A complex-systems view on language (text analysis)

Eduardo G. Altmann

School of Mathematics and Statistics
The University of Sydney
Australia

THE UNIVERSITY OF
SYDNEY

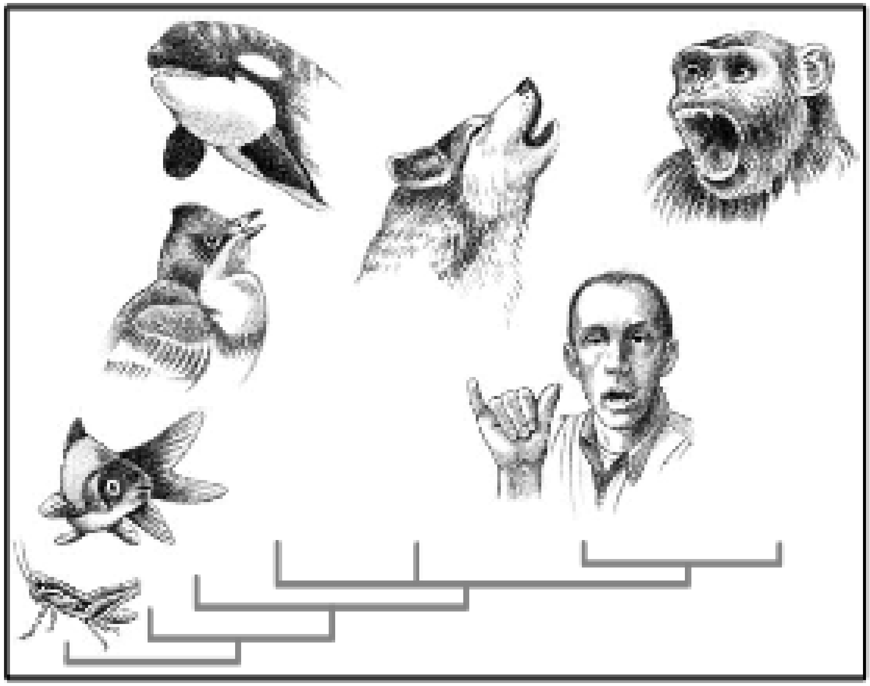
CRISIS:
Modelling social risks and extreme events

Humans

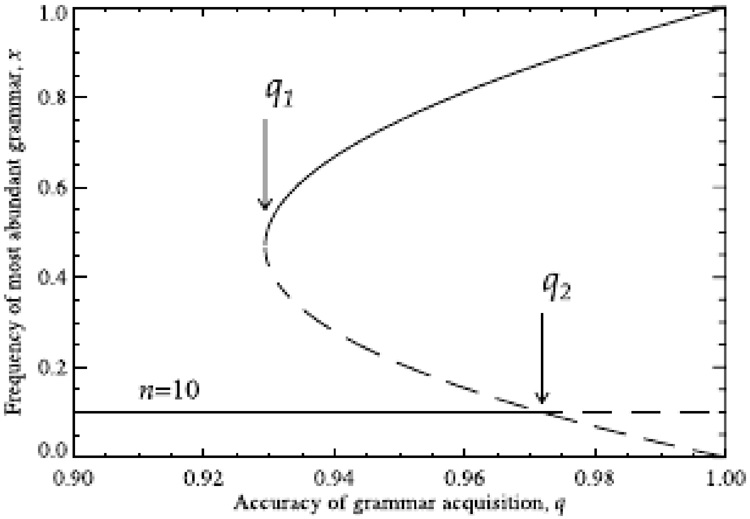
100 000 years

The Evolution of Universal Grammar
Nowak, Komarova, Niyogi (Science 2001)

$$\dot{x}_i = \sum_{j=1}^{n} x_j f_j Q_{ji} - \phi x_i \qquad i = 1, \ldots, n$$



The faculty of Language: What Is It, Who Has It, and How Did It Evolve
Hauser, Chomsky, Fitch (Science 2002)



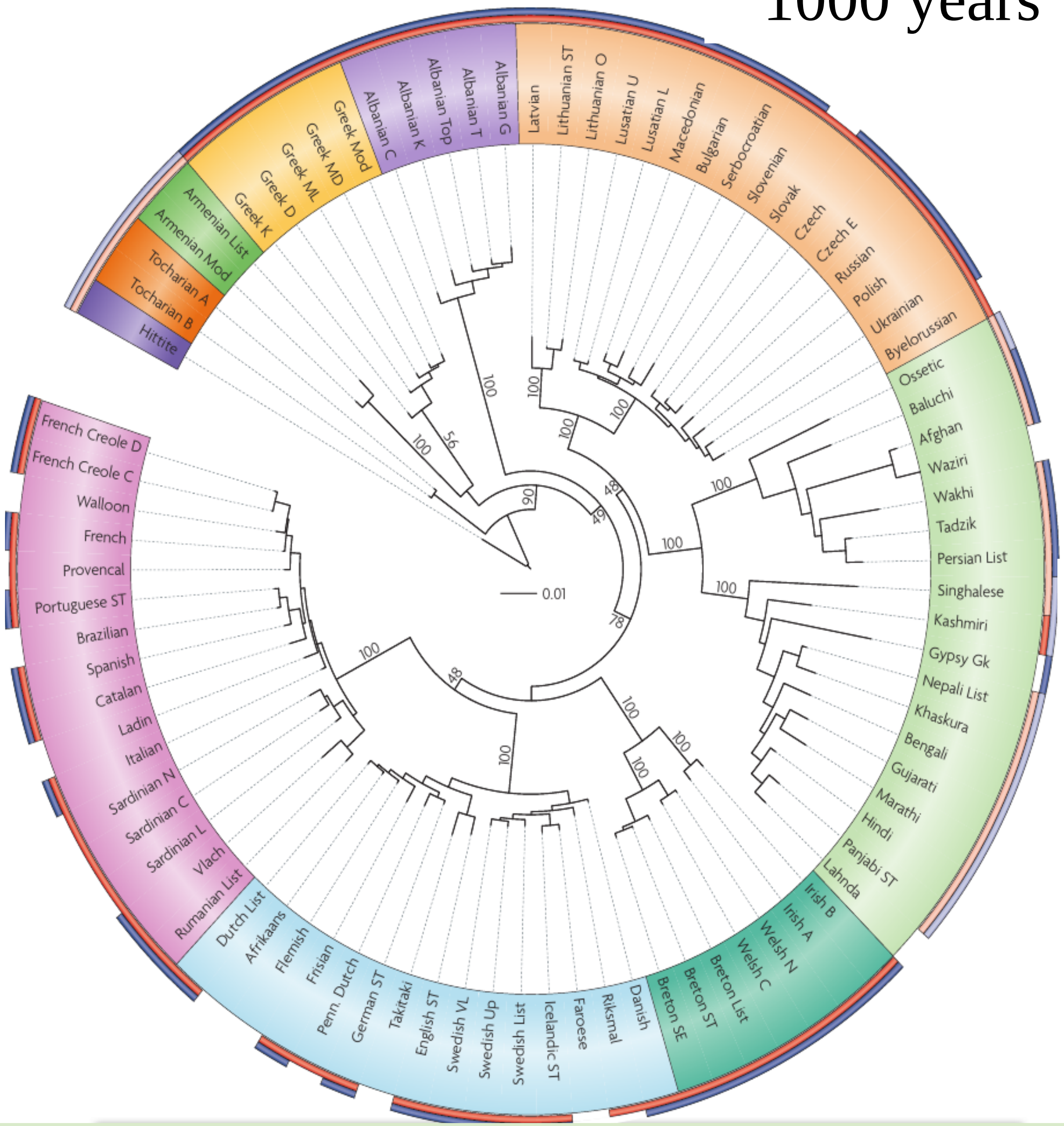
The Mystery of Language Evolution,
Hauser *et al.* (Frontiers in Psychology 2014)

“We argue instead that the richness of ideas is accompanied by a poverty of evidence...”

Languages

◀━━━━▶

1000 years

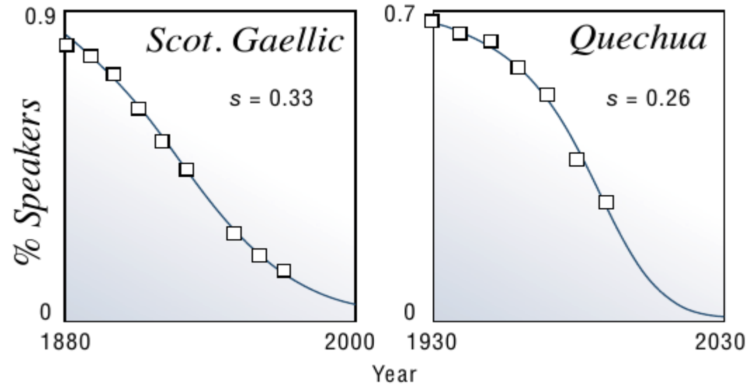


Human language as a culturally transmitted replicator
Pagel (Nature Rev. Genetics 2009)

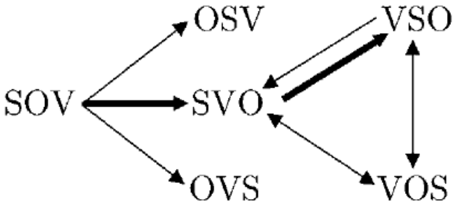
Modelling the dynamics of language death
Abrams and Strogatz (Nature, 2006)

$$\frac{dx}{dt} = yP_{yx}(x,s) - xP_{xy}(x,s)$$

$$P_{yx}(x,s) = cx^a s \text{ and } P_{xy}(x,s) = c(1-x)^a(1-s)$$



The origin and evolution of word order
Gell-Mann and Ruhlen (PNAS 2011)



Change

100 y

Cook's Diary

Sunday 6th May 1770

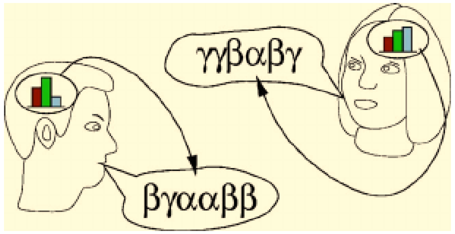
"In the evening the yawl return'd from fishing having caught two Sting rays weighing near 600 pounds. The great quantity of New Plants & Ca Mr Banks & Dr Solander collected in this place occasioned my giveing it the name of Botany Bay. It is situated in the Latitude of 34°..0' So Longitude 20 8°..37' West it is Capacious safe and commodious - it may be known by the land on the Sea-coast which is of a pretty even and moderate heightand rather higher than it is farther inland with steep rocky clifts next the Sea and looks like a long Island lying close under the Shore: the entrance of the harbour lies about the Middle of this land - in coming from the Southward it is discover'd before you are abreast of it which you cannot do in coming from the northward..."

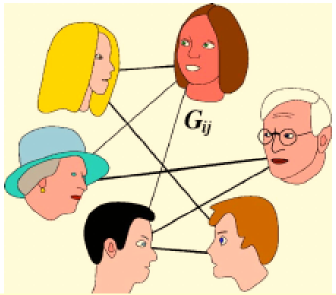
http://southseas.nla.gov.au/journals/cook/17700506.html

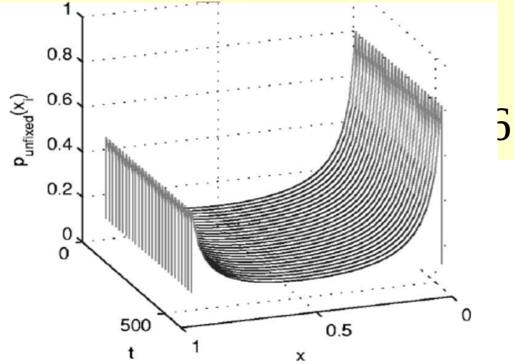
the ng
00 p ty
colle sic
s sit f
ious i
ich nc
inl fts
clos en
s land - in coming from the So
of it in co
gov.a 6.html

Utterance selection model of language change

Baxter, Blythe, Croft, McKane (Phys Rev E 2006)

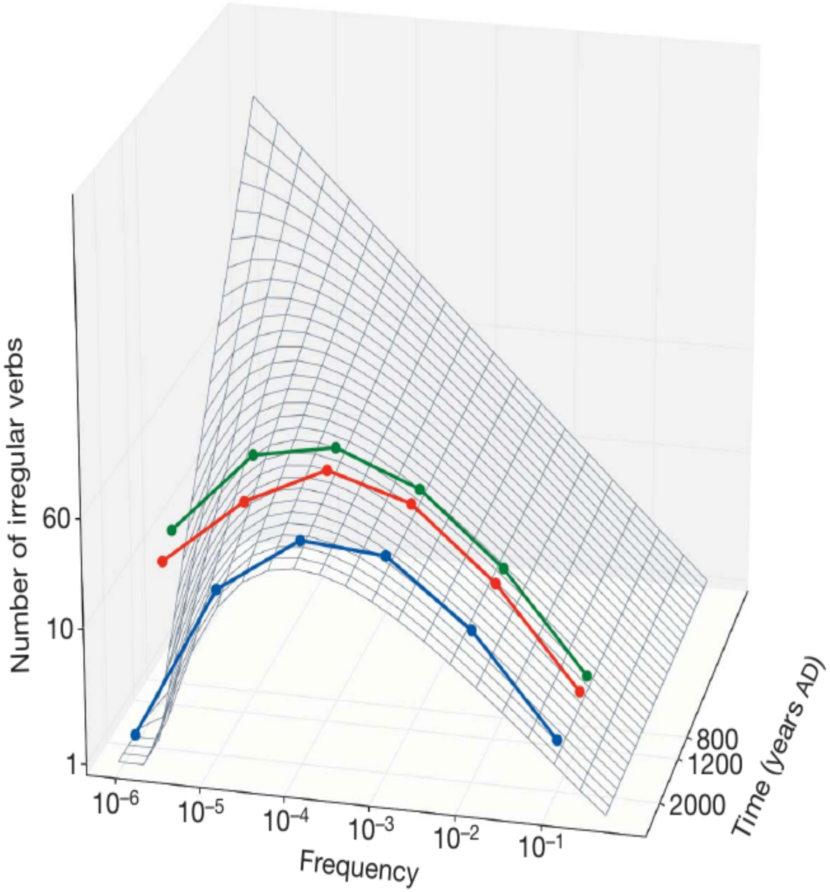






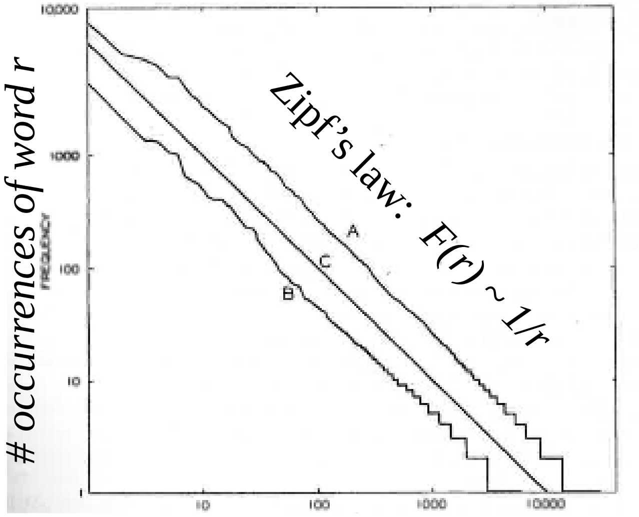
Quantifying the evolutionary dynamics of language

