Language statistics at different temporal, geographical, and grammatical scales

October 31, 2019

Abstract

***

1 Introduction

The statistical study of languages is not new [1], but recent data availability has allowed the possibility of analyzing how languages and their use has changed in time [2, 3, 4, 5]. These studies have considered language change during years and centuries. What can we say about change — not so much of language itself, but of its use — during hours and days? To address this question, we use geolocated Twitter data to compare languages at short timescales, at different spatial scales, and at different "grammatical scales".

2 Methods and Data

Explain rank diversity, include an illustrative spaghetti figure, including marks for Δt. [5, 6, 7, 8] We have found that rank diversity curves for six different Indoeuropean languages are very similar, as they can be fitted with a sigmoid curve with small differences between languages.

Explain data: how it was collected (Alfredo), papers where this data has been used [9, 10, 11]

Data statistics (how many tweets/words/Ngrams), languages, how it was processed (to keep same number of tweets), etc.

Explain scales: temporal, geographical, grammatical.

already studies of grammatical scale for Google Books [7]. We have found that the grammatical scale varies language statistics (rank diversity, change probability, rank entropy and rank complexity) more than changes of language. In other words, a change of grammatical scale implies a greater change in the statistics than a change of language.

3 Results

Explain figures

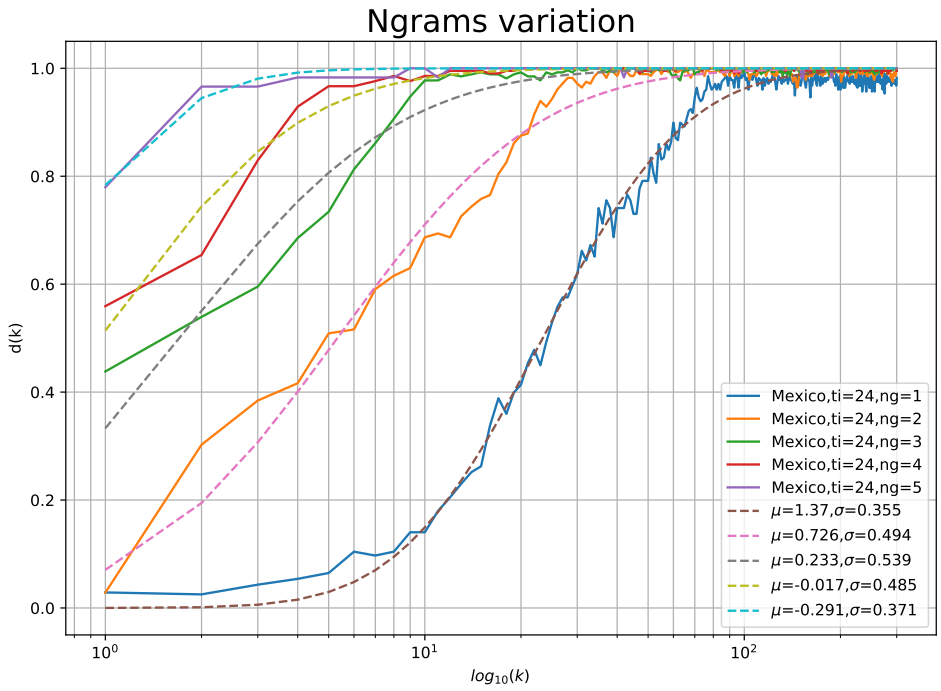

Figure 1: Rank diversities from Mexico for different grammatical scales, $\Delta t = 24$hr.

