entropy

Information Theory and Language

Edited by

Łukasz Dębowski and Christian Bentz

Printed Edition of the Special Issue Published in *Entropy*

www.mdpi.com/journal/entropy

MDPI

Information Theory and Language

Information Theory and Language

Special Issue Editors

Łukasz Dębowski
Christian Bentz

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade

Special Issue Editors

Łukasz Dębowski
Polish Academy of Sciences
Poland

Christian Bentz
University of Zürich
Switzerland
University of Tübingen
Germany

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Entropy* (ISSN 1099-4300) from 2019 to 2020 (available at: https://www.mdpi.com/journal/entropy/special_issues/inf_theory_Lang).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, *Article Number*, Page Range.

ISBN 978-3-03936-026-0 (Hbk)
ISBN 978-3-03936-027-7 (PDF)

Contents

About the Special Issue Editors . . . . **vii**

Łukasz Dębowski and Christian Bentz
Information Theory and Language
Reprinted from: *Entropy* **2020**, *22*, 435, doi:10.3390/e22040435 . . . . **1**

Álvaro Corral and Isabel Serra
The Brevity Law as a Scaling Law, and a Possible Origin of Zipf's Law for Word Frequencies
Reprinted from: *Entropy* **2020**, *22*, 224, doi:10.3390/e22020224 . . . . **7**

Lida Ahmadi and Mark Daniel Ward
Asymptotic Analysis of the kth Subword Complexity
Reprinted from: *Entropy* **2020**, *22*, 207, doi:10.3390/e22020207 . . . . **21**

Luís F Seoane and Ricard Solé
Criticality in Pareto Optimal Grammars?
Reprinted from: *Entropy* **2020**, *22*, 165, doi:10.3390/e22020165 . . . . **55**

Martin Gerlach and Francesc Font-Clos
A Standardized Project Gutenberg Corpus for Statistical Analysis of Natural Language and Quantitative Linguistics
Reprinted from: *Entropy* **2020**, *22*, 126, doi:10.3390/e22010126 . . . . **71**

Maja Linke and Michael Ramscar
How the Probabilistic Structure of Grammatical Context Shapes Speech
Reprinted from: *Entropy* **2020**, *22*, 90, doi:10.3390/e22010090 . . . . **85**

Łukasz Dębowski
Approximating Information Measures for Fields
Reprinted from: *Entropy* **2020**, *22*, 79, doi:10.3390/e22010079 . . . . **108**

Ximena Gutierrez-Vasques and Victor Mijangos
Productivity and Predictability for Measuring Morphological Complexity
Reprinted from: *Entropy* **2020**, *22*, 48, doi:10.3390/e22010048 . . . . **118**

Geng Ren, Shuntaro Takahashi, Kumiko Tanaka-Ishii
Entropy Rate Estimation for English via a Large Cognitive Experiment Using Mechanical Turk
Reprinted from: *Entropy* **2019**, *21*, 1201, doi:10.3390/e21121201 . . . . **137**

Noortje J. Venhuizen, Matthew W. Crocker and Harm Brouwer
Semantic Entropy in Language Comprehension
Reprinted from: *Entropy* **2019**, *21*, 1159, doi:10.3390/e21121159 . . . . **152**

Antoni Hernández-Fernández, Iván G. Torre, Juan-María Garrido and Lucas Lacasa
Linguistic Laws in Speech: The Case of Catalan and Spanish
Reprinted from: *Entropy* **2019**, *21*, 1153, doi:10.3390/e21121153 . . . . **173**

Michael Hahn and Richard Futrell
Estimating Predictive Rate–Distortion Curves via Neural Variational Inference
Reprinted from: *Entropy* **2019**, *21*, 640, doi:10.3390/e21070640 . . . . **189**

Alexander Koplenig, Sascha Wolfer and Carolin Müller-Spitzer
Studying Lexical Dynamics and Language Change via Generalized Entropies: The Problem of Sample Size
Reprinted from: *Entropy* **2019**, *21*, 464, doi:10.3390/e21050464 . . . . **216**

About the Special Issue Editors

Łukasz Dębowski obtained his MS.c. degree in Theoretical Physics from Warsaw University in 1994, Ph.D. degree in Computer Science from the Polish Academy of Sciences in 2005, and Habilitation in Computer Science from the Polish Academy of Sciences in 2015. He visited the Institute of Formal and Applied Linguistics at Charles University in 2001, Santa Fe Institute in 2002, the School of Computer Science and Engineering at the University of New South Wales in 2006, the Centrum Wiskunde & Informatica from 2008 to 2009, and the Department of Advanced Information Technology at Kyushu University in 2015. He is currently an Associate Professor at the Institute of Computer Science of the Polish Academy of Sciences. His research interests include information theory, statistics, linguistics, and natural language engineering. He is a member of the International Quantitative Linguistics Association.

Christian Bentz holds a Ph.D. in Computation, Cognition and Language from the University of Cambridge. He is currently Assistant Professor (Akademischer Rat a. Z.) at the University of Tübingen. Prior to this appointment, he was a Postdoctoral Researcher at URPP Language and Space, University of Zürich, and a Research Fellow at the DFG Center for Advanced Studies "Words, Bones, Genes, Tools" at University of Tübingen. His research focuses on the application of information theory to natural languages and other symbolic systems.

entropy

MDPI

Editorial

Information Theory and Language

Łukasz Dębowski [1,*] and Christian Bentz [2,3]

[1] Institute of Computer Science, Polish Academy of Sciences, ul. Jana Kazimierza 5, 01-248 Warszawa, Poland
[2] URPP Langauge and Space, University of Zürich, Freierstrasse 16, CH-8032 Zürich, Switzerland; chris@christianbentz.de
[3] DFG Center for Advanced Studies, University of Tübingen, Rümelinstraße, D-72070 Tübingen, Germany
* Correspondence: ldebowsk@ipipan.waw.pl; Tel.: +48-22-380-0553

Received: 8 April 2020; Accepted: 9 April 2020; Published: 11 April 2020

Keywords: entropy; mutual information; natural language; statistical language models; statistical language laws; semantics; syntax; complexity; criticality; language resources

Human language is a system of communication. Communication, in turn, consists primarily of information transmission. Writing about the interactions between information and natural language, we cannot fail to mention that information theory has originated with statistical investigations of English text in the turn of the 1940s and 1950s [1,2]. While initially, there were some common interests between information theory and linguistics, for instance, understanding distributional properties of elements in natural language, e.g., [3,4], the following decades brought a growing divide between the fields. They went down separate research paths until the end of the 20th century. Whereas information theory embraced probabilities, also in disguise of algorithms [5], the influential Chomskyan formal theory of syntax deemed the question of probabilities in language as scientifically largely irrelevant [6]. It was only in the 1990s that the gap between information theory and formal language studies started to be bridged by the rapid progress of computational linguistics [7,8]. For a detailed account of this development see also [9]. Presently, this progress has resulted in large-scale neural statistical language models such as the much publicized GPT-2 [10], which is capable of generating surreal but understandable short stories.

To use an information theoretic metaphor, the communication channel between the divergent research traditions is reopening. Looking back at independent discoveries of probabilistic and non-probabilistic accounts of natural language, we deem that the divide might have been necessary to focus attention on particular areas of scientific investigation. However, the time is ripe to integrate the established disjoint scholarships, and to cross-fertilize research. We believe that the frameworks of information theory and linguistics are fully compatible in spite of some historical reservations and different academic curricula.

This Special Issue consists of twelve contributions that cover various recent research areas at the interface of information theory and linguistics. They concern in particular:

- applications of information theoretic concepts to the research of natural languages;
- mathematical work in information theory inspired by natural language phenomena;
- empirical and theoretical investigation of quantitative laws of natural language;
- empirical and theoretical investigation of statistical and neural language models.

We believe that the selection of authors and topics in this Special issue reflects the state of the art of interdisciplinary research. In fact, the formal disciplines of the contributing authors range from linguistics and cognitive science to computer science, mathematics, and physics. Since the various research perspectives cannot be easily arranged in an obvious linear order, we have decided to present the papers in the order of their publication.

The Contributions

- Koplenig, A., Wolfer, S., and Müller-Spitzer, C., *Studying Lexical Dynamics and Language Change via Generalized Entropies: The Problem of Sample Size* [11].

Dependence on sample size is a recurrent problem in quantitative linguistics. This also holds for accounts harnessing, for instance, the entropy of word frequency distributions. Koplenig, Wolfer, and Müller-Spitzer systematically investigate this issue based on a corpus compiled from a weekly news magazine in German, which spans seven decades, and contains more than 200 million word tokens. In particular, they employ the generalized Tsallis entropies, which allow for weighting parts of the frequency spectrum more or less heavily in entropy calculations. It turns out that correlations between the estimated entropies and respective sample sizes are only broken if a heavy bias towards highly frequent words is introduced. In particular, the standard Shannon entropies display a strong dependence on sample size. In an application investigating lexical change over several decades, the authors further propose and illustrate a "litmus test". This entails calculating entropy divergences between parts of the corpus over historical time, and comparing these with entropy divergences calculated for texts in random order. Their results suggest that it is the growing sample size over time which leads to systematic patterns in entropy divergences, potentially independent of genuine lexical change.

- Hahn, M. and Futrell, R., *Estimating Predictive Rate–Distortion Curves via Neural Variational Inference* [12].

The predictive rate-distortion curve quantifies the trade-off between compressing information about the past of a stochastic process and predicting its future accurately. Hence it is a more detailed characteristic of the process complexity than its excess entropy or statistical complexity. Hahn and Futrell study estimation of predictive rate-distortion curves for complex stochastic processes, aimed to be applied for natural language. The authors' method of estimation consists in upper bounding the correct curve by means of a neural network approximation of the investigated process. The method is validated on examples of processes for which the predictive rate-distortion curve is known analytically. Moreover, the authors provide an estimate of the predictive rate-distortion curve for text corpora in five natural languages (English, Russian, Arabic, Japanese, and Chinese). The experiments universally indicate that the excess entropy and statistical complexity for natural language are infinite.

- Hernández-Fernández, A., Torre, I.G., Garrido, J.M., and Lacasa, L., *Linguistic Laws in Speech: The Case of Catalan and Spanish* [13].

There is a hypothesis in quantitative linguistics, called the physical hypothesis, that statistical linguistic laws in written texts are a byproduct of more exact laws present in the acoustic signals of oral communication. In contrast to earlier works, Hernández-Fernández et al. investigate and verify the physical hypothesis using a large oral text corpus, the Glissando Corpus of spoken Catalan and Spanish. The studied quantitative linguistic laws include Zipf's law, Herdan's law, the brevity law, Menzerath–Altmann's law, the log-normality law, and the size-rank law. By aligning the acoustic signal with the speech transcripts, they measure and compare the agreement of each of these laws when measured in both physical and symbolic units. The conclusion of this experiment is that quantitative linguistic laws are satisfied indeed more accurately for the acoustic signal than for the speech transcript.

- Venhuizen, N.J., Crocker, M.W., and Brouwer, H., *Semantic Entropy in Language Comprehension* [14].

The link between information and meaning has been a controversial topic ever since Shannon's work. The alleged disconnection between the two was posed as a main argument against analyzing natural language in the light of information theory. Venhuizen, Crocker and Brouwer illustrate that information theoretic concepts might be fruitfully applied to both linguistic signals, and the points they denote in meaning space. In their experiments, they combine formal semantic tools with neural network technology. Based on a set of training sentences, their neural network learns to map linguistic signals onto meaning vectors representing propositional truth values. This setup allows the authors to trace the semantic expectations of the network in word-by-word online processing. In this context, they tease apart surprisal and entropy reduction, two concepts which were previously often seen as strongly intertwined. Surprisal is calculated based on word-by-word transitions in meaning space, whereas entropy is calculated over meaning vectors which identify a unique semantic model of the world. Given these definitions, suprisal and entropy reduction are not strongly correlated. The authors explain this by pointing out that surprisal is inherently more sensitive to frequency effects in the linguistic signal, while entropy reduction is more strongly influenced by knowledge of the model theoretic world.

- Ren, G., Takahashi, S., and Tanaka-Ishii, K., *Entropy Rate Estimation for English via a Large Cognitive Experiment Using Mechanical Turk* [15].

The entropy rate of a sequence reflects the amount of information conveyed per unit, e.g., characters or words in natural language. It has been proposed also as a measure of the complexity of a sequence. However, estimating the entropy rate of natural languages has proven a challenging endeavor due to the problem of finite sample sizes and long-range dependence. Ren, Takahashi, and Tanaka-Ishii revive an idea going back to Shannon's experiments [2], namely, estimating the entropy rate by using human subjects to predict the next character in a linguistic sequence. They collect more than 100,000 character predictions for English texts by 683 different subjects. Across all subjects and trials, they estimate the entropy rate to around 1.4 bits per character. Using trials selected for high performance (i.e., correctly guessing characters) reduces the estimate to around 1.22 bits per character. In their discussion, the authors point out that this is lower than Shannon's original value of 1.3 bits per character. On the other hand, it is higher than entropy rates estimated with current state-of-the-art neural language models, which are just above 1 bit per character. This suggests that neural language models outperform human subjects in character guessing games.

- Gutierrez-Vasques, X. and Mijangos, V., *Productivity and Predictability for Measuring Morphological Complexity* [16].

There is a recent rise of interest in measuring the morphological complexity of typologically diverse languages. The findings of this research have implications for both theoretical and applied linguistics, especially in the domain of natural language processing. Gutierrez-Vasques and Mijangos propose to apply the information-theoretic concept of entropy rate to word internal structure. Their data sets contain parallel texts for 47 and 133 typologically diverse languages respectively. Using a neural language model they estimate the difficulty of predicting character unigrams and trigrams within words for different languages and writing systems. These estimates are then contrasted with more traditional measures of morphological complexity, such as the type-token ratio for words. It turns out that word internal predictability is only weakly correlated with the type-token ratio, and hence measures a new and independent dimension of morphological complexity.

- Dębowski, Ł., *Approximating Information Measures for Fields* [17].

Motivated by some theoretical problems of statistical modeling of natural language, Dębowski reconsiders the classical problem of generalizing entropy and mutual information from discrete random variables (finite partitions, in more abstract formulation) to arbitrary random variables (fields and σ-fields, respectively). Having noticed a mistake in his paper from 2009, he supplies corrected proofs of the invariance of completion and the chain rule for conditional entropy and mutual information. In the final section, he also discusses how the generalized calculus of conditional entropy and mutual information is useful in particular for studying the ergodic decomposition of strongly non-ergodic stationary processes and its links with statistical modeling of natural language, which possibly should be modeled by a strongly non-ergodic process.

- Linke, M. and Ramscar, M., *How the Probabilistic Structure of Grammatical Context Shapes Speech* [18].

Frequencies of occurrence are a central concept in quantitative linguistics. They are often used to measure the informativeness of units (i.e., characters, words, etc.) in written language. Linke and Ramscar point out several caveats with this approach. Firstly, written language is not a direct reflection of speech. As a remedy, they use a corpus of conversational English of more than 200,000 word tokens with phonetic labels, and compare their results to studies using written language. Secondly, frequencies of occurrence abstract away from co-occurrence patterns at different levels of language structure, e.g., n-grams for words and parts-of-speech, as well as subword structure. The authors argue that grammatical context often predicts usage patterns in speech better than mere frequencies. Thirdly, distributions of frequencies are mostly analyzed over entire texts, for instance, when power law like patterns such as Zipf's law are assessed. However, the authors illustrate that there are systematic differences between the distributions of frequencies for words of different parts-of-speech. Namely, while open class items such as nouns and verbs follow power laws, function words rather follow geometric distributions. In fact, the authors further argue that power law like behavior in aggregate distributions might well be the outcome of mixing distributions which are by themselves geometric.

- Gerlach, M. and Font-Clos, F., *A Standardized Project Gutenberg Corpus for Statistical Analysis of Natural Language and Quantitative Linguistics* [19].

Studies on information theoretic properties of natural languages—and analyses in quantitative linguistics more generally—stand and fall with availability of textual data. The universality of linguistic laws, for instance, can only be ascertained given openly available, cross-linguistic, and transparently processed data. To this end, Gerlach and Font-Clos contribute a standardized version of the Project Gutenberg Corpus, which contains more than 50,000 books in over 20 languages. They give a detailed description of the data acquisition, processing, and metadata annotation procedures. Furthermore, they illustrate how this corpus can be used to measure the topical variability between texts associated with different genres via so-called "bookshelf" labels, and how authors are distinguishable by the Jensen-Shannon divergence applied to their works.

- Seoane, L.F. and Solé, R., *Criticality in Pareto Optimal Grammars?* [20].

Seoane and Solé propose a computational methodology to inspect corpora of texts in order to extract salient levels of linguistic description. Their methodology is grounded in the bottleneck method from information theory, Pareto optimality from multi-objective optimization, and concepts from statistical physics such as energy, entropy, phase transitions and criticality. Their working example concerns extracting the Pareto optimal grammars from 49 newspaper articles taken from the Corpus of Contemporary American English preprocessed by the Natural Language Toolkit (NLTK). The numerical results indicate a critical point in the description of human language. As the authors write, the critical point is the worst case in terms of description since there is no relatively small model which can capture the whole phenomenology at any level of linguistic description.

- Ahmadi, L. and Ward, M.D., *Asymptotic Analysis of the kth Subword Complexity* [21].

The subword complexity is a function which counts how many distinct substrings of a given length appear in a given string. It is a simple characteristic of a string that yields an insight whether the string is periodic, random, or something in between—like a text in natural language. In particular, the subword complexity divided by the string length equals to the type-token ratio investigated in quantitative linguistics. Ahmadi and Ward study some properties of subword complexity from a mathematical point of view. Namely, they investigate the asymptotic behavior of the subword complexity for sequences of independent identically distributed random variables. They derive expressions for the expectation (first moment) and the variance (second moment) of subword complexity. Their methodology involves complex analysis, analytical poissonization and depoissonization, the Mellin transform, and saddle point analysis.

- Corral, Á. and Serra, I., *The Brevity Law as a Scaling Law, and a Possible Origin of Zipf's Law for Word Frequencies* [22].

Corral and Serra study the joint distribution of lengths and frequencies of words, whose marginals are described by the brevity law and Zipf's law for frequencies of frequencies, called also Lotka's law. The investigated corpus is the English subcorpus of the Standardized Project Gutenberg Corpus, introduced in contribution [19]. The authors observe that the marginal distribution of word length is better described by the gamma distribution than by the previously proposed log-normal distribution. Moreover, the conditional frequency distributions at a fixed length exhibit a universal power-law decay and a scaling law analogous to those found in the thermodynamics of critical phenomena. In conclusion, the authors present a four-parameter model for the joint distribution of lengths and frequencies of words.

Acknowledgments: We express our thanks to the authors of the above contributions, the reviewers for their feedback on the manuscripts, and to the journal Entropy and MDPI for their support during this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *30*, 379–423. [CrossRef]
2. Shannon, C. Prediction and entropy of printed English. *Bell Syst. Tech. J.* **1951**, *30*, 50–64. [CrossRef]
3. Harris, Z. *Mathematical Structures of Language*; Interscience Publishers: New York, NY, USA, 1968.
4. Harris, Z. *A Theory of Language and Information: A Mathematical Approach*; Clarendon Press: Oxford, UK, 1991.
5. Kolmogorov, A.N. Three approaches to the quantitative definition of information. *Probl. Inf. Transm.* **1965**, *1*, 1–7. [CrossRef]
6. Chomsky, N. *Syntactic Structures*; Mouton & Co: The Hague, The Netherlands, 1957.
7. Jelinek, F. *Statistical Methods for Speech Recognition*; The MIT Press: Cambridge, MA, USA, 1997.
8. Manning, C.D.; Schütze, H. *Foundations of Statistical Natural Language Processing*; The MIT Press: Cambridge, MA, USA, 1999.
9. Pereira, F. Formal grammar and information theory: together again? *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **2000**, *358*, 1239–1253. [CrossRef]
10. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. Available online: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf (accessed on 9 April 2020).
11. Koplenig, A.; Wolfer, S.; Müller-Spitzer, C. Studying Lexical Dynamics and Language Change via Generalized Entropies: The Problem of Sample Size. *Entropy* **2019**, *21*, 464. [CrossRef]
12. Hahn, M.; Futrell, R. Estimating Predictive Rate–Distortion Curves via Neural Variational Inference. *Entropy* **2019**, *21*, 640. [CrossRef]
13. Hernández-Fernández, A.; Torre, I.G.; Garrido, J.M.; Lacasa, L. Linguistic Laws in Speech: The Case of Catalan and Spanish. *Entropy* **2019**, *21*, 1153. [CrossRef]
14. Venhuizen, N.J.; Crocker, M.W.; Brouwer, H. Semantic Entropy in Language Comprehension. *Entropy* **2019**, *21*, 1159. [CrossRef]
15. Ren, G.; Takahashi, S.; Tanaka-Ishii, K. Entropy Rate Estimation for English via a Large Cognitive Experiment Using Mechanical Turk. *Entropy* **2019**, *21*, 1201. [CrossRef]
16. Gutierrez-Vasques, X.; Mijangos, V. Productivity and Predictability for Measuring Morphological Complexity. *Entropy* **2019**, *22*, 48. [CrossRef]
17. Dębowski, Ł. Approximating Information Measures for Fields. *Entropy* **2020**, *22*, 79. [CrossRef]

18. Linke, M.; Ramscar, M. How the Probabilistic Structure of Grammatical Context Shapes Speech. *Entropy* **2020**, *22*, 90. [CrossRef]
19. Gerlach, M.; Font-Clos, F. A Standardized Project Gutenberg Corpus for Statistical Analysis of Natural Language and Quantitative Linguistics. *Entropy* **2020**, *22*, 126. [CrossRef]
20. Seoane, L.F.; Solé, R. Criticality in Pareto Optimal Grammars? *Entropy* **2020**, *22*, 165. [CrossRef]
21. Ahmadi, L.; Ward, M.D. Asymptotic Analysis of the kth Subword Complexity. *Entropy* **2020**, *22*, 207. [CrossRef]
22. Corral, A.; Serra, I. The Brevity Law as a Scaling Law, and a Possible Origin of Zipf's Law for Word Frequencies. *Entropy* **2020**, *22*, 224. [CrossRef]

entropy

Communication

The Brevity Law as a Scaling Law, and a Possible Origin of Zipf's Law for Word Frequencies

Álvaro Corral [1,2,3,4,*] and Isabel Serra [1,5]

[1] Centre de Recerca Matemàtica, Edifici C, Campus Bellaterra, E-08193 Barcelona, Spain
[2] Departament de Matemàtiques, Facultat de Ciències, Universitat Autònoma de Barcelona, E-08193 Barcelona, Spain
[3] Barcelona Graduate School of Mathematics, Edifici C, Campus Bellaterra, E-08193 Barcelona, Spain
[4] Complexity Science Hub Vienna, Josefstädter Straβe 39, 1080 Vienna, Austria
[5] Barcelona Supercomputing Center-Centro Nacional de Supercomputación, Jordi Girona 29, E-08034 Barcelona, Spain
* Correspondence: acorral@crm.es

Received: 31 December 2019; Accepted: 12 February 2020; Published: 17 February 2020

Abstract: An important body of quantitative linguistics is constituted by a series of statistical laws about language usage. Despite the importance of these linguistic laws, some of them are poorly formulated, and, more importantly, there is no unified framework that encompasses all them. This paper presents a new perspective to establish a connection between different statistical linguistic laws. Characterizing each word type by two random variables—length (in number of characters) and absolute frequency—we show that the corresponding bivariate joint probability distribution shows a rich and precise phenomenology, with the type-length and the type-frequency distributions as its two marginals, and the conditional distribution of frequency at fixed length providing a clear formulation for the brevity-frequency phenomenon. The type-length distribution turns out to be well fitted by a gamma distribution (much better than with the previously proposed lognormal), and the conditional frequency distributions at fixed length display power-law-decay behavior with a fixed exponent $\alpha \simeq 1.4$ and a characteristic-frequency crossover that scales as an inverse power $\delta \simeq 2.8$ of length, which implies the fulfillment of a scaling law analogous to those found in the thermodynamics of critical phenomena. As a by-product, we find a possible model-free explanation for the origin of Zipf's law, which should arise as a mixture of conditional frequency distributions governed by the crossover length-dependent frequency.

Keywords: quantitative linguistics; brevity law; abbreviation law; power laws; scaling; Zipf's law

1. Introduction

The usage of language, both in its written and oral expressions (texts and speech), follows very strong statistical regularities. One of the goals of quantitative linguistics is to unveil, analyze, explain, and exploit those linguistic statistical laws. Perhaps the clearest example of a statistical law in language usage is Zipf's law, which quantifies the frequency of occurrence of words in such written and oral forms [1–6], establishing that there is no unarbitrary way to distinguish between rare and common words (due to the absence of a characteristic scale in "rarity"). Surprisingly, Zipf's law is not only a linguistic law, but seems to be a rather common phenomenon in complex systems where discrete units self-organize into groups, or types (persons into cities, money into persons, etc. [7]).

Zipf's law can be considered as the "tip of the iceberg" of text statistics. Another well-known pattern of this sort is Herdan's law, also called Heaps' law [2,8,9], which states that the growth of vocabulary with text length is sublinear (however, the precise mathematical dependence has been debated [10]). Herdan's law has been related to Zipf's law, sometimes with too simple arguments, although rigorous connections have been established as well [8,10]. The authors of [11] provide another example of relations between linguistic laws, but, in general, no general framework encompassing all laws exists.

Two other laws—the law of word length and the so-called Zipf's law of abbreviation or brevity law— are of particular interest in this work. As far as we know, and in contrast to the Zipf's law of word frequency, these two laws do not have non-linguistic counterparts. The law of word length finds that the length of words (measured in number of letter tokens, for instance) is lognormally distributed [12,13], whereas the brevity law determines that more frequent words tend to be shorter, and rarer words tend to be longer. This is usually quantified between a negative correlation between word frequency and word length [14].

Very recently, Torre et al. [13] parameterized the dependence between mean frequency and length, obtaining (using a speech corpus) that the frequency averaged for fixed length decays exponentially with length. This is in contrast with a result suggested by Herdan (to the best of our knowledge not directly supported by empirical analysis), who proposed a power-law decay, with exponent between 2 and 3 [12]. This result probably arose from an analogy with the word-frequency distribution derived by Simon [15], with an exponential tail that was neglected.

The purpose of our paper is to put these three important linguistic laws (Zipf's law of word frequency, the word-length law, and the brevity law) into a broader context. By means of considering word frequency and word length as two random variables associated to word types, we will see how the bivariate distribution of those two variables is the appropriate framework to describe the brevity-frequency phenomenon. This leads us to several findings: (i) a gamma law for the word-length distribution, in contrast to the previously proposed lognormal shape; (ii) a well-defined functional form for the word-frequency distributions conditioned to fixed length, where a power-law decay with exponent α for the bulk frequencies becomes dominant; (iii) a scaling law for those distributions, apparent as a collapse of data under rescaling; (iv) an approximate power-law decay of the characteristic scale of frequency as a function of length, with exponent δ; and (v) a possible explanation for Zipf's law of word frequency as arising from the mixture of conditional distributions of frequency at different lengths, where Zipf's exponent is determined by the exponents α and δ.

2. Preliminary Considerations

Given a sample of natural language (a text, a fragment of speech, or a corpus, in general), any word type (i.e., each unique word) has an associated word length, which we measure in number of characters (as we deal with a written corpus), and an associated word absolute frequency, which is the number of occurrences of the word type on the corpus under consideration (i.e., the number of tokens of the type). We denote these two random variables as ℓ and n, respectively.