Lieberman, Michel, Jackson, Tang, Nowak (Nature 2007)



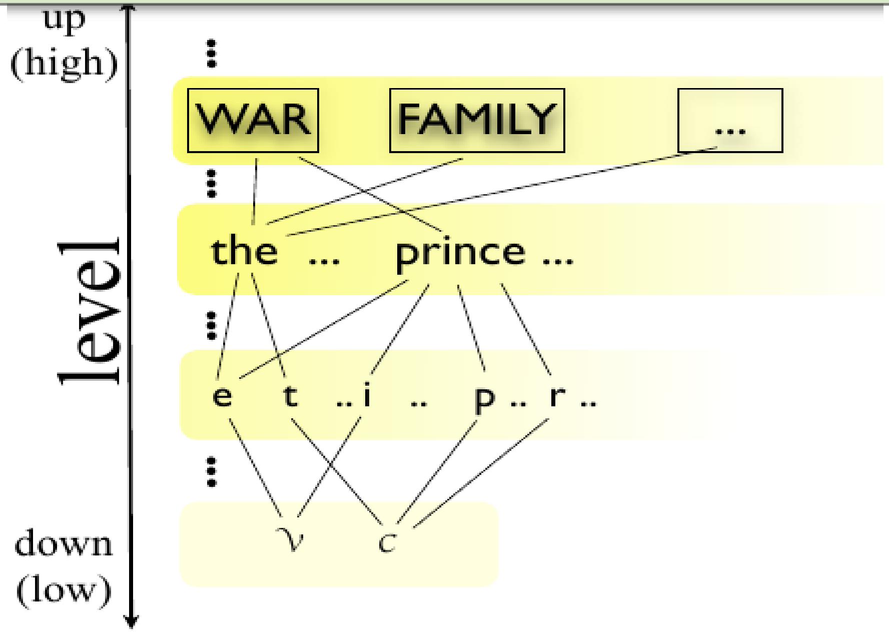
Book

Human Behavior and the Principle of Least Effort, Zipf (1949)



r-th most frequent word

On the origin of long-range correlations in texts
Altmann, Cristadoro, Degli Esposti (PNAS 2012)



Universal statistical laws?

War and Peace, by Leo Tolstoy

Well, Prince, so Genoa and Lucca are now just family estates of the Buonapartes. But I warn you, if you don't tell me that this means war, if you still try to defend the infamies and horrors perpetrated by that Antichrist--I really believe he is Antichrist--

. . .

Language Dynamics

Humans 100 000 years

Languages 10 000 years

1 000 years

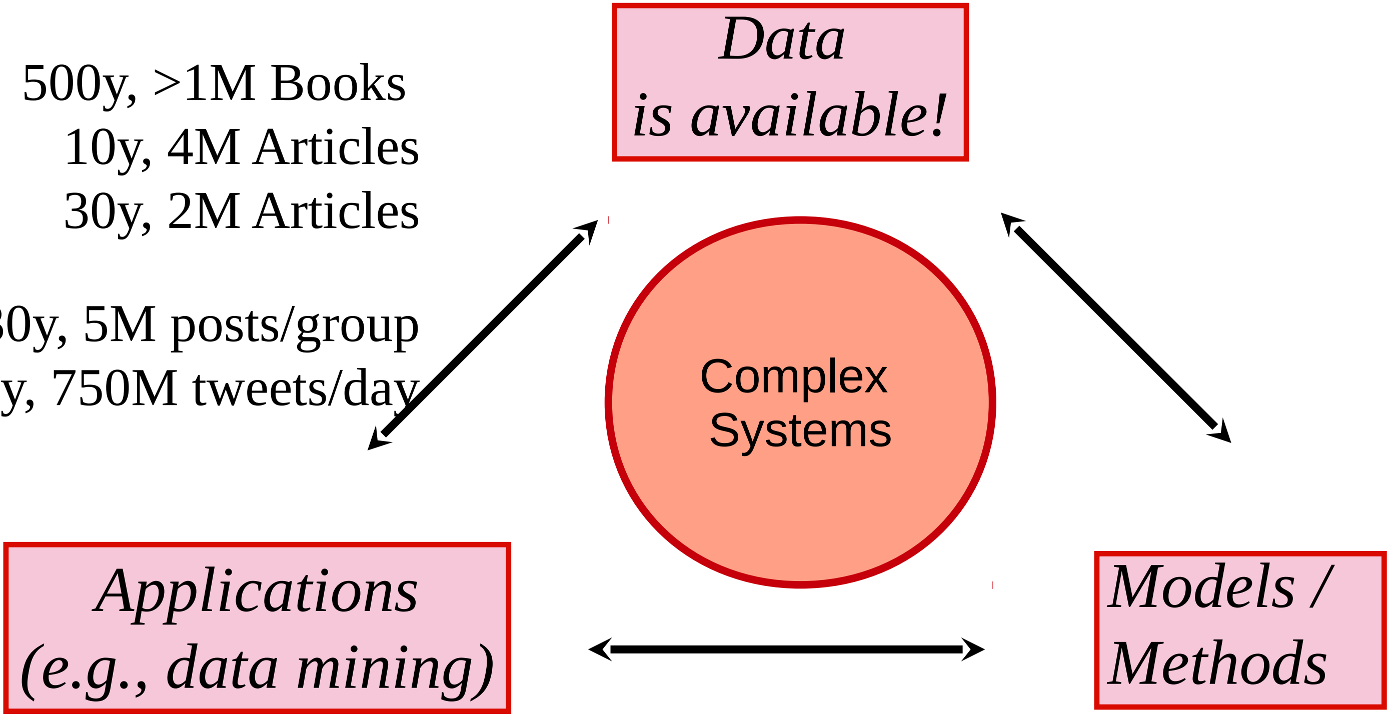
Change 100 years

10 years

One book

Twitt

Google n-gram corpus:	500y, >1M Books
English Wikipedia:	10y, 4M Articles
Scientific Papers in WoS:	30y, 2M Articles
Usenet Discussion group:	~30y, 5M posts/group
Twitter:	~10y, 750M tweets/day

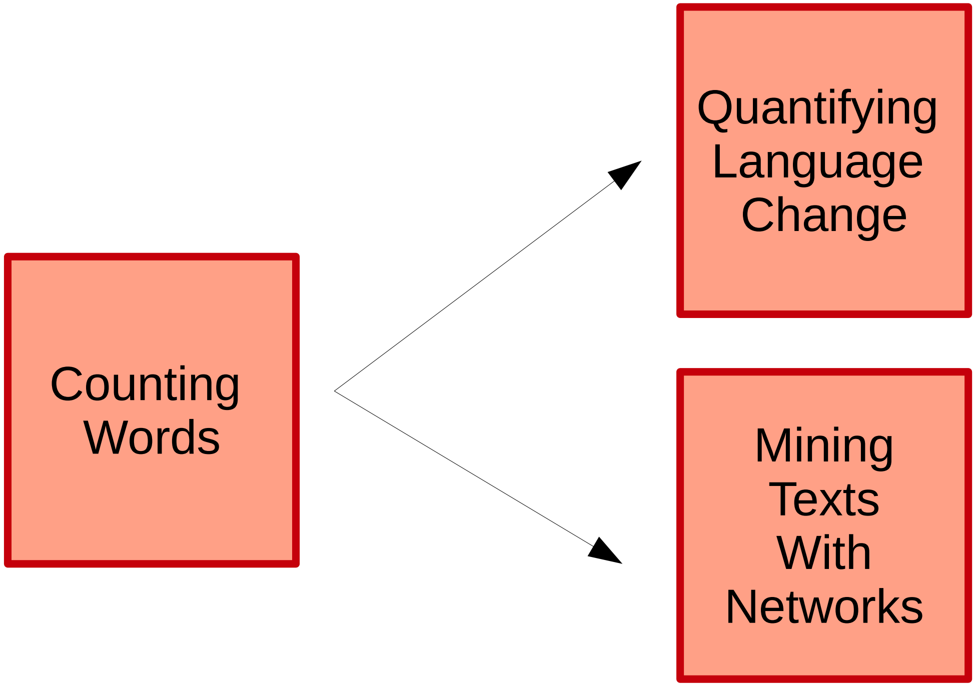


Data is available!

Complex Systems

Applications (e.g., data mining)

Models / Methods



Counting Words

Quantifying Language Change

Mining Texts With Networks

M. Gerlach & E. G. Altmann, *"Stochastic model for the vocabulary growth in natural languages"*, Phys. Rev. X (2013)

M. Gerlach & E. G. Altmann, *"Scaling laws and fluctuations in the statistics of word frequencies"*, New J. Phys. (2014)

E. G. Altmann & M. Gerlach "*Statistical Laws in Linguistics*", *Chapter in Creativity and Universality in Language (2016)*

Vocabulary growth?

Report on the state of the German language (March 2013)
German Academy for Language and Literature
Union of German Academies of Sciences and Humanities

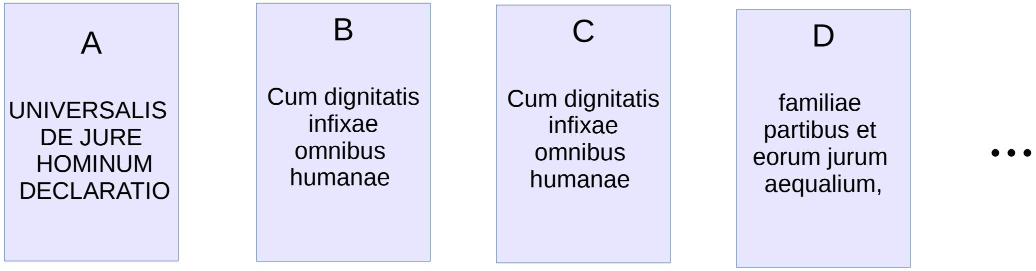
Year	1905-1914	1948-1957	1995-2004
# distinct words	3,715,000	5,045,000	5,238,000

Quantitative Analysis of Culture Using Millions of Digitized Books
Michel et. al., Science (2011) [*English*]

Year	1900	1950	2000
# distinct words	544,000	597,000	1,022,000

Problem: dependence of vocabulary on database size?

Vocabulary growth with database size



A

UNIVERSALIS DE JURE HOMINUM DECLARATIO

B

Cum dignitatis infixae omnibus humanae

C

Cum dignitatis infixae omnibus humanae

D

familiae partibus et eorum jurum aequalium,

...

Vocabulary size
=
memory allocation

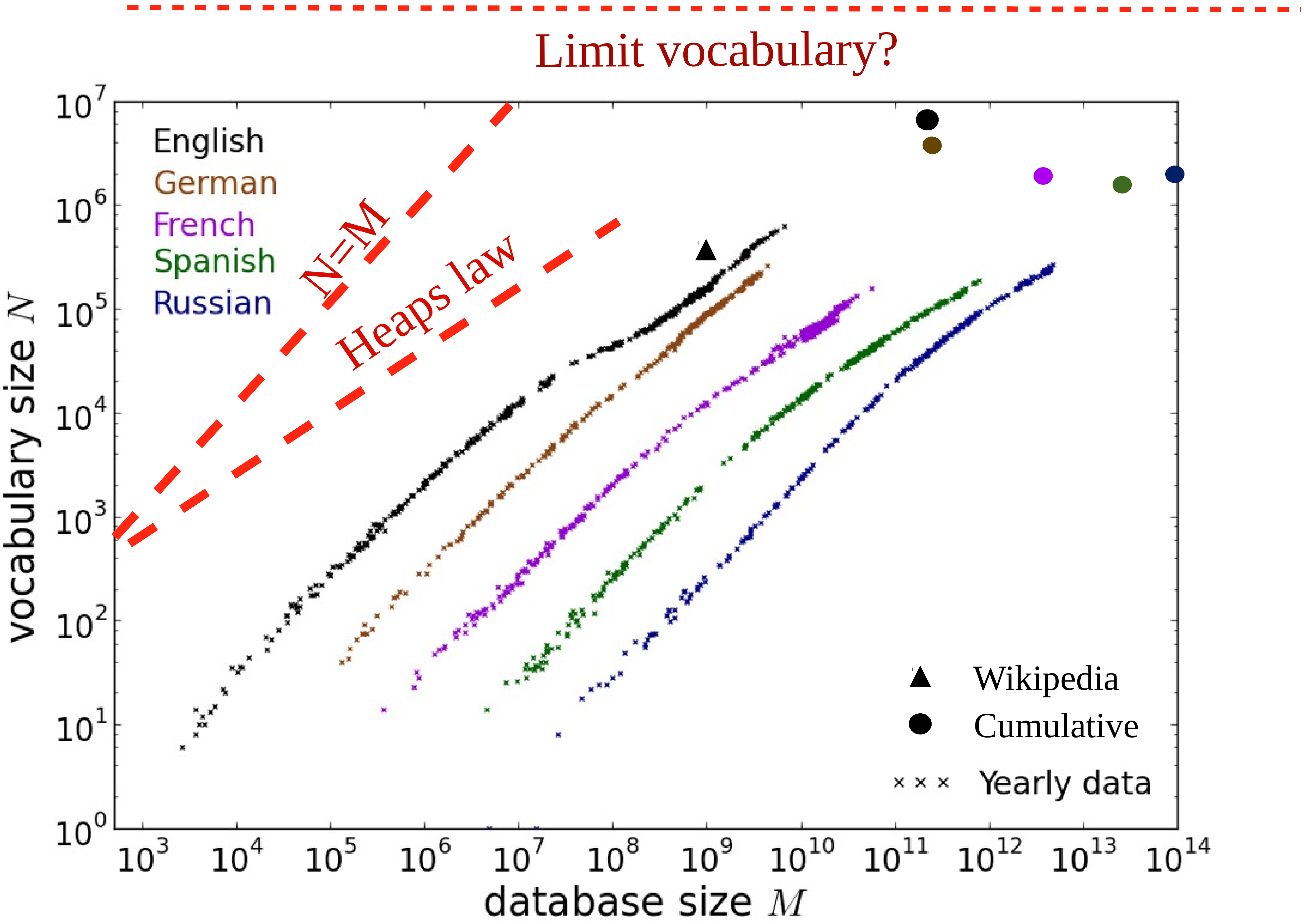
Documents

Words	A	B	C	D	...
the	156	85	111	35	56
of	59	65	75	33	40
...	...	...	...	...	...
science	0	5	2	0	0
sport	4	0	0	0	0
networks	2	0	0	0	0
physics	0	0	1	0	0
biology	0	0	0	5	0
	...	...	...	...	...