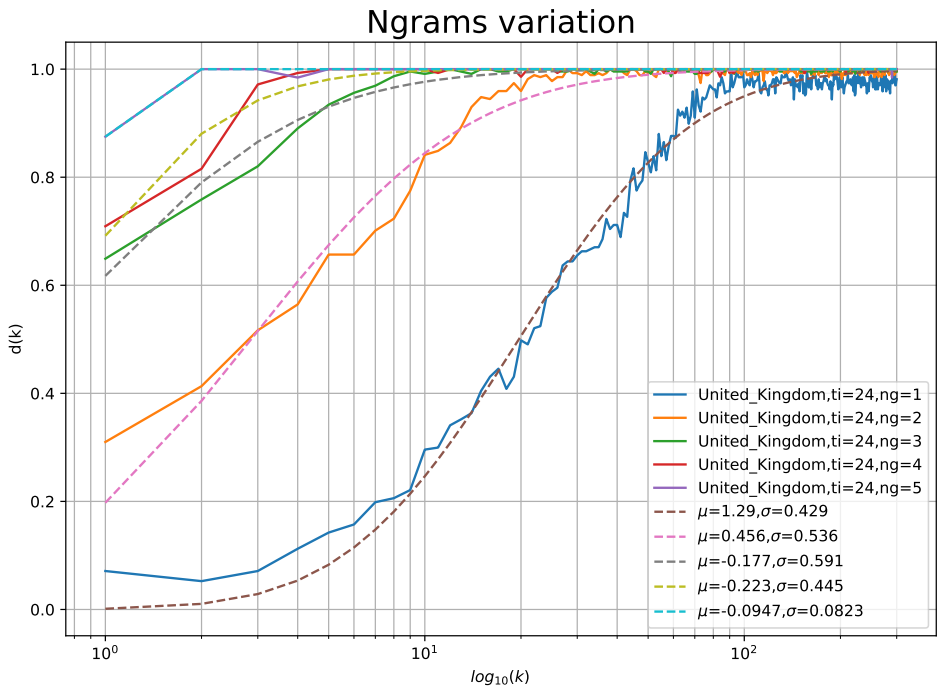

Figure 2: Rank diversities from the United Kingdom for different grammatical scales, $\Delta t =$ 24hr.

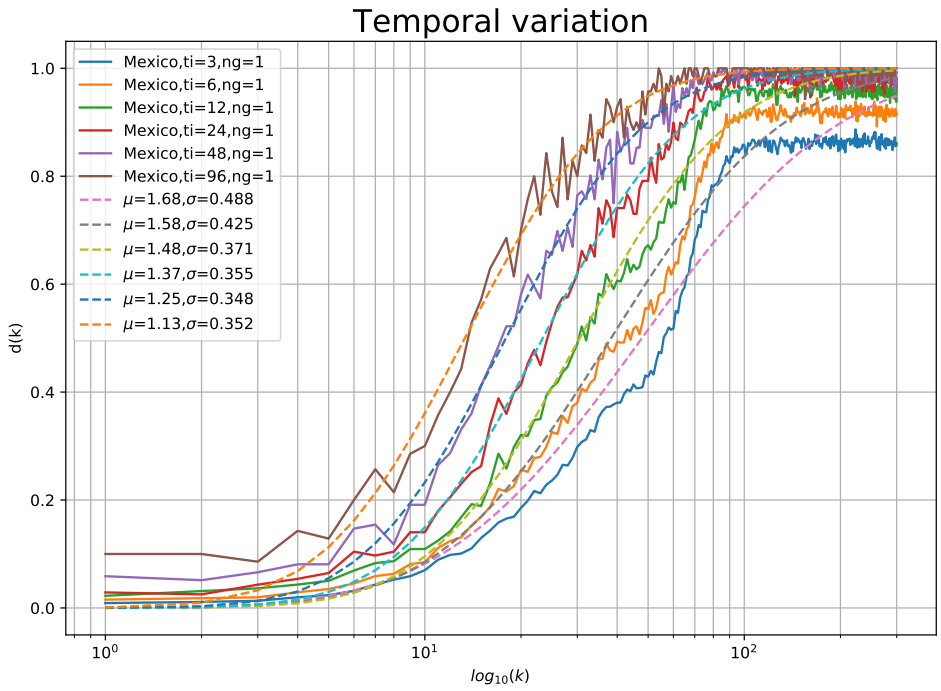

Figure 3: Rank diversities from Mexico for different temporal scales, $N = 1$.

