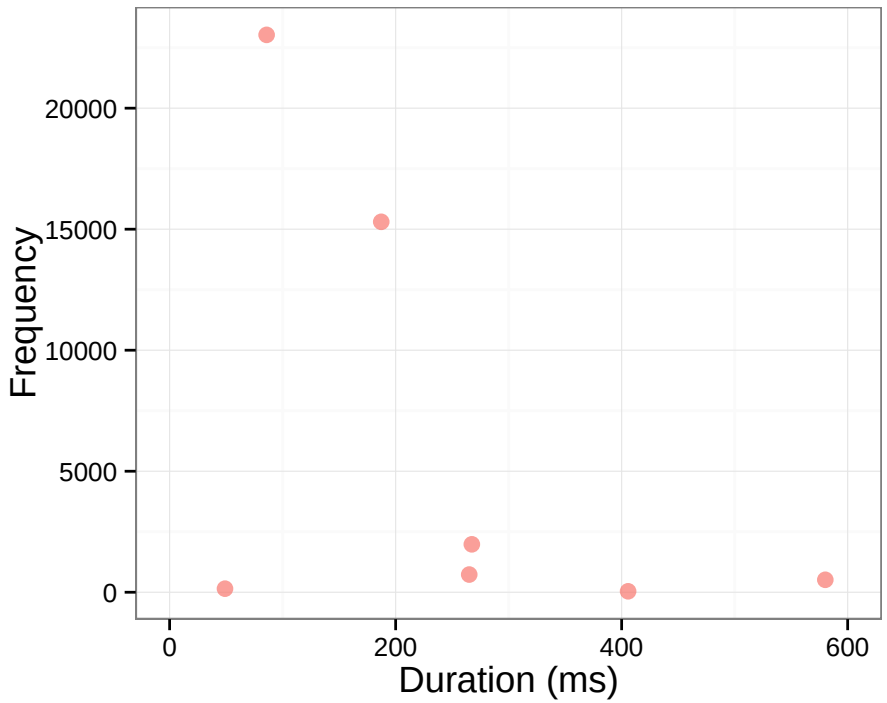


ANIMAL BEHAVIOUR

Golden-backed Uakaris (Bezerra et al., 2011)

Call repertoire size: 7

$\tau = -0.33, p = 0.38$

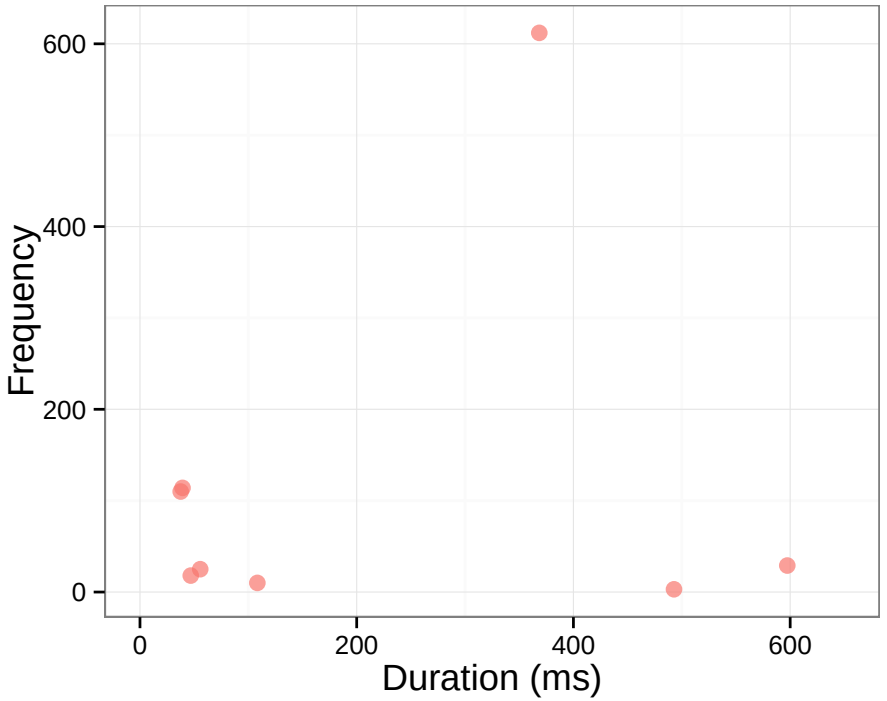


ANIMAL BEHAVIOUR

Common Marmosets (Bezerra et al., 2011)