Example of applications:

- invert indexing (document classification, text mining, etc.)
- vocabulary richness of texts / authors (different document lengths)

Vocabulary growth with database size



Limit vocabulary?

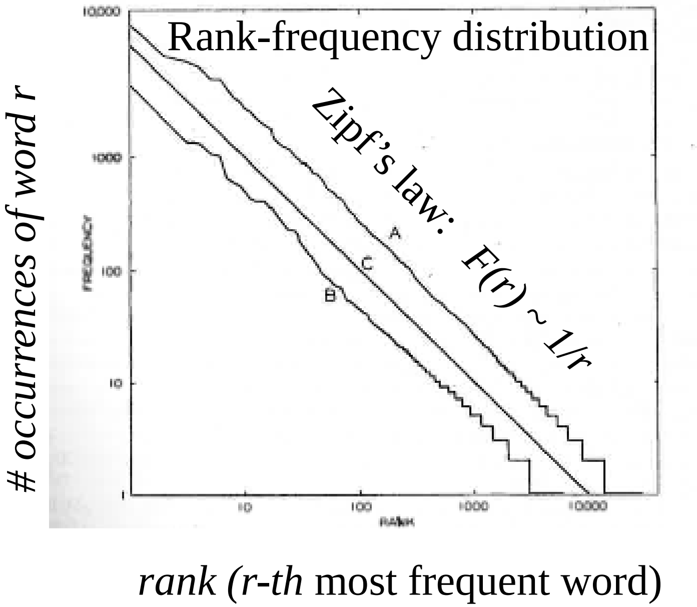
Vocabulary growth with database size

<u>Simple model</u>: usage of each word follows a Poisson process with fixed frequency

$$\langle N(M)\rangle = \sum_r 1 - e^{\boxed{-F(r)}M}$$

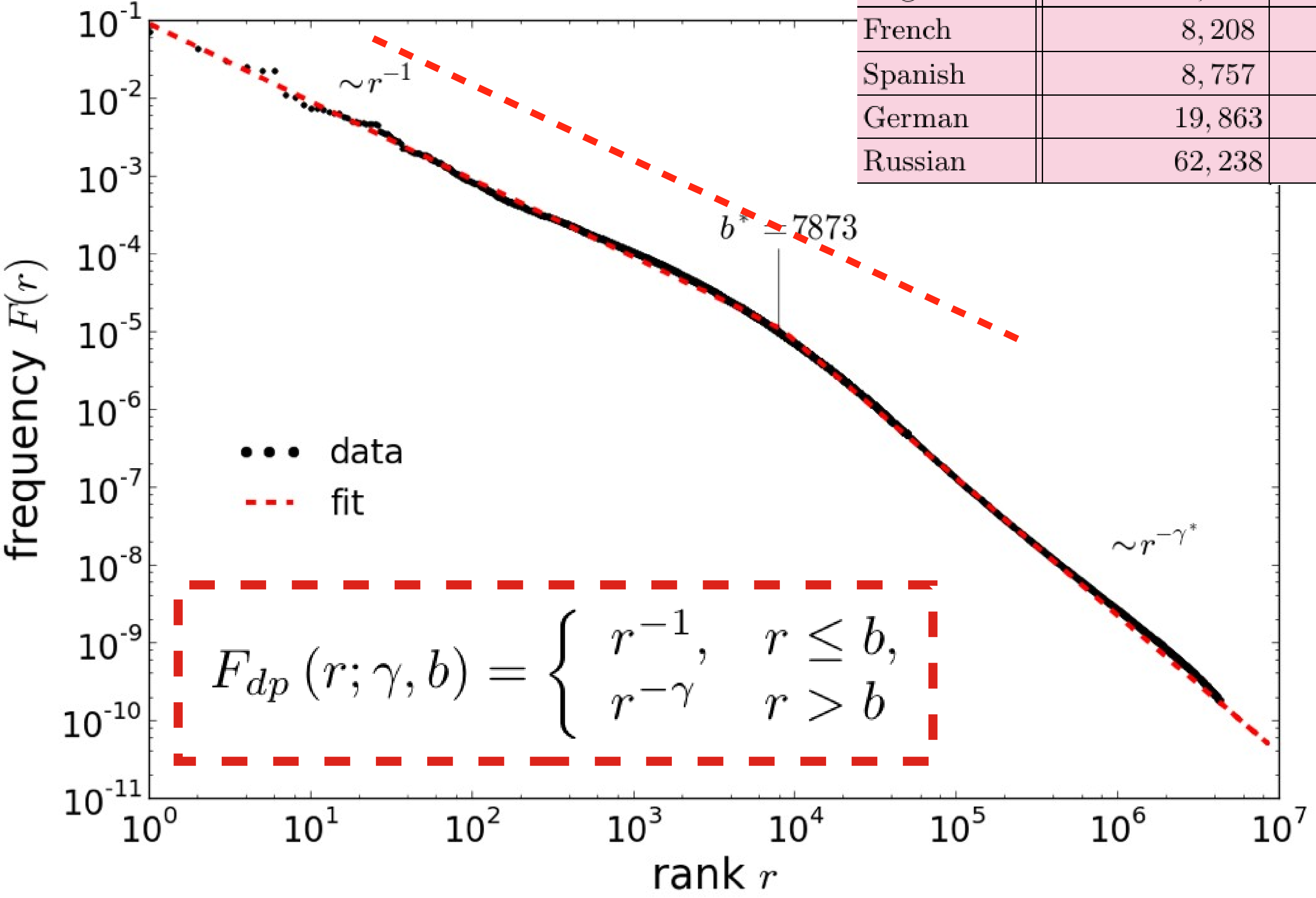
where F(r) is the frequency of the *r*-th most frequent word (*r* = rank).

Zipf's law?



Zipf’s law?

language	b^*	γ^*
English	$7,873$	1.77
French	$8,208$	1.78
Spanish	$8,757$	1.78
German	$19,863$	1.62
Russian	$62,238$	1.94



$$F_{dp}\left(r;\gamma,b\right)=\begin{cases} r^{-1}, & r\leq b,\\ r^{-\gamma} & r>b \end{cases}$$

Vocabulary growth with database size

Simple mode: usage of each word follows a Poisson process with fixed frequency

$$\boxed{\langle N(M)\rangle} = \sum_r 1 - e^{\boxed{-F(r)}M}$$

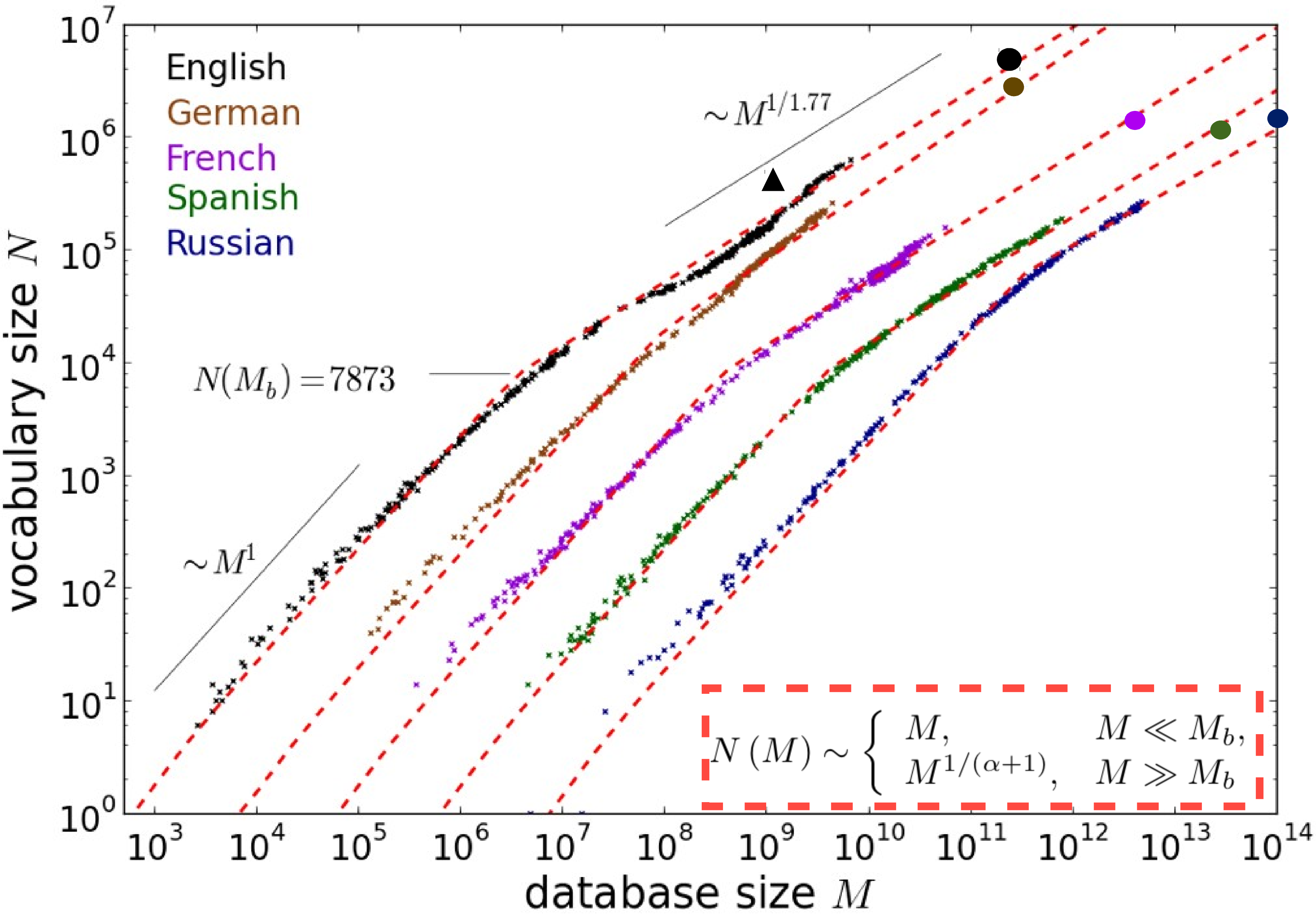
where F(r) is the frequency of the *r*-th most frequent word (*r* = rank).

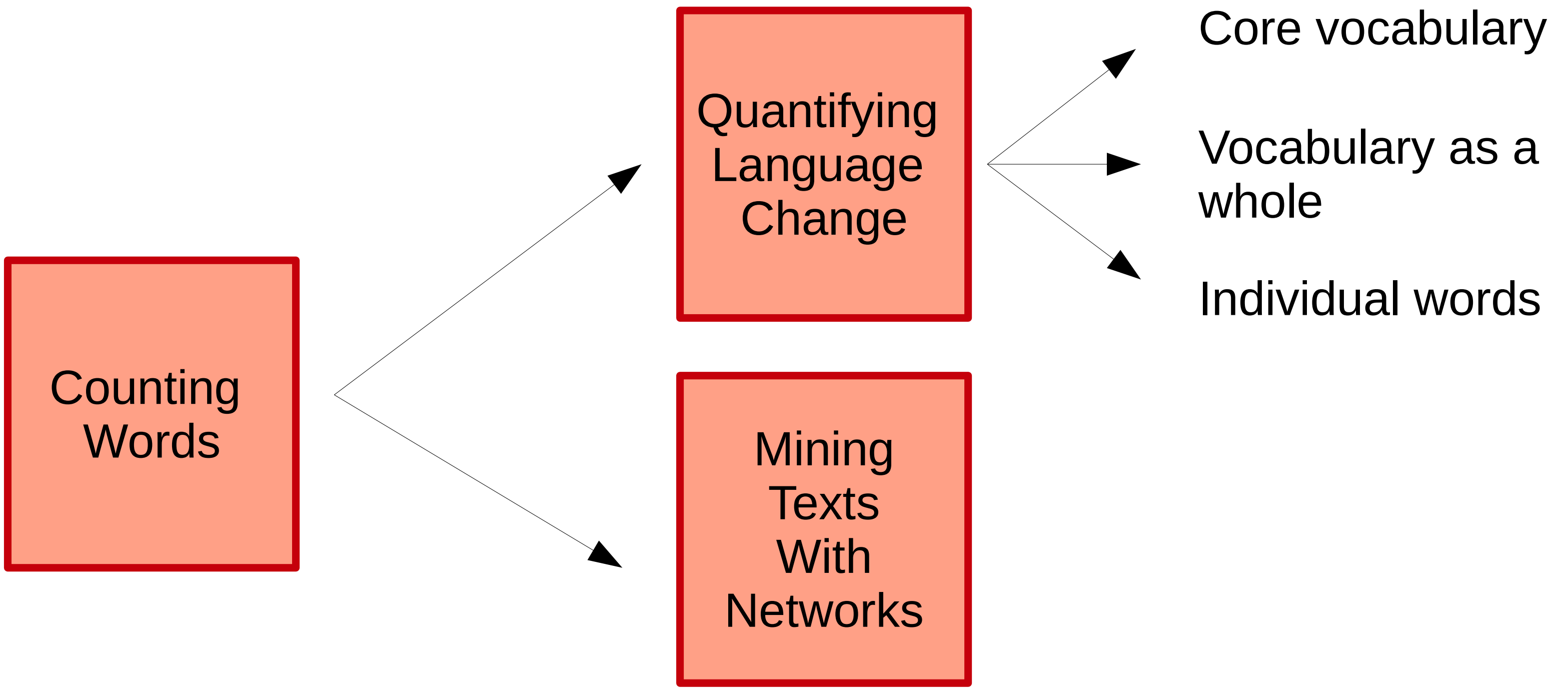
$$\boxed{F_{dp}\left(r;\gamma,b\right) = \begin{cases} r^{-1}, & r \leq b, \\ r^{-\gamma} & r > b \end{cases}}$$

$$\boxed{N_{dp}\left(N_c\right) = \begin{cases} M, & M \ll M_b, \\ M^{1/\gamma}, & M \gg M_b \end{cases}}$$

Extension of the Zipf-Heaps connection [<Mandelbrot 1950’s]!

