Stochastic model for phonemes

Armen Allahverdyan and Weibing Deng

(*Yerevan Physics Institute*) (*Complexity Science Center, Wuhan*)

-- Introduction: What is phoneme ?

-- Ideal gas model for phoneme frequencies in a text

-- Author dependency: each one uses phonemes in his/her own way
PLOS ONE (2016)

-- Other work done by physicists on language

http://www.pks.mpg.de/mpi-doc/sodyn/physicist-language/

Phoneme: a group of sounds; minimal unit related to meaning in a given language

(Dufriche-Desgenettes, B de Courtenay, Trubetskoy, Scherba, Sapir)

r and *l*

USA: [roʊ] [loʊ]

row and *low*

British: [rəʊ] [ləʊ]

Varying sound within phoneme → no changes in meaning

Changing one phoneme to another **can** change the meaning

44 English phonemes; some of them are specific for English

θ (the), ð (think)

The concept of phoneme emerged independently in Greek and Indian traditions together with ideas of atomisms (*Aristotle, Panini*)

Staal 2006, Skoyles 1990

Atom-morpheme metaphor

You pronounce sound not phoneme --> no psychological reality?

(psycholinguistics, experimental studies)

Ear hears sound, brain hears phoneme ???

Analogy with the debate on the meaning of the wave-function in quantum mechanics: epistemic or ontologic ?

Checking the atom-morpheme metaphor

$f_1, \ldots, f_n$ Phoneme frequencies extracted from words of a given text

We extracted phoneme frequencies from texts by native-English authors:
Ch.Darwin, H. Spencer, J. Austen, C. Dickens, J. Tolkien,…….

The idea of statistical physics observables: $f_1, \ldots, f_n$ come from averaging over simple probability density → **ideal gas**

Dirichlet density

$$\mathcal{D}_\beta(\theta_1, ..., \theta_n) \propto \delta\left(\sum_{k=1}^{n} \theta_k - 1\right) \prod_{k=1}^{n} \theta_k^{\beta-1}$$

$$e^{(\beta-1)\ln \theta_k}$$

subset $(\theta_1, ..., \theta_m)$ $(m < n)$ of probabilities $(\theta_1, ..., \theta_n)$.

$$\widehat{\theta}_k = \frac{\theta_k}{\sum_{i=1}^{m} \theta_i}, \quad k = 1, ..., m.$$

Independence defines the Dirichlet

$$\mathcal{P}(\widetilde{\theta}_1, ..., \widetilde{\theta}_n) = \mathcal{D}_\beta(\hat{\theta}_1, ..., \hat{\theta}_m)\, \mathcal{X}(\theta_{m+1}, ..., \theta_n)$$

similar to inverse temperature of ideal (non-interacting atoms) gas

$\beta > 0$

Rank-frequency relation for phonemes.

$$f_1 \geq f_2 \geq \ldots \geq f_n$$

observed ranked frequencies in each text

$$\theta_1,\ldots,\theta_n \longrightarrow \theta_{(1)} > \ldots > \theta_{(n)}$$