Zipf's law of word frequency is written as a power-law relation between $f(n)$ and n [6], i.e.,

$$f(n) \propto \frac{1}{n^{\beta}} \text{ for } n \geq c,$$

where $f(n)$ is the empirical probability mass function of the word frequency n, the symbol $\propto$ denotes proportionality, β is the power-law exponent, and c is a lower cut-off below which the law losses its validity (so, Zipf's law is a high-frequency phenomenon). The exponent β takes values typically close to 2. When very large corpora are analyzed (made from many different texts an authors) another (additional) power-law regime appears at smaller frequencies [16,17],

$$f(n) \propto \frac{1}{n^{\alpha}} \text{ for } a \leq n \leq b,$$

with α a new power law exponent smaller than β, and a and b lower and upper cut-offs, respectively (with $a < b < c$). This second power law is not identified with Zipf's law.

On the other hand, the law of word lengths [12] proposes a lognormal distribution for the empirical probability mass function of word lengths, that is,

$$f(\ell) \sim \mathrm{LN}(\mu, \sigma^2),$$

where LN denotes a lognormal distribution, whose associated normal distribution has mean μ and variance σ^2 (note that with the lognormal assumption it would seem that one is taking a continuous approximation for $f(\ell)$; nevertheless, discreteness of $f(\ell)$ is still possible just redefining the normalization constant). The present paper challenges the lognormal law for $f(\ell)$. Finally, the brevity law [14] can be summarized as

$$\mathrm{corr}(\ell, n) < 0,$$

where $\mathrm{corr}(\ell, n)$ is a correlation measure between ℓ and n, as, for instance, Pearson correlation, Spearman correlation, or Kendall correlation.

We claim that a more complete approach to the relationship between word length and word frequency can be obtained from the joint probability distribution $f(\ell, n)$ of both variables, together with the associated conditional distributions $f(n|\ell)$. To be more precise, $f(\ell, n)$ is the joint probability mass function of type length and frequency, and $f(n|\ell)$ is the probability mass function of type frequency conditioned to fixed length. Naturally, the word-frequency distribution $f(n)$ and the word-length distribution $f(\ell)$ are just the two marginal distributions of $f(\ell, n)$.

The relationships between these quantities are

$$f(\ell) = \sum_{n=1}^{\infty} f(\ell, n),$$

$$f(n) = \sum_{\ell=1}^{\infty} f(\ell, n),$$

$$f(\ell, n) = f(n|\ell) f(\ell).$$

Note that we will not use in this paper the equivalent relation $f(\ell, n) = f(\ell|n) f(n)$, for sampling reasons ($n$ takes many more different values than ℓ; so, for fixed values of n one may find there is not enough statistics to obtain $f(\ell|n)$). Obviously, all probability mass functions fulfil normalization,

$$\sum_{\ell=1}^{\infty} \sum_{n=1}^{\infty} f(\ell, n) = \sum_{n=1}^{\infty} f(n|\ell) = \sum_{\ell=1}^{\infty} f(\ell) = \sum_{n=1}^{\infty} f(n) = 1.$$

We stress that, in our framework, each type yields one instance of the bivariate random variable (ℓ, n), in contrast to another equivalent approach for which it is each token that gives one instance of the (perhaps-different) random variables, see [7]. The use of each approach has important consequences for the formulation of Zipf's law, as it is well known [7], and for the formulation of the word-length law (as it is not so well known [12]). Moreover, our bivariate framework is certainly different to the that in [18], where the frequency was understood as a four-variate distribution with the random variables taking 26 values from a to z, and also to the generalization in [19].

3. Corpus and Statistical Methods

We investigate the joint probability distribution of word-type length and frequency empirically, using all English books in the recently presented Standardized Project Gutenberg Corpus [20], which comprises more than 40,000 books in English, with a total number of tokens equal to 2,016,391,406 and a total number of types of 2,268,043. We disregard types with $n < 10$ (relative frequency below 5×10^{-9}) and also those not composed exclusively by the 26 usual letters from a to z (previously, capital letters were transformed to lower-case). This sub-corpus is further reduced by the elimination of types with length above 20 characters; to avoid typos and "spurious" words (among the eliminated types with $n \geq 10$ we only find three true English words: *incomprehensibilities*, *crystalloluminescence*, and *nitrosodimethylaniline*). This reduces the numbers of tokens and types, respectively, to 2,010,440,020 and 391,529. Thus, all we need for our study is the list of all types (a dictionary) including their absolute frequencies n and their lengths ℓ (measured in terms of number of characters).

Power-law distributions are fitted to the empirical data by using the version for discrete random variables of the method for continuous distributions outlined in [21] and developed in Refs. [22,23], which is based on maximum-likelihood estimation and the Kolmogorov–Smirnov goodness-of-fit test. Acceptable (i.e., non-rejectable) fits require p-values not below 0.20, which are computed with 1000 Monte Carlo simulations. Complete details in the discrete case are available in Refs. [6,24]. This method is similar in spirit to the one by Clauset et al. [25], but avoiding some of the important problems that the latter presents [26,27]. Histograms are drawn to provide visual intuition for the shape of the empirical probability mass functions and the adequacy of fits; in the case of $f(n|\ell)$ and $f(n)$, we use logarithmic binning [22,28]. Nevertheless, the computation of the fits does not make use of the graphical representation of the distributions.

On the other side, the theory of scaling analysis, following the authors of [21,29], allows us to compare the shape of the conditional distributions $f(n|\ell)$ for different values of ℓ. This theory has revealed a very powerful tool in quantitative linguistics, allowing in previous research to show that the shape of the word-frequency distribution does not change as a text increases its length [30,31].

4. Results

First, let us examine the raw data, looking at the scatter plot between frequency and length in Figure 1, where each point is a word type represented by an associated value of n and an associated value of ℓ (note that several or many types can overlap at the same point, if they share their values of ℓ and n, as these are discrete variables). >From the tendency of decreasing maximum n with increasing ℓ, clearly visible in the plot, one could arrive to an erroneous version of the brevity law. Naturally, brevity would be apparent if the scatter plot were homogenously populated (i.e., if $f(\ell, n)$ would be uniform in the domain occupied by the points). However, of course, this is not the case, as we will quantify later. On the contrary, if $f(\ell, m)$ were the product of two independent exponentials, with $m = \ln n$, the scatter plot would be rather similar to the real one (Figure 1), but the brevity law would not hold (because of the independence of ℓ and m, that is, of ℓ and n). We will see that exponentials distributions play an important role here, but not in this way.

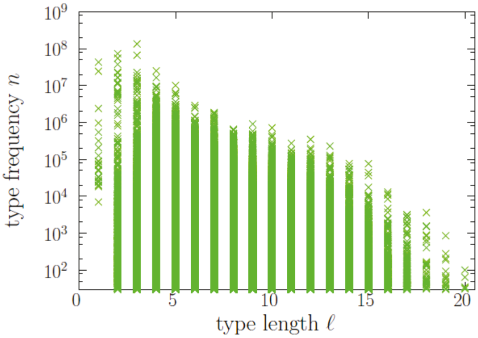


Figure 1. Illustration of the dataset by means of the scatter plot between word-type frequency and length. Frequencies below 30 are not shown.

A more acceptable approach to the brevity-frequency phenomenon is to calculate the correlation between ℓ and n. For the Pearson correlation, our dataset yields $\text{corr}(\ell, n) = -0.023$, which, despite looking very small, is significantly different from zero, with a p-value below 0.01 for 100 reshufflings of the frequency (all the values obtained after reshuffling the frequencies keeping the lengths fixed are between -0.004 and 0.006). If, instead, we calculate the Pearson correlation between ℓ and the logarithm m of the frequency we get $\text{corr}(\ell, m) = -0.083$, again with a p-value below 0.01. Nevertheless, as neither the underlying joint distributions $f(\ell, n)$ or $f(\ell, m)$ resemble a Gaussian at all, nor the correlation seems to be linear (see Figure 1), the meaning of the Pearson correlation is difficult to interpret. We will see below that the analysis of the conditional distributions $f(n|\ell)$ provides more useful information.

4.1. Marginal Distributions

Let us now study the word-length distribution, $f(\ell)$, shown in Figure 2. The distribution is clearly unimodal (with its maximum at $\ell = 7$), and although it has been previously modeled as a lognormal [12], we get a nearly perfect fit using a gamma distribution,

$$f(\ell) = \frac{\lambda}{\Gamma(\gamma)} (\lambda\ell)^{\gamma-1} e^{-\lambda\ell}, \tag{1}$$

with shape parameter $\gamma = 11.10 \pm 0.02$ and inverted scale parameter $\lambda = 1.439 \pm 0.003$ (where the uncertainty corresponds to one standard deviation, and $\Gamma(\gamma)$ denotes the gamma function). Notice then that, for large lengths, we would get an exponential decay (asymptotically, strictly speaking). However, there is an important difference between the lognormal distribution proposed in [13] and the gamma distribution found here, which is that the former case refers to the length of tokens, whereas in our case we deal with the length of types (of course, length of tokens and length of types is the same length, but the relative number of tokens and types is different, depending on length). This was already distinguished by Herdan [12], who used the terms occurrence distribution and dictionary distribution, and proposed that both of them were lognormal. In the caption of Figure 2 we provide the log-likelihoods of both the gamma and lognormal fits, concluding that the gamma distribution yields a better fit for the "dictionary distribution" of word lengths. The fit is specially good in the range $\ell > 2$.

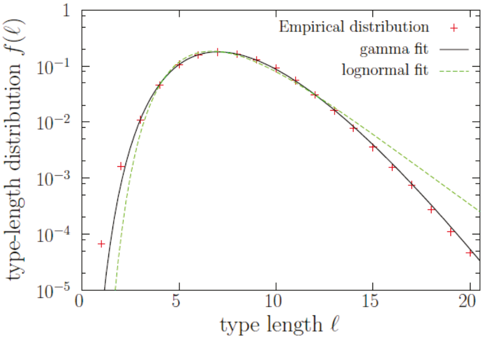


Figure 2. Probability mass function $f(\ell)$ of type length, together with gamma and lognormal fits. Note that the majority of types are those with lengths between 4 and 13, and that $f(\ell)$ is roughly constant between 5 and 10. The superiority of the gamma fit is visually apparent, and this is confirmed by log-likelihood equal to $-872{,}175.2$ in front of the value $-876{,}535.1$ for the lognormal (a discrete gamma distribution slightly improves the fit, but the simple continuous case here is enough for our purposes). The parameters resulting for the gamma fit are given in the text, and those for the lognormal are $\mu = 1.9970 \pm 0.0005$ and $\sigma = 0.3081 \pm 0.0003$.

Regarding the other marginal distribution, which is the word-frequency distribution $f(n)$ represented in Figure 3, we get that, as expected, Zipf's law is fulfilled with $\beta = 1.94 \pm 0.03$ for

$n \geq 1.9 \times 10^5$ (this is almost three orders of magnitude), see Table 1. Another power-law regime in the bulk, as in [16], is found to hold for one order of magnitude and a half (only), from $a \simeq 400$ to $b \simeq 14{,}000$, with exponent $\alpha = 1.41 \pm 0.005$, see Table 2. Note that although the truncated power law for the bulk part of the distribution is much shorter than the one for the tail (1.5 orders of magnitude in front of almost 3), the former contains many more data (50,000 in front of ~1000), see Tables 1 and 2 for the precise figures. Note also that the two power-law regimes for the frequency translate into two exponential regimes for m (the logarithm of n).

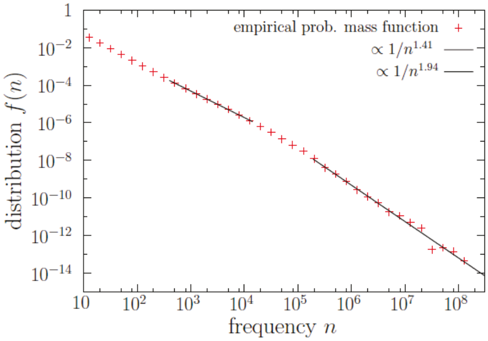


Figure 3. Probability mass function $f(n)$ of type frequency (this is a marginal distribution with respect $f(\ell, n)$). The results of the power-law fits are also shown. The fit of a truncated continuous power law, maximizing number of data, yields $\alpha = 1.41$; the fit of a untruncated discrete power law yields $\beta = 1.94$.

Table 1. Results of the fitting of an discrete untruncated power law to the conditional distributions $f(n|\ell)$, denoted by a fixed ℓ, and to the marginal distribution $f(n)$, denoted by the range $1 \leq \ell \leq 20$. N is total number of types, n_{max} is the frequency of the most frequent type, c is the lower cuf-off of the fit, ordermag is $\log_{10}(n_{max}/c)$, N_c is the number of types with $n \geq c$, β is the resulting fitting exponent, σ_β is its standard deviation, and p is the p-value of the fit. For the conditional distributions, the possible fits are restricted to the range $n > \langle n^2|\ell\rangle / \langle n|\ell\rangle$. The fit proceeds by sweeping 50 values of c per order of magnitude and using 1000 Monte Carlo simulations for the calculation of p. Of all the fits with $p \geq 0.20$ for a given ℓ, the one with smaller c is selected. Outside the range $5 \leq \ell \leq 14$, the number of types in the tail (below 10) is too low to yield a meaningful fit.

ℓ	N	n_{max} ($\times 10^5$)	c ($\times 10^5$)	Ordermag	N_c	$\beta \pm \sigma_\beta$	p
5	41,773	101	15.8	0.80	19	2.75 ± 0.46	0.97
6	62,277	29.0	3.80	0.88	60	2.79 ± 0.24	0.31
7	69,653	18.6	2.88	0.81	55	2.51 ± 0.21	0.32
8	63,574	6.55	1.10	0.78	133	2.82 ± 0.17	0.25
9	50,595	9.12	1.10	0.92	79	2.82 ± 0.21	0.25
10	35,679	7.16	0.83	0.93	69	2.90 ± 0.24	0.75
11	21,536	2.73	0.40	0.84	83	3.03 ± 0.23	0.58
12	11,973	3.49	0.46	0.88	34	2.78 ± 0.33	0.65
13	6240	2.28	0.44	0.72	13	2.57 ± 0.52	0.27
14	3035	0.77	0.24	0.51	12	2.67 ± 0.56	0.22
≤20	391,529	1341	1.91	2.85	927	1.94 ± 0.03	0.44

Table 2. Results of the fitting of a truncated power law to the conditional distributions $f(n|\ell)$, denoted by a fixed ℓ, and to the marginal distribution $f(n)$, denoted by the range $1 \leq \ell \leq 20$. N is total number of types; a and b are the lower and upper cut-offs of the fit, respectively; N_{ab} is the number of types with $a \leq n \leq b$; α is the resulting fitting exponent; σ_α is its standard deviation; and p is the p-value of the fit. The fit of a continuous power law is attempted in the range $n < 0.1\langle n^2|\ell\rangle / \langle n|\ell\rangle$, sweeping 20 values of a and b per order of magnitude and using 1000 Monte Carlo simulations for the calculation of p. Of all the fits with $p \geq 0.20$, for a given ℓ, the one with larger b/a is selected, except for $f(n)$, where the largest N_{ab} is used.

ℓ	N	a ($\times 10^2$)	b ($\times 10^3$)	Ordermag	N_{ab}	$\alpha \pm \sigma_\alpha$	p
1	26	126	2510	2.30	23	1.391 ± 0.155	0.24
2	636	20	2510	3.10	188	1.486 ± 0.045	0.24
3	4282	7.94	4470	3.75	1171	1.428 ± 0.016	0.30
4	17,790	0.40	398	4.00	10,618	1.402 ± 0.005	0.20
5	41,773	5.62	178	2.50	5747	1.426 ± 0.009	0.37
6	62,277	3.98	39.8	2.00	8681	1.421 ± 0.009	0.27
7	69,653	2.00	28.2	2.15	13,392	1.449 ± 0.007	0.25
8	63,574	2.51	11.2	1.65	9849	1.417 ± 0.010	0.41
9	50,595	2.00	10.0	1.70	8850	1.400 ± 0.010	0.25
10	35,679	1.12	8.91	1.90	8454	1.428 ± 0.010	0.21
11	21,536	0.56	1.41	1.40	6227	1.469 ± 0.015	0.22
12	11,973	0.63	5.01	1.90	3866	1.411 ± 0.013	0.51
13	6240	0.56	3.98	1.85	2144	1.396 ± 0.019	0.90
14	3035	0.25	2.24	1.95	1567	1.496 ± 0.022	0.27
15	1384	0.22	2.00	1.95	777	1.488 ± 0.031	0.59
16	612	0.28	0.45	1.20	256	1.569 ± 0.082	0.22
17	296	0.13	0.14	1.05	205	1.784 ± 0.110	0.24
18	107	0.11	0.16	1.15	79	2.008 ± 0.172	0.28
≤20	391,529	3.98	14.1	1.55	51,972	1.413 ± 0.005	0.21

4.2. Power Laws and Scaling Law for the Conditional Distributions

As mentioned, the conditional word-frequency distributions $f(n|\ell)$ are of substantial relevance. In Figure 4, we display some of those functions, and it turns out that n is broadly distributed for each value of ℓ (roughly in the same qualitative way it happens without conditioning to the value of ℓ). Remarkably, the results of a scaling analysis [21,29], depicted in Figure 5, show that all the different $f(n|\ell)$ (for $3 \leq \ell \leq 14$) share a common shape, with a scale determined by a scale parameter in frequency. Indeed, rescaling n as $n\langle n|\ell\rangle / \langle n^2|\ell\rangle$ and $f(n|\ell)$ as $f(n|\ell)\langle n^2|\ell\rangle^2 / \langle n|\ell\rangle^3$, where the first and second empirical moments, $\langle n|\ell\rangle$ and $\langle n^2|\ell\rangle$, are also conditioned to the value of ℓ, we obtain an impressive data collapse, valid for ~7 orders of magnitude in n, which allows us to write the scaling law

$$f(n|\ell) \simeq \frac{\langle n|\ell\rangle^3}{\langle n^2|\ell\rangle^2} g\left(\frac{n\langle n|\ell\rangle}{\langle n^2|\ell\rangle}\right) \text{ for } 3 \leq \ell \leq 14,$$

where the key point is that the scaling function $g(...)$ is the same function for any value of ℓ. For $\ell > 14$ the statistics is low and the fulfilment of the scaling law becomes uncertain. Defining the scale parameter $\theta(\ell) = \langle n^2|\ell\rangle / \langle n|\ell\rangle$, we get alternative expressions for the same scaling law,

$$f(n|\ell) \simeq \frac{\langle n|\ell\rangle}{\theta^2(\ell)} g\left(\frac{n}{\theta(\ell)}\right) \propto \frac{1}{\theta^\alpha(\ell)} g\left(\frac{n}{\theta(\ell)}\right) \text{ for } 3 \leq \ell \leq 14,$$

where constants of proportionality have been reabsorbed into g, and the scale parameter has to be understood as proportional to a characteristic scale of the conditional distributions (i.e., θ is the characteristic scale, up to a constant factor; it is the relative change of θ what will be important for us). The reason for the fulfillment of these relations is the power-law dependence between the moments and the scale parameter when a scaling law holds, this power-law dependence is $\langle n|\ell\rangle \propto \theta^{2-\alpha}$ and $\langle n^2|\ell\rangle \propto \theta^{3-\alpha}$ for $1 < \alpha < 2$, see [21,29].

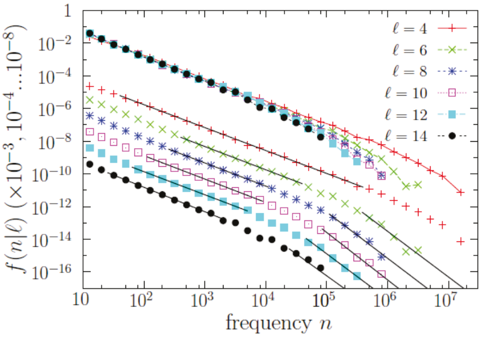


Figure 4. Probability mass functions $f(n|\ell)$ of frequency n conditioned to fixed value of length ℓ, for several values of ℓ. Distributions are shown twice: all together and individually, displaced in the vertical axis by diverse factors 10^{-3}, 10^{-4} up to 10^{-8}, for clarity sake of the power-law fits, represented by dark continuous lines.

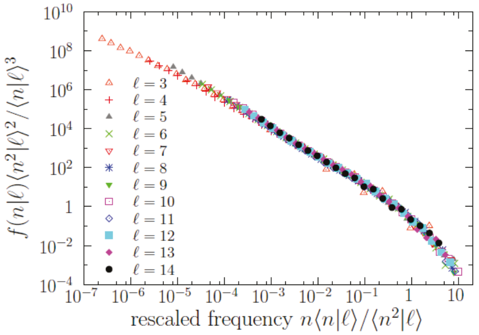


Figure 5. Word frequency probability mass functions $f(n|\ell)$ conditioned to fixed value of length rescaled by the ratio of powers of moments, as a function as rescaled frequency, for all values of length from 3 to 14. The data collapse guarantees the fulfilment of a scaling law.

The data collapse also unveils more clearly the functional form of the scaling function g, allowing to fit its power-law shape in two different ranges. The scaling function turns out to be compatible with a double power-law distribution, i.e., a (long) power law for $n/\theta < 0.1$ with exponent α at ~1.4 and another (short) power law for $n/\theta > 1$ with exponent β at ~2.75; in one formula,

$$g(y) \propto \begin{cases} 1/y^{1.4} & \text{for } y \ll 1, \\ 1/y^{2.75} & \text{for } y > 1, \end{cases} \tag{2}$$

for $y = n/\theta$. In other words, there is a (smooth) change of exponent (a change of log-log slope) at a value of $n \simeq C\theta(\ell)$, with the proportionality constant C taking some value in between 0.1 and 1 (as the transition from one regime to the other is smooth there is not a well defined value of C that separates both). Fitting power laws to those ranges we get the results shown in Tables 1 and 2. Note that $C\theta(\ell)$ can be understood as the characteristic scale of $f(n|\ell)$ mentioned before, and can be also called a frequency crossover.

Nevertheless, although the power-law regime for intermediate frequencies ($n < 0.1\theta$) is very clear, the validity of the other power law (the one for large frequencies) is questionable, in the sense that the power law provides an "acceptable" fit but other distributions could do the same good job, due to the limited range spanned by the tail (less than one order of magnitude). Our main reason to fit a power law to the large-frequency regime is the comparison with Zipf's law ($\beta \simeq 2$), and, as we see, the resulting value of β for $f(n|\ell)$ turns out to be rather large (the results of β for all $f(n|\ell)$ turn out to be statistically compatible with $\beta = 2.75$). In addition, we will show in the next subsection that the high-frequency behavior of the conditional distributions (power law or not) has nothing to do with Zipf's law.

4.3. Brevity Law and Possible Origin of Zipf's Law

Coming back to the scaling law, its fulfillment has an important consequence: it is the scale parameter $\theta(\ell)$ and not the conditional mean $\langle n|\ell\rangle$ what sets the scale of the conditional distributions $f(n|\ell)$. Figure 6 represents the brevity law in terms of the scale parameter as a function of ℓ (the conditional mean value is also shown, for comparison, overimposed to maps of $f(n,\ell)$ and $f(n|\ell)$). Note that the authors of [13] dealt with the conditional mean, finding an exponential decay $\langle n|\ell\rangle \propto 26^{-0.6\ell}$. Using our corpus (which is certainly different), we find that such an exponential decay for the mean is valid in a range of ℓ between 1 and 5, approximately. In contrast, the scale parameter θ shows an approximate power-law decay from about $\ell = 6$ to 15, with an exponent δ around 3 (or 2.8, to be more precise), i.e.,

$$\theta(\ell) \propto \frac{1}{\ell^{\delta}}$$

(note that Herdan assumed this exponent to be 2.4, with no clear empirical support [12]). Beyond $\ell = 15$, the decay of $\theta(\ell)$ is much faster. Nevertheless, these results are somewhat qualitative.

With these limitations, we could write a new version of the scaling law as

$$f(n|\ell) \simeq \ell^{\delta\alpha} g\left(\ell^{\delta} n\right) \tag{3}$$

where the proportionality constant between θ and ℓ^{δ} has been reabsorbed in the scaling function g. The corresponding data collapse is shown in Figure 7, for $5 \le \ell \le 14$. Despite the rough approximation provided by the power-law decay of $\theta(\ell)$, the data collapse in terms of scaling law (3) is nearly excellent for $\delta = 2.8$. This version of the scaling law provides a clean formulation of the brevity law: the characteristic scale of the distribution of n conditioned to the value of ℓ decays with increasing ℓ as $1/\ell^{\delta}$; i.e., the larger ℓ, the shorter the conditional distribution $f(n|\ell)$, quantified by the exponent δ.

However, in addition to a new understanding of the brevity law, the scaling law in terms of ℓ provides, as a by-product, an empirical explanation of the origin of Zipf's law. In the regime of ℓ in which the scaling law is approximately valid, i.e., for $\ell_1 \le \ell \le \ell_2$, we can obtain the distribution of frequency as a mixture of conditional distributions (by the law of total probability),

$$f(n) = \int_{\ell_1}^{\ell_2} f(n|\ell) f(\ell) d\ell$$

(where we take a continuous approximation, replacing sum over ℓ by integration; this is essentially a mathematical rephrasing). Substituting the scaling law and introducing the change of variables $x = \ell^{\delta} n$ we get

$$f(n) = \int_{\ell_1}^{\ell_2} \ell^{\delta\alpha} g\left(\ell^{\delta} n\right) f(\ell) d\ell \propto \int_{\ell_1^{\delta} n}^{\ell_2^{\delta} n} \left(\frac{x}{n}\right)^{\alpha} g(x) \frac{x^{-1+1/\delta}}{n^{1/\delta}} dx$$
$$= \frac{1}{n^{\alpha+1/\delta}} \int_{\ell_1^{\delta} n}^{\ell_2^{\delta} n} x^{\alpha-1+1/\delta} g(x) dx$$

where we also have taken advantage of the fact that, in the region of interest, $f(\ell)$ can be considered (in a rough approximation) as constant.

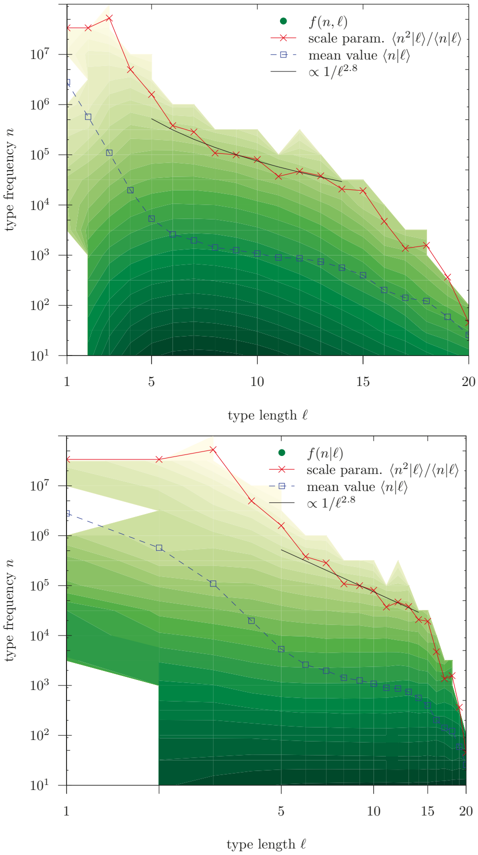


Figure 6. Estimated value of the scale parameter θ of the frequency conditional distributions ($\theta = \langle n^2|\ell\rangle/\langle n|\ell\rangle$) as a function of type length ℓ, together with conditional mean value $\langle n|\ell\rangle$. A decaying power law with exponent 2.8, shown as a guide to the eye, is close to the values of the scale parameter for $6 \leq \ell \leq 13$. The curves are overimposed to the values of the joint distribution $f(n,\ell)$ in the (**top panel**) and to the conditional distribution $f(n|\ell)$ in the (**bottom panel**). Notice that in the last case both axes are logarithmic. The shadower the green color, the higher the value of $f(n,\ell)$ and $f(n|\ell)$.