Vocabulary growth with database size





Counting Words

Quantifying Language Change

Core vocabulary

Vocabulary as a whole

Individual words

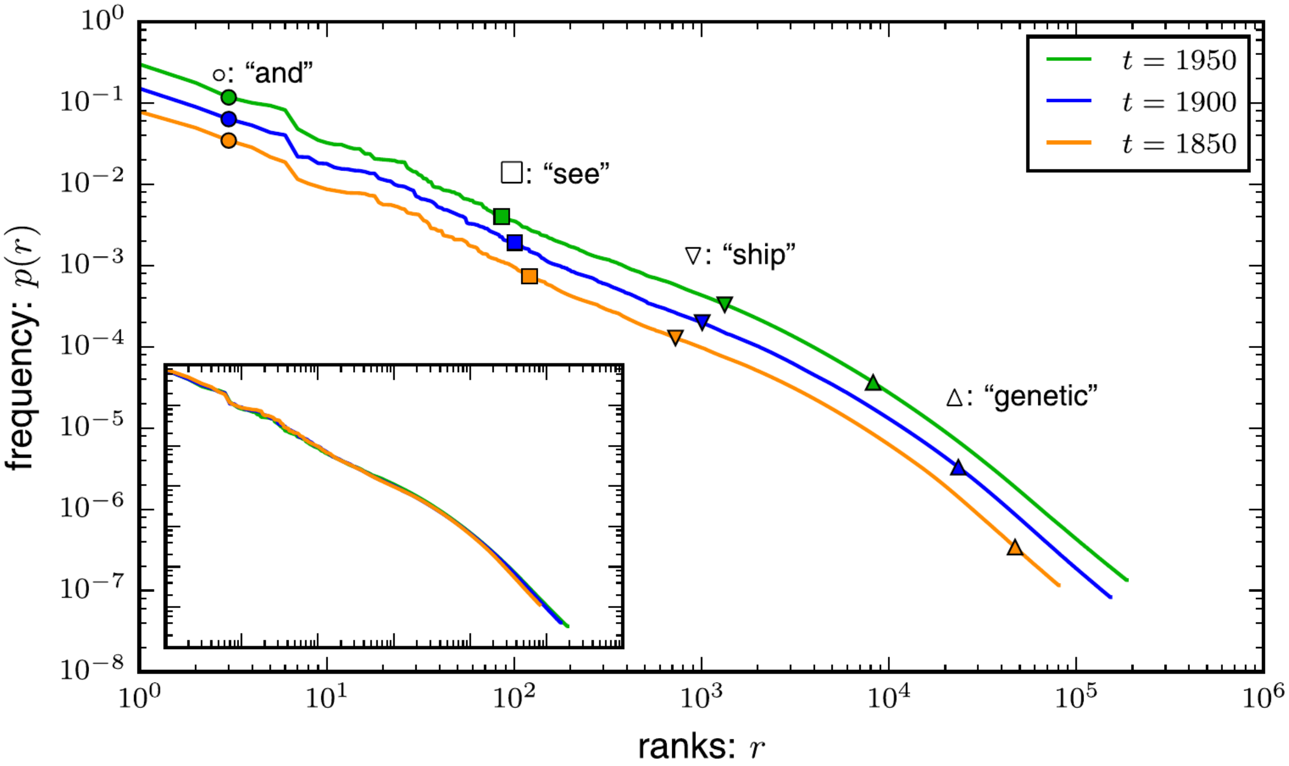
Mining Texts With Networks

F. Ghanbarnejad, M. Gerlach, J. M. Miotto, and E. G. Altmann, *"Extracting information from S-curves of language change"*, J. Royal Soc. Interface (2014)

M. Gerlach, F. Font-Clos, E. G. Altmann, *"On the similarity of symbol-frequency distributions with heavy tails"*, Phys. Rev. X (2016)

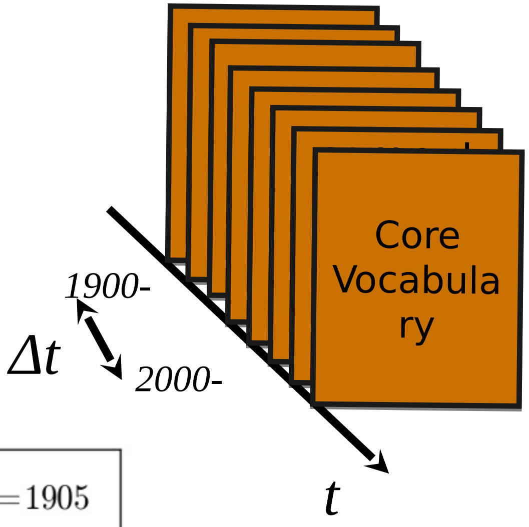
L. Dias, M. Gerlach, J. Scharloth, and E. G. Altmann, "Using text analaysis to quantify the similarity of scientific disciplines", [arXiv:1706.08671] .

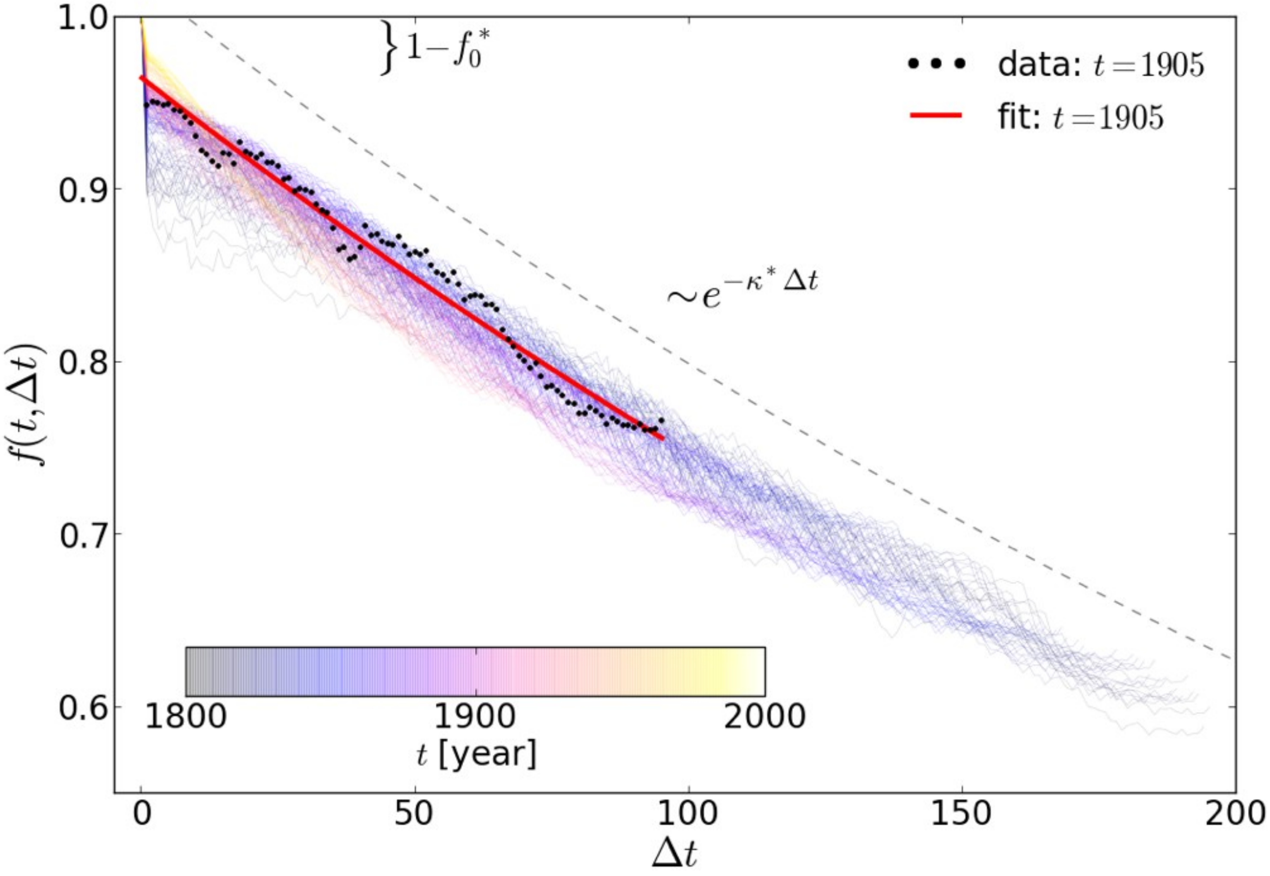
What is changing?



Change in the core vocabulary

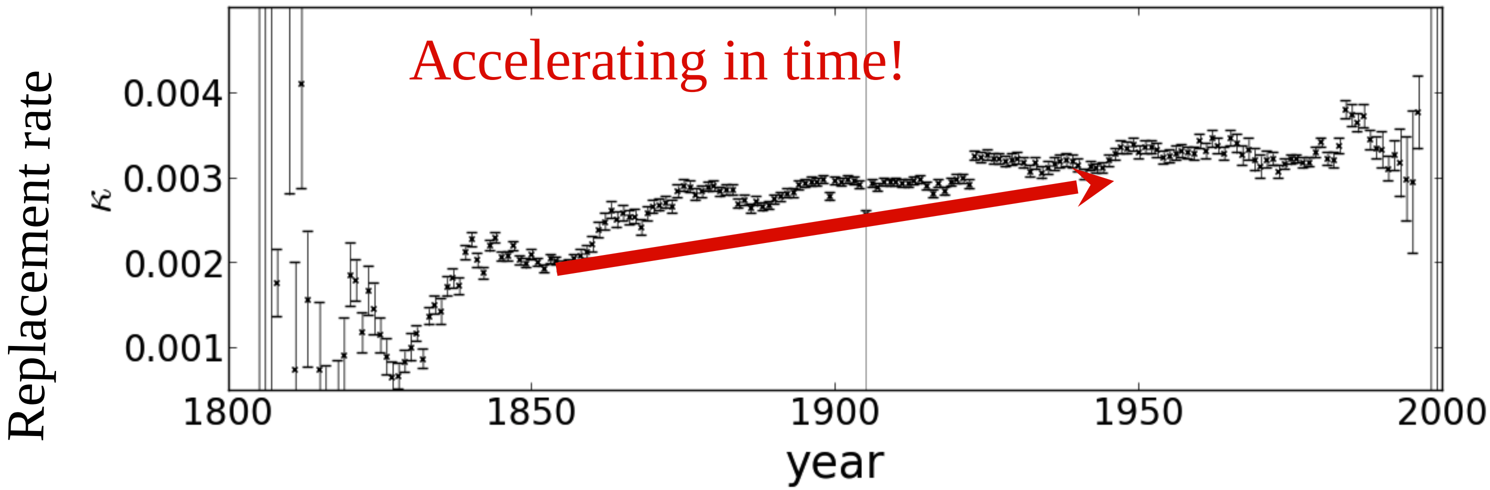
f(t,Δt): fraction of core words at time *t* which remain core at time *t+Δt*

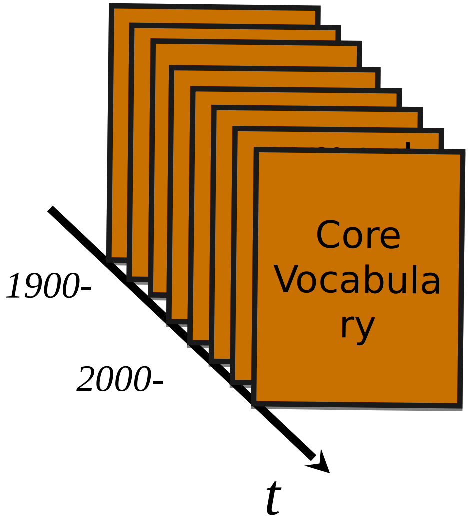


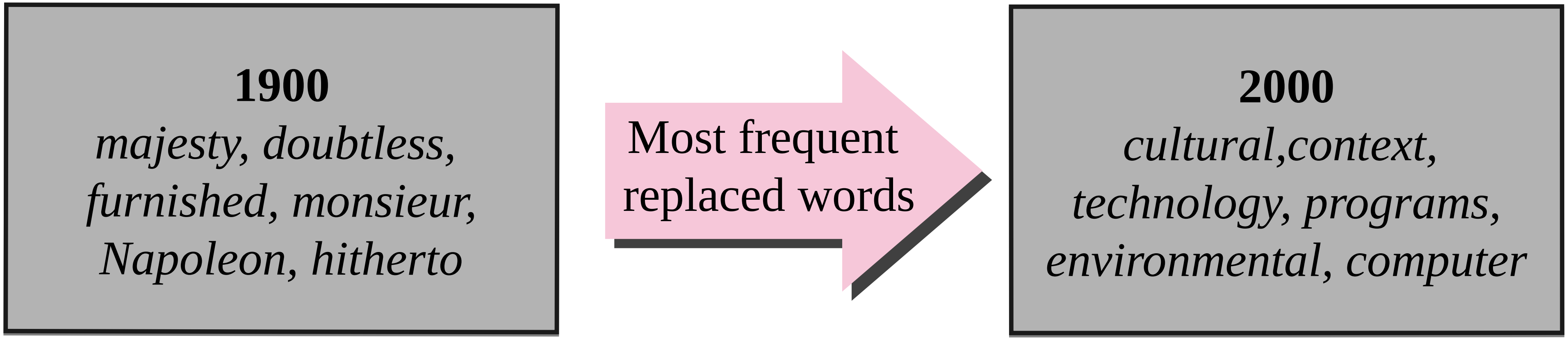


Change in the core vocabulary

Replacement in the core vocabulary:
Nc/κ≈30 words/year


Accelerating in time!