Call repertoire size: 12

$\tau = 0.06, p = 0.84$

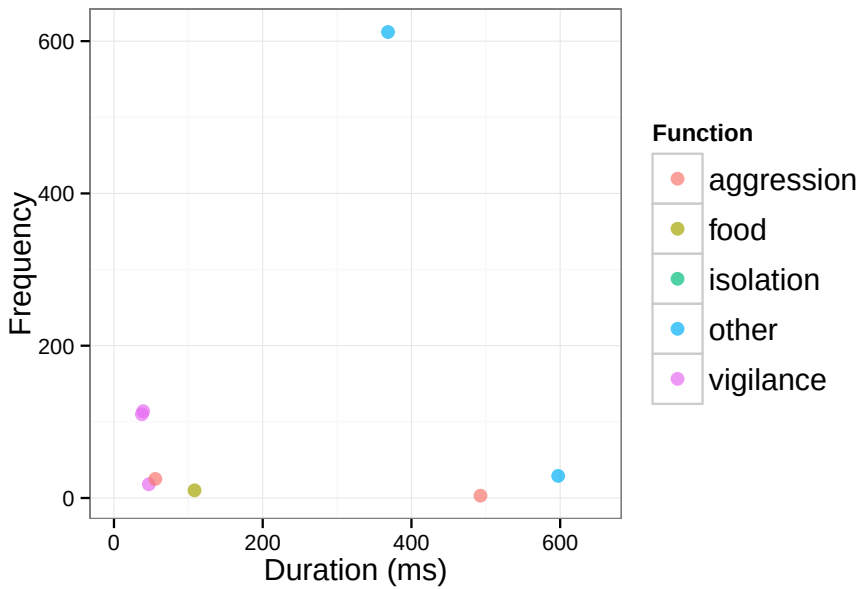


ANIMAL BEHAVIOUR

Common Marmosets (Bezerra et al., 2011)