From here, we can see that in the case where the frequency is small ($n \ll \theta(\ell_2)$), the integration limits are also small, and then the last integral scales with n as $n^{1/\delta}$ (because we have that $g(x) \propto 1/x^{\alpha}$), which implies that we recover a power law with exponent α for $f(n)$, i.e., $f(n) \propto 1/n^{\alpha}$. However, for larger frequencies (n above $\theta(\ell_2)$ but below $\theta(\ell_1)$), the integral does not scale with n but can be considered instead as constant and then we get Zipf's law as

$$f(n) \propto n^{-\left(\alpha+\frac{1}{\delta}\right)}.$$

This means that Zipf's exponent can be obtained from the values of the intermediate-frequency power-law conditional exponent α and the brevity exponent δ as

$$\beta_z = \alpha + \frac{1}{\delta},$$

where we have introduced a subscript z in β to stress that this is the β exponent appearing in Zipf's law, corresponding to the marginal distribution $f(n)$, and to distinguish it from the one of the conditional distributions, that we may call β_c. Note then that β_c plays no role in the determination of β_z, and, in fact, the scaling function does not need to have a power-law tail to obtain Zipf's law. This sort of argument is similar to the one used in statistical seismology [32], but in that case the scaling law was elementary (i.e., $\theta = \langle n|\ell\rangle$).

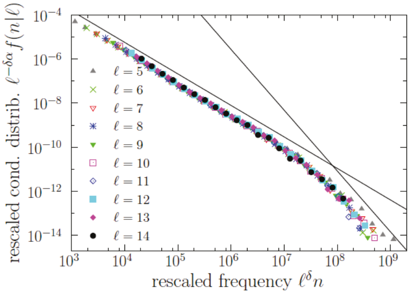


Figure 7. Same as Figure 5, from $\ell = 5$ to 14, changing the scale factors from combination of powers of moments ($\langle n|\ell\rangle$ and $\langle n^2|\ell\rangle$) to powers of length (in concrete, $\ell^{-\delta}$ and $\ell^{\delta\alpha}$). The collapse signals the fulfilment of a scaling law. Two decreasing power laws with exponents 1.43 and 2.76 are shown as straight lines, for comparison.

We can check the previous exponent relation using the empirical values of the exponent. We do not have a unique measure of α, but from Table 2, we see that its value for the different $f(n|\ell)$ is quite well defined. Taking the harmonic mean between the values $4 \le \ell \le 14$ we get $\bar{\alpha} = 1.43$, which together with $\delta = 2.8$ leads to $\beta_z \simeq 1.79$, not far from the ideal Zipf's value $\beta_z = 2$ and closer to the empirical value $\beta_z = 1.94$. The reason to calculate the harmonic mean of the exponents comes from the fact that it is the maximum-likelihood outcome when untruncated power-law datasets are put together [33]; when the power laws are truncated, the result is closer to the untruncated case when the range b/a is large.

5. Conclusions

Using a large corpus of English texts, we have seen how three important laws of quantitative linguistics, which are the type-length law, Zipf's law of word frequency, and the brevity law, can be put into a unified framework just considering the joint distribution of length and frequency.

Straightforwardly, the marginals of the joint distribution provide both the type-length distribution and the word-frequency distribution. We reformulate the type-length law, finding that the gamma distribution provides an excellent fit of type lengths for values larger than 2, in contrast to the previously proposed lognormal distribution [12] (although some previous research was dealing not with type length but with token length [13]). For the distribution of word frequency, we confirm the well-known Zipf's law, with an exponent $\beta_z = 1.94$; we also confirm the second intermediate power-law regime that emerges in large corpora [16], with an exponent $\alpha = 1.4$.

The advantages of the perspective provided by considering the length-frequency joint distribution become apparent when dealing with the brevity phenomenon. In concrete, this property arises very clearly when looking at the distributions of frequency conditioned to fixed length. These show a well-defined shape, characterized by a power-law decay for intermediate frequencies followed by a faster decay, which is well modeled by a second power law, for larger frequencies. The exponent α for the intermediate regime turns out to be the same as the one for the usual (marginal) distribution of frequency, $\alpha \simeq 1.4$. However, the exponent for higher frequencies β_c turns out to be larger than 2 and unrelated to Zipf's law.

At this point, scaling analysis reveals as a very powerful tool to explore and formulate the brevity law. We observe that the conditional frequency distributions show scaling for different values of length, i.e., when the distributions are rescaled by a scale parameter (proportional to the characteristic scale of each distribution), these distributions collapse into a unique curve, showing that they share a common shape (although at different scales). The characteristic scale of the distributions turns out to be well described by the scale parameter (given by the ratio of moments $\langle n^2|\ell\rangle / \langle n|\ell\rangle$), instead than by the mean value ($\langle n|\ell\rangle$). This is the usual case when the distributions involved have a power-law shape (with exponent $\alpha > 1$) close to the origin [29]. This also highlights the importance of looking at the whole distribution and not to mean values when one is dealing with complex phenomena.

Going further, we obtain that the characteristic scale of the conditional frequency distributions decays, approximately, as a power law of the type length, with exponent δ, which allows us to rewrite the scaling law in a form that is reminiscent to the one used in the theory of phase transitions and critical phenomena. Despite that the power-law behavior for the characteristic scale of frequency is rather rough, the derived scaling law shows an excellent agreement with the data. Note that taking together the marginal length distribution, Equation (1), and the scaling law for the conditional frequency distribution, Equation (3), we can write for the joint distribution

$$f(\ell, n) \propto \lambda^{\gamma} \ell^{\delta\alpha+\gamma-1} g(\ell^{\delta} n) e^{-\lambda\ell},$$

with the scaling function $g(x)$ given by Equation (2), up to proportionality factors.

Finally, the fulfilment of a scaling law of this form allows us to obtain a phenomenological (model free) explanation of Zipf's law as a mixture of the conditional distributions of frequencies. In contrast to some accepted explanations of Zipf's law, which put the origin of the law outside the linguistic realm (such as Simon's model [15], where only the reinforced growth of the different types counts; other explanations are in [19,34]), our approach indicates that the origin of Zipf's law can be fully linguistic, as it depends crucially on the length of the words (and the length is a purely linguistic attribute). Thus, at fixed length, each (conditional) frequency distribution shows a scale-free (power-law) behavior, up to a characteristic frequency where the power law (with exponent α) breaks down. This breaking-down frequency depends on length through the exponent δ. The mixture of different power laws, with exponent α and cut at a scale governed by the exponent δ, yields a Zipf's exponent $\beta_z = \alpha + \delta^{-1}$. Strictly speaking, our explanation of Zipf's law does not fully explain Zipf's law, but transfers the explanation to the existence of a power law with a smaller exponent ($\alpha \simeq 1.4$) as well as to the crossover frequency that depends on length as $\ell^{-\delta}$. Clearly, more research is necessary to explain the shape of the conditional distributions. It is noteworthy that a similar phenomenology for Zipf's law (in general) was proposed in [34], using the concept of "underlying unobserved variables", which in the case of word frequencies were associated (without quantification) to part of speech (grammatical categories). From our point of view, the "underlying unobserved variables" in the case of word frequencies would be instead word (type) lengths.

Although our results are obtained using a unique English corpus, we believe they are fully representative of this language, at least when large corpora are used. Naturally, further investigations are needed to confirm the generality of our results. Of course, a necessary extension of our work is the use of corpora on other languages, to establish the universality of our results, as done, e.g., in [14]. The length of words is simply measured in number of characters, but nothing precludes the use of number of phonemes or mean time duration of types (in speech, as in [13]). At the end, the goal of this kind of research is to pursue a unified theory of linguistic laws, as proposed in [35]. The line of research shown in this paper seems to be a promising one.

Author Contributions: Methodology, Á.C. and I.S.; writing, Á.C.; visualization, I.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Spanish MINECO, through grants FIS2012-31324, FIS2015-71851-P, PGC-FIS2018-099629-B-I00, and MDM-2014-0445 (María de Maeztu Program). I.S. was funded through the Collaborative Mathematics Project from La Caixa Foundation.

Acknowledgments: We are indebted to Francesc Font-Clos for providing the valuable corpus released in [20]. Our interest in the brevity law arose from our interaction with Ramon Ferrer-i-Cancho, in particular from the reading of [35,36].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zipf, G.K. *Human Behavior and the Principle of Least Effort*; Addison-Wesley: Boston, MA, USA, 1949.
2. Baayen, R.H. *Word Frequency Distributions*; Kluwer: Dordrecht, The Netherlands, 2001.
3. Baroni, M. Distributions in text. In *Corpus linguistics: An International Handbook*; Lüdeling, A., Kytö, M., Eds.; Mouton de Gruyter: Berlin, Germany, 2009; Volume 2, pp. 803–821.
4. Zanette, D. Statistical patterns in written language. *arXiv* **2014**, arXiv:1412.3336v1.
5. Piantadosi, S.T. Zipf's law in natural language: A critical review and future directions. *Psychon. Bull. Rev.* **2014**, *21*, 1112–1130. [CrossRef] [PubMed]
6. Moreno-Sánchez, I.; Font-Clos, F.; Corral, A. Large-scale analysis of Zipf's law in English texts. *PLoS ONE* **2016**, *11*, e0147073. [CrossRef] [PubMed]
7. Corral, A.; Serra, I.; Ferrer-i-Cancho, R. The distinct flavors of Zipf's law in the rank-size and in the size-distribution representations, and its maximum-likelihood fitting. *arXiv* **2019**, arXiv:1908.01398.
8. Mandelbrot, B. On the theory of word frequencies and on related Markovian models of discourse. In *Structure of Language and its Mathematical Aspects*; Jakobson, R., Ed.; American Mathematical Society: Providence, RI, USA, 1961; pp. 190–219.
9. Heaps, H.S. *Information retrieval: Computational and Theoretical Aspects*; Academic Press: Cambridge, MA, USA, 1978.
10. Font-Clos, F.; Corral, A. Log-log convexity of type-token growth in Zipf's systems. *Phys. Rev. Lett.* **2015**, *114*, 238701. [CrossRef]
11. Altmann, E.G.; Gerlach, M. Statistical laws in linguistics. In *Creativity and Universality in Language. Lecture Notes in Morphogenesis*; Esposti, M.D., Altmann, E.G., Pachet, F., Eds.; Springer: Berlin/Heidelberger, Germany, 2016.
12. Herdan, G. The Relation Between the Dictionary Distribution and the Occurrence Distribution of Word Length and its Importance for the Study of Quantitative Linguistics. *Biometrika* **1958**, *45*, 222–228. [CrossRef]
13. Torre, I.G.; Luque, B.; Lacasa, L.; Kello, C.T.; Hernández-Fernández, A. On the physical origin of linguistic laws and lognormality in speech. *R. Soc. Open Sci.* **2019**, *6*, 191023. [CrossRef]
14. Bentz, C.; Ferrer-i-Cancho, R. Zipf's law of abbreviation as a language universal. In *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*; Bentz, C., Jäger, G., Yanovich, I., Eds.; University of Tübingen: Tübingen, Germany, 2016.
15. Simon, H.A. On a class of skew distribution functions. *Biometrika* **1955**, *42*, 425–440. [CrossRef]
16. Ferrer i Cancho, R.; Solé, R.V. Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited. *J. Quant. Linguist.* **2001**, *8*, 165–173. [CrossRef]
17. Williams, J.R.; Bagrow, J.P.; Danforth, C.M.; Dodds, P.S. Text mixing shapes the anatomy of rank-frequency distributions. *Phys. Rev. E* **2015**, *91*, 052811. [CrossRef]
18. Stephens, G.J.; Bialek, W. Statistical mechanics of letters in words. *Phys. Rev. E* **2010**, *81*, 066119. [CrossRef] [PubMed]
19. Corral, A.; García del Muro, M. From Boltzmann to Zipf through Shannon and Jaynes. *Entropy* **2020**, *22*, 179. [CrossRef]
20. Gerlach, M.; Font-Clos, F. A standardized Project Gutenberg Corpus for statistical analysis of natural language and quantitative linguistics. *Entropy* **2020**, *22*, 126. [CrossRef]
21. Peters, O.; Deluca, A.; Corral, A.; Neelin, J.D.; Holloway, C.E. Universality of rain event size distributions. *J. Stat. Mech.* **2010**, *11*, P11030. [CrossRef]
22. Deluca, A.; Corral, A. Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions. *Acta Geophys.* **2013**, *61*, 1351–1394. [CrossRef]

23. Corral, A.; González, A. Power law distributions in geoscience revisited. *Earth Space Sci.* **2019**, *6*, 673–697. [CrossRef]
24. Corral, A.; Boleda, G.; Ferrer-i-Cancho, R. Zipf's law for word frequencies: Word forms versus lemmas in long texts. *PLoS ONE* **2015**, *10*, e0129031. [CrossRef]
25. Clauset, A.; Shalizi, C.R.; Newman, M.E.J. Power-law distributions in empirical data. *SIAM Rev.* **2009**, *51*, 661–703. [CrossRef]
26. Corral, A.; Font, F.; Camacho, J. Non-characteristic half-lives in radioactive decay. *Phys. Rev. E* **2011**, *83*, 066103. [CrossRef]
27. Voitalov, I.; van der Hoorn, P.; van der Hofstad, R.; Krioukov, D. Scale-free networks well done. *Phys. Rev. Res.* **2019**, *1*, 033034. [CrossRef]
28. Deluca, A.; Corral, A. Scale invariant events and dry spells for medium-resolution local rain data. *Nonlinear Proc. Geophys.* **2014**, *21*, 555–567. [CrossRef]
29. Corral, A. Scaling in the timing of extreme events. *Chaos Solitons Fract.* **2015**, *74*, 99–112. [CrossRef]
30. Font-Clos, F.; Boleda, G.; Corral, A. A scaling law beyond Zipf's law and its relation to Heaps' law. *New J. Phys.* **2013**, *15*, 093033. [CrossRef]
31. Corral, A.; Font-Clos, F. Dependence of exponents on text length versus finite-size scaling for word-frequency distributions. *Phys. Rev. E* **2017**, *96*, 022318. [CrossRef]
32. Corral, A. Statistical features of earthquake temporal occurrence. In *Modelling Critical and Catastrophic Phenomena in Geoscience*; Bhattacharyya, P., Chakrabarti, B.K., Eds.; Springer: Berlin/Heidelberger, Germany, 2007.
33. Navas-Portella, V.; Serra, I.; Corral, A.; Vives, E. Increasing power-law range in avalanche amplitude and energy distributions. *Phys. Rev. E* **2018**, *97*, 022134. [CrossRef]
34. Aitchison, L.; Corradi, N.; Latham, P.E. Zipf's law arises naturally when there are underlying, unobserved variables. *PLoS Comput. Biol.* **2016**, *12*, e1005110. [CrossRef]
35. Ferrer-i-Cancho, R. Compression and the origins of Zipf's law for word frequencies. *Complexity* **2016**, *21*, 409–411. [CrossRef]
36. Ferrer-i-Cancho, R.; Bentz, C.; Seguin, C. Compression and the origins of Zipf's law of abbreviation. *arXiv* **2015**, arXiv:1504.04884.

entropy

MDPI

Article

Asymptotic Analysis of the kth Subword Complexity

Lida Ahmadi [1,*,†] **and Mark Daniel Ward** [2]

[1] Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA
[2] Department of Statistics, Purdue University, West Lafayette, IN 47907, USA; mdw@purdue.edu
[*] Correspondence: lida.ahmadi@csusb.edu
[†] Current address: Department of Mathematics, 5500 University Parkway, San Bernardino, CA 92407, USA.

Received: 25 December 2019; Accepted: 4 February 2020; Published: 12 February 2020

Abstract: Patterns within strings enable us to extract vital information regarding a string's randomness. Understanding whether a string is random (Showing no to little repetition in patterns) or periodic (showing repetitions in patterns) are described by a value that is called the kth Subword Complexity of the character string. By definition, the kth Subword Complexity is the number of distinct substrings of length k that appear in a given string. In this paper, we evaluate the expected value and the second factorial moment (followed by a corollary on the second moment) of the kth Subword Complexity for the binary strings over memory-less sources. We first take a combinatorial approach to derive a probability generating function for the number of occurrences of patterns in strings of finite length. This enables us to have an exact expression for the two moments in terms of patterns' auto-correlation and correlation polynomials. We then investigate the asymptotic behavior for values of $k = \Theta(\log n)$. In the proof, we compare the distribution of the kth Subword Complexity of binary strings to the distribution of distinct prefixes of independent strings stored in a trie. The methodology that we use involves complex analysis, analytical poissonization and depoissonization, the Mellin transform, and saddle point analysis.

Keywords: subword complexity; asymptotics; generating functions; saddle point method; probability; the Mellin transform; moments

1. Introduction

Analyzing and understanding occurrences of patterns in a character string is helpful for extracting useful information regarding the nature of a string. We classify strings to low-complexity and high-complexity, according to their level of randomness. For instance, we take the binary string $X = 10101010...$, which is constructed by repetitions of the pattern $w = 10$. This string is periodic, and therefore has low randomness. Such periodic strings are classified as low-complexity strings, whereas strings that do not show periodicity are considered to have high complexity. An effective way of measuring a string's randomness is to count all distinct patterns that appear as contiguous subwords in the string. This value is called the Subword Complexity. The name is given by Ehrenfeucht, Lee, and Rozenberg [1], and initially was introduced by Morse and Hedlund in 1938 [2]. The higher the Subword Complexity, the more complex the string is considered to be.

Assessing information about the distribution of the Subword Complexity enables us to better characterize strings, and determine atypically random or periodic strings that have complexities far from the average complexity [3]. This type of string classification has applications in fields such as data compression [4], genome analysis (see [5–9]), and plagiarism detection [10]. For example, in data compression, a data set is considered compressible if it has low complexity, as consists of repeated subwords. In computational genomics, Subword Complexity (known as k-mers) is used in detection of repeated sequences and DNA barcoding [11,12]. k-mers are composed of A, T, G, and C nucleotides. For instance, 7-mers for a DNA sequence GTAGAGCTGT is four, meaning that there are 4-hour distinct substrings of length 7 in the given DNA sequence. Counting k-mers becomes challenging for longer DNA sequences. Our results can be easily extended to the alphabet $\{A, T, G, C\}$ and directly applied in theoretical analysis of the genomic k-mer distributions under the Bernoulli probabilistic model, particularly when the length n of the sequence approaches infinity.

There are two variations for the definition of the Subword Complexity: the one that counts all distinct subwords of a given string (also known as Complexity Index and Sequence Complexity [13]), and the one that only counts the subwords of the same length, say k, that appear in the string. In our work, we analyze the latter, and we call it the kth Subword Complexity to avoid any confusion.

Throughout this work, we consider the kth Subword Complexity of a random binary string of length n over a memory-less source, and we denote it by $X_{n,k}$. We analyze the first and second factorial moments of $X_{n,k}$ (1) for the range $k = \Theta(\log n)$, as $n \to \infty$. More precisely, will divide the analysis into three ranges as follows.

i. $\frac{1}{\log q^{-1}} \log n < k < \frac{2}{\log q^{-1} + \log p^{-1}} \log n,$

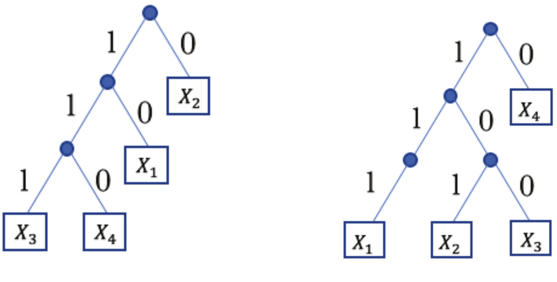
ii. $\frac{2}{\log q^{-1} + \log p^{-1}} \log n < k < \frac{1}{q \log q^{-1} + p \log p^{-1}} \log n$, and

iii. $\frac{1}{q \log q^{-1} + p \log p^{-1}} \log n < k < \frac{1}{\log p^{-1}} \log n.$

Our approach involves two major steps. First, we choose a suitable model for the asymptotic analysis, and afterwards we provide proofs for the derivation of the asymptotic expansion of the first two factorial moments.

1.1. Part I

This part of the analysis is inspired by the earlier work of Jacquet and Szpankowski [14] on the analysis of suffix trees by comparing them to independent tries. A trie, first introduced by René de la Briandais in 1959 (see [15]), is a search tree that stores n strings, according to their prefixes. A suffix tree, introduced by Weiner in 1973 (see [16]), is a trie where the strings are suffixes of a given string. An example of these data structures are given in Figure 1.



(**a**) Suffix Tree (**b**) Trie

Figure 1. The suffix tree in (**a**) is built over the first four suffixes of string $X = 101110...$, and the trie in (**b**) is build over strings $X_1 = 111...$, $X_2 = 101...$, $X_3 = 100$, and $X_4 = 010....$

A direct asymptotic analysis of the moments is a difficult task, as patterns in a string are not independent from each other. However, we note that each pattern in a string can be regarded as a prefix of a suffix of the string. Therefore, the number of distinct patterns of length k in a string is actually the number of nodes of the suffix tree at level k and lower. It is shown by I. Gheorghiciuc and M. D. Ward [17] that the expected value of the k-th Subword Complexity of a Bernoulli string of length n is asymptotically comparable to the expected value of the number of nodes at level k of a trie built over n independent strings generated by a memory-less source.

We extend this analysis to the desired range for k, and we prove that the result holds for when k grows logarithmically with n. Additionally, we show that asymptotically, the second factorial moment of the k-th Subword Complexity can also be estimated by admitting the same independent model generated by a memory-less source. The proof of this theorem heavily relies on the characterization of the overlaps of the patterns with themselves and with one another. Autocorrelation and correlation polynomials explicitly describe these overlaps. The analytic properties of these polynomials are key to understanding repetitions of patterns in large Bernoulli strings. This, in conjunction with Cauchy's integral formula (used to compare the generating functions in the two models) and the residue theorem, provides solid verification that the second factorial moment in the Subword Complexity behaves the same as in the independent model.

To make this comparison, we derive the generating functions of the first two factorial moments in both settings. In a paper published by F. Bassino, J. Clément, and P. Nicodème in 2012 [18], the authors provide a multivariate probability generating function $f(z,x)$ for the number of occurrences of patterns in a finite Bernoulli string. That is, given a pattern w, the coefficient of the term $z^n x^m$ in $f(z,x)$ is the probability in the Bernoulli model that a random string of size n has exactly m occurrences of the pattern w. Following their technique, we derive the exact expression for the generating functions of the first two factorial moments of the kth Subword Complexity. In the independent model, the generating functions are obtained by basic probability concepts.

1.2. Part II

This part of the proof is analogous to the analysis of profile of tries [19]. To capture the asymptotic behavior, the expressions for the first two factorial moments in the independent trie are further improved by means of a Poisson process. The poissonized version yields generating functions in the form of harmonic sums for each of the moments. The Mellin transform and the inverse Mellin transforms of these harmonic sums establish a connection between the asymptotic expansion and singularities of the transformed function. This methodology is sufficient for when the length k of the patterns are fixed. However, allowing k to grow with n, makes the analysis more challenging. This is because for large k, the dominant term of the poissonized generating function may come from the term involving k, and singularities may not be significant compared to the growth of k. This issue is treated by combining the singularity analysis with a saddle point method [20]. The outcome of the analysis is a precise first-order asymptotics of the moments in the poissonized model. Depoissonization theorems are then applied to obtain the desired result in the Bernoulli model.

2. Results

For a binary string $X = X_1X_2...X_n$, where X_i's ($i = 1, ..., n$) are independent and identically distributed random variables , we assume that $\mathbf{P}(X_i = 1) = p$, $\mathbf{P}(X_i = 0) = q = 1 - p$, and $p > q$. We define the kth Subword Complexity, $X_{n,k}$, to be the number of distinct substrings of length k that appear in a random string X with the above assumptions. In this work, we obtain the first order asymptotics for the average and the second factorial moment of $X_{n,k}$. The analysis is done in the range $k = \Theta(\log n)$. We rewrite this range as $k = a \log n$, and by performing a saddle point analysis, we will show that

$$1/\log q^{-1} < a < 1/\log p^{-1} \tag{1}$$

In the first step, we compare the kth Subword Complexity to an independent model constructed in the following way: We store a set of n independently generated strings by a memory-less source in a trie. This means that each string is a sequence of independent and identically distributed Bernoulli random variables from the binary alphabet $\mathcal{A} = \{0, 1\}$, with $\mathbf{P}(1) = p$, $\mathbf{P}(0) = q = 1 - p$. We denote the number of distinct prefixes of length k in the trie by $\hat{X}_{n,k}$, and we call it *the kth prefix complexity*.

Before proceeding any further, we remind that factorial moments of a random variable are defined as following.

Definition 1. *The jth factorial moment of a random variable X is defined as*

$$\mathbf{E}[(X)_j] = \mathbf{E}[(X)(X-1)(X-2)...(X-j+1)], \tag{2}$$

where j = 1, 2, ... will show that the first and second factorial moments of $X_{n,k}$ are asymptotically comparable to those of $\hat{X}_{n,k}$, when $k = \Theta(\log n)$. We have the following theorems.

Theorem 1. *For large values of n, and for $k = \Theta(\log n)$, there exists $M > 0$ such that*

$$\mathbf{E}[X_{n,k}] - \mathbf{E}[\hat{X}_{n,k}] = O(n^{-M}).$$

We also prove a similar result for the second factorial moments of the *k*th Subword Complexity and the *k*th Prefix Complexity:

Theorem 2. *For large values of n, and for $k = \Theta(\log n)$, there exists $\epsilon > 0$ such that*

$$\mathbf{E}[(X_{n,k})_2] - \mathbf{E}[(\hat{X}_{n,k})_2] = O(n^{-\epsilon}).$$

In the second part of our analysis, we derive the first order asymptotics of the *k*th Prefix Complexity. The methodology used here is analogous to the analysis of profile of tries [19]. The rate of the asymptotic growth depends on the location of the value *a* as seen in (1) . For instance, for the average *k*th Subword Complexity ,$\mathbf{E}[X_{n,k}]$, we have the following observations.

i. For the range $I_1 : \dfrac{1}{\log q^{-1}} < a < \dfrac{2}{\log q^{-1} + \log p^{-1}}$, the growth rate is of order $O(2^k)$,

ii. in the range $I_2 : \dfrac{2}{\log q^{-1} + \log p^{-1}} < a < \dfrac{1}{q \log q^{-1} + p \log p^{-1}}$, we observe some oscillations with *n*, and

iii. in the range $I_3 : \dfrac{1}{q \log q^{-1} + p \log p^{-1}} < a < \dfrac{1}{\log p^{-1}}$, the average has a linear growth $O(n)$.

The above observations will be discussed in depth in the proofs of the following theorems.

Theorem 3. *The average of the kth Prefix Complexity has the following asymptotic expansion*

i. *For $a \in I_1$,*

$$\mathbf{E}[\hat{X}_{n,k}] = 2^k - \Phi_1((1+\log p)\log_{p/q} n)\frac{n^{\nu}}{\sqrt{\log n}}\left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right), \tag{3}$$

where $\nu = -r_0 + a\log(p^{-r_0} + q^{-r_0})$, and

$$\Phi_1(x) = \frac{(p/q)^{-r_0/2} + (p/q)^{r_0/2}}{\sqrt{2\pi}\log p/q}\sum_{j\in\mathbb{Z}}\Gamma(r_0 + it_j)e^{-2\pi ijx}$$

is a bounded periodic function.

ii. *For $a \in I_2$,*

$$\mathbf{E}[\hat{X}_{n,k}] = \Phi_1((1+\log p)\log_{p/q} n)\frac{n^{\nu}}{\sqrt{\log n}}\left(1+O\left(\frac{1}{\sqrt{\log n}}\right)\right).$$

iii. *For $a \in I_3$*

$$\mathbf{E}[\hat{X}_{n,k}] = n + O(n^{\nu_0}),$$

for some $\nu_0 < 1$.