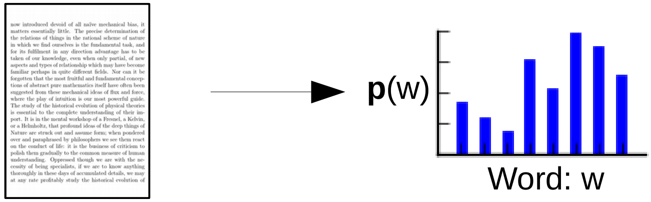


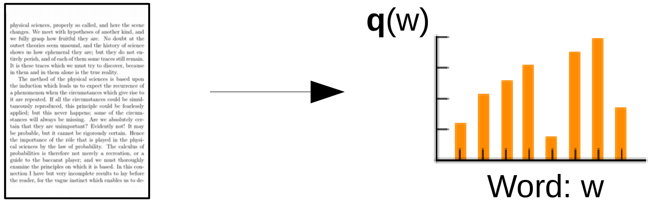

1900
majesty, doubtless, furnished, monsieur, Napoleon, hitherto

Most frequent replaced words

2000
cultural,context, technology, programs, environmental, computer

Vocabulary Distance





Generalized Jensen-Shannon Divergence

Jensen-Shannon Divergence

$$H_\alpha(H(\boldsymbol{p}) = \frac{1}{-}\sum_i p_i \log p_i 1 \Big)$$

$$H(\boldsymbol{p}) = -\sum_i p_i \log p_i$$

$$H_{\alpha=1}(\boldsymbol{p}) = H(\boldsymbol{p})$$

Havrda&Chrvát, Kybernetika (1967)

$$D_\alpha(\boldsymbol{p},\boldsymbol{q}) = H_\alpha\left(\frac{\boldsymbol{p}+\boldsymbol{q}}{2}\right) - \frac{1}{2}H_\alpha(\boldsymbol{p}) - \frac{1}{2}H_\alpha(\boldsymbol{q})$$

$$D(\boldsymbol{p},\boldsymbol{q}) = H\left(\frac{\boldsymbol{p}+\boldsymbol{q}}{2}\right) - \frac{1}{2}H(\boldsymbol{p}) - \frac{1}{2}H(\boldsymbol{q})$$

$$D_{\alpha=1}(\boldsymbol{p},\boldsymbol{q}) = D(\boldsymbol{p},\boldsymbol{q})$$

Burbea&Rao, IEEE TIT (1982)

→ $\sqrt{D_\alpha}$ is a metric for $\alpha \in [0,2]$

→ $\sqrt{D}$ is a metric

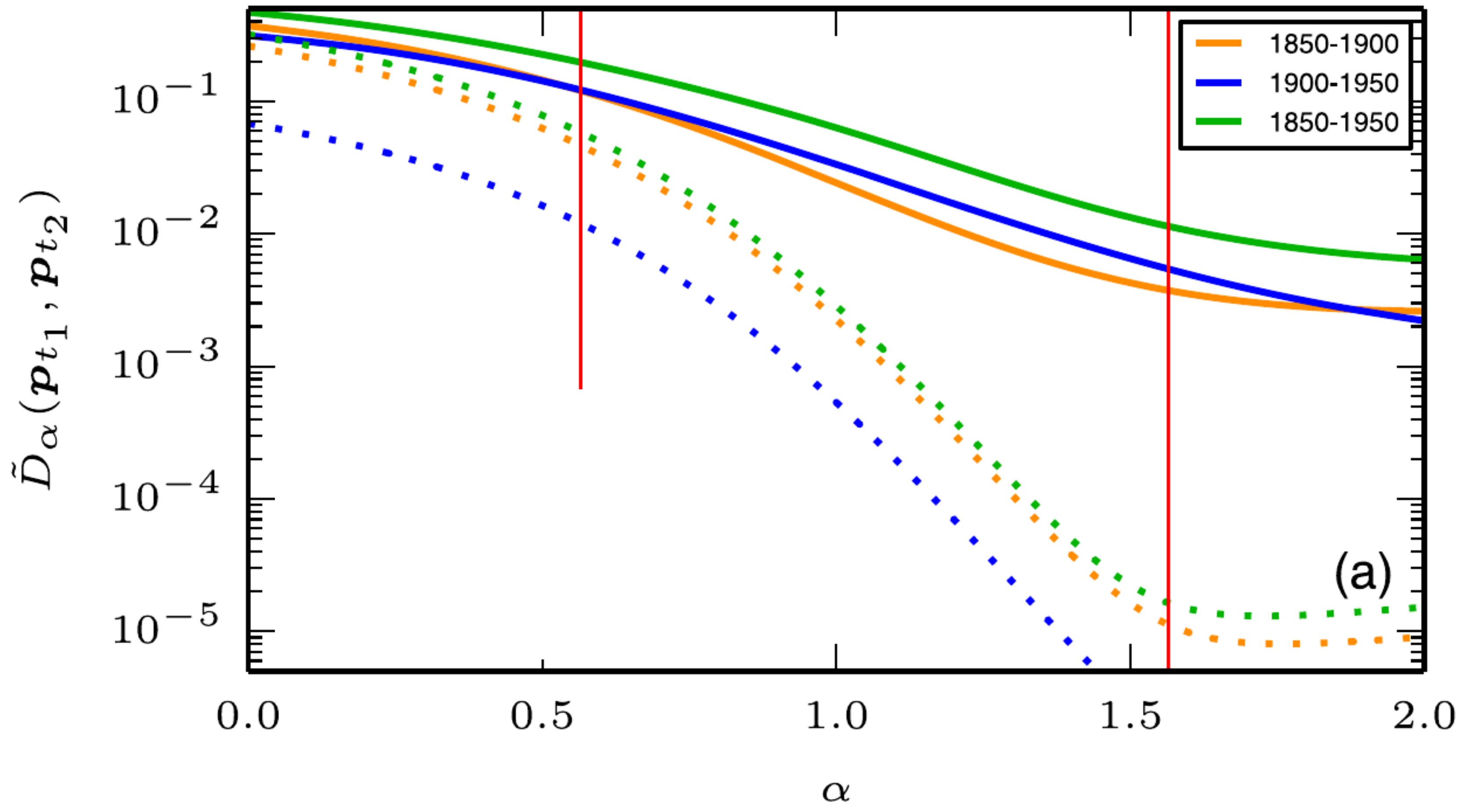
Briet *et al.*, Phys Rev A (2009)

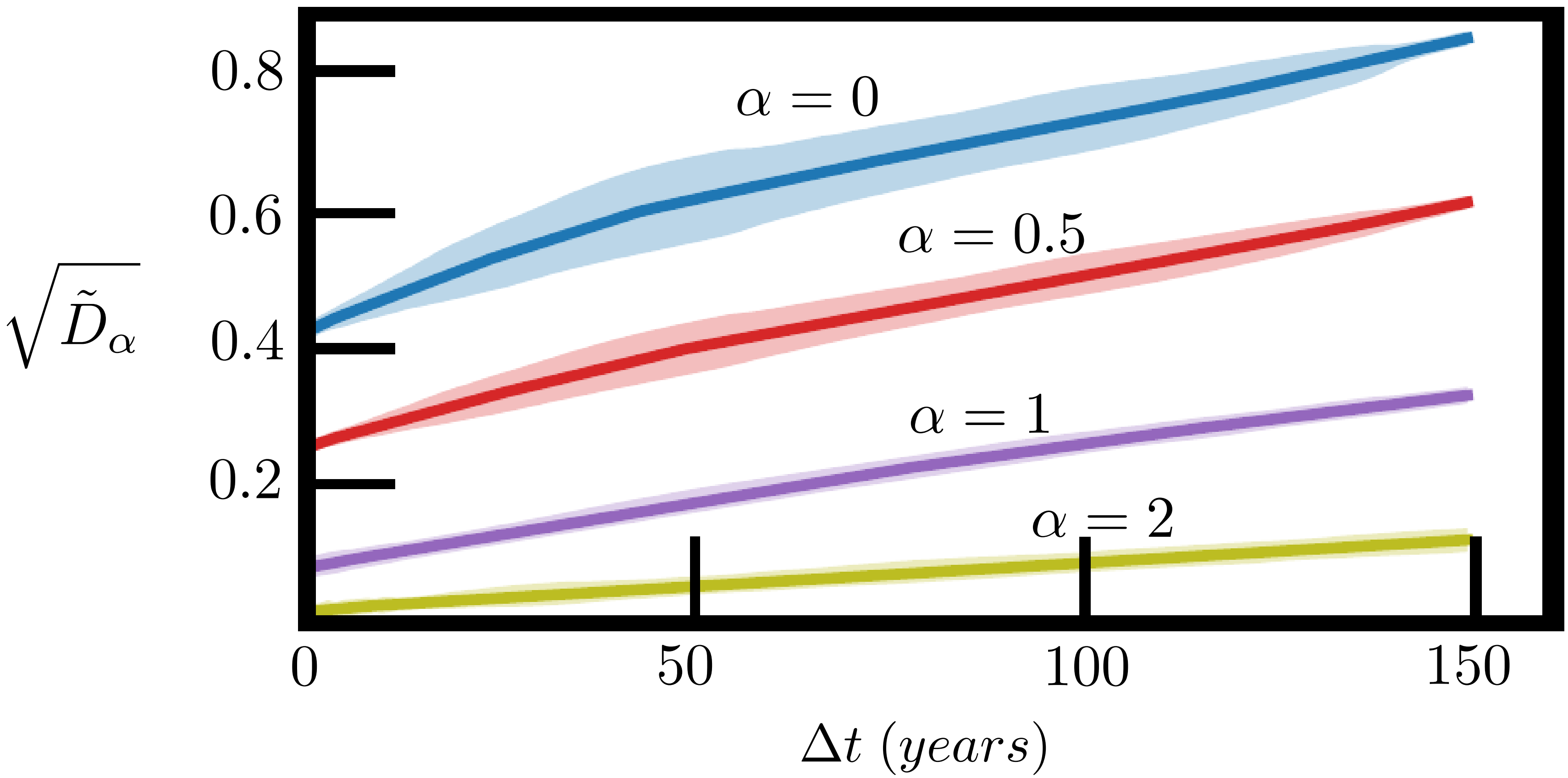
Slow convergence of statistical estimators due to Zipf's law: $F_r \sim r^{-\gamma}$

	H_α	D_α, $\tilde{D}_\alpha(\boldsymbol{p} \neq \boldsymbol{q})$	D_α, $\tilde{D}_\alpha(\boldsymbol{p} = \boldsymbol{q})$
Bias:	$V^{(\alpha)}/N$	$V^{(\alpha)}/N$	$V^{(\alpha)}/N$
Fluctuations:	$V^{(2\alpha)}/N$	$V^{(2\alpha)}/N$	$V^{(2\alpha-1)}/N^2$

$$V^{(\alpha)} \propto \begin{cases} N^{-\alpha+1+1/\gamma} & \alpha < 1+1/\gamma \\ \text{constant} & \alpha > 1+1/\gamma, \end{cases}$$

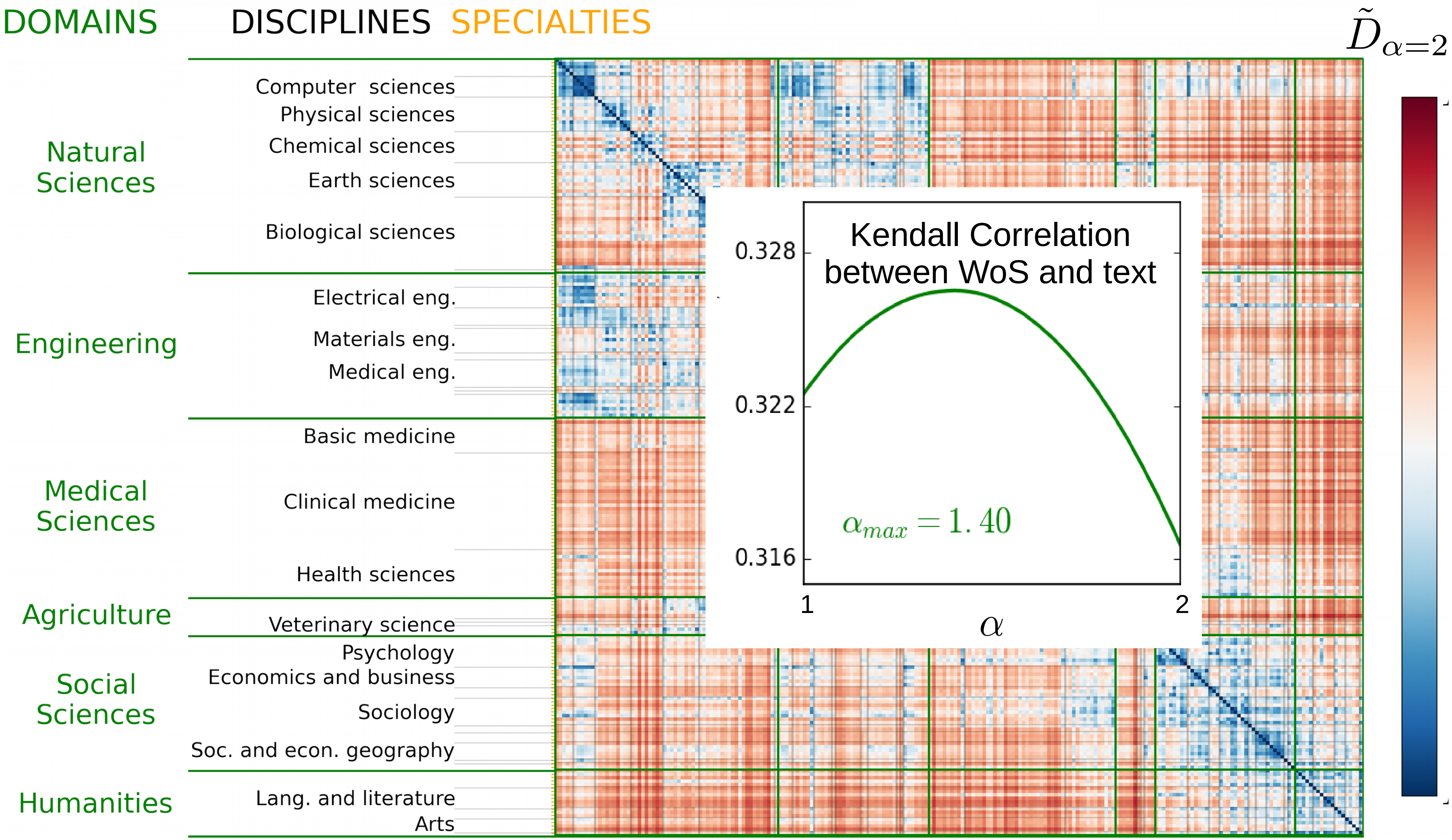
Change of English
(Google n-gram database 1520-2010)