ordered distribution from Dirichlet

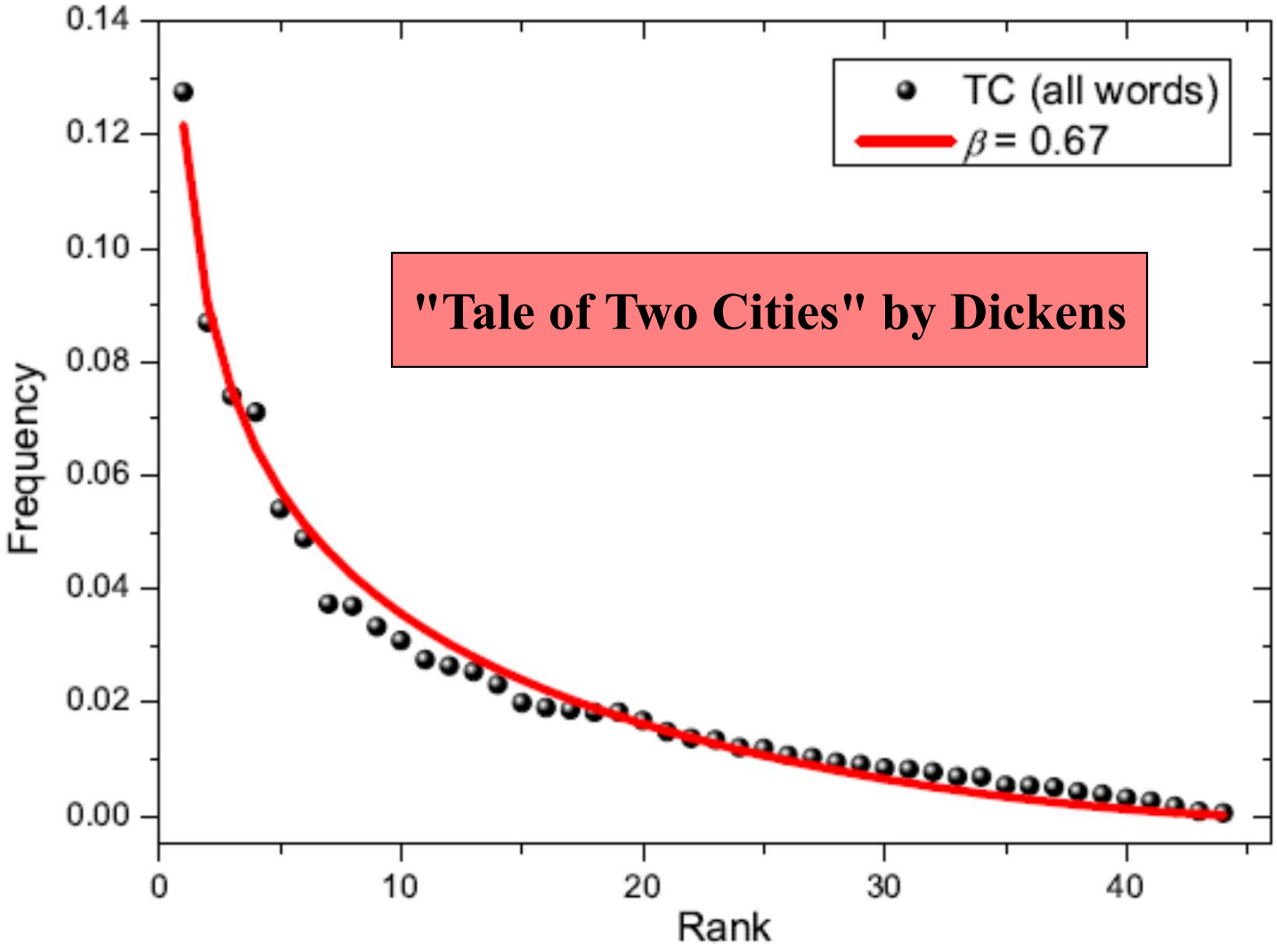
$$f_r \approx \left\langle \theta_{(r)} \right\rangle$$

sought relation

$$\frac{r}{n} = 1 - \varphi\left(\left\langle \theta_{(r)} \right\rangle n\beta \right)$$

theoretical rank-frequency relation

$$\varphi(y) = \frac{1}{\Gamma[\beta]} \int_0^y \mathrm{d}x\, x^{\beta-1} e^{-x}$$



β → **is found from the best fit**

Each author has its own "temperature" β that is stable across his/her texts

C. Lyell	0.798	0.785	0.792
A. R. Wallace	0.744	0.756	0.739
C. Darwin	0.817	0.810	0.822
H. Spenser	0.646	0.658	0.650
H. G. Wells	0.737	0.735	0.724

$$\min[|\beta_D - \beta_S|] > \max[|\beta_D - \beta_{D'}|]$$

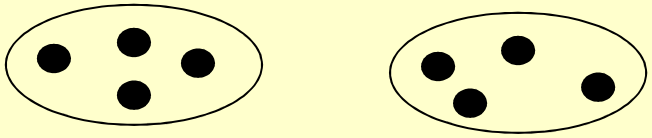
Word frequencies hold the universal Zipf's law

J. Estoup 1916, G. Zipf 1949

No author-dependency

$$f_r \propto r^{-1}$$

Clustering for distances



$$\min[\,\rho_{DS}\,] > \max[\,\rho_{D\,D'}\,]$$

$$\rho_{i\,j} = \frac{1}{2}\sum_{k=1}^{n} |\, f_k[i] - f_k[j]\,|$$

Variational distance
for frequencies

The effect holds for other types of distances and for phoneme frequencies
extracted from all or different words of the text

Caused by the same vocabulary employed by an author in his/her texts ?

$$\rho_{ij} = \frac{1}{2}\sum_{k=1}^{n} | f_k[i] - f_k[j] |$$

When comparng two texts extract phoneme frequencies from their <u>different words only</u>

$$\min[\rho_{DS}] > \max[\rho_{DD'}]$$

The effect survives and does not become weaker

Summary

Phoneme frequencies are described by the ideal-gas (Dirichlet) model

The atom-phoneme metaphor works

The parameter ("temperature") shows author-dependency

Confirmed by other means (distances)

The effect is not due to vocabulary and is intrinsic to phonemes

Points towards psychological reality of phonemes

Can be employed for authorship attribution