Theorem 4. *The second factorial moment of the kth Prefix Complexity has the following asymptotic expansion.*

i. *For $a \in I_1$,*

$$\mathbf{E}[(\hat{X}_{n,k})_2] = \left(2^k - \Phi_1(\log_{p/q} n(1+\log p))\frac{n^{\nu}}{\sqrt{\log n}}\left(1+O\left(\frac{1}{\sqrt{\log n}}\right)\right)\right)^2.$$

ii. *For $a \in I_2$,*

$$\mathbf{E}[(\hat{X}_{n,k})_2] = \Phi_1^2(\log_{p/q} n(1+\log p))\frac{n^{2\nu}}{\log n}\left(1+O\left(\frac{1}{\log n}\right)\right).$$

iii. *For $a \in I_3$,*

$$\mathbf{E}[(\hat{X}_{n,k})_2] = n^2 + O(n^{2\nu_0}).$$

The periodic function $\Phi_1(x)$ in Theorems 3 and 4 is shown in Figure 2.

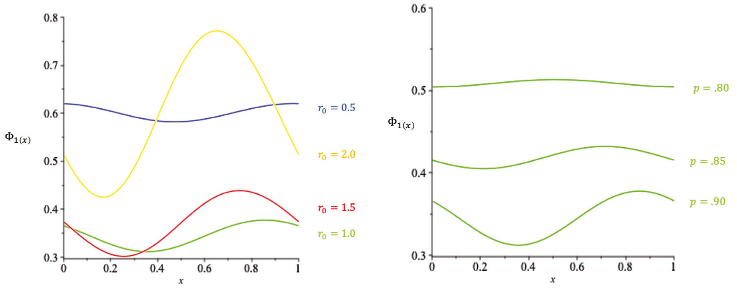


Figure 2. Left: $\Phi_1(x)$ at $p = 0.90$, and various levels of r_0. The amplitude increases as r_0 increases. **Right**: $\Phi_1(x)$ at $r_0 = 1$, and various levels of p. The amplitude tends to zero as $p \to 1/2^+$.

The results in Theorem 4 will follow for the second moment of the kth Subword Complexity as the analysis can be easily extended from the second factorial moment to the second moment. The variance however, as seen in Figure 3, does not show the same asymptotic behavior as the variance of kth Subword Complexity.

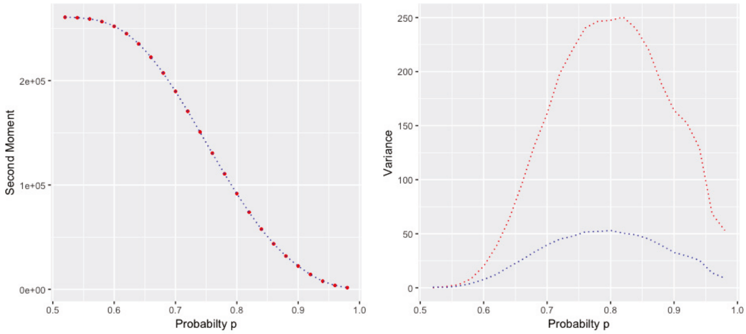


Figure 3. Approximated second moments (**left**), and variances (**right**) of the kth Subword Complexity (**red**), and the kth Prefix Complexity (**blue**), for $n = 4000$, at different probability levels, averaged over 10,000 iterations.

3. Proofs and Methods

3.1. Groundwork

We first introduce a few terminologies and lemmas regarding overlaps of patterns and their number of occurrences in texts. Some of the notations we use in this work are borrowed from [18] and [21].

Definition 2. *For a binary word $w = w_1 ... w_k$ of length k, The autocorrelation set $\mathcal{S}_w$ of the word w is defined in the following way.*

$$\mathcal{S}_w = \{w_{i+1} ... w_k \,|\, w_1 ... w_i = w_{k-i+1} ... w_k\}. \tag{4}$$

The autocorrelation index set is

$$\mathcal{P}(w) = \{i \,|\, w_1 ... w_i = w_{k-i+1} ... w_k\}, \tag{5}$$

And the autocorrelation polynomial is

$$S_w(z) = \sum_{i \in \mathcal{P}(w)} \mathbf{P}(w_{i+1} ... w_k) z^{k-i}. \tag{6}$$

Definition 3. *For the distinct binary words $w = w_1 ... w_k$ and $w' = w'_1 ... w'_{k'}$, the correlation set $\mathcal{S}_{w,w'}$ of the words w and w' is*

$$\mathcal{S}_{w,w'} = \{w'_{i+1} ... w'_k \,|\, w'_1 ... w'_i = w_{k-i+1} ... w_k\}. \tag{7}$$

The correlation index set is

$$\mathcal{P}(w, w') = \{i \,|\, w'_1 ... w'_i = w_{k-i+1} ... w_k\}, \tag{8}$$

The correlation polynomial is

$$S_{w,w'}(z) = \sum_{i \in \mathcal{P}(w,w')} \mathbf{P}(w'_{i+1} ... w'_k) z^{k-i}. \tag{9}$$

The following two lemmas present the probability generating functions for the number of occurrences of a single pattern and a pair of distinct pattern, respectively, in a random text of length n. For a detailed dissection on obtaining such generating functions, refer to [18].

Lemma 1. *The Occurrence probability generating function for a single pattern w in a binary text over a memoryless source is given by $F_w(z, x-1)$, where*

$$F_w(z,t) = \frac{1}{1 - A(z) - \dfrac{t\mathbf{P}(w)z^k}{1 - t(S_w(z) - 1)}}, \tag{10}$$

The coefficient $[z^n x^m]F_w(z, x-1)$ is the probability that a random binary string of length n has m occurrences of the pattern w.

Lemma 2. *The Occurrence PGF for two distinct Patterns of length k in a Bernoulli random text is given by $F_{w,w'}(z, x_1 - 1, x_2 - 1)$ where,*

$$F_{w,w'}(z, t_1, t_2) = \frac{1}{1 - A(z) - M(z, t_1, t_2)}, \tag{11}$$

and

$$M(z, t_1, t_2) = \begin{pmatrix} \mathbf{P}(w)z^k t_1 & \mathbf{P}(w')z^k t_2 \end{pmatrix} \left(\mathbb{I} - \begin{pmatrix} (S_w(z) - 1)t_1 & S_{w,w'}(z)t_2 \\ S_{w',w}(z)t_1 & (S_{w'}(z) - 1)t_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The coefficient $[z^n x_1^{m_1} x_2^{m_2}]F_{w,w'}(z, x_1 - 1, x_2 - 1)$ is the probability that there are m_1 occurrences of w and m_2 occurrences of w' in a random string of length n.

The above results will be used to find the generating functions for the first two factorial moments of the kth Subword Complexity in the following section.

3.2. Derivation of Generating Functions

Lemma 3. *For generating functions $H_k(z) = \sum_{n \geq 0} \mathbf{E}[X_{n,k}]z^n$ and $G_k(z) = \sum_{n \geq 0} \mathbf{E}[(X_{n,k})_2]z^n$, we have*

i.

$$H_k(z) = \sum_{w \in \mathcal{A}^k} \left(\frac{1}{1-z} - \frac{S_w(z)}{D_w(z)} \right), \tag{12}$$

where $D_w(z) = \mathbf{P}(w)z^k + (1-z)S_w(z)$, and

ii.

$$G_k(z) = \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \left(\frac{1}{1-z} - \frac{S_w(z)}{D_w(z)} - \frac{S_{w'}(z)}{D_{w'}(z)} + \frac{S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)} \right), \tag{13}$$

where

$$\begin{aligned} D_{w,w'}(z) = {} & (1-z)(S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)) \\ & + z^k \left(\mathbf{P}(w)(S_{w'}(z) - S_{w,w'}(z)) + \mathbf{P}(w')(S_w(z) - S_{w',w}(z)) \right). \end{aligned} \tag{14}$$

Proof.
i. We define

$$X_{n,k}^{(w)} = \begin{cases} 1 & \text{if } w \text{ appears at least once in string } X \\ 0 & \text{otherwise.} \end{cases}$$

This yields

$$\begin{aligned} \mathbf{E}[X_{n,k}^{(w)}] &= \mathbf{P}(X_{n,k}^{(w)} = 1) \\ &= 1 - P(X_{n,k}^{(w)} = 0) \\ &= 1 - [z^n x^0] F_w(z, x). \end{aligned} \tag{15}$$

We observe that $[z^n x^0] F_w(z,x) = [z^n] F_w(z,0)$. By defining $f_w(z) = F_w(z,0)$ and from (10), we obtain

$$f_w(z) = \frac{S_w(z)}{\mathbf{P}(w) z^k + (1-z) S_w(z)}. \tag{16}$$

Having the above function, we derive the following result.

$$\begin{aligned} H(z) &= \sum_{n \geq 0} \mathbf{E}[X_{n,k}] z^n \\ &= \sum_{n \geq 0} \sum_{w \in \mathcal{A}^k} (1 - [z^n] f_w(z)) z^n \\ &= \sum_{w \in \mathcal{A}^k} \left(\frac{1}{1-z} - f_w(z) \right) \\ &= \sum_{w \in \mathcal{A}^k} \left(\frac{1}{1-z} - \frac{S_w(z)}{D_w(z)} \right). \end{aligned} \tag{17}$$

ii. For this part, we first note that

$$\begin{aligned} \mathbf{E}[(X_{n,k})_2] &= \mathbf{E}[X_{n,k}^2] - \mathbf{E}[X_{n,k}] \\ &= \mathbf{E}\left[(X_{n,k}^{(w)} + \ldots + X_{n,k}^{(w^{(r)})})^2 \right] - \mathbf{E}\left[X_{n,k}^{(w)} + \ldots + X_{n,k}^{(w^{(r)})} \right] \\ &= \sum_{w \in \mathcal{A}^k} \mathbf{E}\left[(X_{n,k}^{(w)})^2 \right] + \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \mathbf{E}\left[X_{n,k}^{(w)} X_{n,k}^{(w')} \right] - \sum_{w \in \mathcal{A}^k} \mathbf{E}\left[X_{n,k}^{(w)} \right] \\ &= \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \mathbf{E}\left[X_{n,k}^{(w)} X_{n,k}^{(w')} \right]. \end{aligned} \tag{18}$$

Due to properties of indicator random variables, we observe that the expected value of the second factorial moment has only one term:

$$\mathbf{E}[(X_{n,k})_2] = \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \mathbf{E}\left[X_{n,k}^{(w)} X_{n,k}^{(w')} \right]. \tag{19}$$

We proceed by defining a second indicator variable as following.

$$X_{n,k}^{(w)} X_{n,k}^{(w')} = \begin{cases} 1 & \text{if } X_{n,k}^{(w)} = X_{n,k}^{(w')} = 1 \\ 0 & \text{otherwise,} \end{cases}$$

This gives

$$\begin{aligned}\mathbf{E}[X_{n,k}^{(w)}X_{n,k}^{(w')}] &= \mathbf{P}\left(X_{n,k}^{(w)}=1, X_{n,k}^{(w')}=1\right)\\ &= 1-\mathbf{P}\left(X_{n,k}^{(w)}=0\cup X_{n,k}^{(w')}=0\right)\\ &= 1-\mathbf{P}\left(X_{n,k}^{(w)}=0\right)-\mathbf{P}\left(X_{n,k}^{(w')}=0\right)+\mathbf{P}\left(X_{n,k}^{(w)}=0, X_{n,k}^{(w')}=0\right).\end{aligned}$$

Finally, we are able to express $\mathbf{E}[(X_{n,k})_2]$ in the following

$$\mathbf{E}[(X_{n,k})_2] = \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} \left(1-[z^n]\,f_w(z)-[z^n]f_{w'}(z)+[z^n]f_{ww'}(z)\right), \tag{20}$$

where $f_{w,w'}(z)=F_{w,w'}(z,0,0)$ and $[z^n]F_{w,w'}(z,0,0)=[z^nx_1^0x_2^0]F_{w,w'}(z,x_1,x_2)$. By (11) we have

$$f_{w,w'}(z)=\frac{S_w(z)S_{w'}(z)-S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)} \tag{21}$$

Having the above expression, we finally obtain

$$\begin{aligned}G_k(z) &= \sum_{n\geq 0}\mathbf{E}[(X_{n,k})_2]z^n\\ &= \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\sum_{n\geq 0}\left(1-[z^n]f_w(z)-[z^n]f_{w'}(z)+[z^n]f_{w,w'}(z)\right)z^n\\ &= \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\left(\frac{1}{1-z}-f_w(z)-f_{w'}(z)+f_{w,w'}(z)\right)\\ &= \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\left(\frac{1}{1-z}-\frac{S_w(z)}{D_w(z)}-\frac{S_{w'}(z)}{D_{w'}(z)}+\frac{S_w(z)S_{w'}(z)-S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)}\right).\end{aligned} \tag{22}$$

□

In the following lemma, we present the generating functions for the first two factorial moments for the kth Prefix Complexity in the independent model.

Lemma 4. *For $\hat{H}_k(z)=\sum_{n\geq 0}\mathbf{E}[\hat{X}_{n,k}]z^n$ and $\hat{G}_k(z)=\sum_{n\geq 0}\mathbf{E}[(\hat{X}_{n,k})_2]z^n$, which are the generating functions for $\mathbf{E}[\hat{X}_{n,k}]$ and $\mathbf{E}[(\hat{X}_{n,k})_2]$ respectively, we have*

i.

$$\hat{H}_k(z)=\sum_{w\in\mathcal{A}^k}\left(\frac{1}{1-z}-\frac{1}{1-(1-\mathbf{P}(w))z}\right). \tag{23}$$

ii.

$$\begin{aligned}\hat{G}_k(z) = &\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\left(\frac{1}{1-z}-\frac{1}{1-(1-\mathbf{P}(w))z}-\frac{1}{1-(1-\mathbf{P}(w'))z}\right)\\ &+\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\frac{1}{1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z}.\end{aligned} \tag{24}$$

Proof.
i. We define the indicator variable $\hat{X}_{n,k}^{(w)}$ as follows.

$$\hat{X}_{n,k}^{(w)} = \begin{cases} 1 & \text{if } w \text{ is a prefix of at least one string in } P \\ 0 & \text{otherwise.} \end{cases}$$

For each $\hat{X}_{n,k}^{(w)}$, we have

$$\begin{aligned} \mathbf{E}[\hat{X}_{n,k}^{(w)}] &= \mathbf{P}(\hat{X}_{n,k}^{(w)} = 1) \\ &= 1 - P(\hat{X}_{n,k}^{(w)} = 0) \\ &= 1 - (1 - \mathbf{P}(w))^n . \end{aligned} \tag{25}$$

Summing over all words w of length k, determines the generating function $\hat{H}(z)$:

$$\begin{aligned} \hat{H}(z) &= \sum_{n \geq 0} \mathbf{E}[\hat{X}_{n,k}] z^n \\ &= \sum_{w \in \mathcal{A}^k} \left(\frac{1}{1-z} - \frac{1}{1-(1-\mathbf{P}(w))z} \right) . \end{aligned} \tag{26}$$

ii. Similar to in (18) and (20), we obtain

$$\begin{aligned} \mathbf{E}[(\hat{X}_{n,k})_2] &= \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \mathbf{E}[\hat{X}_{n,k}^{(w)} \hat{X}_{n,k}^{(w')}] \\ &= \sum_{\substack{w,w^{'} \in \mathcal{A}^k \\ w \neq w'}} \left(1 - (1-\mathbf{P}(w))^n - (1-\mathbf{P}(w'))^n + (1 - \mathbf{P}(w) - \mathbf{P}(w'))^n \right) . \end{aligned} \tag{27}$$

Subsequently, we obtain the generating function below.

$$\begin{aligned} \hat{G}(z) &= \sum_{n \geq 0} \mathbf{E}[(\hat{X}_{n,k})_2] z^n \\ &= \sum_{\substack{w,w^{'} \in \mathcal{A}^k \\ w \neq w'}} \sum_{n \geq 0} \left(1 - (1-\mathbf{P}(w))^n - (1-\mathbf{P}(w'))^n + (1 - \mathbf{P}(w) - \mathbf{P}(w'))^n \right) z^n \\ &= \sum_{\substack{w,w^{'} \in \mathcal{A}^k \\ w \neq w'}} \left(\frac{1}{1-z} - \frac{1}{1-(1-\mathbf{P}(w))z} - \frac{1}{1-(1-\mathbf{P}(w'))z} \right) \\ &\quad + \sum_{\substack{w,w^{'} \in \mathcal{A}^k \\ w \neq w'}} \frac{1}{1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z} . \end{aligned} \tag{28}$$

□

Our first goal is to compare the coefficients of the generating functions in the two models. The coefficients are expected to be asymptotically equivalent in the desired range for k. To compare the coefficients, we need more information on the analytic properties of these generating functions. This will be discussed in Section 3.3.

3.3. Analytic Properties of the Generating Functions

Here, we turn our attention to the smallest singularities of the two generating functions given in Lemma 3. It has been shown by Jacquet and Szpankowski [21] that $D_w(z)$ has exactly one root in the disk $|z| \leq \rho$. Following the notations in [21], we denote the root within the disk $|z| \leq \rho$ of $D_w(z)$ by A_w, and by bootstrapping we obtain

$$A_w = 1 + \frac{1}{S_w(1)}\mathbf{P}(w) + O\left(\mathbf{P}(w)^2\right). \tag{29}$$

We also denote the derivative of $D_w(z)$ at the root A_w, by B_w, and we obtain

$$B_w = -S_w(1) + \left(k - \frac{2S'_w(1)}{S_w(1)}\mathbf{P}(w)\right) + O\left(\mathbf{P}(w)^2\right). \tag{30}$$

In this paper, we will prove a similar result for the polynomial $D_{w,w'}(z)$ through the following work.

Lemma 5. *If w and w' are two distinct binary words of length k and $\delta = \sqrt{p}$, there exists $\rho > 1$, such that $\rho\delta < 1$ and*

$$\sum_{w \in \mathcal{A}^k} [\![|S_{w,w'}(\rho)| \leq (\rho\delta)^k \theta]\!] \mathbf{P}(w) \geq 1 - \theta\delta^k. \tag{31}$$

Proof. If the minimal degree of $S_{w,w'}(z)$ is greater than $> \lfloor k/2 \rfloor$, then

$$|S_{w,w'}(\rho)| \leq (\rho\delta)^k \theta. \tag{32}$$

for $\theta = (1-p)^{-1}$. For a fixed w', we have

$$\begin{aligned}
&\sum_{w \in \mathcal{A}^k} [\![S_{w,w'}(z) \text{ has minimal degree} \leq \lfloor k/2 \rfloor]\!]\, \mathbf{P}(w) \\
&\quad = \sum_{i=1}^{\lfloor k/2 \rfloor} \sum_{w \in \mathcal{A}^k} [\![S_{w,w'}(z) \text{ has minimal degree } = i]\!]\, \mathbf{P}(w) \\
&\quad = \sum_{i=1}^{\lfloor k/2 \rfloor} \sum_{w_1 \ldots w_i \in \mathcal{A}^i} \mathbf{P}(w_1 \ldots w_i) \\
&\qquad\qquad \sum_{w_{i+1} \ldots w_k \in \mathcal{A}^{k-i}} [\![S_{w,w'}(z) \text{ has minimal degree } = i]\!] \mathbf{P}(w_{i+1} \ldots w_k) \\
&\quad \leq \sum_{i=1}^{\lfloor k/2 \rfloor} \sum_{w_1 .. w_i \in \mathcal{A}^i} \mathbf{P}(w_{i+1} \ldots w_k) p^{k-i} \\
&\quad = \sum_{i=1}^{\lfloor k/2 \rfloor} p^{k-i} \sum_{w_1 .. w_i \in \mathcal{A}^i} \mathbf{P}(w_1 \ldots w_i) \\
&\quad = \sum_{i=1}^{\lfloor k/2 \rfloor} p^{k-i} \leq \frac{p^{k - \lfloor k/2 \rfloor}}{1-p}.
\end{aligned} \tag{33}$$

This leads to the following

$$\begin{aligned}\sum_{w\in\mathcal{A}^k} &[\![\text{ every term of } S_{w,w'}(z) \text{ is of degree} > \lfloor k/2\rfloor]\!]\mathbf{P}(w)\\ &= 1 - \sum_{w\in\mathcal{A}^k}[\![S_{w,w'}(z) \text{ has a term of degree} \leq \lfloor k/2\rfloor]\!]\mathbf{P}(w)\\ &\geq 1-\frac{p^{\lceil k/2\rceil}}{1-p} \geq 1-\theta\delta^k. \end{aligned} \tag{34}$$

□

Lemma 6. *There exist $K' > 0$, and $\rho > 1$ such that $p\rho < 1$, and such that, for every pair of distinct words w, and w' of length $k \geq K'$, and for $|z| \leq \rho$, we have*

$$|S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)| > 0. \tag{35}$$

In other words, $S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)$ does not have any roots in $|z| \leq \rho$.

Proof. There are three cases to consider:
Case *i*. When either $S_w(z) = 1$ or $S_{w'}(z) = 1$, then every term of $S_{w,w'}(z)S_{w',w}(z)$ has degree k or larger, and therefore

$$|S_{w,w'}(z)S_{w',w}(z)| \leq k\frac{(p\rho)^k}{1-p\rho}. \tag{36}$$

There exists $K_1 > 0$, such that for $k > K_1$, we have $\lim_{k\to\infty} k\dfrac{(p\rho)^k}{1-p\rho} = 0$. This yields

$$\begin{aligned}|S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)| &\geq |S_w(z)S_{w'}(z)| - |S_{w,w'}(z)S_{w',w}(z)|\\ &\geq 1-k\frac{(p\rho)^k}{1-p\rho} > 0.\end{aligned} \tag{37}$$

Case *ii*. If the minimal degree for $S_w(z) - 1$ or $S_{w'}(z) - 1$ is greater than $\lfloor k/2\rfloor$, then every term of $S_{w,w'}(z)S_{w',w}(z)$ has degree at least $k/2$. We also note that, by Lemma 9, $|S_w(z)S_{w'}(z)| > 0$. Therefore, there exists $K_2 > 0$, such that

$$\begin{aligned}|S_w(z)S_{w'}(z) - S_{w',w}(z)S_{w,w'}(z)| &\geq |S_w(z)S_{w'}(z)| - |S_{w',w}(z)S_{w,w'}(z)|\\ &> 0 \quad \text{for } k > K_2.\end{aligned} \tag{38}$$

Case *iii*. The only remaining case is where the minimal degree for $S_w(z) - 1$ and $S_{w'}(z) - 1$ are both less than or equal to $\lfloor k/2\rfloor$. If $w = w_1...w_k$, then $w' = uw_1...w_{k-m}$, where u is a word of length $m \geq 1$. Then we have

$$S_{w',w}(z) = \mathbf{P}(w_{k-m+1}...w_k)z^m\left(S_w(z) - O\left((pz)^{k-m}\right)\right). \tag{39}$$

There exists $K_3 > 0$, such that

$$\begin{aligned}|S_{w',w}(z)| &\leq (p\rho)^m\left(|S_w(z)| + O\left((p\rho)^{k-m}\right)\right.\\ &= (p\rho)^m|S_w(z)| + O\left((p\rho)^k\right)\\ &< |S_w(z)| \quad \text{for } k > K_3\,.\end{aligned} \tag{40}$$

Similarly, we can show that there exists K_3', such that $|S_{w,w'}(z)| < |S_{w'}(z)|$. Therefore, for $k > K_3'$ we have

$$\begin{aligned} |S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)| &\geq |S_w(z)||S_{w'}(z)| - |S_{w,w'}(z)||S_{w',w}(z)| \\ &> |S_w(z)||S_{w'}(z)| - |S_w(z)||S_{w'}(z)| = 0. \end{aligned} \tag{41}$$

We complete the proof by setting $K' = \max\{K_1, K_2, K_3, K_3'\}$. □

Lemma 7. *There exist $K_{w,w'} > 0$ and $\rho > 1$ such that $p\rho < 1$, and for every word w and w' of length $k \geq K_{w,w'}$, the polynomial*

$$\begin{aligned} D_{w,w'}(z) = &(1-z)(S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)) \\ &+ z^k \left(\mathbf{P}(w)(S_{w'}(z) - S_{w,w'}(z)) + \mathbf{P}(w')(S_w(z) - S_{w',w}(z))\right), \end{aligned} \tag{42}$$

has exactly one root in the disk $|z| \leq \rho$.

Proof. First note that

$$\begin{aligned} |S_w(z) - S_{w',w}(z)| &\leq |S_w(z)| + |S_{w',w}(z)| \\ &\leq \frac{1}{1-p\rho} + \frac{p\rho}{1-p\rho} = \frac{1+p\rho}{1-p\rho}. \end{aligned} \tag{43}$$

This yields

$$\begin{aligned} &\left| z^k \left(\mathbf{P}(w)(S_{w'}(z) - S_{w,w'}(z)) + \mathbf{P}(w')(S_w(z) - S_{w',w}(z))\right)\right| \\ &\quad \leq (p\rho)^k \left(|S_w(z) - S_{w',w}(z)| + |S_{w'}(z) - S_{w,w'}(z)|\right) \\ &\quad \leq (p\rho)^k \left(\frac{2(1+p\rho)}{1-p\rho}\right). \end{aligned} \tag{44}$$

There exist K', K'' large enough, such that, for $k > K'$, we have

$$|(S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z))| \geq \beta > 0,$$

and for $k > K''$,

$$(p\rho)^k \left(\frac{2(1+p\rho)}{1-p\rho}\right) < (\rho - 1)\beta.$$

If we define $K_{w,w'} = \max\{K', K''\}$, then we have, for $k \geq K_{w,w'}$,

$$\begin{aligned} (p\rho)^k \left(\frac{2(1+p\rho)}{1-p\rho}\right) &< (\rho-1)\beta \\ &< |(1-z)(S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z))|. \end{aligned} \tag{45}$$

by Rouché's theorem, as $(1-z)(S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z))$ has only one root in $|z| \leq \rho$, then also $D_{w,w'}(z)$ has exactly one root in $|z| \leq \rho$. □

We denote the root within the disk $|z| \leq \rho$ of $D_{w,w'}(z)$ by $\alpha_{w,w'}$, and by bootstrapping we obtain

$$\begin{aligned} \alpha_{w,w'} = 1 &+ \frac{S_{w'}(1) - S_{w,w'}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w) \\ &+ \frac{S_w(1) - S_{w',w}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w') + O(p^{2k}). \end{aligned} \tag{46}$$

We also denote the derivative of $D_{w,w'}(z)$ at the root $\alpha_{w,w'}$, by $\beta_{w,w'}$, and we obtain

$$\beta_{w,w'} = S_{w,w'}(1)S_{w',w}(1) - S_w(1)S_{w'}(1) + O(kp^k). \tag{47}$$

We will refer to these expressions in the residue analysis that we present in the next section.

3.4. Asymptotic Difference

We begin this section by the following lemmas on the autocorrelation polynomials.

Lemma 8 (Jacquet and Szpankowski, 1994)**.** *For most words w, the autocorrelation polynomial $S_w(z)$ is very close to 1, with high probably. More precisely, if w is a binary word of length k and $\delta = \sqrt{p}$, there exists $\rho > 1$, such that $\rho\delta < 1$ and*

$$\sum_{w\in\mathcal{A}^k} [\![|S_w(\rho) - 1| \leq (\rho\delta)^k\theta]\!]\mathbf{P}(w) \geq 1 - \theta\delta^k, \tag{48}$$

where $\theta = (1-p)^{-1}$. We use Iverson notation

$$[\![A]\!] = \begin{cases} 1 & \textit{if A holds} \\ 0 & \textit{otherwise} \end{cases}$$

Lemma 9 (Jacquet and Szpankowski, 1994)**.** *There exist $K > 0$ and $\rho > 1$, such that $p\rho < 1$, and for every binary word w with length $k \geq K$ and $|z| \leq \rho$, we have*

$$|S_w(z)| > 0. \tag{49}$$

In other words, $S_w(z)$ does not have any roots in $|z| \leq \rho$.