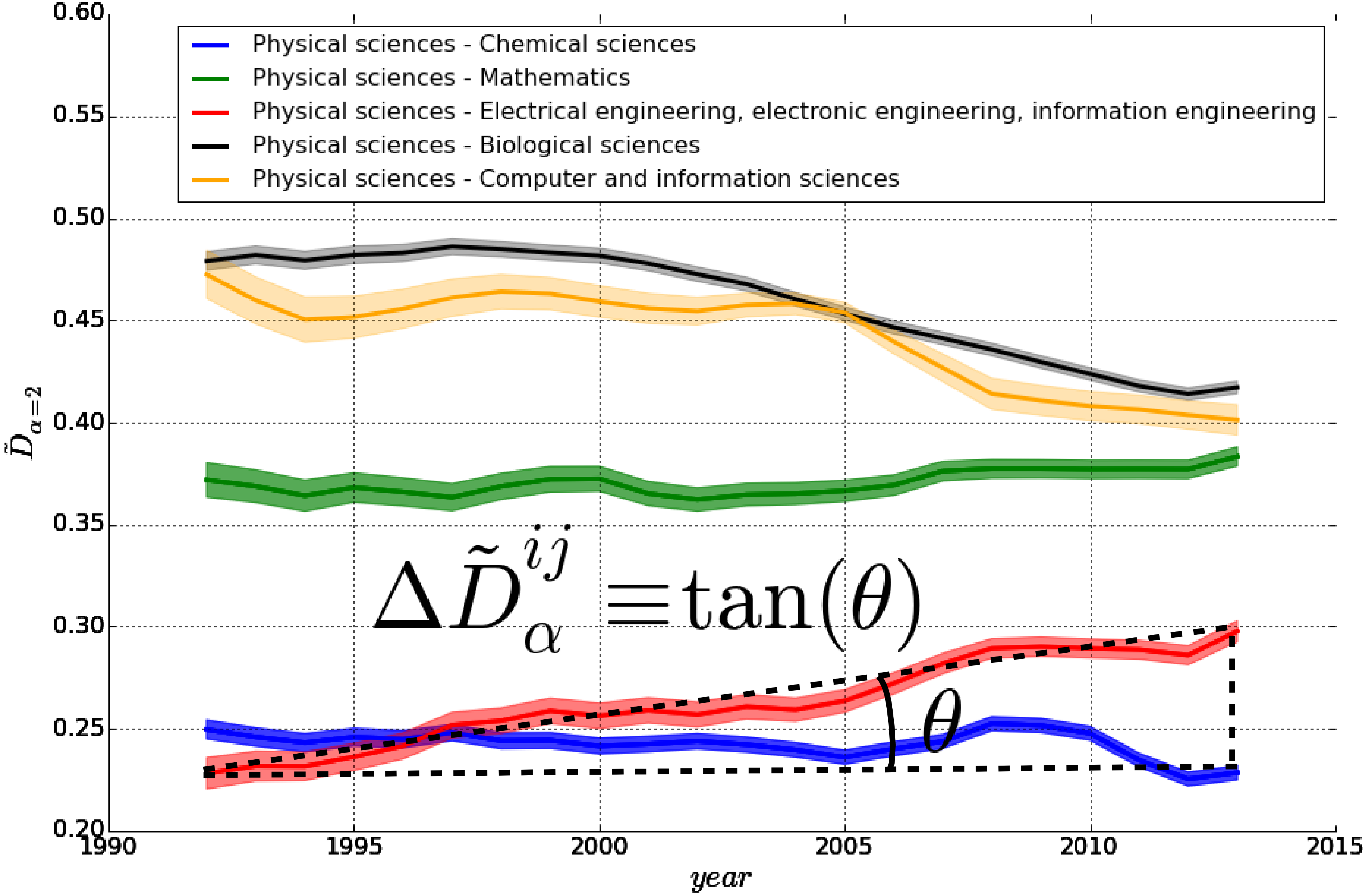
Change of English
(Google n-gram database 1520-2010)

Similarity of Scientific Disciplines
(title and abstract of all Web of Science papers 1990-2014)



Similarity of Scientific Disciplines

$\langle \Delta \tilde{D}_{\alpha}^{(i,j)} = 2 \rangle_{i,j} \approx 0$ Not significantly different from zero (T-test, p = 0.056; Wilcoxon test p = 0.17)

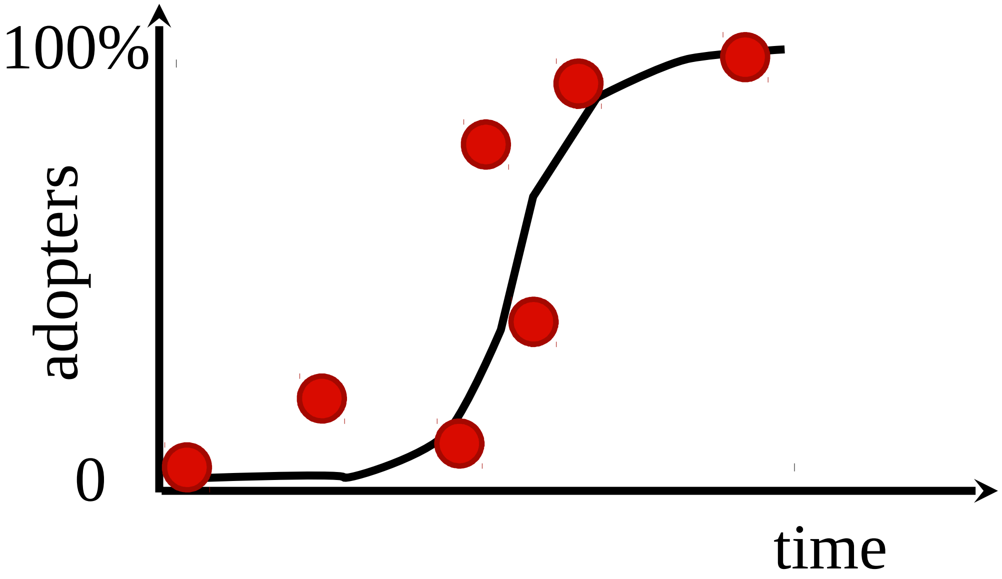


$$\Delta \tilde{D}_{\alpha}^{ij} \equiv \tan(\theta)$$

Adoption of new words

“The progress of language change through a community follows a lawful course, an ***S-curve*** *from minority to majority to totality.”*

Weinreich, Labov, Herzog, (1968)
Empirical foundations for a theory of language change



What is the empirical support?

“...up to a dozen points for a single change”

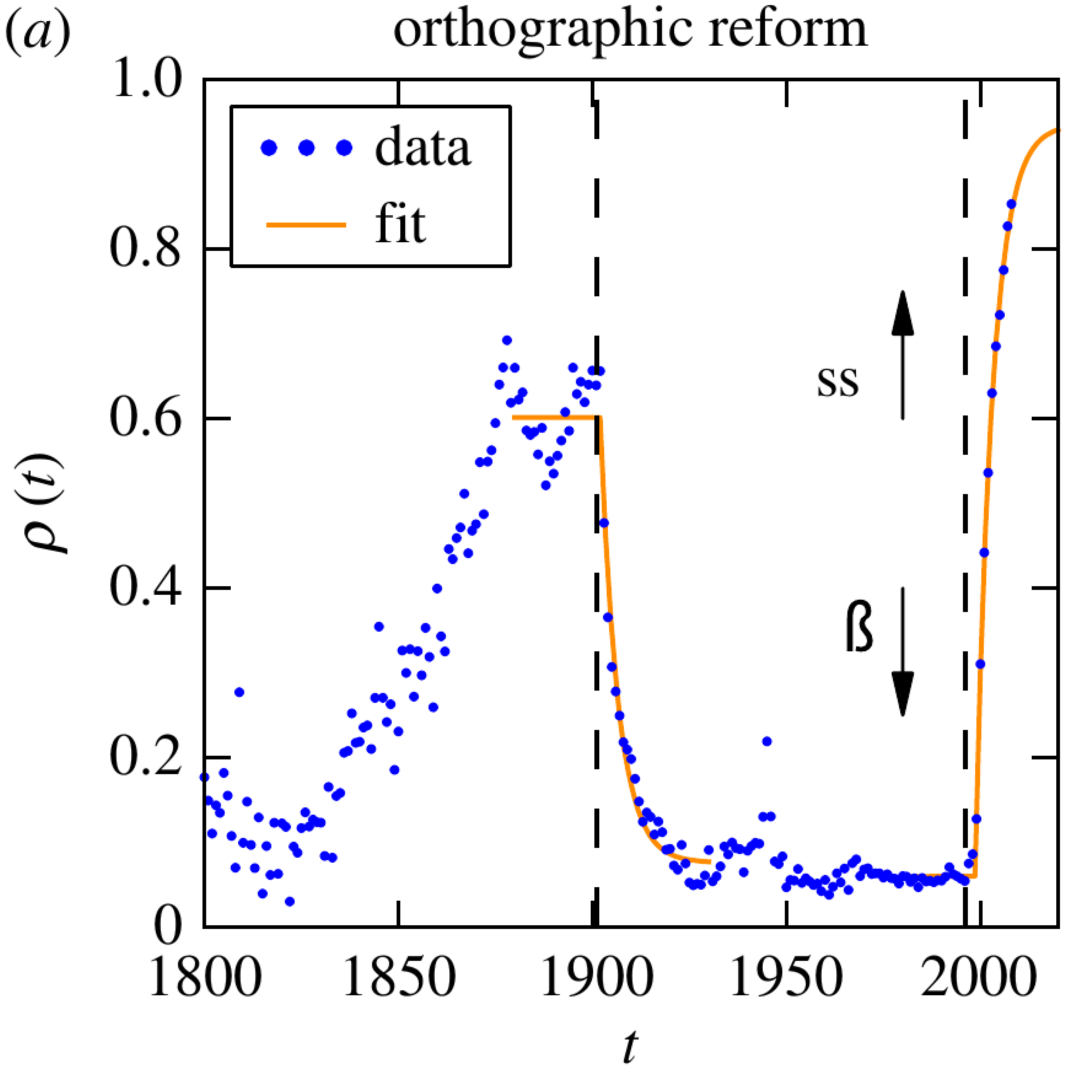
R. A. Blythe and W. Croft,
Language 88, 269 (2012)

- Are all changes following S-curves? No!
- Are all S-curves the same? No!
- Can we extract from S-curves information about the process of change? Yes!

Adoption of new words

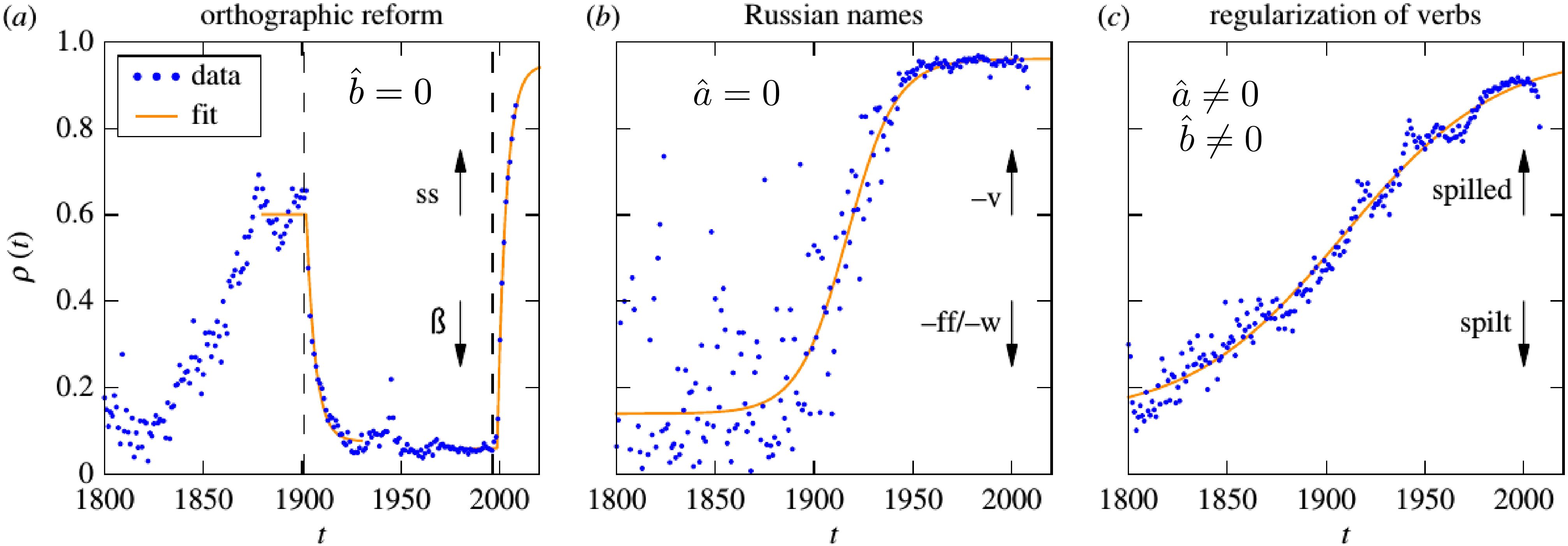
Ortography reform (1996): ß → ss

2,000 different words (e.g., Kongreß → Kongress)