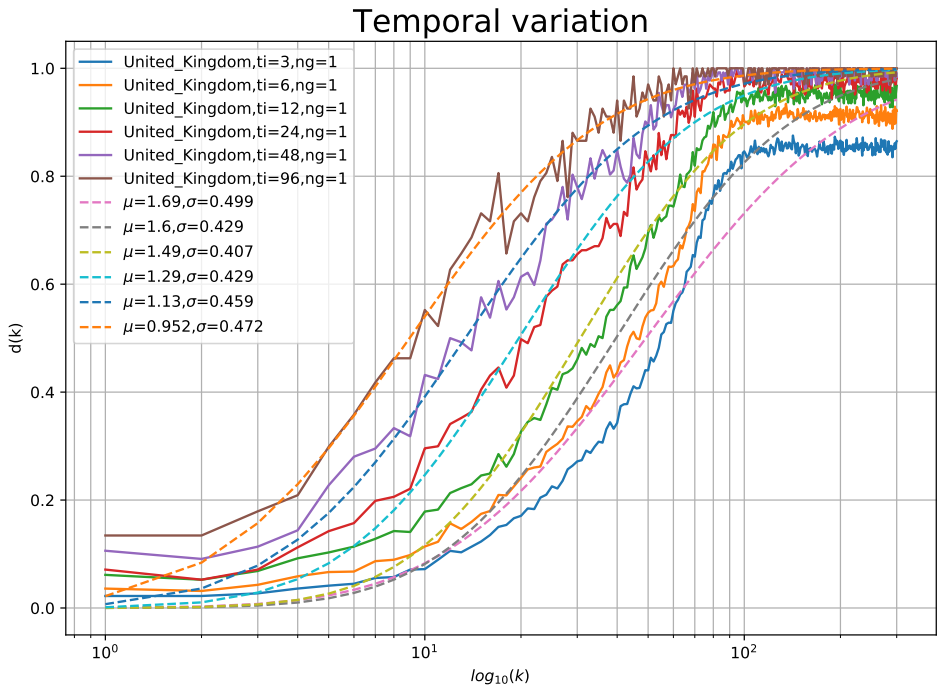

Figure 4: Rank diversities from the United Kingdom for different temporal scales, $N = 1$.

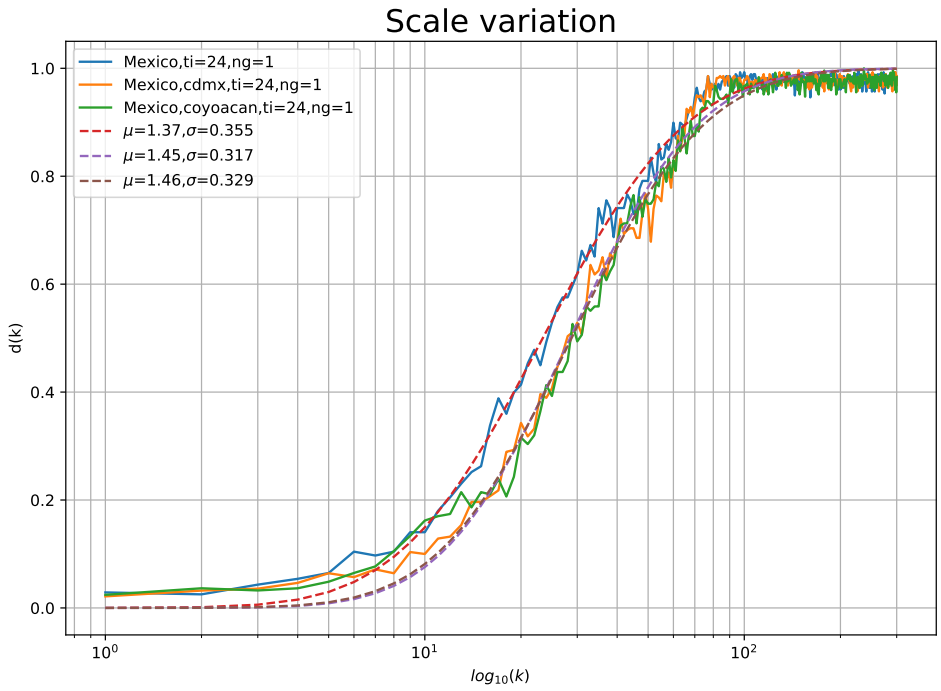

Figure 5: Rank diversities from Mexico for different geographical scales, $N = 1$, $\Delta t = 24$hr.