Lemma 10. *With high probability, for most distinct pairs $\{w, w'\}$, the correlation polynomial $S_{w,w'}(z)$ is very close to 0. More precisely, if w and w' are two distinct binary words of length k and $\delta = \sqrt{p}$, there exists $\rho > 1$, such that $\rho\delta < 1$ and*

$$\sum_{w\in\mathcal{A}^k} [\![|S_{w,w'}(\rho)| \leq (\rho\delta)^k\theta]\!]\mathbf{P}(w) \geq 1 - \theta\delta^k \tag{50}$$

We will use the above results to prove that the expected values in the Bernoulli model and the model built over a trie are asymptotically equivalent. We now prove Theorem 1 below.

Proof of Theorem 1. From Lemmas 3 and 4, we have

$$H(z) = \sum_{w\in\mathcal{A}^k} \left(\frac{1}{1-z} - \frac{S_w(z)}{D_w(z)}\right),$$

and

$$\hat{H}(z) = \sum_{w\in\mathcal{A}^k} \left(\frac{1}{1-z} - \frac{1}{1-(1-\mathbf{P}(w))z}\right).$$

subtracting the two generating functions, we obtain

$$H(z) - \hat{H}(z) = \sum_{w\in\mathcal{A}^k} \left(\frac{1}{1-(1-\mathbf{P}(w))z} - \frac{S_w(z)}{D_w(z)}\right). \tag{51}$$

We define

$$\Delta_w(z) = \frac{1}{1-(1-\mathbf{P}(w))z} - \frac{S_w(z)}{D_w(z)}. \tag{52}$$

Therefore, by Cauchy integral formula (see [20]), we have

$$[z^n]\Delta_w(z) = \frac{1}{2\pi i}\oint \Delta_w(z)\frac{dz}{z^{n+1}} = \mathrm{Res}_{z=0}\,\Delta_w(z)\frac{dz}{z^{n+1}}, \tag{53}$$

where the path of integration is a circle about zero with counterclockwise orientation. We note that the above integrand has poles at $z=0$, $z=\dfrac{1}{1-\mathbf{P}(w)}$, and $z=A_w$ (refer to expression (29)). Therefore, we define

$$I^w(\rho) := \frac{1}{2\pi i}\int_{|z|=\rho} \Delta_w(z)\frac{dz}{z^{n+1}}, \tag{54}$$

where the circle of radius ρ contains all of the above poles. By the residue theorem, we have

$$\begin{aligned} I^w(\rho) &= \mathrm{Res}_{z=0}\frac{\Delta_w(z)}{z^{n+1}} + \mathrm{Res}_{z=A_w}\frac{\Delta_w(z)}{z^{n+1}} + \mathrm{Res}_{z=1/1-\mathbf{P}(w)}\frac{\Delta_w(z)}{z^{n+1}} \\ &= [z^n]\Delta_w(z) - \mathrm{Res}_{z=A_w}\frac{H_w(z)}{z^{n+1}} + \mathrm{Res}_{z=1/1-\mathbf{P}(w)}\frac{\hat{H}_w(z)}{z^{n+1}} \end{aligned} \tag{55}$$

We observe that

$$\begin{aligned} &\mathrm{Res}_{z=A_w}\frac{\Delta_w(z)}{z^{n+1}} = \frac{S_w(A_w)}{B_w\,A_w^{n+1}}, \quad \text{where } B_w \text{ is as in (30)} \\ &\mathrm{Res}_{z=1/1-\mathbf{P}(w)}\frac{\hat{H}_w(z)}{z^{n+1}} = -(1-\mathbf{P}(w))^{n+1}. \end{aligned}$$

Then we obtain

$$[z^n]\Delta_w = I^w(\rho) - \frac{S_w(A_w)}{B_w\,A_w^{n+1}} - (1-\mathbf{P}(w))^{n+1}, \tag{56}$$

and finally, we have

$$\begin{aligned} [z^n](H(z)-\hat{H}(z)) &= \sum_{w\in\mathcal{A}^k}[z^n]\Delta_w \\ &= \sum_{w\in\mathcal{A}^k} I_n^w(\rho) - \sum_{w\in\mathcal{A}^k}\left(\frac{S_w(A_w)}{B_w\,A_w^{n+1}} + (1-\mathbf{P}(w))^{n+1}\right). \end{aligned} \tag{57}$$

First, we show that, for sufficiently large n, the sum $\sum_{w\in\mathcal{A}^k}\left(\dfrac{S_w(A_w)}{B_w\,A_w^{n+1}} + (1-\mathbf{P}(w))^{n+1}\right)$ approaches zero. □

Lemma 11. *For large enough n, and for $k=\Theta(\log n)$, there exists $M>0$ such that*

$$\sum_{w\in\mathcal{A}^k}\left(\frac{S_w(A_w)}{B_w\,A_w^{n+1}} + (1-\mathbf{P}(w))^{n+1}\right) = O(n^{-M}). \tag{58}$$

Proof. We let

$$r_w(z) = (1-\mathbf{P}(w))^z + \frac{S_w(A_w)}{B_w A_w^z}. \tag{59}$$

The Mellin transform of the above function is

$$r_w^*(s) = \Gamma(s)\log^{-s}\left(\frac{1}{1-\mathbf{P}(w)}\right) - \frac{S_w(A_w)}{B_w}\Gamma(s)\log^{-s}(A_w). \tag{60}$$

We define

$$C_w = \frac{S_w(A_w)}{B_w} = \frac{S_w(A_w)}{-S_w(1)+O(k\mathbf{P}(w))}, \tag{61}$$

which is negative and uniformly bounded for all w. Also, for a fixed s, we have

$$\begin{aligned}\ln^{-s}\left(\frac{1}{1-\mathbf{P}(w)}\right) &= \ln^{-s}\left(1+\mathbf{P}(w)+O\left(\mathbf{P}(w)^2\right)\right)\\ &= \left(\mathbf{P}(w)+O\left(\mathbf{P}(w)^2\right)\right)^{-s}\\ &= \mathbf{P}(w)^{-s}\left(1+O\left(\mathbf{P}(w)\right)\right)^{-s}\\ &= \mathbf{P}(w)^{-s}\left(1+O\left(\mathbf{P}(w)\right)\right),\end{aligned} \tag{62}$$

$$\begin{aligned}\ln^{-s}(A_w) = \ln^{-s}&\left(1-\left(-\frac{\mathbf{P}(w)}{S_w(1)}+O\left(\mathbf{P}(w)^2\right)\right)\right)\\ &= \left(\frac{\mathbf{P}(w)}{S_w(1)}+O\left(\mathbf{P}(w)^2\right)\right)^{-s}\\ &= \left(\frac{\mathbf{P}(w)}{S_w(1)}\right)^{-s}\left(1+O\left(\mathbf{P}(w)\right)\right)^{-s}\\ &= \left(\frac{\mathbf{P}(w)}{S_w(1)}\right)^{-s}\left(1+O\left(\mathbf{P}(w)\right)\right),\end{aligned} \tag{63}$$

and therefore, we obtain

$$r_w^*(s) = \Gamma(s)\mathbf{P}(w)^{-s}\left(1-\frac{1}{S_w(1)^{-s}}\right)O(1). \tag{64}$$

From this expression, and noticing that the function has a removable singularity at $s=0$, we can see that the Mellin transform $r_w^*(s)$ exists on the strip where $\Re(s) > -1$. We still need to investigate the Mellin strip for the sum $\sum_{w\in\mathcal{A}^k} r_w^*(s)$. In other words, we need to examine whether summing $r_w^*(s)$ over all words of length k (where k grows with n) has any effect on the analyticity of the function. We observe that

$$\begin{aligned}\sum_{w\in\mathcal{A}^k}|r_w^*(s)| &= \sum_{w\in\mathcal{A}^k}|\Gamma(s)\mathbf{P}(w)^{-s}\left(1-\frac{1}{S_w(1)^{-s}}\right)O(1)|\\ &\le |\Gamma(s)|\sum_{w\in\mathcal{A}^k}\mathbf{P}(w)^{-\Re(s)}\left(1-\frac{1}{S_w(1)^{-\Re(s)}}\right)O(1)\\ &= (q^k)^{-\Re(s)-1}|\Gamma(s)|\sum_{w\in\mathcal{A}^k}\mathbf{P}(w)(1-S_w(1)^{\Re(s)})O(1).\end{aligned}$$

Lemma 8 allows us to split the above sum between the words for which $S_w(1) \le 1+O(\delta^k)$ and words that have $S_w(1) > 1+O(\delta^k)$.

Such a split yields the following

$$\sum_{w\in\mathcal{A}^k}|r_w^*(s)| = (q^k)^{-\Re(s)-1}|\Gamma(s)|O(\delta^k). \tag{65}$$

This shows that $\sum_{w\in\mathcal{A}^k} r_w^*(s)$ is bounded above for $\Re(s) > -1$ and, therefore, it is analytic. This argument holds for $k = \Theta(\log n)$ as well, as $(q^k)^{-\Re(s)-1}$ would still be bounded above by a constant $M_{s,k}$ that depends on s and k.

We would like to approximate $\sum_{w\in\mathcal{A}^k} r_w^*(s)$ when $z \to \infty$. By the inverse Mellin transform, we have

$$\sum_{w\in\mathcal{A}^k} r_w(z) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \Big(\sum_{w\in\mathcal{A}^k} r_w^*(s)\Big) z^{-s} ds. \tag{66}$$

We choose $c \in (-1, M)$ for a fixed $M > 0$. Then by the direct mapping theorem [22], we obtain

$$\sum_{w\in\mathcal{A}^k} r_w(z) = O(z^{-M}). \tag{67}$$

and subsequently, we get

$$\sum_{w\in\mathcal{A}^k} \left(\frac{S_w(A_w)}{B_w\, A_w^{n+1}} + (1-\mathbf{P}(w))^{n+1}\right) = O(n^{-M}). \tag{68}$$

□

We next prove the asymptotic smallness of $I_n^w(\rho)$ in (54).

Lemma 12. *Let*

$$I_n^w(\rho) = \frac{1}{2\pi i}\int_{|z|=\rho} \left(\frac{1}{1-(1-\mathbf{P}(w))z} - \frac{S_w(z)}{D_w(z)}\right)\frac{dz}{z^{n+1}}. \tag{69}$$

For large n and $k = \Theta(\log n)$, we have

$$\sum_{w\in\mathcal{A}^k} I_n^w(\rho) = O\left(\rho^{-n}(\rho\delta)^k\right). \tag{70}$$

Proof. We observe that

$$|I_n^w(\rho)| \le \frac{1}{2\pi}\int_{|z|=\rho} \left| \frac{\mathbf{P}(w)z\left(z^{k-1} - S_w(z)\right)}{D_w(z)(1-(1-\mathbf{P}(w))z)}\frac{1}{z^{n+1}}\right| dz. \tag{71}$$

For $|z| = \rho$, we show that the denominator in (71) is bounded away from zero.

$$\begin{aligned} |D_w(z)| &= |(1-z)S_w(z) + \mathbf{P}(w)z^k| \\ &\ge |1-z||S_w(z)| - \mathbf{P}(w)|z^k| \\ &\ge (\rho-1)\alpha - (p\rho)^k, \quad \text{where } \alpha > 0 \text{ by Lemma 9}. \\ &> 0, \qquad \text{we assume } k \text{ to be large enough such that } (p\rho)^k < \alpha(\rho-1). \end{aligned} \tag{72}$$

To find a lower bound for $|1-(1-\mathbf{P}(w))z|$, we can choose K_w large enough such that

$$\begin{aligned} |1-(1-\mathbf{P}(w))z| &\ge |1-(1-\mathbf{P}(w))|z|| \\ &\ge |1-\rho(1-p^{K_w})| \\ &> 0. \end{aligned} \tag{73}$$

We now move on to finding an upper bound for the numerator in (71), for $|z| = \rho$.

$$\begin{aligned}|z^{k-1} - S_w(z)| &\leq |S_w(z) - 1| + |1 - z^{k-1}| \\ &\leq (S_w(\rho) - 1) + (1 + \rho^{k-1}) \\ &= (S_w(\rho) - 1) + O(\rho^k). \end{aligned} \tag{74}$$

Therefore, there exists a constant $\mu > 0$ such that

$$\begin{aligned}|I_n^w| &\leq \mu\rho\mathbf{P}(w)\left((S_w(\rho) - 1) + O(\rho^k)\right)\frac{1}{\rho^{n+1}} \\ &= O(\rho^{-n})\left(\mathbf{P}(w)(S_w(\rho) - 1) + \mathbf{P}(w)O(\rho^k)\right). \end{aligned} \tag{75}$$

Summing over all patterns w, and applying Lemma 8, we obtain

$$\begin{aligned}\sum_{w\in\mathcal{A}^k} |I_n^w(\rho)| &= O(\rho^{-n}) \sum_{w\in\mathcal{A}^k} \mathbf{P}(w)(S_w(\rho) - 1) + O(\rho^{-n+k}) \sum_{w\in\mathcal{A}^k} \mathbf{P}(w) \\ &= O(\rho^{-n})\left(\theta(\rho\delta)^k + \frac{p\rho}{1 - p\rho}\theta\delta^k\right) + O(\rho^{-n+k}) \\ &= O(\rho^{-n}(\rho\delta)^k), \end{aligned} \tag{76}$$

which approaches zero as $n \to \infty$ and $k = \Theta(\log n)$. This completes the proof of of Theorem 1. □

Similar to Theorem 1, we provide a proof to show that the second factorial moments of the kth Subword Complexity and the kth Prefix Complexity, have the same first order asymptotic behavior. We are now ready to state the proof of Theorem 2.

Proof of Theorem 2. As discussed in Lemmas 3 and 4, the generating functions representing $\mathbf{E}[(X_{n,k})_2]$ and $\mathbf{E}[(\hat{X}_{n,k})_2]$ respectively, are

$$G(z) = \sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} \left(\frac{1}{1-z} - \frac{S_w(z)}{D_w(z)} - \frac{S_{w'}(z)}{D_{w'}(z)} + \frac{S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)}\right),$$

and

$$\begin{aligned}\hat{G}(z) = &\sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} \left(\frac{1}{1-z} - \frac{1}{1-(1-\mathbf{P}(w))z} - \frac{1}{1-(1-\mathbf{P}(w'))z}\right) \\ &+ \sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} \frac{1}{1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z}. \end{aligned}$$

Note that

$$G(z) - \hat{G}(z) = \sum_{\substack{w'\in\mathcal{A}^k \\ w\neq w'}} \sum_{w\in\mathcal{A}^k} \left(\frac{1}{1-(1-\mathbf{P}(w))z} - \frac{S_w(z)}{D_w(z)}\right) \tag{77}$$

$$+ \sum_{\substack{w\in\mathcal{A}^k \\ w\neq w'}} \sum_{w'\in\mathcal{A}^k} \left(\frac{1}{1-(1-\mathbf{P}(w'))z} - \frac{S_{w'}(z)}{D_{w'}(z)}\right) \tag{78}$$

$$+ \sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} \left(\frac{1}{1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z} - \frac{S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)}\right) \tag{79}$$

In Theorem 1, we proved that for every $M > 0$ (which does not depend on n or k), we have

$$H(z) - \hat{H}(z) = \sum_{w \in \mathcal{A}^k} \left(\frac{1}{1-(1-\mathbf{P}(w))z} - \frac{S_w(z)}{D_w(z)} \right) = O(n^{-M}).$$

Therefore, both (77) and (78) are of order $(2^k - 1)O(n^{-M}) = O(n^{-M + a\log 2})$ for $k = a \log n$. Thus, to show the asymptotic smallness, it is enough to choose $M = a \log 2 + \epsilon$, where ϵ is a small positive value. Now, it only remains to show (79) is asymptotically negligible as well. We define

$$\Delta_{w,w'}(z) = \frac{1}{1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z} - \frac{S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)}. \tag{80}$$

Next, we extract the coefficient of z^n

$$[z^n]\Delta_{w,w'}(z) = \frac{1}{2\pi i} \oint \Delta_{w,w'}(z) \frac{dz}{z^{n+1}}, \tag{81}$$

where the path of integration is a circle about the origin with counterclockwise orientation. We define

$$I_n^{w,w'}(\rho) = \frac{1}{2\pi i} \int_{|z|=\rho} \Delta_{w,w'}(z) \frac{dz}{z^{n+1}}, \tag{82}$$

The above integrand has poles at $z = 0$, $z = \alpha_{w,w'}$ (as in (46)), and $z = \frac{1}{1-\mathbf{P}(w)-\mathbf{P}(w')}$. We have chosen ρ such that the poles are all inside the circle $|z| = \rho$. It follows that

$$I_n^{w,w'}(\rho) = \mathrm{Res}_{z=0} \frac{\Delta_{w,w'}(z)}{z^{n+1}} + \mathrm{Res}_{z=\alpha_{w,w'}} \frac{\Delta_{w,w'}(z)}{z^{n+1}} + \mathrm{Res}_{z=\frac{1}{1-\mathbf{P}(w)-\mathbf{P}(w')}} \frac{\Delta_w(z)}{z^{n+1}}, \tag{83}$$

and the residues give us the following.

$$\mathrm{Res}_{z=\frac{1}{1-\mathbf{P}(w)-\mathbf{P}(w')}} \frac{1}{1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z z^{n+1}} = -(1-\mathbf{P}(w)-\mathbf{P}(w'))^{n+1},$$

and

$$\mathrm{Res}_{z=\alpha_{w,w'}} \frac{S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)} = \frac{S_w(\alpha_{w,w'})S_{w'}(\alpha_{w,w'}) - S_{w,w'}(\alpha_{w,w'})S_{w',w}(\alpha_{w,w'})}{\beta_{w,w'}\alpha_{w,w'}^{n+1}},$$

where $\beta_{w,w'}$ is as in (47). Therefore, we get

$$\begin{aligned}\sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} [z^n]\Delta_{w,w'}(z) = &\sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} I_n^{w,w'}(\rho) \\ &- \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \left(\frac{S_w(\alpha_{w,w'})S_{w'}(\alpha_{w,w'}) - S_{w,w'}(\alpha_{w,w'})S_{w',w}(\alpha_{w,w'})}{\beta_{w,w'}\alpha_{w,w'}^{n+1}} \right. \\ &\left. + (1-\mathbf{P}(w)-\mathbf{P}(w'))^{n+1} \right).\end{aligned} \tag{84}$$

We now show that the above two terms are asymptotically small. □

Lemma 13. *There exists $\epsilon > 0$ where the sum*

$$\sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \left(\frac{S_w(\alpha_{w,w'})S_{w'}(\alpha_{w,w'}) - S_{w,w'}(\alpha_{w,w'})S_{w',w}(\alpha_{w,w'})}{\beta_{w,w'}\alpha_{w,w'}^{n+1}} + (1 - \mathbf{P}(w) - \mathbf{P}(w'))^{n+1} \right)$$

is of order $O(n^{-\epsilon})$.

Proof. We define

$$r_{w,w'}(z) = \frac{S_w(\alpha_{w,w'})S_{w'}(\alpha_{w,w'}) - S_{w,w'}(\alpha_{w,w'})S_{w',w}(\alpha_{w,w'})}{\beta_{w,w'}\alpha_{w,w'}^{z}} + (1 - \mathbf{P}(w) - \mathbf{P}(w'))^{z}.$$

The Mellin transform of the above function is

$$r^*_{w,w'}(s) = \Gamma(s) \log^{-s}\left(\frac{1}{1 - \mathbf{P}(w) - \mathbf{p}(w')}\right) + C_{w,w'}\Gamma(s)\log^{-s}(\alpha_{w,w'}), \tag{85}$$

where $C_{w,w'} = \dfrac{S_w(\alpha_{w,w'})S_{w'}(\alpha_{w,w'}) - S_{w,w'}(\alpha_{w,w'})S_{w',w}(\alpha_{w,w'})}{\beta_{w,w'}}$. We note that $C_{w,w'}$ is negative and uniformly bounded from above for all $w, w' \in \mathcal{A}^k$.For a fixes s, we also have,

$$\begin{aligned}
\ln^{-s}\left(\frac{1}{1 - \mathbf{P}(w) - \mathbf{P}(w')}\right) &= \ln^{-s}\left(1 + \mathbf{P}(w) + \mathbf{P}(w') + O\left(p^{2k}\right)\right) \\
&= \left(\mathbf{P}(w) + \mathbf{P}(w') + O\left(p^{2k}\right)\right)^{-s} \\
&= (\mathbf{P}(w) + \mathbf{P}(w'))^{-s}\left(1 + O\left(p^{k}\right)\right)^{-s} \\
&= (\mathbf{P}(w) + \mathbf{P}(w'))^{-s}\left(1 + O\left(p^{k}\right)\right),
\end{aligned} \tag{86}$$

and

$$\begin{aligned}
\ln^{-s}(\alpha_{w,w'}) = \Big(&\frac{S_{w'}(1) - S_{w,w'}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w) \\
&+ \frac{S_w(1) - S_{w',w}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w') + O(p^{2k})\Big)^{-s} \\
= \Big(&\frac{S_{w'}(1) - S_{w,w'}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w) \\
&+ \frac{S_w(1) - S_{w',w}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w')\Big)^{-s}\left(1 + O(p^k)\right).
\end{aligned} \tag{87}$$

Therefore, we have

$$\begin{aligned}
r^*_{w,w'}(s) = \Gamma(s)&\left(\mathbf{P}(w) + \mathbf{P}(w')\right)^{-s}(1 + O(p^k)) \\
- \Gamma(s)&\Bigg(\frac{S_{w'}(1) - S_{w,w'}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w) \\
&+ \frac{S_w(1) - S_{w',w}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w')\Bigg)^{-s}\left(1 + O(p^k)\right)O(1).
\end{aligned} \tag{88}$$

To find the Mellin strip for the sum $\sum_{w\in\mathcal{A}^k} r^*_{w,w'}(s)$, we first note that

$$(x+y)^a \le x^a + y^a, \quad \text{for any real } x, y > 0 \text{ and } a \le 1.$$

Since $-\Re(s) < 1$, we have

$$\big(\mathbf{P}(w) + \mathbf{P}(w')\big)^{-\Re(s)} \le \mathbf{P}(w)^{-\Re(s)} + \mathbf{P}(w')^{-\Re(s)}, \tag{89}$$

and

$$\begin{aligned}
&\left(\frac{S_{w'}(1) - S_{w,w'}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w)\text{ß}n + \frac{S_w(1) - S_{w',w}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w')\right)^{-\Re(s)} \\
&\qquad \le \left(\frac{S_{w'}(1) - S_{w,w'}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w)\right)^{-\Re(s)} \\
&\qquad\qquad + \left(\frac{S_w(1) - S_{w',w}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w')\right)^{-\Re(s)}.
\end{aligned} \tag{90}$$

Therefore, we get

$$\begin{aligned}
\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} |r^*_{w,w'}(s)| \le |\Gamma(s)|O(1)& \\
\left(\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} \mathbf{P}(w)^{-\Re(s)} \left(1 - \left(\frac{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}{S_{w'}(1) - S_{w,w'}(1)}\right)^{\Re(s)}\right)\right.& \\
\left. + \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} \mathbf{P}(w')^{-\Re(s)} \left(1 - \left(\frac{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}{S_w(1) - S_{w',w}(1)}\right)^{\Re(s)}\right)\right)& \\
\le (q^k)^{-\Re(s)-1}|\Gamma(s)|O(1)&
\end{aligned}$$

$$\left(\sum_{\substack{w'\in\mathcal{A}^k\\ w\neq w'}} \sum_{w\in\mathcal{A}^k} \mathbf{P}(w)\left(1 - (S_w(1))^{\Re(s)}\left(1 - \frac{S_{w,w'}(1)}{S_{w'}(1)}\right)^{-\Re(s)}\right)\right. \tag{91}$$

$$+ \sum_{\substack{w'\in\mathcal{A}^k\\ w\neq w'}} \sum_{w\in\mathcal{A}^k} \mathbf{P}(w)S_{w,w'}(1)^{\Re(s)}\left(\frac{S_{w'}(1) - S_{w,w'}(1)}{S_{w',w}(1)}\right)^{-\Re(s)} \tag{92}$$

$$+ \sum_{\substack{w\in\mathcal{A}^k\\ w\neq w'}} \sum_{w'\in\mathcal{A}^k} \mathbf{P}(w')\left(1 - (S_{w'}(1))^{\Re(s)}\left(1 - \frac{S_{w',w}(1)}{S_w(1)}\right)^{-\Re(s)}\right) \tag{93}$$

$$\left. + \sum_{\substack{w\in\mathcal{A}^k\\ w\neq w'}} \sum_{w'\in\mathcal{A}^k} \mathbf{P}(w')S_{w',w}(1)^{\Re(s)}\left(\frac{S_w(1) - S_{w',w}(1)}{S_{w,w'}(1)}\right)^{-\Re(s)}\right). \tag{94}$$

By Lemma 10, with high probability, a randomly selected w has the property $S_{w,w'}(1) = O(\delta^k)$, and thus

$$\left(1 - \frac{S_{w,w'}(1)}{S_{w'}(1)}\right)^{-\Re(s)} = 1 + O(\delta^k).$$

With that and by Lemma 8, for most words w,

$$1 - S_w(1)^{\Re(s)}(1 + O(\delta^k)) = O(\delta^k).$$

Therefore, both sums (91) and (93) are of the form $(2^k - 1)O(\delta^k)$. The sums (92) and (94) are also of order $(2^k - 1)O(\delta^k)$ by Lemma 10. Combining all these terms we will obtain

$$\sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} |r^*_{w,w'}(s)| \leq (2^k - 1)(q^k)^{-\Re(s)-1}|\Gamma(s)|O(\delta^k)O(1). \tag{95}$$

By the inverse Mellin transform, for $k = a \log n$, $M = a \log 2 + \epsilon$ and $c \in (-1, M)$, we have

$$\begin{aligned}\sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} r_{w,w'}(z) &= \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\Big(\sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} r^*_{w,w'}(s)\Big)z^{-s}ds = O(z^{-M})O(2^k) \\ &= O(z^{-\epsilon}).\end{aligned} \tag{96}$$

□

In the following lemma we show that the first term in (85) is asymptotically small.