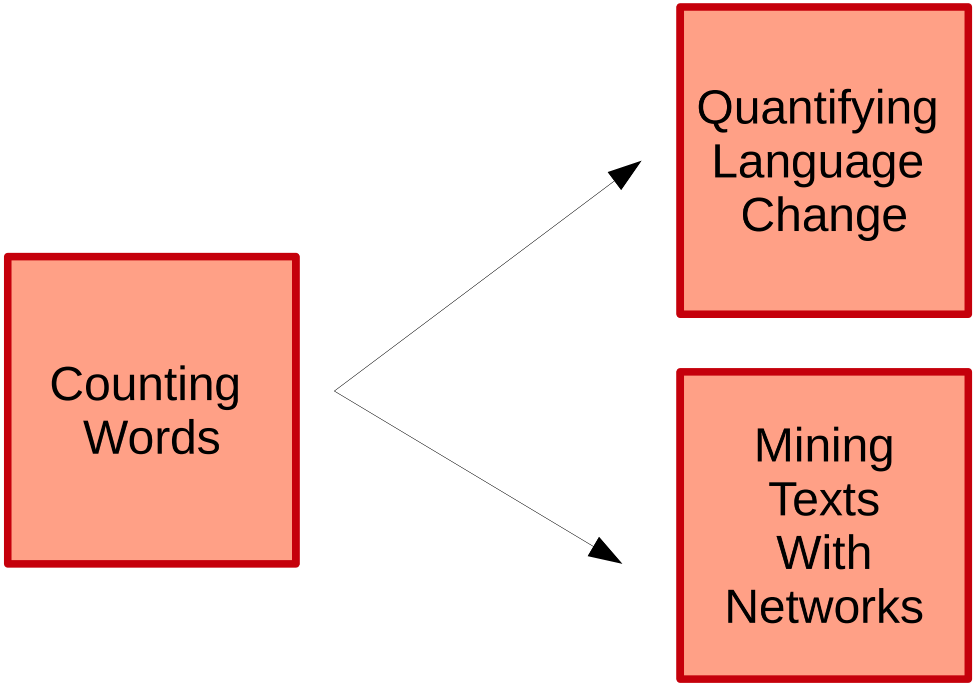


Best fit
=
exponential

Adoption of new words



$$\frac{d\rho(t)}{dt} = (a + b\,\rho(t))\,(1 - \rho(t)) \begin{cases} b = 0 \Rightarrow \rho(t) = \text{ exponential} \\ a = 0 \Rightarrow \rho(t) = \text{ symmetric S-curve} \end{cases}$$



Counting Words

Quantifying Language Change

Mining Texts With Networks

M. Gerlach, T. Peixoto, E. G. Altmann, *"A network approach to topic models", [arXiv:1708.01677]* .

Text mining

A

UNIVERSALIS DE JURE HOMINUM DECLARATIO

B

Cum dignitatis infixae omnibus humanae

C

Cum dignitatis infixae omnibus humanae

D

familiae partibus et eorum jurum aequalium,

…

Documents

Words

	A	B	C	D	…
the	156	85	111	35	56
of	59	65	75	33	40
…	...	...	...	...	...
science	0	5	2	0	0
sport	4	0	0	0	0
networks	2	0	0	0	0
physics	0	0	1	0	0
biology	0	0	0	5	0
	...	...	...	...	...

Topic Models

Documents

Words

	A	B	C	D	...
the	156	85	111	35	56
of	59	65	75	33	40
...	...	...	...	...	...
science	0	5	2	0	0
sport	4	0	0	0	0
networks	2	0	0	0	0
physics	0	0	1	0	0
biology	0	0	0	5	0
	...	...	...	...	...

$A_{\omega,d}$

=

Topics

Words

	1	2	3	K
the	2%	3%	2%	2
of	1%	0.2%		0.4%
...	...	...	...	...
science	0.05%	0	0.04%	0
sport	0	0.1%	0	0
networks	0.05%	0	0	0
physics	0.1%	0	0.005%	0
biology	0.001%	0	0.1%	0
	...	...	...	...

$\varphi_{j,w}$

*

Documents

Topics

	A	B	C	D	.
1		50 %	90 %	20 %	
2	80 %				
3	10 %	50 %		80 %	
K	10 %		10 %		

$\theta_{d,j}$

Latent Dirichlet Allocation (LDA)

Blei, Ng, Jordan (Journal of Machine Learning 2003), >20k citations
Implementation: McCallum's MALLET (http://mallet.cs.umass.edu)

- Fixed number of topics K
- Dirichlet Priors
- Inference problem:

$$P(Model|Data) = P(Data|Model)\frac{P(Model)}{P(Data)}$$

$$Data = A_{\omega,d}$$

$$Model = \{\varphi_{j,w}, \theta_{d,j}\}$$

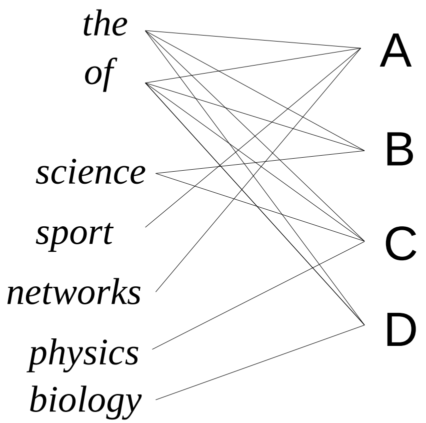
$$P(Model) = Prior = \begin{cases} \varphi_{j,w} \sim Dir(\beta) \\ \theta_{d,j} \sim Dir(\alpha) \end{cases}$$

Communities in Networks

Documents

Words	A	B	C	D	...
the	156	85	111	35	56
of	59	65	75	33	40
...	...	...	...	...	...
science	0	5	2	0	0
sport	4	0	0	0	0
networks	2	0	0	0	0
physics	0	0	1	0	0
biology	0	0	0	5	0
	...	...	...	...	...

=



=

	A	B	C	D	...	*the*	*of*	...
A	0	0	0	0		156	59	
B	0	0	0	0		85	65	
C	0	0	0	0		111	75	
D	0	0	0	0		35	33	
...					...			
the	156	85	111	35		0	0	
of	59	65	75	33		0	0	
...								

Connections to topic models: Ball, Karrer, Newman (2011), Lancichinetti et al (PRX 2014)

Stochastic Block Models (SBM)

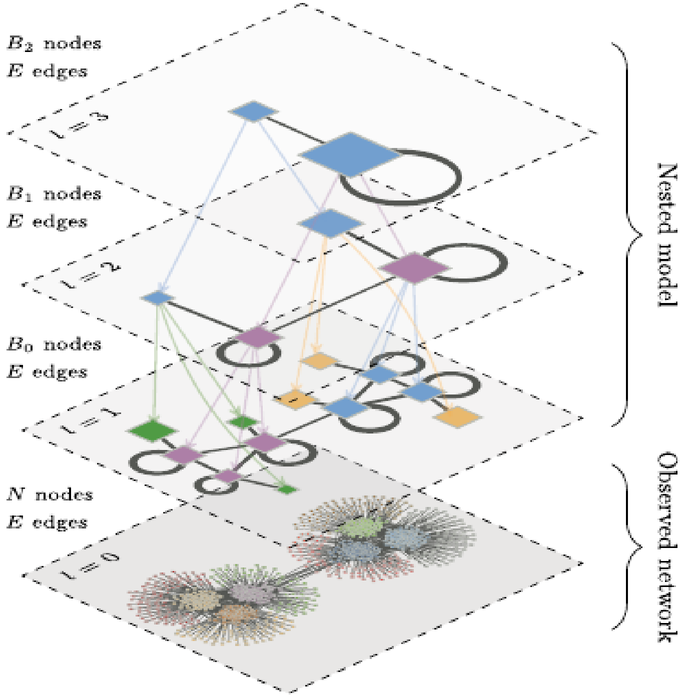
Holland, Laskey, Leinhardt (Social Networks 1983)

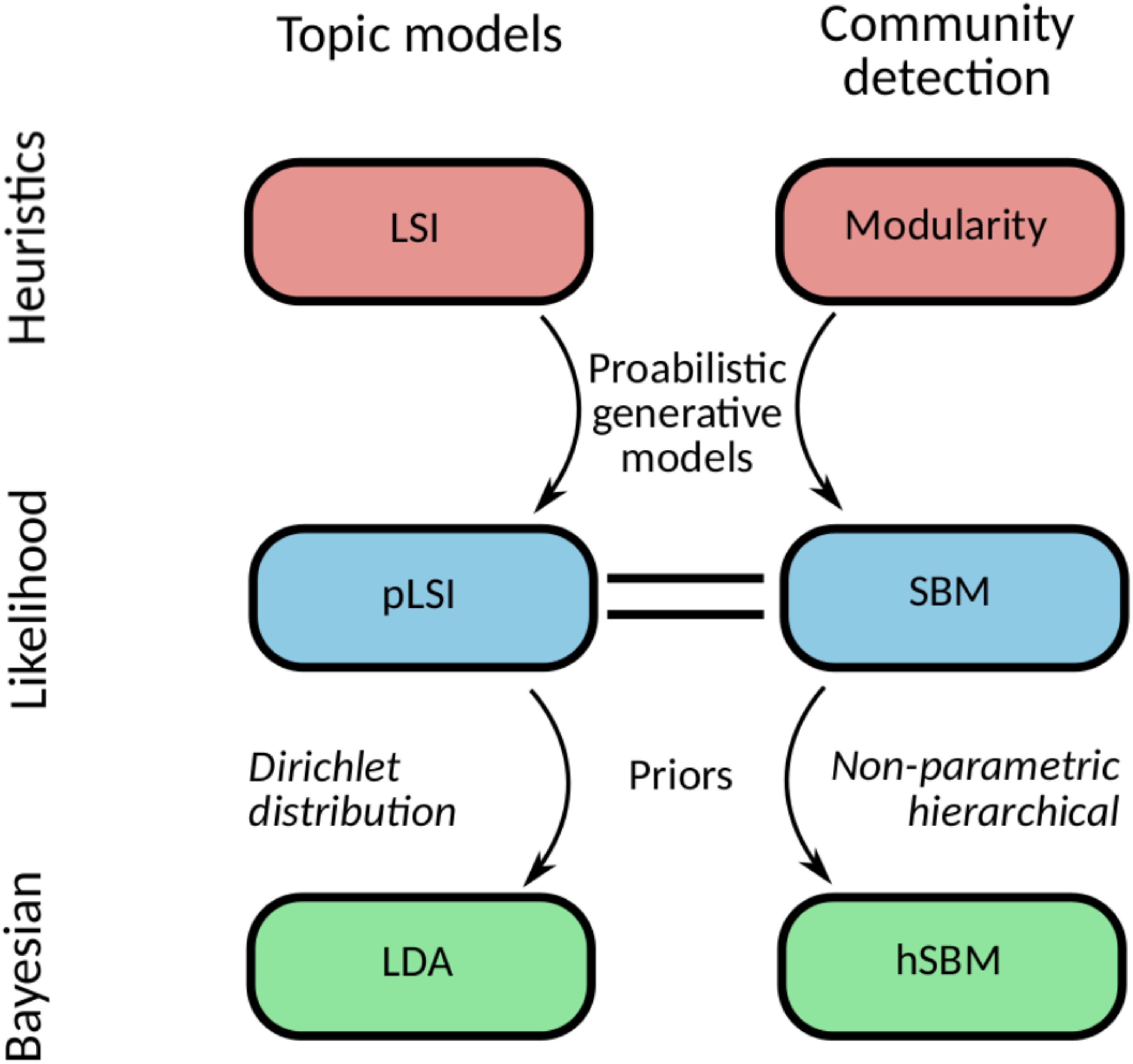
- Probability of connection between nodes depends on the blocks they belong
- Number of Blocks << Number of nodes (links)

Generative model: non-parametric hierarchical SBM

Peixoto (PRX 2014, PRX 2015, http://graph-tool.skewed.de**)**

- number of blocks (topics) not fixed
- prior at one level is set by the upper hierarchy level
- each link (word token in a document) is assigned to a pair of blocks





Topic models

Community detection

Heuristics

Likelihood

Bayesian

LSI

Modularity

Proabilistic generative models

pLSI

SBM

Dirichlet distribution

Priors

Non-parametric hierarchical

LDA

hSBM

Model Comparison
(between LDA and SBM)

Which model compacts better the data in terms of coding or description length (DL)?