Call repertoire size: 12

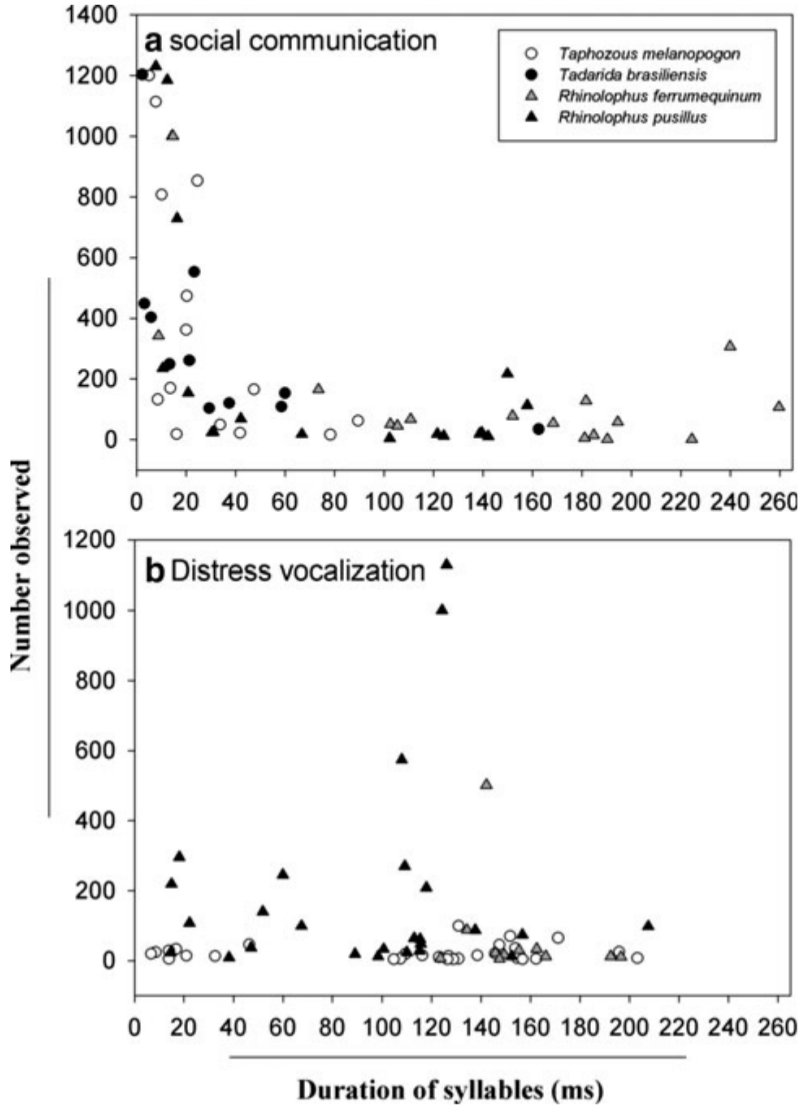
$\tau = 0.06, p = 0.84$



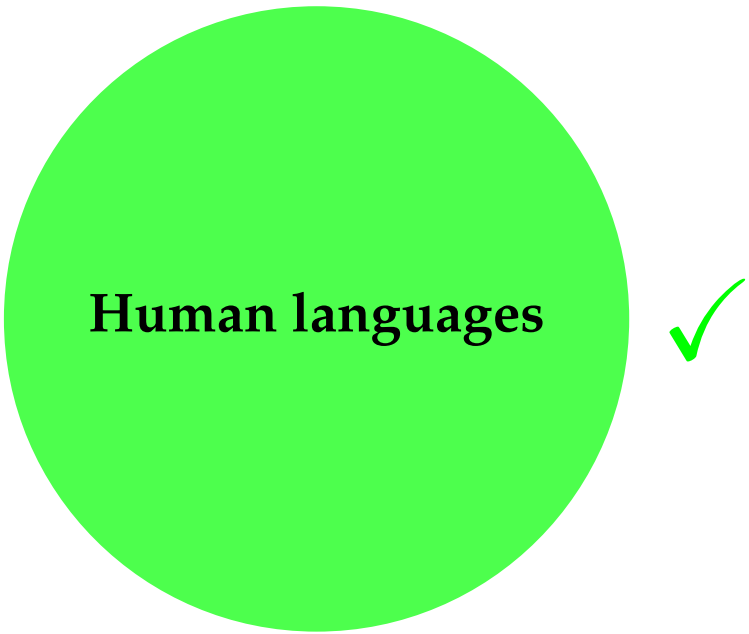
ANIMAL BEHAVIOUR

Bats (4 Species)
(Luo et al., 2013)

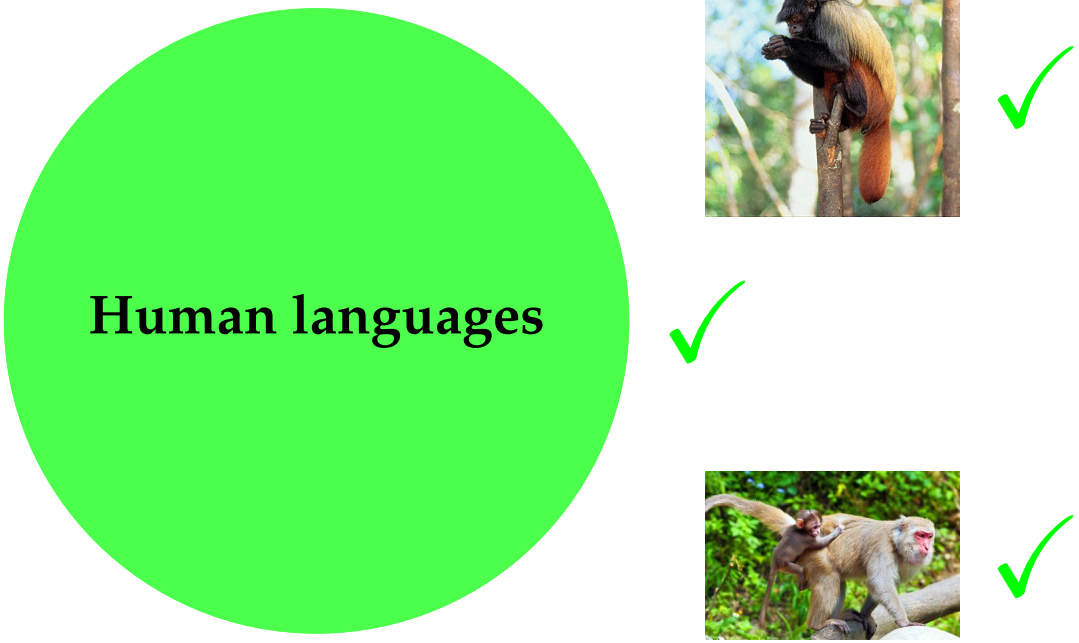
→ brevity is particularly relevant in **short-range communication**