Lemma 14. *Recall that*

$$I_n^{w,w'}(\rho) = \frac{1}{2\pi i}\int_{|z|=\rho} \Delta_{w,w'}(z)\frac{dz}{z^{n+1}}.$$

We have

$$\sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} I_n^{w,w'}(\rho) = O\left(\rho^{-n+2k}\delta^k\right). \tag{97}$$

Proof. First note that

$$\begin{aligned}\Delta_{w,w'}(z) &= \frac{1}{1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z} - \frac{S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)} \\ &= \frac{z\mathbf{P}(w)\left(S_{w,w'}(z)S_{w',w}(z) - S_w(z)S_{w'}(z) + z^{k-1}S_{w'}(z) - z^{k-1}S_{w,w'}(z)\right)}{(1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z)\,D_{w,w'}(z)} \\ &\quad + \frac{z\mathbf{P}(w')\left(S_{w',w}(z)S_{w,w'}(z) - S_{w'}(z)S_w(z) + z^{k-1}S_w(z) - z^{k-1}S_{w',w}(z)\right)}{(1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z)\,D_{w,w'}(z)}.\end{aligned} \tag{98}$$

We saw in (73) that $|1-(1-\mathbf{P}(w'))z| \geq c_2$, and therefore, it follows that

$$|1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z| \geq c_1 \tag{99}$$

For $z = \rho$, $|D_{w,w'}(z)|$ is also bounded below as the following

$$\begin{aligned}|D_{w,w'}(z)| = \big|&(1-z)(S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)) \\ &+ z^k\left(\mathbf{P}(w)(S_{w'}(z) - S_{w,w'}(z)) + \mathbf{P}(w')(S_w(z) - S_{w',w}(z))\right)\big| \\ \geq\ & |(1-z)(S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z))| \\ & - \left|z^k\right| \left|(\mathbf{P}(w)(S_{w'}(z) - S_{w,w'}(z)) + \mathbf{P}(w')(S_w(z) - S_{w',w}(z)))\right| \\ \geq\ & (\rho-1)\beta - (p\rho)^k\left(\frac{2(1+p\rho)}{1-p\rho}\right), \end{aligned} \tag{100}$$

which is bounded away from zero by the assumption of Lemma 7. Additionally, we show that the numerator in (98) is bounded above, as follows

$$\begin{aligned}&|S_{w,w'}(z)S_{w',w}(z) - S_w(z)S_{w'}(z) + z^{k-1}S_{w'}(z) - z^{k-1}S_{w,w'}(z)| \leq \\ &\quad |S_{w'}(z)(z^{k-1} - S_w(z))| + |S_{w,w'}(z)(S_{w',w}(z) - z^{k-1})| \\ &\quad \leq S_{w'}(\rho)\left((S_w(\rho)-1) + O(\rho^k)\right) + S_{w,w'}(\rho)\left(S_{w',w}(\rho) + O(\rho^k)\right). \end{aligned} \tag{101}$$

This yields

$$\begin{aligned}\sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} |I_n^{w,w'}| \leq O(\rho^{-n}) &\sum_{\substack{w'\in\mathcal{A}^k \\ w\neq w'}} S_{w'}(\rho) \sum_{w\in\mathcal{A}^k} \mathbf{P}(w)\left((S_w(\rho)-1) + O(\rho^k)\right) \\ &+ O(\rho^{-n}) \sum_{\substack{w'\in\mathcal{A}^k \\ w\neq w'}} \sum_{w\in\mathcal{A}^k} \mathbf{P}(w)S_{w,w'}(\rho)\left(S_{w',w}(\rho) + O(\rho^k)\right). \end{aligned} \tag{102}$$

By (75), the first term above is of order $(2^k-1)O(\rho^{-n+k})$ and by Lemma 10 and an analysis similar to (75), the second term yields $(2^k-1)O(\rho^{-n+k})$ as well. Finally, we have

$$\sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} |I_n^{w,w'}| \leq O(\rho^{-n+2k}\delta^k).$$

Which goes to zero asymptotically, for $k = \Theta(\log n)$. □

This lemma completes our proof of Theorem 2.

3.5. Asymptotic Analysis of the kth Prefix Complexity

We finally proceed to analyzing the asymptotic moments of the kth Prefix Complexity. The results obtained hold true for the moments of the kth Subword Complexity. Our methodology involves poissonization, saddle point analysis (the complex version of Laplace's method [23]), and depoissonization.

Lemma 15 (Jacquet and Szpankowski, 1998)**.** *Let $\tilde{G}(z)$ be the Poisson transform of a sequence g_n. If $\tilde{G}(z)$ is analytic in a linear cone S_θ with $\theta < \pi/2$, and if the following two conditions hold:*
(I) For $z \in S_\theta$ and real values $B, r > 0, \nu$

$$|z| > r \to |\tilde{G}(z)| \leq B|z^\nu|\Psi(|z|), \tag{103}$$

where $\Psi(x)$ *is such that, for fixed* t, $\lim_{x\to\infty} \dfrac{\Psi(tx)}{\Psi(x)} = 1$;
(II) For $z \notin S_\theta$ *and* $A, \alpha < 1$

$$|z| > r \to |\tilde{G}(z)e^z| \le Ae^{\alpha|z|}. \tag{104}$$

Then, for every non-negative integer n, *we have*

$$g_n = \tilde{G}(n) + O(n^{\nu-1}\Psi(n)).$$

On the Expected Value: To transform the sequence of interest, $(\mathbf{E}[\hat{X}_{n,k}])_{n\ge 0}$, into a Poisson model, we recall that in (25) we found

$$\mathbf{E}[\hat{X}_{n,k}] = \sum_{w\in\mathcal{A}^k} \left(1 - (1 - \mathbf{P}(w))^n\right).$$

Thus, the Poisson transform is

$$\begin{aligned}
\tilde{E}_k(z) &= \sum_{n=0}^{\infty} \mathbf{E}[\hat{X}_{n,k}]\frac{z^n}{n!}e^{-z} \\
&= \sum_{n=0}^{\infty}\sum_{w\in\mathcal{A}^k} \left(1 - (1 - \mathbf{P}(w))^n\right)\frac{z^n}{n!}e^{-z} \\
&= \sum_{w\in\mathcal{A}^k} \left(1 - e^{-z\mathbf{P}(w)}\right).
\end{aligned} \tag{105}$$

To asymptotically evaluate this harmonic sum, we turn our attention to the Mellin Transform once more. The Mellin transform of $\tilde{E}_k(z)$ is

$$\begin{aligned}
\tilde{E}_k^*(s) &= -\Gamma(s)\sum_{w\in\mathcal{A}^k} P(w)^{-s} \\
&= -\Gamma(s)(p^{-s} + q^{-s})^k,
\end{aligned} \tag{106}$$

which has the fundamental strip $s \in \langle -1, 0\rangle$. For $c \in (-1, 0)$, the inverse Mellin integral is the following

$$\begin{aligned}
\tilde{E}_k(z) &= \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \tilde{E}_k^*(s)\cdot z^{-s}ds \\
&= \frac{-1}{2\pi i}\int_{c-i\infty}^{c+i\infty} z^{-s}\Gamma(s)(p^{-s} + q^{-s})^k ds \\
&= \frac{-1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \Gamma(s)e^{-k(s\frac{\log z}{k} - \log(p^{-s}+q^{-s}))}ds \\
&= \frac{-1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \Gamma(s)e^{-kh(s)}ds,
\end{aligned} \tag{107}$$

where we define $h(s) = \dfrac{s}{a} - \log(p^{-s} + q^{-s})$ for $k = a\log z$. We emphasize that the above integral involves k, and k grows with n. We evaluate the integral through the saddle point analysis. Therefore, we choose the line of integration to cross the saddle point r_0. To find the saddle point r_0, we let $h'(r_0) = 0$, and we obtain

$$(p/q)^{-r_0} = \frac{a\log p^{-1} - 1}{1 - a\log q^{-1}}, \tag{108}$$

and therefore,

$$r_0 = \frac{-1}{\log p/q} \log \left(\frac{a \log q^{-1} - 1}{1 - a \log p^{-1}} \right), \tag{109}$$

where $\frac{1}{\log q^{-1}} < a < \frac{1}{\log p^{-1}}$.

By (108) and the fact that $(p/q)^{it_j} = 1$ for $t_j = \frac{2\pi j}{\log p/q}$ and $j \in \mathbb{Z}$, we can see that there are actually infinitely many saddle points z_j of the form $r_0 + it_j$ on the line of integration.

We remark that the location of r_0 depends on the value of a. We have $r_0 \to \infty$ as $a \to \frac{1}{\log q^{-1}}$, and $r_0 \to -\infty$ as $a \to \frac{1}{\log p^{-1}}$. We divide the analysis into three parts, for the three ranges $r_0 \in (0, \infty)$, $r_0 \in (-1, 0)$, and $r_0 \in (-\infty, -1)$.

In the first range, which corresponds to

$$\frac{1}{\log q^{-1}} < a < \frac{2}{\log q^{-1} + \log p^{-1}}, \tag{110}$$

we perform a residue analysis, taking into account the dominant pole at $s = -1$. In the second range, we have

$$\frac{2}{\log q^{-1} + \log p^{-1}} < a < \frac{1}{q \log q^{-1} + p \log p^{-1}}, \tag{111}$$

and we get the asymptotic result through the saddle point method. The last range corresponds to

$$\frac{1}{q \log q^{-1} + p \log p^{-1}} < a < \frac{1}{\log p^{-1}}, \tag{112}$$

and we approach it with a combination of residue analysis at $s = 0$, and the saddle point method. We now proceed by stating the proof of theorem 3.

Proof of Theorem 3. We begin with proving part *ii* which requires a saddle point analysis. We rewrite the inverse Mellin transform with integration line at $\Re(s) = r_0$ as

$$\begin{aligned} \tilde{E}_k(z) &= \frac{-1}{2\pi} \int_{-\infty}^{\infty} z^{-(r_0+it)} \Gamma(r_0 + it)(p^{-(r_0+it)} + q^{-(r_0+it)})^k dt \\ &= \frac{-1}{2\pi} \int_{-\infty}^{\infty} \Gamma(r_0 + it) e^{-k((r_0+it)\frac{\log z}{k} - \log(p^{-(r_0+it)} + q^{-(r_0+it)}))} dt. \end{aligned} \tag{113}$$

Step one: Saddle points' contribute to the integral estimation

First, we are able to show those saddle points with $|t_j| > \sqrt{\log n}$ do not have a significant asymptotic contribution to the integral. To show this, we let

$$T_k(z) = \int_{|t| > \sqrt{\log n}} z^{-r_0 - it} \Gamma(r_0 + it)(p^{-r_0 - it} + q^{-r_0 - it})^k dt. \tag{114}$$

Since $|\Gamma(r_0 + it)| = O(|t|^{r_0 - \frac{1}{2}} e^{\frac{-\pi|t|}{2}})$ as $|t| \to \pm\infty$, we observe that

$$\begin{aligned}T_k(z) &= O\left(z^{-r_0}(p^{-r_0}+q^{-r_0})^k\int_{\sqrt{\log n}}^{\infty} t^{r_0/2-1/2}e^{-\pi t/2}dt\right)\\ &= O\left(z^{-r_0}(p^{-r_0}+q^{-r_0})^k(\log n)^{r_0/4-1/4}\int_{\sqrt{\log n}}^{\infty} e^{-\pi t/2}dt\right)\\ &= O\left(z^{-r_0}(p^{-r_0}+q^{-r_0})^k(\log n)^{r_0/4-1/4}e^{-\pi\sqrt{\log n}/2}\right)\\ &= O\left((\log n)^{r_0/4-1/4}e^{-\pi\sqrt{\log n}/2}\right),\end{aligned} \tag{115}$$

which is very small for large n. Note that for $t \in (\sqrt{\log n}, \infty)$, $t^{r_0/2-1/2}$ is decreasing, and bounded above by $(\log n)^{r_0/4-1/4}$.

Step two: Partitioning the integral

There are now only finitely many saddle points to work with. We split the integral range into sub-intervals, each of which contains exactly one saddle point. This way, each integral has a contour traversing a single saddle point, and we will be able to estimate the dominant contribution in each integral from a small neighborhood around the saddle point. Assuming that j^* is the largest j for which $\dfrac{2\pi j}{\log p/q} \leq \sqrt{\log n}$, we split the integral $\tilde{E}_k(z)$ as following

$$\begin{aligned}\tilde{E}_k(z) = &-\frac{1}{2\pi}\left(\sum_{|j|<j^*}\int_{|t-t_j|\leq\frac{\pi}{\log p/q}} z^{-r_0+it}\Gamma(r_0+it)(p^{-r_0-it}+q^{-r_0-it})^k dt\right)\\ &-\frac{1}{2\pi}\int_{\frac{\pi}{\log p/q}\leq|t_j^*|<\sqrt{\log n}}\Gamma(r+it)z^{-r_0+it}(p^{-r_0-it}+q^{-r_0-it})^k dt.\end{aligned} \tag{116}$$

By the same argument as in (115), the second term in (116) is also asymptotically negligible. Therefore, we are only left with

$$\tilde{E}_k(z) = \sum_{|j|<j^*} S_j(z), \tag{117}$$

where $S_j(z) = -\dfrac{1}{2\pi}\int_{|t-t_j|\leq\frac{\pi}{\log p/q}} z^{-r_0+it}\Gamma(r_0+it)(p^{-r_0-it}+q^{-r_0-it})^k dt)$.

Step three: Splitting the saddle contour

For each integral S_j, we write the expansion of $h(t)$ about t_j, as follows

$$h(t) = h(t_j) + \frac{1}{2}h''(t_j)(t-t_j)^2 + O((t-t_j)^3). \tag{118}$$

The main contribution for the integral estimate should come from an small integration path that reduces $kh(t)$ to its quadratic expansion about t_j. In other words, we want the integration path to be such that

$$k(t-t_j)^2 \to \infty, \qquad \text{and} \qquad k(t-t_j)^3 \to 0. \tag{119}$$

The above conditions are true when $|t - t_j| \gg k^{-1/2}$ and $|t - tj| \ll k^{-1/3}$. Thus, we choose the integration path to be $|t - t_j| \le k^{-2/5}$. Therefore, we have

$$S_j(z) = -\frac{1}{2\pi}\int_{|t-t_j|\le k^{-2/5}} z^{-r_0+it}\Gamma(r_0+it)(p^{-r_0-it}+q^{-r_0-it})^k dt$$
$$-\frac{1}{2\pi}\int_{k^{-2/5}<|t-t_j|<\frac{\pi}{\log p/q}} z^{-r_0+it}\Gamma(r_0+it)(p^{-r_0-it}+q^{-r_0-it})^k dt. \tag{120}$$

Saddle Tails Pruning.

We show that the integral is small for $k^{-2/5} < |t - t_j| < \dfrac{\pi}{\log p/q}$. We define

$$S_j^{(1)}(z) = -\frac{1}{2\pi}\int_{k^{-2/5}<|t-t_j|<\frac{\pi}{\log p/q}} z^{-r_0+it}\Gamma(r_0+it)(p^{-r_0-it}+q^{-r_0-it})^k dt. \tag{121}$$

Note that for $|t - t_j| \le \dfrac{\pi}{\log p/q}$, we have

$$\begin{aligned}
|p^{-r_0-it}+q^{-r_0-it}| &= (p^{-r_0}+q^{-r_0})\sqrt{1-\frac{2p^{-r_0}q^{-r_0}}{(p^{-r_0}+q^{-r_0})^2}(1-\cos(t\log p/q))} \\
&\le (p^{-r_0}+q^{-r_0})\left(1-\frac{p^{-r_0}q^{-r_0}}{(p^{-r_0}+q^{-r_0})^2}(1-\cos\left(t-t_j\right)\log p/q)\right) \\
&\qquad \text{since}\sqrt{1-x}\le 1-\frac{x}{2} \text{ for } x\in[0,1] \\
&\le (p^{-r_0}+q^{-r_0})\left(1-\frac{2p^{-r_0}q^{-r_0}}{\pi^2(p^{-r_0}+q^{-r_0})^2}((t-t_j)\log p/q)^2\right) \\
&\qquad \text{since } 1-\cos x \ge \frac{2x^2}{\pi^2} \text{ for } |x|\le\pi \\
&\le (p^{-r_0}+q^{-r_0})e^{-\gamma(t-t_j)^2},
\end{aligned} \tag{122}$$

where $\gamma = \dfrac{2p^{-r_0}q^{-r_0}\log^2 p/q}{\pi^2(p^{-r_0}+q^{-r_0})^2}$. Thus,

$$\begin{aligned}
S_j^{(1)}(z) &= O\left(z^{-r_0}|\Gamma(r_0+it)|\int_{k^{-2/5}<|t-t_j|<\frac{\pi}{\log p/q}}|p^{-r_0-it}+q^{-r_0-it}|dt\right) \\
&= O\left(z^{-r_0}(p^{-r_0}+q^{-r_0})^k\int_{k^{-2/5}}^{\infty}e^{-\gamma k u^2}du\right) \\
&= O\left(z^{-r_0}(p^{-r_0}+q^{-r_0})^k k^{-3/5}e^{-\gamma k^{1/5}}\right), \text{since } \operatorname{erf}(x) = O\left(e^{-x^2}/x\right).
\end{aligned} \tag{123}$$

Central Approximation.

Over the main path, the integrals are of the form

$$\begin{aligned}
S_j^{(0)}(z) &= -\frac{1}{2\pi}\int_{|t-t_j|\le k^{-2/5}}\Gamma(r_0+it)z^{-r_0+it}(p^{-r_0-it}+q^{-r_0-it})^k dt \\
&= -\frac{1}{2\pi}\int_{|t-t_j|\le k^{-2/5}}\Gamma(r_0+it)e^{-kh(t)}dt.
\end{aligned}$$

We have

$$h^{''}(t_j) = \frac{\log^2 p/q}{((p/q)^{-r_0/2} + (p/q)^{r_0/2})^2}, \tag{124}$$

and

$$p^{-r_0 - it_j} + q^{-r_0 - it_j} = p^{-it_j}(p^{-r_0} + q^{-r_0}). \tag{125}$$

Therefore, by Laplace's theorem (refer to [22]) we obtain

$$\begin{aligned} S_j^{(0)}(z) &= \frac{1}{\sqrt{2\pi k h''(t_j)}} \Gamma(r_0 + it_j) e^{-kh(t_j)}(1 + O(k^{-1/2})) \\ &= \frac{(p/q)^{-r_0/2} + (p/q)^{r_0/2}}{\sqrt{2\pi} \log p/q} \\ &\qquad \times z^{-r_0}(p^{-r_0} + q^{-r_0})^k \Gamma(r_0 + it_j) z^{-it_j} p^{-ikt_j} k^{-1/2} \left(1 + O\left(\frac{1}{\sqrt{k}}\right)\right). \end{aligned} \tag{126}$$

We finally sum over all j ($|j| < j^*$), and we get

$$\begin{aligned} \tilde{E}_k(z) = & \frac{(p/q)^{-r_0/2} + (p/q)^{r_0/2}}{\sqrt{2\pi} \log p/q} \\ & \times \sum_{|j|<j^*} z^{-r_0}(p^{-r_0} + q^{-r_0})^k \Gamma(r_0 + it_j) z^{-it_j} p^{-ikt_j} k^{-1/2} \left(1 + O\left(\frac{1}{\sqrt{k}}\right)\right). \end{aligned} \tag{127}$$

We can rewrite $\tilde{E}_k(z)$ as

$$\tilde{E}_k(z) = \Phi_1((1 + a \log p) \log_{p/q} n) \frac{z^\nu}{\sqrt{\log n}} \left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right), \tag{128}$$

where $\nu = -r_0 + a \log(p^{-r_0} + q^{-r_0})$, and

$$\Phi_1(x) = \frac{(p/q)^{-r_0/2} + (p/q)^{r_0/2}}{\sqrt{2a\pi} \log p/q} \sum_{|j|<j^*} \Gamma(r_0 + it_j) e^{-2\pi ijx}. \tag{129}$$

For part *ii*, we move the line of integration to $r_0 \in (0, \infty)$. Note that in this range, we must consider the contribution of the pole at $s = 0$. We have

$$\tilde{E}_k(z) = \mathrm{Res}_{s=0} \tilde{E}_k^*(s) z^{-s} + \int_{r_0 - i\infty}^{r_0 + i\infty} \tilde{E}_k^*(z) z^{-s} ds. \tag{130}$$

Computing the residue at $s = 0$, and following the same analysis as in part *i* for the above integral, we arrive at

$$\tilde{E}_k(z) = 2^k - \Phi_1((1 + a \log p) \log_{p/q} n) \frac{z^\nu}{\sqrt{\log n}} \left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right). \tag{131}$$

For part *iii*. of Theorem 3, we shift the line of integration to $c_0 \in (-2,-1)$, then we have

$$\begin{aligned}\tilde{E}_k(z) &= \mathrm{Res}_{s=-1}\tilde{E}_k^*(s)z^{-s} + \int_{c-i\infty}^{c+i\infty} \tilde{E}_k^*(z)z^{-s}ds \\ &= z + O\left(z^{-c_0}(p^{-c_0}+q^{-c_0})^k\right) \\ &= z^{a\log 2} + O(z^{\nu_0}), \end{aligned} \tag{132}$$

where $\nu_0 = -c_0 + a\log(p^{-c_0}+q^{-c_0}) < 1$.

Step four: Asymptotic depoissonization

To show that both conditions in (15) hold for $\tilde{E}_k(z)$, we extend the real values z to complex values $z = ne^{i\theta}$, where $|\theta| < \pi/2$. To prove (103), we note that

$$|e^{-i\theta(r_0+it)}\Gamma(r_0+it)| = O(|t|^{r_0-1/2}e^{t\theta-\pi|t|/2}), \tag{133}$$

and therefore

$$\tilde{E}_k(ne^{i\theta}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\theta(r_0+it)}n^{-r_0-it}\Gamma(r_0+it)(p^{-r_0-it}+q^{-r_0-it})^k dt \tag{134}$$

is absolutely convergent for $|\theta| < \pi/2$. The same saddle point analysis applies here and we obtain

$$|\tilde{E}_k(z)| \le B\frac{|z^{\nu}|}{\sqrt{\log n}}, \tag{135}$$

where $B = |\Phi_1((1+a\log p)\log_{p/q} n)|$, and ν is as in (128). Condition (103) is therefore satisfied. To prove condition (104) We see that for a fixed k,

$$\begin{aligned}|\tilde{E}_k(z)e^z| &\le \sum_{w\in\mathcal{A}^k} |e^z - e^{z(1-\mathbf{P}(w))}| \\ &\le 2^{k+1}e^{|z|\cos(\theta)}. \end{aligned} \tag{136}$$

Therefore, we have

$$\mathbf{E}[\hat{X}_{n,k}] = \tilde{E}(n) + O\left(\frac{n^{\nu-1}}{\sqrt{\log n}}\right). \tag{137}$$

This completes the proof of Theorem 3. □

On the Second Factorial Moment: We poissonize the sequence $(\mathbf{E}[(\hat{X}_{n,k})_2])_{n\ge 0}$ as well. By the analysis in (27),

$$\mathbf{E}[(\hat{X}_{n,k})_2] = \sum_{\substack{w,w'\in\mathcal{A}^k \\ w\ne w'}} \left(1-(1-\mathbf{P}(w))^n - (1-\mathbf{P}(w'))^n + (1-\mathbf{P}(w)-\mathbf{P}(w'))^n\right),$$

which gives the following poissonized form

$$\begin{aligned}
\tilde{G}(z) &= \sum_{n\geq 0} \mathbf{E}[(\hat{X}_{n,k})_2]\frac{z^n}{n!}e^{-z} \\
&= \sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} 1 - e^{-\mathbf{P}(w)z} - e^{-\mathbf{P}(w')z} + e^{-(\mathbf{P}(w)+\mathbf{P}(w'))z} \\
&= \sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} \left(1 - e^{-\mathbf{P}(w')z}\right)\left(1 - e^{-\mathbf{P}(w)z}\right) \\
&= \left(\sum_{w\in\mathcal{A}^k}\left(1 - e^{-\mathbf{P}(w)z}\right)\right)^2 - \sum_{w\in\mathcal{A}^k}\left(1 - e^{-\mathbf{P}(w)z}\right)^2 \\
&= (\tilde{E}_k(z))^2 - \sum_{w\in\mathcal{A}^k}\left(1 - e^{-\mathbf{P}(w)z}\right)^2 \\
&= (\tilde{E}_k(z))^2 - \sum_{w\in\mathcal{A}^k}\left(1 - 2e^{-\mathbf{P}(w)z} + e^{-2\mathbf{P}(w)z}\right). \qquad (138)
\end{aligned}$$

We show that in all ranges of a the leftover sum in (138) has a lower order contribution to $\tilde{G}_k(z)$ compared to $(\tilde{E}_k(z))^2$. We define

$$\tilde{L}_k(z) = \sum_{w\in\mathcal{A}^k}\left(1 - 2e^{-\mathbf{P}(w)z} + e^{-2\mathbf{P}(w)z}\right). \qquad (139)$$

In the first range for k, we take the Mellin transform of $\tilde{L}_k(z)$, which is

$$\begin{aligned}
\tilde{L}_k^*(s) &= -2\Gamma(s)\sum_{w\in\mathcal{A}^k}\mathbf{P}(w)^{-s} + \Gamma(s)\sum_{w\in\mathcal{A}^k}(2\mathbf{P}(w))^{-s} \\
&= -2\Gamma(s)(p^{-s}+q^{-s})^k + \Gamma(s)2^{-s}(p^{-s}+q^{-s})^k \\
&= \Gamma(s)(p^{-s}+q^{-s})^k(2^{-s-1}-1), \qquad (140)
\end{aligned}$$

and we note that the fundamental strip for this Mellin transform of is $\langle -2, 0\rangle$ as well. The inverse Mellin transform for $c \in (-2, 0)$ is

$$\begin{aligned}
\tilde{L}_k(z) &= \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\tilde{L}_k^*(s)z^{-s}ds \\
&= \frac{1}{\pi i}\int_{c-i\infty}^{c+i\infty}\Gamma(s)(p^{-s}+q^{-s})^k(2^{-s-1}-1)z^{-s}ds \qquad (141)
\end{aligned}$$

We note that this range of r_0 corresponds to

$$\frac{2}{\log q^{-1} + \log p^{-1}} < a < \frac{p^2+q^2}{q^2\log q^{-1} + p^2\log p^{-1}}. \qquad (142)$$

The integrand in (141) is quite similar to the one seen in (107). The only difference is the extra term $2^{-s-1} - 1$. However, we notice that $2^{-s-1} - 1$ is analytic and bounded. Thus, we obtain the same saddle points with the real part as in (109) and the same imaginary parts in the form of $\dfrac{2\pi i j}{\log p/q}$, $j \in \mathbb{Z}$. Thus, the same saddle point analysis for the integral in (107) applies to $\tilde{L}_k(z)$ as well. We avoid repeating the similar steps, and we skip to the central approximation, where by Laplace's theorem (ref. [22]), we get

$$\begin{aligned}\tilde{L}_k(z) = &\frac{(p/q)^{-r_0/2} + (p/q)^{r_0/2}}{\sqrt{2\pi}\log p/q} \\ &\times \sum_{|j|<j^*} z^{-r_0}(p^{-r_0} + q^{-r_0})^k (2^{-r_0-1-it_j} - 1) \\ &\qquad \times \Gamma(r_0 + it_j) z^{-it_j} p^{-ikt_j} k^{-1/2} \left(1 + O\left(\frac{1}{\sqrt{k}}\right)\right), \end{aligned} \tag{143}$$

which can be represented as

$$\tilde{L}_k(z) = \Phi_2((1 + a\log p)\log_{p/q} n)\frac{z^\nu}{\sqrt{\log n}}\left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right), \tag{144}$$

where

$$\Phi_2(x) = \frac{(p/q)^{-r_0/2} + (p/q)^{r_0/2}}{\sqrt{2a\pi}\log p/q} \sum_{|j|<j^*} (2^{-r_0-1-it_j} - 1)\Gamma(r_0 + it_j)e^{-2\pi ijx}. \tag{145}$$

This shows that $\tilde{L}_k(z) = O\left(\frac{z^\nu}{\sqrt{\log n}}\right)$, when

$$\frac{2}{\log q^{-1} + \log p^{-1}} < a < \frac{p^2 + q^2}{q^2 \log q^{-1} + p^2 \log p^{-1}}.$$

Subsequently, for $\frac{1}{\log q^{-1}} < a < \frac{2}{\log q^{-1} + \log p^{-1}}$, we get

$$\tilde{L}_k(z) = 2^k - \Phi_2((1 + a\log p)\log_{p/q} n)\frac{z^\nu}{\sqrt{\log n}}\left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right), \tag{146}$$

and for $\frac{p^2 + q^2}{q^2 \log q^{-1} + p^2 \log p^{-1}} < a < \frac{1}{\log p^{-1}}$, we get

$$\tilde{L}_k(z) = O(n^2). \tag{147}$$

It is not difficult to see that for each range of a as stated above, $\tilde{L}_k(z)$ has a lower order contribution to the asymptotic expansion of $\tilde{G}_k(z)$, compared to $(\tilde{E}_k(z))^2$. Therefore, this leads us to Theorem 4, which will be proved bellow.