Grünwald (*The Minimum Description Length Principle,2007)*

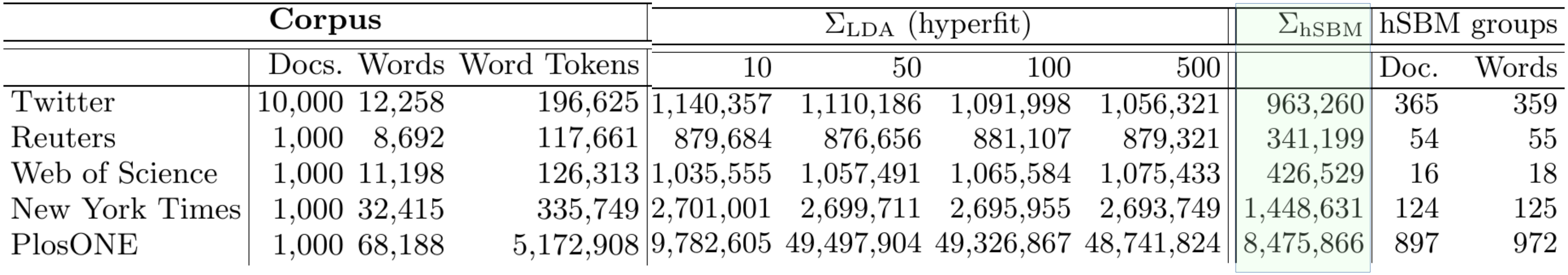
$$\Sigma = DL(\text{data}|\text{model}) + DL(\text{model})$$

Minimum description length (MDL) for probabilistic models:

$$\hat{\Sigma} = -\log P(D|\hat{\theta}) - \log P(\hat{\theta})$$

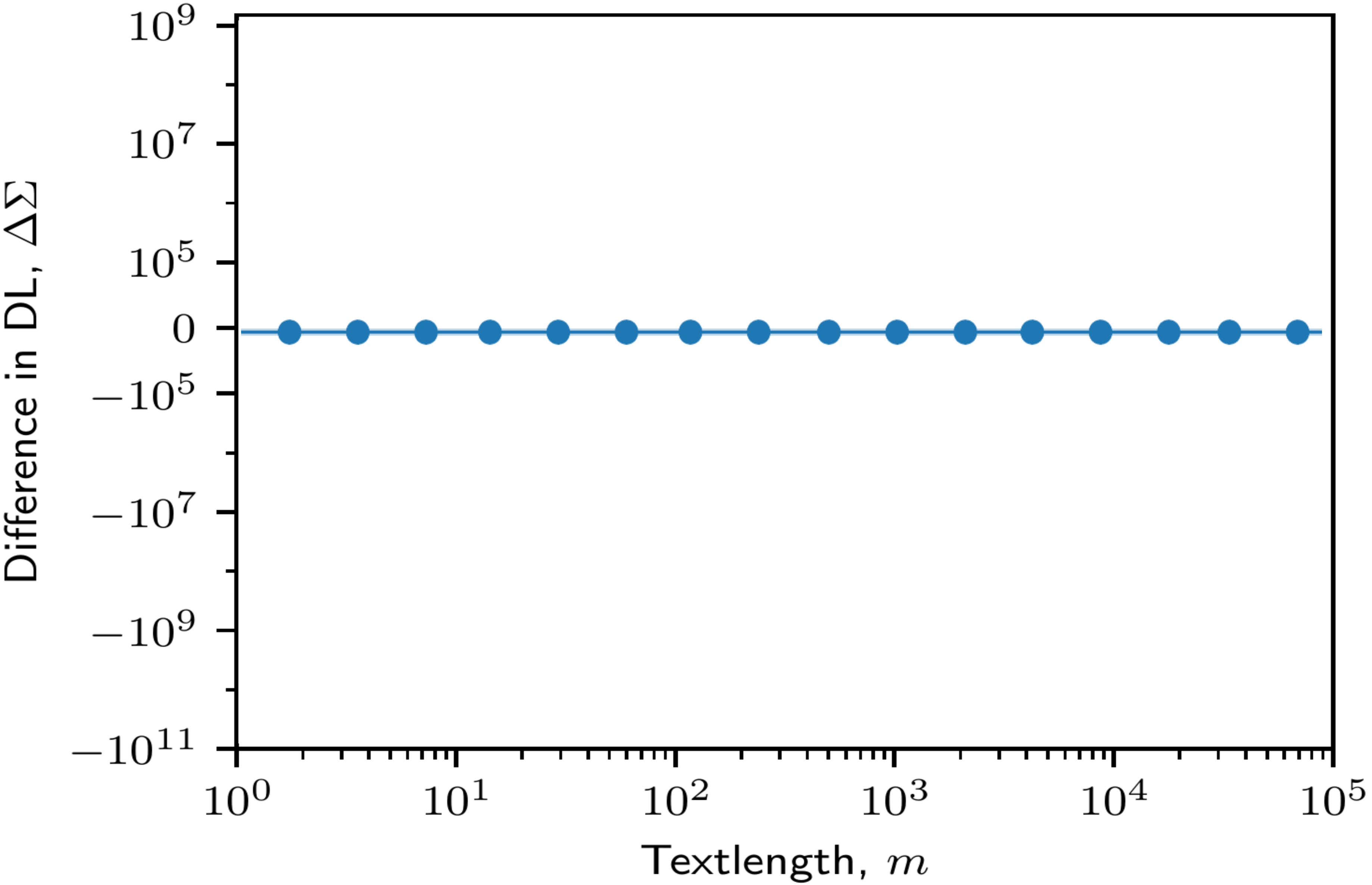
- D= data
- θ = discrete parameters of the model

$$\hat{\theta} = arg\,\underset{\theta}{max}\, P(D|\theta)P(\theta)$$

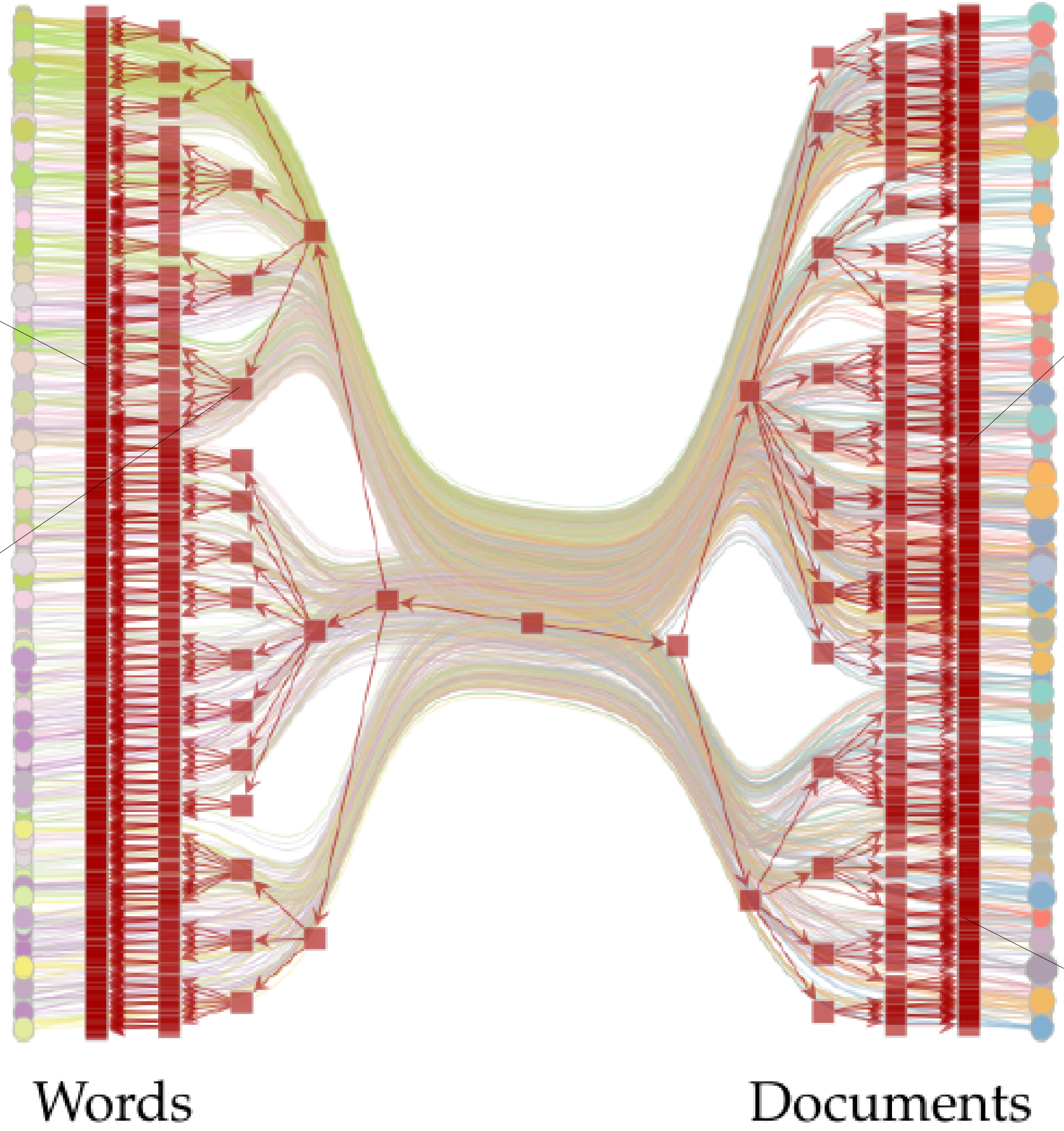


Corpus				Σ_{LDA} (hyperfit)				Σ_{hSBM}	hSBM groups	
	Docs.	Words	Word Tokens	10	50	100	500		Doc.	Words
Twitter	10,000	12,258	196,625	1,140,357	1,110,186	1,091,998	1,056,321	963,260	365	359
Reuters	1,000	8,692	117,661	879,684	876,656	881,107	879,321	341,199	54	55
Web of Science	1,000	11,198	126,313	1,035,555	1,057,491	1,065,584	1,075,433	426,529	16	18
New York Times	1,000	32,415	335,749	2,701,001	2,699,711	2,695,955	2,693,749	1,448,631	124	125
PlosONE	1,000	68,188	5,172,908	9,782,605	49,497,904	49,326,867	48,741,824	8,475,866	897	972

LDA generated documents:
10 topics, 1M documents, following Heaps' and Zipf's laws



Wikipedia Data



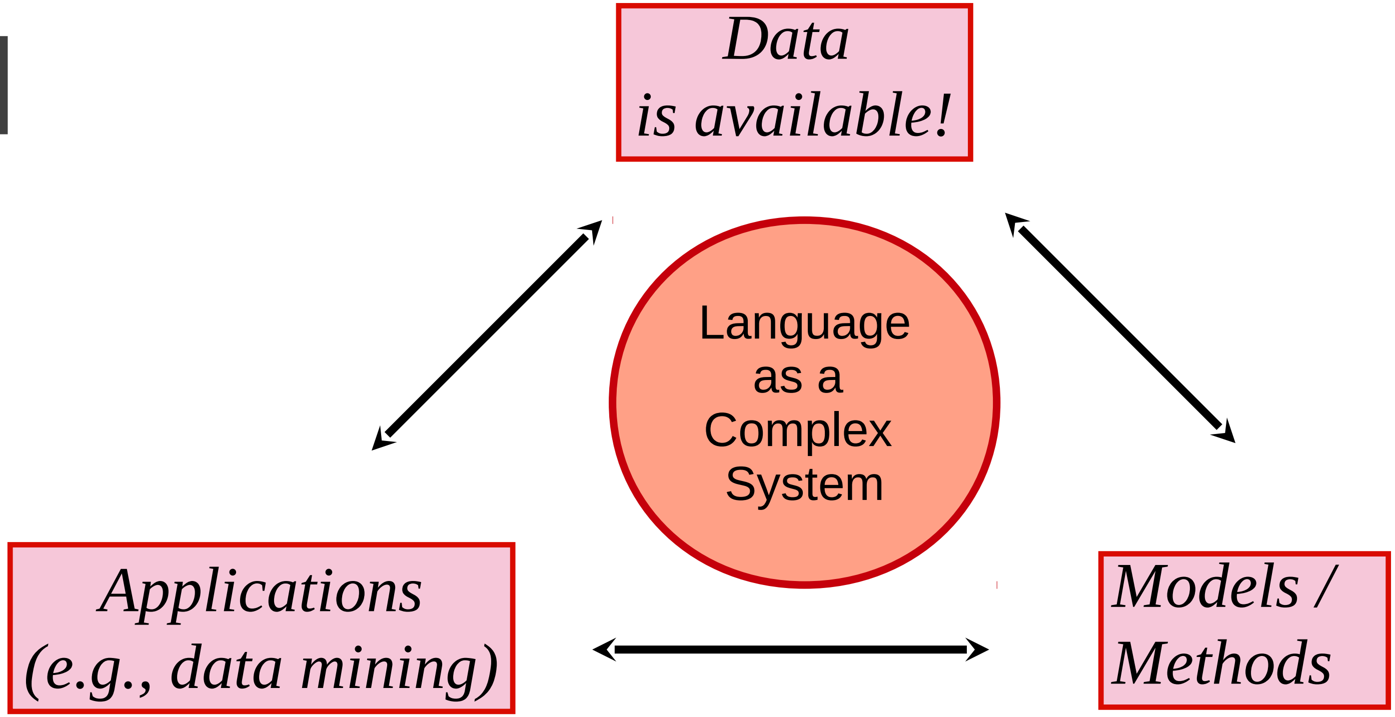
partner
partners
relational
repair
forgiveness
deception
transgression
infidelity
jealousy
transgressions

women
children
culture
person
cultural
psychology
men
music
core
mental

Assibilation
Structural_linguistics
Suffix
Text_simplification
Proprietor
Young’s_Analytical_
_Concordance_to_the_Bible
Loculus_(architecture)
Inverse_copular_constructions
Affection_(linguistics)
International_Nonproprietary_
_Name

Duality_(electricity_and...
Couple_(mechanics)
Invariant_mass
Lorentz_force
Polhode
Bertrand’s_theorem
Versorium
Movement_parameter
Angular_velocity
Gravitation

Conclusions



Data is available!

Language as a Complex System

Applications (e.g., data mining)

Models / Methods

Thank you for your attention!

E. G. Altmann, G. Cristadoro, and M. Degli Esposti, *"On the origin of long-range correlations in texts"*, PNAS (2012)

F. Ghanbarnejad, M. Gerlach, J. M. Miotto, and E. G. Altmann, *"Extracting information from S-curves of language change"*, J. Royal Soc. Interface (2014)

M. Gerlach & E. G. Altmann, *"Scaling laws and fluctuations in the statistics of word frequencies"*, New J. Phys. (2014)

M. Gerlach & E. G. Altmann, *"Stochastic model for the vocabulary growth in natural languages"*, Phys. Rev. X (2013)

E. G. Altmann & M. Gerlach *"Statistical Laws in Linguistics"*, *Chap. in Creativity and Universality in Language (2016)*

M. Gerlach, F. Font-Clos, E. G. Altmann, *“On the similarity of symbol-frequency distributions with heavy tails",* Phys. Rev. X (2016)

L. Dias, M. Gerlach, J. Scharloth, and E. G. Altmann, “Using text analaysis to quantify the similarity of scientific disciplines”, [arXiv:1706.08671] .

M. Gerlach, T. Peixoto, E. G. Altmann, *"A network approach to topic models”, [arXiv:1708.01677]* .