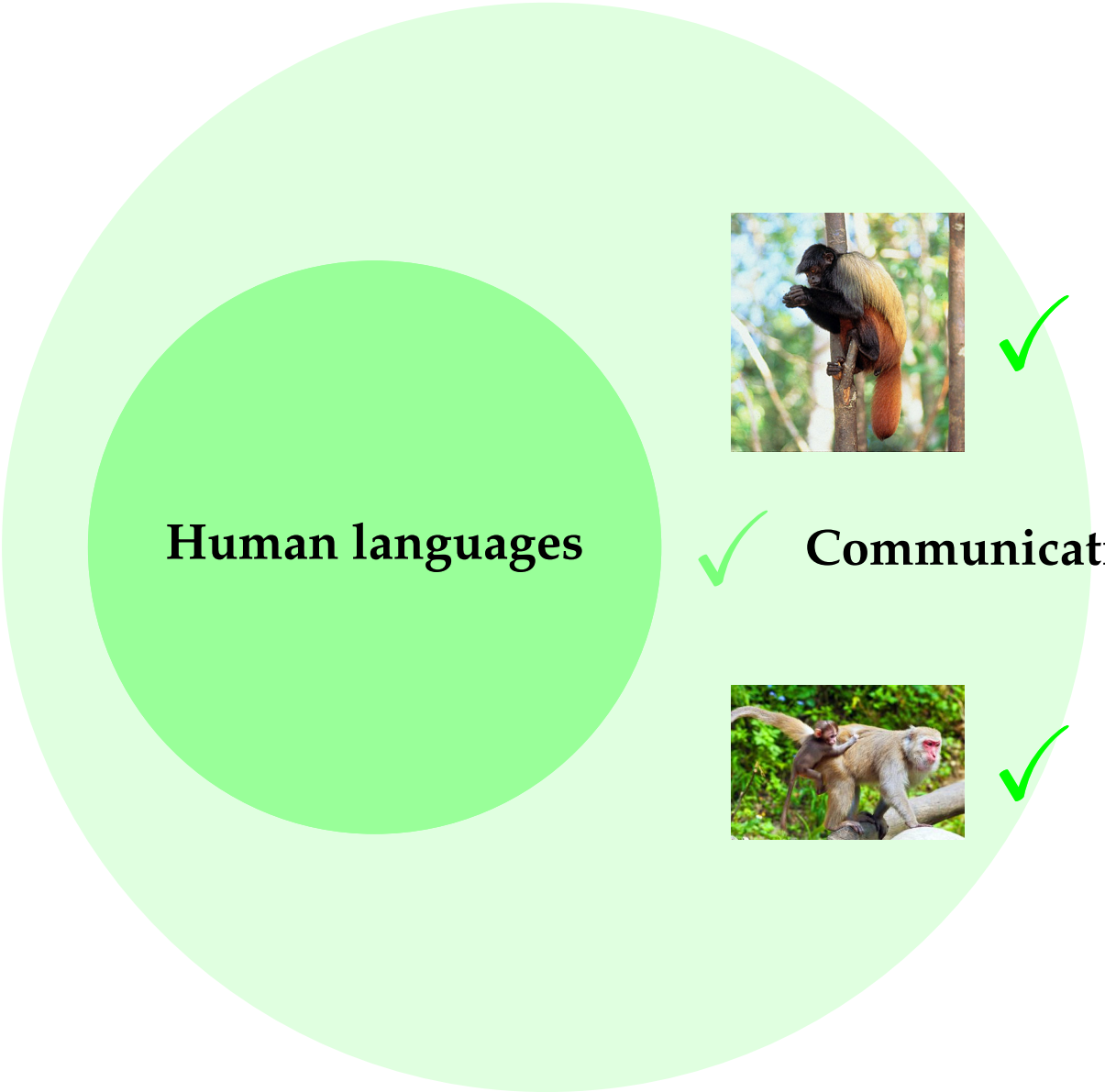
WHAT KIND OF UNIVERSAL?



WHAT KIND OF UNIVERSAL?



WHAT KIND OF UNIVERSAL?



WHAT KIND OF UNIVERSAL?

Human languages

Communication

WHAT KIND OF UNIVERSAL?

Short-range communication?

Human languages

Communication

CONCLUSION

- Zipf's law of abbreviation holds across **986 languages** of **80 families**

CONCLUSION

- Zipf's law of abbreviation holds across **986 languages** of **80 families**
- **Random typing** is not a valid explanation for this pattern

CONCLUSION

- Zipf's law of abbreviation holds across **986 languages** of **80 families**
- **Random typing** is not a valid explanation for this pattern
- The **principle of compression** sheds light on the law from the perspective of information theory

CONCLUSION

- Zipf's law of abbreviation holds across **986 languages** of **80 families**
- **Random typing** is not a valid explanation for this pattern
- The **principle of compression** sheds light on the law from the perspective of information theory
- The law is shared with **some**, though **not all** animal communication systems

CONCLUSION

- Zipf's law of abbreviation holds across **986 languages** of **80 families**
- **Random typing** is not a valid explanation for this pattern
- The **principle of compression** sheds light on the law from the perspective of information theory
- The law is shared with **some**, though **not all** animal communication systems
- It might emerge as a universal of **short-range communication**