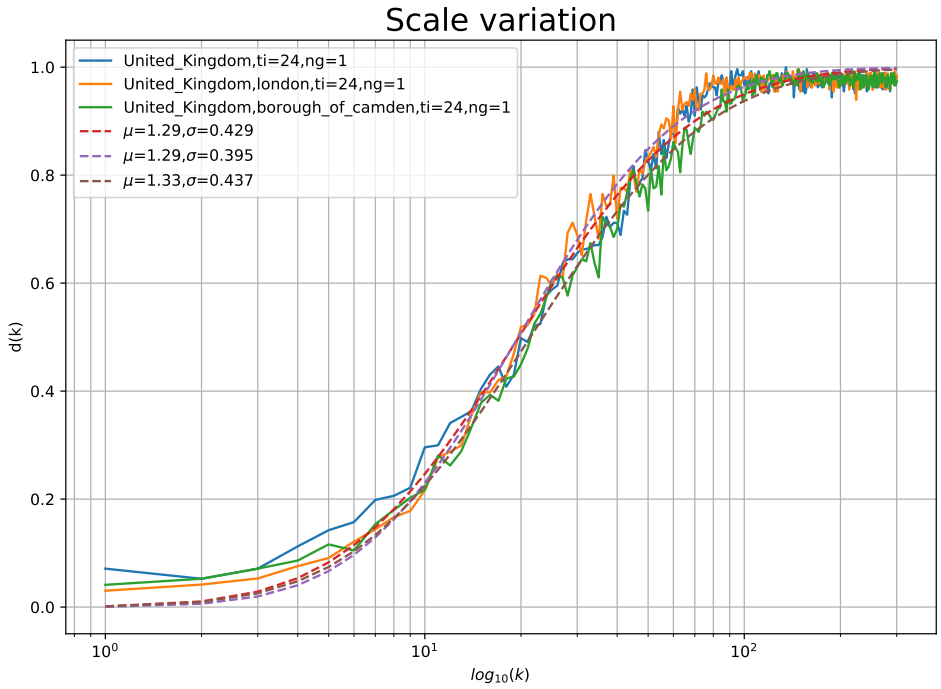

Figure 6: Rank diversities from the United Kingdom for different geographical scales, $N = 1$, $\Delta t = 24$hr.

scales vs. μ's

4 Conclusions

Which scales are more relevant? All of them, but grammatical > temporal > geographical. In general, rank diversity grows faster at higher scales in all three cases.

Acknowledgments

References

[1] G. K. Zipf. *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press, Cambridge, MA, USA, 1932.

[2] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[3] Matjaž Perc. Evolution of the most common English words and phrases over the centuries. *Journal of The Royal Society Interface*, 9(77):3323–3328, 2012.

[4] Martin Gerlach and Eduardo G. Altmann. Stochastic model for the vocabulary growth in natural languages. *Phys. Rev. X*, 3:021006, May 2013.

[5] Germinal Cocho, Jorge Flores, Carlos Gershenson, Carlos Pineda, and Sergio Sánchez. Rank diversity of languages: Generic behavior in computational linguistics. *PLoS ONE*, 10(4):e0121898, 04 2015.

[6] José A. Morales, Sergio Sánchez, Jorge Flores, Carlos Pineda, Carlos Gershenson, Germinal Cocho, Jerónimo Zizumbo, Rosalío F. Rodríguez, and Gerardo Iñiguez. Generic temporal features of performance rankings in sports and games. *EPJ Data Science*, 5(1):33, 2016.

[7] José A. Morales, Ewan Colman, Sergio Sánchez, Fernanda Sánchez-Puig, Carlos Pineda, Gerardo Iñiguez, Germinal Cocho, Jorge Flores, and Carlos Gershenson. Rank dynamics of word usage at multiple scales. *Frontiers in Physics*, 6:45, 2018.

[8] Germinal Cocho, Rosalío F. Rodríguez, Sergio Sánchez, Jorge Flores, Carlos Pineda, and Carlos Gershenson. Rank-frequency distribution of natural languages: A difference of probabilities approach. *Physica A: Statistical Mechanics and its Applications*, 532:121795, 2019.

[9] A. J. Morales, J. Borondo, J. C. Losada, and R. M. Benito. Efficiency of human activity on information spreading on twitter. *Social Networks*, 39:1–11, 2014.

[10] A. J. Morales, J. Borondo, J. C. Losada, and R. M. Benito. Measuring political polarization: Twitter shows the two sides of venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(3):033114, 2015.

[11] Alfredo J. Morales, Vaibhav Vavilala, Rosa M. Benito, and Yaneer Bar-Yam. Global patterns of synchronization in human communications. *Journal of The Royal Society Interface*, 14(128):20161048, 2017.