Proof of Theorem 4. It is only left to show that the two depoissonization conditions hold: For condition (103) in Theorem 15, from (135) we have

$$|\tilde{G}_k(z)| \le B^2 \frac{|z^{2\nu}|}{\log n}, \tag{148}$$

and for condition (104), we have, for fixed k,

$$\begin{aligned}|\tilde{G}_k(z)e^z| &\le \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \ne w'}} \left| e^z - e^{(1-\mathbf{P}(w))z} - e^{(1-\mathbf{P}(w'))z} + e^{(1-(\mathbf{P}(w)+\mathbf{P}(w')))z} \right| \\ &\le 4^k e^{|z|\cos\theta}. \end{aligned} \tag{149}$$

Therefore both depoissonization conditions are satisfied and the desired result follows. □

Corollary. A Remark on the Second Moment and the Variance

For the second moment we have

$$\begin{aligned}\mathbf{E}\left[(\hat{X}_{n,k})^2\right] &= \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} \mathbf{E}\left[\hat{X}^{(w)}_{n,k}\hat{X}^{(w')}_{n,k}\right] + \sum_{w\in\mathcal{A}^k}\mathbf{E}[\hat{X}^{(w)}_{n,k}]\\ &= \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} \left(1-(1-\mathbf{P}(w))^n-(1-\mathbf{P}(w'))^n+(1-\mathbf{P}(w)-\mathbf{P}(w'))^n\right)\\ &\qquad + \sum_{w\in\mathcal{A}^k}\left(1-(1-\mathbf{P}(w))^n\right). \end{aligned} \tag{150}$$

Therefore, by (105) and (138) the Poisson transform of the second moment, which we denote by $\tilde{G}^{(2)}_k(z)$ is

$$\tilde{G}^{(2)}_k(z) = (\tilde{E}_k(z))^2 + \tilde{E}_k(z) - \sum_{w\in\mathcal{A}^k}\left(1-2e^{-\mathbf{P}(w)z}+e^{-2\mathbf{P}(w)z}\right), \tag{151}$$

which results in the same first order asymptotic as the second factorial moment. Also, it is not difficult to extend the proof in Chapter 6 to show that the second moments of the two models are asymptotically the same. For the variance we have

$$\begin{aligned}\mathrm{Var}[\hat{X}_{n,k}] &= \mathbf{E}\left[(\hat{X}_{n,k})^2\right] - \left(\mathbf{E}\left[\hat{X}_{n,k}\right]\right)^2\\ &= \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} \left(1-(1-\mathbf{P}(w))^n-(1-\mathbf{P}(w'))^n+(1-\mathbf{P}(w)-\mathbf{P}(w'))^n\right)\\ &\qquad + \sum_{w\in\mathcal{A}^k}\left(1-(1-\mathbf{P}(w))^n\right)\\ &\quad - \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} \left(1-(1-\mathbf{P}(w))^n-(1-\mathbf{P}(w'))^n+(1-\mathbf{P}(w)-\mathbf{P}(w'))^n\right)\\ &\qquad - \sum_{w\in\mathcal{A}^k}\left(1-(1-\mathbf{P}(w))^n-(1-\mathbf{P}(w))^n+(1-\mathbf{P}(w))^{2n}\right)\\ &= \sum_{w\in\mathcal{A}^k}\left((1-\mathbf{P}(w))^n-(1-\mathbf{P}(w))^{2n}\right). \end{aligned} \tag{152}$$

Therefore the Poisson transform, which we denote by $\tilde{G}^{\mathrm{var}}_k(z)$ is

$$\tilde{G}^{\mathrm{var}}_k(z) = \sum_{w\in\mathcal{A}^k}\left(e^{-\mathbf{P}(w)z} - e^{-(2\mathbf{P}(w)+(\mathbf{P}(w))^2)z}\right). \tag{153}$$

The Mellin transform of the above function has the following form

$$\tilde{G}*^{\mathrm{var}}_k(z) = \Gamma(s)(p^{-s}+q^{-s})^k(-1+O(\mathbf{P}(w))). \tag{154}$$

This is quite similar to what we saw in (106), which indicates that the variance has the same asymptotic growth as the expected value. But the variance of the two models do not behave in the same way (cf. Figure 2).

4. Summary and Conclusions

We studied the first-order asymptotic growth of the first two (factorial) moments of the kth Subword Complexity. We recall that the kth Subword Complexity of a string of length n is denoted by $X_{n,k}$, and is defined as the number of distinct subwords of length k, that appear in the string. We are interested in the asymptotic analysis for when k grows as a function of the string's length. More specifically, we conduct the analysis for $k = \Theta(\log n)$, and as $n \to \infty$.

The analysis is inspired by the earlier work of Jacquet and Szpankowski on the analysis of suffix trees, where they are compared to independent tries (cf. [14]). In our work, we compare the first two moments of the kth Subword Complexity to the kth Prefix Complexity over a random trie built over n independently generated binary strings. We recall that we define the kth Prefix Complexity as the number of distinct prefixes that appear in the trie at level k and lower.

We obtain the generating functions representing the expected value and the second factorial moments as their coefficients, in both settings. We prove that the first two moments have the same asymptotic growth in both models. For deriving the asymptotic behavior, we split the range for k into three intervals. We analyze each range using the saddle point method, in combination with residue analysis. We close our work with some remarks regarding the comparison of the second moment and the variance to the kth Prefix Complexity.

5. Future Challenges

The intervals' endpoints for a in Theorems 3 and 4 are not investigated in this work. The asymptotic analysis of the end points can be studied using van der Waerden saddle point method [24].

The analogous results are not (yet) known in the case where the underlying probability source has Markovian dependence or in the case of dynamical sources.

Author Contributions: This paper is based on a Ph.D. dissertation conducted by the L.A. under the supervision of the M.D.W. All authors have read and agreed to the published version of the manuscript.

Funding: M.D.W. Ward's research is supported by FFAR Grant 534662, by the USDA NIFA Food and Agriculture Cyberinformatics and Tools (FACT) initiative, by NSF Grant DMS-1246818, by the NSF Science & Technology Center for Science of Information Grant CCF-0939370, and by the Society Of Actuaries.

Acknowledgments: The authors thank Wojciech Szpankowski and Mireille Régnier for insightful conversations on this topic.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PGF	Probabilty Generating Function
P	Probability
E	Expected value
Var	Variance
$\mathbf{E}[(X_{n,k})_2]$	The second factorial moment of $X_{n,k}$

References

1. Ehrenfeucht, A.; Lee, K.; Rozenberg, G. Subword complexities of various classes of deterministic developmental languages without interactions. *Theor. Comput. Sci.* **1975**, *1*, 59–75. [CrossRef]
2. Morse, M.; Hedlund, G.A. Symbolic Dynamics. *Am. J. Math.* **1938**, *60*, 815–866. [CrossRef]
3. Jacquet, P.; Szpankowski, W. *Analytic Pattern Matching: From DNA to Twitter*; Cambridge University Press: Cambridge, UK, 2015.
4. Bell, T.C.; Cleary, J.G.; Witten, I.H. *Text Compression*; Prentice-Hall: Upper Saddle River, NJ, USA, 1990.

5. Burge, C.; Campbell, A.M.; Karlin, S. Over-and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 1358–1362. [CrossRef] [PubMed]
6. Fickett, J.W.; Torney, D.C.; Wolf, D.R. Base compositional structure of genomes. *Genomics* **1992**, *13*, 1056–1064. [CrossRef]
7. Karlin, S.; Burge, C.; Campbell, A.M. Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Res.* **1992**, *20*, 1363–1370. [CrossRef] [PubMed]
8. Karlin, S.; Mrázek, J.; Campbell, A.M. Frequent Oligonucleotides and Peptides of the Haemophilus Influenzae Genome. *Nucleic Acids Res.* **1996**, *24*, 4263–4272. [CrossRef] [PubMed]
9. Pevzner, P.A.; Borodovsky, M.Y.; Mironov, A.A. Linguistics of Nucleotide Sequences II: Stationary Words in Genetic Texts and the Zonal Structure of DNA. *J. Biomol. Struct. Dyn.* **1989**, *6*, 1027–1038. [CrossRef] [PubMed]
10. Chen, X.; Francia, B.; Li, M.; Mckinnon, B.; Seker, A. Shared information and program plagiarism detection. *IEEE Trans. Inf. Theory* **2004**, *50*, 1545–1551. [CrossRef]
11. Chor, B.; Horn, D.; Goldman, N.; Levy, Y.; Massingham, T. Genomic DNA k-mer spectra: models and modalities. *Genome Biol.* **2009**, *10*, R108. [CrossRef]
12. Price, A.L.; Jones, N.C.; Pevzner, P.A. De novo identification of repeat families in large genomes. *Bioinformatics* **2005**, *21*, i351–i358. [CrossRef]
13. Janson, S.; Lonardi, S.; Szpankowski, W. On the Average Sequence Complexity. In *Annual Symposium on Combinatorial Pattern Matching*; Springer: Berlin/Heidelberger, Germany, 2004; pp. 74–88.
14. Jacquet, P.; Szpankowski, W. Autocorrelation on words and its applications: Analysis of suffix trees by string-ruler approach. *J. Comb. Theory Ser. A* **1994**, *66*, 237–269. [CrossRef]
15. Liang, F.M. *Word Hy-phen-a-tion by Com-put-er*; Technical Report; Stanford University: Stanford, CA, USA, 1983.
16. Weiner, P. Linear pattern matching algorithms. In Proceedings of the 14th Annual Symposium on Switching and Automata Theory (swat 1973), Iowa City, IA, USA, 15–17 October 1973; pp. 1–11.
17. Gheorghiciuc, I.; Ward, M.D. On correlation Polynomials and Subword Complexity. *Discrete Math. Theor. Comput. Sci.* **2007**, *7*, 1–18.
18. Bassino, F.; Clément, J.; Nicodème, P. Counting occurrences for a finite set of words: Combinatorial methods. *ACM Trans. Algorithms* **2012**, *8*, 31. [CrossRef]
19. Park, G.; Hwang, H.K.; Nicodème, P.; Szpankowski, W. Profile of Tries. In *Latin American Symposium on Theoretical Informatics*; Springer: Berlin/Heidelberger, Germany, 2008; pp. 1–11.
20. Flajolet, P.; Sedgewick, R. *Analytic Combinatorics*; Cambridge University Press: Cambridge, UK, 2009.
21. Lothaire, M. *Applied Combinatorics on Words*; Cambridge University Press: Cambridge, UK, 2005; Volume 105.
22. Szpankowski, W. *Average Case Analysis of Algorithms on Sequences*; John Wiley & Sons: Chichester, UK, 2011; Volume 50.
23. Widder, D.V. *The Laplace Transform (PMS-6)*; Princeton University Press: Princeton, NJ, USA, 2015.
24. van der Waerden, B.L. On the method of saddle points. *Appl. Sci. Res.* **1952**, *2*, 33–45. [CrossRef]

entropy

MDPI

Article

Criticality in Pareto Optimal Grammars?

Luís F Seoane [1,]*and Ricard Solé [2,3,4,]*

[1] Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), Campus UIB, 07122 Palma de Mallorca, Spain
[2] ICREA-Complex Systems Lab, Universitat Pompeu Fabra (GRIB), Dr Aiguader 80, 08003 Barcelona, Spain
[3] Institut de Biologia Evolutiva, CSIC-UPF, Pg Maritim de la Barceloneta 37, 08003 Barcelona, Spain
[4] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
* Correspondence: seoane@ifisc.uib-csic.es (L.F.S.) ricard.sole@upf.edu (R.S.)

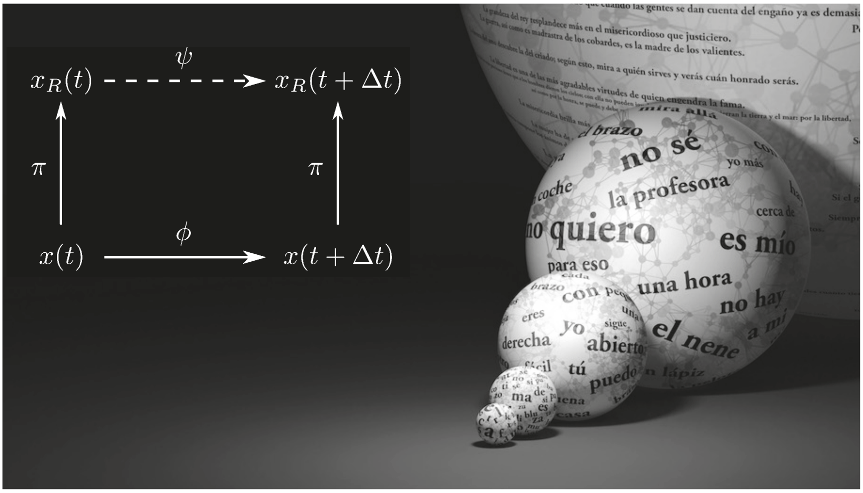
Received: 31 December 2019; Accepted: 27 January 2020; Published: 31 January 2020

Abstract: What are relevant levels of description when investigating human language? How are these levels connected to each other? Does one description yield smoothly into the next one such that different models lie naturally along a hierarchy containing each other? Or, instead, are there sharp transitions between one description and the next, such that to gain a little bit accuracy it is necessary to change our framework radically? Do different levels describe the same linguistic aspects with increasing (or decreasing) accuracy? Historically, answers to these questions were guided by intuition and resulted in subfields of study, from phonetics to syntax and semantics. Need for research at each level is acknowledged, but seldom are these different aspects brought together (with notable exceptions). Here, we propose a methodology to inspect empirical corpora systematically, and to extract from them, blindly, relevant phenomenological scales and interactions between them. Our methodology is rigorously grounded in information theory, multi-objective optimization, and statistical physics. Salient levels of linguistic description are readily interpretable in terms of energies, entropies, phase transitions, or criticality. Our results suggest a critical point in the description of human language, indicating that several complementary models are simultaneously necessary (and unavoidable) to describe it.

Keywords: syntax; Pareto-optimality; bottleneck method; phase transitions; statistical mechanics

1. Introduction

What is the "right" level of description for the faculty of human language? What would allow us to properly describe how it operates given the multiple scales involved—from letters and words to whole sentences? This nested character of language organization (Figure 1) pervades the great challenge of understanding how it originated and how we could generate it artificially. The standard answer to these and similar questions is given by rules of thumb that have helped us, historically, to navigate the linguistic complexities. We have identified salient aspects (e.g., phonetics, formal grammars, etc.) to which whole fields are devoted. In adopting a level of description, we hope to encapsulate a helpful snippet of knowledge. To guide these choices we must broadly fulfill two goals: (i) the system under research (human language) must be somehow simplified and (ii) despite that simplification we must still capture as many relevant, predictive features about our system's unfolding as possible. Some simplifications work better than others. In general, opting for a specific level does not mean that another one is not informative.



$$\begin{tikzcd} x_R(t) \arrow[r, dashed, "\psi"] & x_R(t+\Delta t) \\ x(t) \arrow[u, "\pi"] \arrow[r, "\phi"] & x(t+\Delta t) \arrow[u, "\pi"] \end{tikzcd}$$

Figure 1. Different levels of grammar. Language contains several layers of complexity that can be gauged using different kinds of measures and are tied to different kinds of problems. The background picture summarizes the enormous combinatorial potential connecting different levels, from the alphabet (smaller sphere) to grammatically correct sentences (larger sphere). On top of this, it is possible to describe each layer by means of a coarse-grained symbolic dynamics approach. One particularly relevant level is the one associated to the way syntax allows generating grammatically correct strings $x(t)$. As indicated in the left diagram, symbols succeed each other following some rules ϕ. A coarse-graining π groups up symbols in a series of classes such that the names of these classes; $x_R(t)$ also generate some symbolic dynamics whose rules are captured by ψ. How much information can the dynamics induced by ψ recover about the original dynamics induced by ϕ? Good choices of π and ψ will preserve as much information as possible despite being relatively simple.

A successful approach to explore human language is through networks. Nodes of a language web can be letters, syllables, or words; links can represent co-occurrences, structural similarity, phonology, or syntactic or semantic relations [1–7]. Are these different levels of description nested parsimoniously into each other? Or do sharp transitions exist that establish clear phenomenological realms? Most of the network-level topological analyses suggest potential paths to understand linguistic processing and hint at deeper features of language organization. However, the connection between different levels are seldom explored, with few exceptions based on purely topological patterns [8]; or some ambitious attempts to integrate all linguistic scales from the evolutionary one to the production of phonemes [9,10].

In this paper, we present a methodology to tackle this problem in linguistics: When are different levels of description pertinent? When can we forgo some details and focus on others? For example, when do we need to attend to syntactic constraints, and when do we need to pay attention to phonology? How do the descriptions at different levels come together? This interplay can be far from trivial: note, e.g., how phonetics dictates the grammatical choice of the determiner form "a" or "an" in English. Similarly, phonetic choices with no grammatical consequence can evolve into rigid syntactic rules in the long term. Is the description at a higher level always grounded in all previous stages, or do descriptions exist that do not depend on details from other scales? Likely, these are not all or nothing question. Therefore, rather, how many details in a given description do we need to carry on to the next one?

To exemplify how these questions can be approached, we look at written corpora as symbolic series. There are many ways in which a written corpus can be considered a symbolic series. For example, we can study the succession of letters in a text. Then, the available vocabulary consists of all letters in the alphabet (often including punctuations marks):

$$\chi^{letters} \equiv \{a, b, \dots, z, !, ?, \dots\}. \tag{1}$$

Alternatively, we can consider words as indivisible. In such cases, our vocabulary (χ^{words}) would consist of all entries in a dictionary. We can study even simpler symbolic dynamics, e.g., if we group together all words of each given grammatical class and consider words within a class equal to each other. From this point of view, we do not gain much by keeping explicit words in our corpora. We can just substitute each one by its grammatical class, for example,

$$\text{green} \quad \text{colorless} \quad \text{ideas} \quad \text{sleep} \quad \text{furiously} \longrightarrow adj \;\; adj \;\; noun \;\; verb \;\; adv. \tag{2}$$

After this, we can study the resulting series that have as symbols elements of the coarse-grained vocabulary:

$$\chi^{grammar} \equiv \{noun, verb, adj, adv, prep, \dots\}. \tag{3}$$

Further abstractions are possible. For example, we can introduce a mapping that retains the difference between nouns and verbs, and groups all other words in an abstract third category:

$$\text{adj} \quad \text{adj} \quad \text{noun} \quad \text{verb} \quad \text{adv} \longrightarrow cat_3 \;\; cat_3 \;\; noun \;\; verb \;\; cat_3. \tag{4}$$

It is fair to ask which of these descriptions are more useful, when to stop our abstractions, whether different levels define complementary or redundant aspects of language, etc. Each of these descriptions introduces an operation that maps the most fine-grained vocabulary into less detailed ones, for example,

$$\pi : \chi^{words} \rightarrow \chi^{grammar}. \tag{5}$$

To validate the accuracy of this mapping, we need a second element. At the most fundamental level, some unknown rules ϕ exist. They are the ones connecting words to each other in real language and correspond to the generative mechanisms that we would like to unravel. At the level coarse-grained by a mapping π, we can propose a description Ψ (Figure 1) that captures how the less-detailed dynamics advance. How well can we recover the original series depends on our choices of π and Ψ. Particularly good descriptions at different scales conform the answers to the questions raised above. The ϕ and Ψ mappings play roles similar to language grammar, i.e., sets of rules that tell us what words can follow each other. Some rules show up in actual corpora more often than others. Almost every sentence needs to deal with the Subject-Verb-Object (SVO) rule, but only seldom do we find all types of adjectives in a same phrase. If we would infer a grammar empirically by looking at English corpora, we could easily oversee that there is a rule for adjective order too. However, as it can be so easily missed, this might not be as important as SVO to understand how English works.

Here, we investigate grammars, or sets of rules, that are empirically derived from written corpora. We would like to study as many grammars as possible, and to evaluate numerically how well each of them works. In this approach, a wrong rule (e.g., one proposing that sentence order in English is VSO instead of SVO) would perform poorly and be readily discarded. It is more difficult to test descriptive grammars (e.g., a rule that dictates the adjective order), so instead we adopt abstract models that tell us the probability that classes of words follow each other. For example, in English, it is likely to find an adjective or a noun after a determiner, but it is unlikely to find a verb. Our approach is inspired by the information bottleneck method [11–15], rate distortion theory [16,17], and similar techniques [18–22]. In all these studies, arbitrary symbolic dynamics are divided into the observations up to a certain point, $\overleftarrow{x}$, the dynamics from that point onward, $\overrightarrow{x}$, and some coarse-grained model R (which plays the role of our π and Ψ combined) that attempts to conceptualize what has happened in $\overleftarrow{x}$ to predict what will happen in $\overrightarrow{x}$. This scheme allows us to quantify mathematically how good is a choice of $R \equiv \{\pi, \Psi\}$. For example, it is usual to search for models R that maximize the quantity:

$$I(\overleftarrow{x} : R) + \alpha I(\overleftarrow{x} : \overrightarrow{x}|R) \tag{6}$$

for some $\alpha > 0$. The first term captures the information that the model carries about the observed dynamics $\overleftarrow{x}$, the second term captures the information that the past dynamics carry about the future given the filter imposed by the model R, and the metaparameter α weights the importance of each term towards the global optimization.

We will evaluate our probabilistic grammars in a similar (yet slightly different) fashion. For our method of choice, we first acknowledge that we are facing a Pareto, or Multi-Objective Optimization (MOO) problem [23–25]. In this kind of problem we attempt to minimize or maximize different traits of the model simultaneously. Such efforts are often in conflict with each other. In our case, we want to make our models as simple as possible, but in that simplicity we ask that they retain as much of their predictive power as possible. We will quantify how different grammars perform in both these regards, and rank them accordingly. MOO problems rarely present global optima, i.e., we will not be able to find the best grammar. Instead, MOO solutions are usually embodied by Pareto-optimal trade-offs. These are collections of designs that cannot be improved in both optimization targets simultaneously. In our case these will be grammars that cannot be made simpler without losing some accuracy in their description of a text, or that cannot be made more accurate without making them more complicated.

The solutions to MOO problems are connected with statistical mechanics [25–29]. The geometric representation of the optimal trade-off reveals phase transitions (similar to the phenomena of water turning into ice or evaporating promptly with slight variations of temperature around 0 or 100 degrees Celsius) and critical points. In our case, Pareto optimal grammars would give us a collection of linguistic descriptions that simultaneously optimize how simply language rules can become while retaining as much of their explanatory power as possible. The different grammars along a trade-off would become optimal descriptions at different levels, depending on how much detail we wish to track about a corpus. Positive (second order) phase transitions would indicate salient grammars that are adequate descriptions of a corpus at several scales. Negative (first order) phase transitions would indicate levels at which the optimal description of our language changes drastically and very suddenly between extreme sets of rules. Critical points would indicate the presence of somehow irreducible complexity in which different descriptions of a language become simultaneously necessary, and aspects included in one description are not provided by any other. Although critical points seem a worst-case scenario towards describing language, they are a favorite of statistical physics. Systems at a critical point often display a series of desirable characteristics, such as versatility, enhanced computational abilities, and optimal handling of memory [30–38].

In Section 2 we explain how we infer our π and Ψ (i.e., our abstract "grammatical classes" and associated grammars), and the mathematical methods used to quantify how simple and accurate they are. In Section 3, we present some preliminary results, always keeping in mind that this paper is an illustration of the intended methodology. More thorough implementations will follow in the future. In Section 4, we reflect about the insights that we might win with these methods, how they could integrate more linguistic aspects, and how they could be adapted to deal with the complicated, hierarchical nature of language.

2. Methods

2.1. Corpus Description and Preparation

We took a sample of 49 newspaper articles from the Corpus of Contemporary American English [39]. The articles were selected such that they did not contain foreign (non-English) words or symbols. We substituted by a period every punctuation mark that indicated the end of a sentence and removed any other punctuation mark except for the apostrophes indicating a contraction (e.g., "don't") or a genitive (e.g., "someone's"). Ideally, we would like to use raw texts and see Pareto optimal grammars emerging from them. These should also include instructions about how alien symbols or words (loosely speaking, any items that are not proper of English language, e.g., french terms, accent marks, etc.) are treated. However, these are rather minor details. Effective grammars should specify first how its own words are articulated.

Our more basic level of analysis will already be a coarse-grained one. Again, ideally, we would present our methods with texts in which each word is explicitly expelled out. Our blind techniques should then infer grammatical classes (if any were useful) based on how different words correlate. For example, we expect that our blind methods would be able, at some point, to group all nouns together based on their syntactic regularities. While this is possible, it is very time- and resource-consuming for the demonstration intended here. Therefore, we preprocessed our corpus using Python's Natural Language Processing Toolkit [40] to map every word into one of the $N_G = 34$ grammatical classes shown in Table 1. We then substituted every word in the corpus by its grammatical class. The resulting texts constitute the symbolic dynamics that we analyze.

Table 1. Grammatical classes present in the most fine-grained level of our corpora.

Conjunction	Adverb
Cardinal number	Adverb, comparative
Determiner	Adverb, superlative
Existential there	to
Preposition	Interjection
Adjective	Verb, base form
Adjective, comparative	Verb, past tense
Adjective, superlative	Verb, gerund or present participle
Modal	Verb, past participle
Noun, singular	Verb, non-3rd person singular present
Noun, plural	Verb, 3rd person singular present
Proper noun, singular	Wh-determiner
Proper noun, plural	Wh-pronoun
Predeterminer	Possessive wh-pronoun
Possessive ending	Wh-adverb
Personal pronoun	None of the above
Possessive pronoun	'.'

2.2. Word Embeddings and Coarse-Graining

We would like to explore the most general grammars possible. However, as advanced above, to make some headway we restrict ourselves to grammar models that encode a tongue's rules in a probabilistic way, telling us how likely it is that words follow each other in a text. Even in this narrower class there is an inscrutably large number of possibilities depending, e.g., on how far back we look into a sentence to determine the next word's likelihood, on whether we build intermediate phrases to keep track of the symbolic dynamics in a hierarchical way, etc. Here, we only attempt to predict the next word given the current one. We will also restrict ourselves to maximum entropy (*MaxEnt*) models, which are the models that introduce less further assumptions provided a series of observations [37,41–49]. We explain these kind of models in the next subsection. First, we need to introduce some notation and a suitable encoding of our corpus so we can manipulate it mathematically.

We use a one-hot embedding, which substitutes each word in a text by a binary string that consists of all zeros and exactly one 1. The position of the 1 indicates the class of word that we are dealing with. Above, we illustrated several levels of coarse-graining. In a very fundamental one, each word represents a class of its own. Our vocabulary in the simple example sentence "green colorless ideas sleep furiously" consists of

$$\chi^{words} \equiv \{ideas, sleep, green, colorless, furiously\} \tag{7}$$

which in its binary form becomes

$$\tilde{\chi}^{words} = \{10000, 01000, 00100, 00010, 00001\}. \tag{8}$$

We also illustrated a level of coarse-graining in which nouns and verbs are retained, but all other words are grouped together in a third category (Equation (4)). The corresponding vocabulary

$$\chi \equiv \{noun, verb, cat_3\} \tag{9}$$

becomes, through the one-hot embedding:

$$\tilde{\chi} = \{100, 010, 001\}. \tag{10}$$

Throughout this paper, we will note by χ^λ the vocabulary (set of unique symbols) at a description level λ, and we will refer by $\tilde{\chi}^\lambda$ to its one-hot representation. We will name $c_j^\lambda \in \chi^\lambda$, with $j \in \{1, \dots, N^\lambda\}$, to each of the N^λ unique symbols at description level λ. Each of these symbols stands for an abstract class of words, which might or might not correspond to actual grammatical classes in the standard literature. The binary representation of each class is correspondingly noted by $\sigma_j^\lambda \in \tilde{\chi}^\lambda$.

To explore models of different complexity we start with all the grammatical classes outlined in Table 1 and proceed by lumping categories together. We will elaborate a probabilistic grammar for each level of coarse-graining. Later, we will compare the performance of all descriptions. In lumping grammatical classes together there are some choices more effective than others. For example, it seems wise to group comparative and superlative adverbs earlier than nouns and verbs. We expect the former to behave more similarly than the later, and therefore to lose less descriptive power when treating both comparative and superlative adverbs as one class. In future versions of this work, we intend to explore arbitrary lumping strategies. Here, to produce results within a less demanding computational framework, we use an informed shortcut. We build the maximum entropy model of the least coarse-grained category (which, again, in this paper consists of the grammatical classes in Table 1). Through some manipulations explained below, this model allows us to extract correlations between a current word and the next one (illustrated in Figure 2). These correlations allow us to build a dendogram (Figure 3a) based on how similarly different grammatical classes behave.

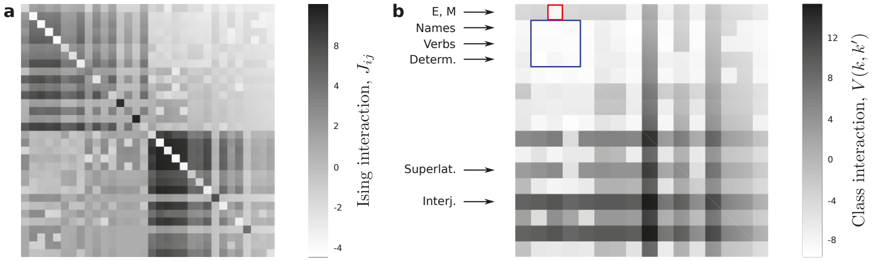


Figure 2. Interactions between spins and word classes. (**a**) A first crude model with spins encloses more information than we need for the kind of calculations that we wish to do right now. (**b**) A reduced version of that model gives us an interaction energy between words or classes of words. These potentials capture some non-trivial features of English syntax, e.g., the existential "there" in "there is" or modal verbs (marked E and M respectively) have a lower interaction energy if they are followed by verbs. Interjections present fairly large interaction energy with any other word, perhaps as a consequence of their independence within sentences.

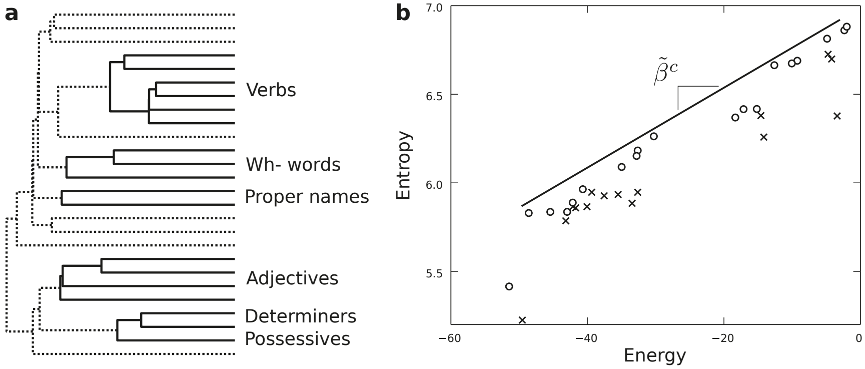


Figure 3. Pareto optimal maximum entropy models of human language. Among all the models that we try out, we prefer those Pareto optimal in energy minimization and entropy maximization. (**a**) These reveal a hierarchy of models in which different word classes group up at different levels. The clustering reveals a series of grammatical classes that belong together owing to the statistical properties of the symbolic dynamics, such as possessives and determiners which appear near to adjectives. (**b**) A first approximation to the Pareto front of the problem. Future implementations will try out more grammatical classes and produce better quality Pareto fronts, establishing whether phase transitions or criticality are truly present.

This dendogram suggests an order in which to merge the different classes, which is just a good guess. There are many reasons why the hierarchy emerging from the dendogram might not be the best coarse-graining. We will explore more exhaustive possibilities in the future. In any case, this scheme defines a series of functions π^{λ} (which play the role of π in Figure 1) that map the elements of the most fine-grained vocabulary $\chi^0 \equiv \chi^{grammar}$ (as defined by the classes in Table 1) into a series of each time more coarse-grained and abstract categories χ^λ, with $\lambda = 1, \dots, N_G - 1$ indicating how many categories have been merged at that level.

2.3. Maximum-Entropy Models

To build the MaxEnt model at a given level λ of coarse-graining, we substitute every word in our corpus by its binary representation. Our text then becomes a binary string. For example, with the coarse-graining in which nouns and verbs are kept, and all other words are abstracted into cat_3, we have

$$\text{green} \quad \text{colorless} \quad \text{ideas} \quad \text{sleep} \quad \text{furiously} \longrightarrow 001 \;\; 001 \;\; 100 \;\; 010 \;\; 001. \tag{11}$$

We indicate the i-th word in a text by $w(i)$. Its grammatical class in the description level λ is noted:

$$c^\lambda(i) \equiv \pi^\lambda(w(i)), \tag{12}$$

and its binary representation:

$$\sigma^\lambda(i) \equiv \tilde{\pi}^\lambda(w(i)). \tag{13}$$

Both mappings π^λ and $\tilde{\pi}^\lambda$ contain the same information, and both of them play the role of π in Figure 1. Note that $c^\lambda(i) = c_j^\lambda$ for some j, and that although $i \in \{1, \dots, N_w\}$ indexes words as they happen in a text (of length N_w), $j \in \{1, \dots, N^\lambda\}$ indexes unique grammatical classes in χ^λ. Each binary representation consists of N^λ bits. When necessary, we will use a subindex k to label $\sigma_{j,k}^\lambda$ as the k-th bit of the j-th class's binary representation at a given coarse-graining level λ.

We next produce binary samples that include each word and the one next to it in a text: $\langle \sigma^\lambda(i) | \sigma^\lambda(i+1) \rangle$, where $\langle \cdot | \cdot \rangle$ indicates concatenation. Thus, the coarse-grained sentence from Equation (11) yields the samples:

$$\{001001, 001100, 100010, 010001\}. \tag{14}$$

Each sample has size $2N^\lambda$ (when needed, the index k over bits will also label positions from 1 to $2N^\lambda$). Large corpora will produce huge collections of such samples. We can summarize these collections by giving the empirical frequency $F\left(\left\langle \sigma_j^\lambda | \sigma_{j'}^\lambda \right\rangle\right)$ with which each of the $\left(N^\lambda\right)^2$ possible bit strings with length $2N^\lambda$ shows up. These collections behave as samples of what is known as spin glasses in statistical mechanics. We have powerful mathematical tools to infer MaxEnt models for spin glasses – therefore all these efforts.

3. Results

Using the methodology described above, we have coarse-grained the words of a written corpus, first, into the 34 grammatical classes shown in Table 1. This process is illustrated by Equation (2). The resulting symbolic series was binarized to create samples akin to spin glasses, a well studied model from statistical mechanics that allows us to use powerful mathematical tools on our problem. This process was then repeated at several levels of coarse graining as words were further lumped into abstract grammatical categories (e.g., as in Equation (4)). At each level of description, the inferred spin glass model plays the role of a grammar that constrains, in a probabilistic fashion, how word classes can follow each other in a text. These mathematical tools from spin glass theory allow us to test grammars from different description levels against each other as will become clear now.

In spin glasses, a collection of little magnets (or spins) is arranged in space. We say that a magnet is in state $\sigma = 1$ if its north pole is pointing upwards and in state $\sigma = -1$ if its pointing downwards (these are equivalent to the 1s and 0s in our word samples). Two of these little magnets interact through their magnetic fields. These fields build up a force that tends to align both spins in the same direction, whichever it is, just as two magnets in your hand try to fall along a specific direction with respect to each other. On top of this, the spins can interact with an external magnetic field—bringing in a much bigger magnet which orientation cannot be controlled. This external field tends to align the little spins along its fixed, preferred direction. Given the spin states σ_1 and σ_2, the energy of their interaction with the external magnetic field and with each other can be written as

$$\begin{aligned} E(\sigma_1, \sigma_2) &= -\frac{1}{2}\left(2h_1\sigma_1 + \sigma_1 J_{12}\sigma_2 + \sigma_2 J_{21}\sigma_1 + 2h_2\sigma_2\right) \\ &= -\frac{1}{2}\left(J_{11}\sigma_1 + \sigma_1 J_{12}\sigma_2 + \sigma_2 J_{21}\sigma_1 + J_{22}\sigma_2\right). \end{aligned} \tag{15}$$

J_{12} and J_{21} (with $J_{12} = J_{21}$) denote the strength of the interaction between the spins, and $J_{11} \equiv 2h_1$ and $J_{22} = 2h_2$ denote the interaction of each spin with the external field. The terms h_1 and h_2 are also known as biases. If the spins are aligned with each other and with the external field, the resulting energy is the lowest possible. Each misalignment increases the energy of the system. In physics, states with less energy are more probable. Statistical mechanics allows us to write precisely the likelihood of finding this system in each of its four ($\{1,1\}$, $\{1,-1\}$, $\{-1,1\}$, and $\{-1,-1\}$) possible states:

$$P(\sigma_1, \sigma_2) = \frac{e^{-\beta E(\sigma_1,\sigma_2)}}{Z}, \tag{16}$$

where $\beta = 1/T$ is the inverse of the temperature. The term

$$\begin{aligned} Z &= e^{-\beta E(1,1)} + e^{-\beta E(1,-1)} + e^{-\beta E(-1,1)} + e^{-\beta E(-1,-1)} \\ &= \sum_{\sigma_1, \sigma_2 = \pm 1} e^{-\beta E(\sigma_1,\sigma_2)} \end{aligned} \tag{17}$$

is known as the partition function and is a normalizing factor that guarantees that the probability distribution in Equation (16) is well defined.

Back to our text corpus in its binary representation, we know the empirical frequency $F\left(\left\langle \sigma_j^\lambda | \sigma_{j'}^\lambda \right\rangle\right)$ with which each of the possible spin configurations shows up—we just need to read it from our corpus. We can treat our collection of 0s and 1s as if they were ± 1 samples of a spin glass, and attempt to infer the β^λ and J^λ which (through a formula similar to Equation (16)) more faithfully reproduce the observed sample frequencies. The superindex in β^λ and J^λ indicates that they will change with the level of coarse-graining. Inferring those β^λ and J^λ amounts to finding the MaxEnt model at that coarse-grained level. As advanced above, MaxEnt models are convenient because they are the models that introduce less extra hypotheses given some observations. In other words, if we infer the MaxEnt model for some λ, any other model with the same coarse-graining would be introducing spurious hypotheses that are not suggested by the data. To infer MaxEnt models, we used Minimum Probability Flow Learning (MPFL [50]), a fast and reliable method that infers the J^λ given a sufficiently large sample.

Each grammatical class is represented by N^λ spins at the λ-th coarse-graining. This implies, as we know, that our samples consists of $2N^\lambda$ spints. MPFL returns a matrix J^λ of size $2N^\lambda \times 2N^\lambda$. This matrix embodies our abstract, probabilistic grammar (and plays the role of Ψ in Figure 1). Each entry $J_{kk'}^\lambda$ of this matrix tells us the interaction energy between the k-th and k'-th bits in a sample (with $k, k' = 1, \ldots, 2N^\lambda$). However, each grammatical class is represented not by one spin, but by a configuration of spins that has only one 1. To obtain the interaction energies between grammatical classes (rather than between spins), we need to compute

$$V^{\lambda}(c_j^{\lambda}, c_{j'}^{\lambda}) \quad = \quad \frac{1}{2}\sum_{k,k'} \sigma_{j,k}^{\lambda} J_{kk'}^{\lambda} \sigma_{j',k'}^{\lambda}. \tag{18}$$

This energy in turn tells us the frequency with which we should observe each pair of words according to the model:

$$P^{\lambda}\left(\left\langle c_j^{\lambda} | c_{j'}^{\lambda} \right\rangle\right) \quad = \quad \frac{1}{Z^{\lambda}} e^{\beta V^{\lambda}(c_j^{\lambda}, c_{j'}^{\lambda})}. \tag{19}$$

We inferred MaxEnt models for the more fine-grained level of description (χ^0 as given by the grammatical classes in Table 1), as well as for every other intermediate level χ^{λ}. Figure 2a shows the emerging spin-spin interactions for $l = 15$, which consists of only 19 (versus the original 34) grammatical classes. This matrix presents a clear box structure:

$$J^{\lambda} \quad = \quad \left[\begin{array}{c|c} 2h^{\lambda} & \overrightarrow{\partial}^{\lambda} \\ ine \overleftarrow{\partial}^{\lambda} & 2\bar{h}^{\lambda} \end{array} \right]. \tag{20}$$

The diagonal blocks ($2h^{\lambda}$ and $2\bar{h}^{\lambda}$) represent the interactions between all spins that define, separately, the first and second words in each sample. As our corpus becomes infinitely large, $h^{\lambda} \to \bar{h}^{\lambda}$. These terms do not capture the interaction between grammatical classes. In the spin-glass analogy, they are equivalent to the interaction of each word with the external magnet that biases the presence of some grammatical classes over others. Such biases affect the frequencies $P^{\lambda}(c_j^{\lambda})$ with which individual classes show up, but not the frequency with which they are paired up. Therefore, the h^{λ} and $\bar{h}^{\lambda}$ are not giving us much syntactic information.

More interesting for us are the interaction terms stored in $\overrightarrow{\partial}^{\lambda}$ and $\overleftarrow{\partial}^{\lambda}$. The inference method used guarantees that $\overrightarrow{\partial}^{\lambda} = (\overleftarrow{\partial}^{\lambda})^T$. It is from these terms that we can compute the part of $V^{\lambda}(c_j^{\lambda}, c_{j'}^{\lambda})$ (shown in Figure 2b) that pertains to pairwise interaction alone (i.e., the energy of the spin system when we discount the interaction with the external field). $V^{\lambda}(c_j^{\lambda}, c_{j'}^{\lambda})$ encodes the energy of two word classes when they are put next to each other in a text. The order in which words appear after each other is relevant, therefore that matrix is not symmetric. These energies reflect some of the rules of English. For example, the first row (labeled "E, M") is a class that has lumped together the existential "there" (as in "there is" and "there are") with all modal verbs. These tend to be followed by a verb in English, thus the matrix entry coding for $\langle ``E, M'' | ``verb'' \rangle$ (marked in red) is much lower than most entries for any other $\langle ``E, M'' | \cdot \rangle$. The blue square encompasses verbs, nouns, and determiners. Although the differences there are very subtle, the energies reflect that it is more likely to see a noun after a determiner and not the other way around, and also that it is less likely to see a verb after a determiner.

It is not straightforward to compare all energies because they are affected by the raw frequency with which pairs of words show up in a text. In that sense, our corpus size might be sampling some pairings insufficiently so that their energies do not reflect proper English use. On the other hand, classes such as nouns, verbs, and determiners happen so often (and so often combined with each other) that they present very low energies as compared with other possible pairs. This makes the comparison more difficult by visual inspection.

It is possible to use $V^{\lambda}(c_j^{\lambda}, c_{j'}^{\lambda})$ to generate a synthetic text $\tilde{T}^{\lambda}$ and evaluate its energy $E^0(\tilde{T}^{\lambda})$ using the most fine-grained model J^0. If the coarse-grained model $V^{\lambda}(c_j^{\lambda}, c_{j'}^{\lambda})$ retains a lot of the original structure, the generated text will fit gracefully in the rules dictated by J^0—just as magnets falling into place. Such texts would present very low energy when evaluated by J^0. If the coarse-grained model has erased much of the original structure, the synthetic text will present odd pairings. These would feel similar to magnets that we are forcing into a wrong disposition, therefore resulting in a large energy when J^0 is used. In other words, this energy reflects how accurate each coarse-grained model is.

That accuracy is one of the targets in our MOO problem, in which we attempt to retain as much information as possible with models as simple as possible. To quantify that second target, simplicity, we turn to entropy. The simplest model possible generates words that fall in either class of χ^0 randomly and uniformly, thus presenting the largest entropy possible. More complex models, in their attempt to remain accurate, introduce constraints as to how the words in the coarse-grained model must be mapped back into the classes available in χ^0. That operation would be the reverse of π^λ. This reverse mapping, however, cannot be undone without error because the coarse-graining erases information. Entropy measures the amount of information that has been erased, and therefore how simple the model has been made.

Figure 3b shows the energy $E^0(T^\lambda)$ and entropy $S^0(T^\lambda)$ for synthetic texts generated with the whole range of coarse-grainings explored. In terms of Pareto optimality, we expect our models to have as low an energy as possible while having the largest entropy compatible with each energy—just as thermodynamic systems do. Such models would simultaneously optimize their simplicity and accuracy. Within the sample, some of these models are Pareto dominated (crosses in Figure 3b) by some others. This means that for each of those models at least some other one exists that is simpler and more accurate at the same time. These models are suboptimal regarding both optimization targets, so we do not need to bother with them.The non-dominated ones (marked by circles in Figure 3b) capture better descriptions in both senses (accuracy and simplicity). They are such that we cannot move from one to another without improving an optimization target and worsening the other. They embody the optimal trade-off possible (of course, limited by all the approximations made in this paper), and we cannot choose a model over the others without introducing some degree of artificial preference either for simplicity or accuracy.

In statistical mechanics the energy and entropy of a system are brought together by the free energy:

$$F = E - \hat{T}S = E - S/\hat{\beta}. \tag{21}$$

Here, $\hat{T}$ plays a role akin to a temperature and $\hat{\beta}$ plays the role of its inverse. We noted $\hat{\beta} \neq \beta$ to indicate that these temperature and inverse temperature are different from the ones in Equation (19). Those temperatures control how often a word shows up given a model, whereas $\hat{\beta}$ controls how appropriate each level of description is. When $\hat{\beta}$ is low (and $\hat{T}$ is large), a minimum free energy in Equation (21) is attained by maximizing the entropy rather than minimizing the energy. This is, low $\hat{\beta}$ selects for simpler descriptions. When $\hat{\beta}$ is large (and $\hat{T}$ is small), we prefer models with lower energy, i.e., higher accuracy.

By varying $\hat{\beta}$ we visit the range of models available, i.e., we visit the collection of Pareto optimal grammars (circles in Figure 3b). In statistical mechanics, by varying the temperature of a system we visit a series of states of matter (this is, we put, e.g., a glass of water at different temperatures and observe how its volume and pressure change). At some relevant points, called phase transitions, the states of matter change radically, e.g., water freezes swiftly at 0 degrees Celsius, and evaporates right at 100 degrees Celsius. The geometry of Pareto optimal states of matter tells us when such transitions occur [25–29].

Similarly, the geometric disposition of Pareto optimal models in Figure 3b tells us when a drastic change in our best description is needed as we vary $\hat{\beta}$. Relevant phase transitions are given by cavities and salient points along the Pareto optimal solutions. In the first approach, we observe several cavities. More interestingly, perhaps, is the possibility that our Pareto optimal models might fall along a straight line; one has been added as a guideline in Figure 3b. Although there are obvious deviations from it, such description might be feasible at large. Straight lines in this plot are interesting because they indicate the existence of special critical points [28,37,46–48]. In the next section, we discuss what criticality might mean in this context.

4. Discussion

In this paper, we study how different hierarchical levels in the description of human language are entangled with each other. Our work is currently at a preliminary stage, and this manuscript aims at presenting overall goals and a possible methodological way to tackle relevant questions. Some interesting results are presented as an illustration and discussed in this section to exemplify the kind of debate that this line of research can spark.

Our work puts forward a rigorous and systematic framework to tackle the questions introduced above, namely, what levels of description are relevant to understand human language and how do these different descriptions interact with each other. Historically, we have answered these questions guided by intuition. Some aspects of language are so salient that they demand a sub-field of their own. Although this complexity and interconnectedness is widely acknowledged, its study is still fairly compartmentalized. The portray of language as a multilayered network system is a recent exception [8], as it is the notable and lasting effort by Christiansen et al. [9,10] to link all scales of language production, development, and evolution in a unified frame.

We generated a collection of models that describe a written English corpus. These models trade optimally a decreasing level of accuracy by increasing simplicity. By doing so, they gradually lose track of variables involved in the description at more detailed levels. For example, as we saw above, the existential "there" is merged with modal verbs. Indeed, these two classes were lumped together before the distinction between all other verbs was erased. Although those grammatical classes are conceptually different, our blind methodology found convenient to merge them earlier in order to elaborate more efficient compact grammars.

Remaining as accurate as possible while becoming as simple as possible is a multi-objective optimization problem. The conflicting targets are captured by the energy and entropy that artificial texts generated by a coarse-grained model have when evaluated at the most accurate level of description. We could have quantified these targets in other ways (e.g., counting the number of grammatical classes to quantify complexity, and measuring similarity between synthetic and real texts for accuracy). Those alternative choices should be explored systematically in the future to understand which options are more informative. Our choices, however, make our results easy to interpret in physical terms. For example, improbable (unnatural) texts have high energies in any good model.

The grammars that optimally trade between accuracy (low energy) and simplicity (high entropy) conform the Pareto front (i.e., the solution) of the MOO problem. Its shape in the energy-entropy plane (Figure 3) is linked to phase transitions [25–29]. According to this framework, we do not find evidence of a positive (second order) phase transition. What could such a transition imply for our system? The presence of a positive phase transition in our data would suggest the existence of a salient level of description capable of capturing a large amount of linguistic structure in relatively simple terms. For example, if a unique grammatical rule would serve to connect words together disregarding of the grammatical classes in which we have split our vocabulary. We would expect that to be the case, e.g., if a single master rule such as merge would serve to generate all the complexity of human language without further constraints arising. This does not seem to be the case. However, this does not rule out the existence of the relevant merge operation, nor does it deny its possible fundamental role. Indeed, Chomsky proposes that merge is the fundamental operation of syntax, but that it leaves the creative process of language underconstrained

[51–53]. As a result, actual implementations (i.e., real languages) see a plethora of further complexities arising in a phenomena akin to symmetry breaking.

The presence of a negative (first order) phase transition would acknowledge several salient levels of description needed to understand human language. These salient descriptions would furthermore present an important gap separating them. This would indicate that discrete approaches would be possible to describe language without missing any detail by ignoring the intermediate possibilities. If that were the case, we would still need to analyze the emerging models and look at similarities between them to understand whether both models capture a same core phenomenology at two relevant (yet distant) scales; or whether each model focuses on a specific, complementary aspect that the other description has no saying about. Some elements in Figure 3b are compatible with this kind of phase transition.

However, the disposition of several Pareto optimal grammars along a seemingly straight line rather suggests the existence of a special kind of critical phenomenon [28,37,46–48]. Criticality is a worst-case scenario in terms of description. It implies that there is no trivial model, nor couple of models, nor relatively small collection that can capture the whole of linguistic phenomenology at any level. A degenerate number of descriptions is simultaneously necessary, and elements trivial in a level can become cornerstones of another. Also, potentially, constraints imposed by a linguistic domain (e.g., phonology) can penetrate all the way and alter the operating rules of other domains (e.g., syntax or semantics). We can list examples of how this happens in several tongues (such as the case of determiners "a" and 'an' in English mentioned above). The kind of criticality suggested by our results would indicate that such intrusions are the norm rather than the exception. Note that this opportunistic view of grammar appears compatible with Christiansen's thesis that language evolved, as an interface, to make itself useful to our species, necessarily exploiting all kinds of hacks along its way [9].

Zipf's law is a notable distribution in linguistics [54,55]. It states that the n-th most abundant word in a text shows up with a frequency that is inversely proportional to that word's rank (i.e., n). The presence of this distribution in linguistic corpora has been linked to an optimal balance between communicative tensions [54,56,57]. It has also been proved mathematically that Zipf's law is an unavoidable feature of open-ended evolving systems [58]. Languages and linguistic creativity are candidates to present open-ended evolution. Could this open-endedness be reflected also in the diversity of grammatical rules that form a language? Could we expect to find a power-law in the distribution of word combinations with a given energy? If that were the case, Bialek et al. [37,47] proved mathematically that the relationship between energy and entropy of such grammars must be linear and therefore critical. In other words, our observation of criticality in this work, if confirmed, would be a strong hint (yet not sufficient) that the relevant Zipf distribution may also be lurking behind grammars derived empirically from written corpora.

Numerous simplifications were introduced to produce the preliminary results in this paper. We started our analysis with words that have already been coarse-grained into 34 grammatical classes, barring the emergence of further intermediate categories dictated, e.g., by semantic use. We know that semantic considerations can condition combinations of words, such as what verbs can be applied to what kinds of agents [59]. The choice of words as units (instead of letters or syllables) is another limiting factor. Words are symbols whose meanings do not depend on physical correlates with the objects signified [60]. In that sense, their association to their constituent letters and phonems is arbitrary. Their meaning is truly emergent and not rooted in their parts. Introducing letters, syllables, and phonetics in our analysis might reveal and allow us to capture that true emergence.

To do this it might be necessary to work with hierarchical models that allow correlations beyond the next and previous words considered here. This kind of hierarchy, in general, is a critical aspect of language [53] that our approach should capture in due time. We have excluded it in this work to attain preliminary results in a reasonable time. Although hierarchical models are likely to be more demanding (in computational terms), they can be parsimoniously incorporated in our framework. A possibility is to use epsilon machines [61–63], which naturally lump together pieces of symbolic dynamics to find out causal states. These causal states act as shielding units that advance a symbolic dynamics in a uniquely determined way—just like phrases or sentences provide a sense of closure at their end, and direct the future of a text in new directions.

Author Contributions: Original conceptualization and data analysis: L.S. Both authors elaborated the research and wrote the manuscript. All authors have read and agree to the published version of the manuscript.

Funding: R.S. and L.S. were both supported by the Botín Foundation by Banco Santander through its Santander Universities Global Division.

Acknowledgments: L.S. wishes to acknowledge logistic and funding support from the Institute for Interdisciplinary Physics anc Complex Systems (IFISC) at the University of the Balearic Islands. The authors thank the Santa Fe Institute for hosting our visit where most of this paper was done at the Cormac McCarthy Library. Special thanks to Ephraim Winslow and Thomas Wake for enlightening comments.

Conflicts of Interest: The authors declare no conflicts of interests.

References

1. Ferrer, I.; Cancho, R.; Riordan, O.; Bollobás, B. The consequences of Zipf's law for syntax and symbolic reference. *Proc. R. Soc. B* **2005**, *272*, 561–565. [CrossRef] [PubMed]
2. Solé, R. Language: Syntax for free? *Nature* **2005**, *434*, 289. [CrossRef] [PubMed]
3. Corominas-Murtra, B.; Valverde, S.; Solé, R. The ontogeny of scale-free syntax networks: Phase transitions in early language acquisition. *Adv. Complex Syst.* **2009**, *12*, 371–392. [CrossRef]
4. Arbesman, S.; Strogatz, S.H.; Vitevitch, M.S. The structure of phonological networks across multiple languages. *Int. J. Bifurcat. Chaos* **2010**, *20*, 679–685. [CrossRef]
5. Solé, R.V.; Corominas-Murtra, B.; Valverde, S.; Steels, L. Language networks: Their structure, function, and evolution. *Complexity* **2010**, *15*, 20–26. [CrossRef]
6. Solé, R.V.; Seoane, L.F. Ambiguity in language networks. *Linguist. Rev.* **2015**, *32*, 5–35. [CrossRef]
7. Seoane, L.F.; Solé, R. The morphospace of language networks. *Sci. Rep.* **2018**, *8*, 1–14. [CrossRef]
8. Martinčić-Ipšić, S.; Margan, D.; Mexsxtrović, A. Multilayer network of language: A unified framework for structural analysis of linguistic subsystems. *Phys. A* **2016**, *457*, 117–128. [CrossRef]
9. Christiansen, M.H.; Chater, N. Language as shaped by the brain. *Behav. Brain Sci.* **2008**, *31*, 489–509. [CrossRef]
10. Christiansen, M.H.; Chater, N. *Creating Language: Integrating Evolution, Acquisition, and Processing*; MIT Press: Cambridge, MA, USA, 2016.
11. Tishby, N.; Pereira, F.C.; Bialek, W. The information bottleneck method. *arXiv* **2000**, arXiv:physics/0004057.
12. Still, S.; Bialek, W.; Bottou, L. Geometric clustering using the information bottleneck method. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2003.
13. Still, S.; Crutchfield, J.P. Structure or Noise? *arXiv* **2007**, arXiv:0708.0654.
14. Still, S.; Crutchfield, J.P.; Ellison, C.J. *Optimal Causal Inference*; Santa Fe Institute Working Paper #2007-08-024; Santa Fe Institute: Santa Fe, NM, USA, 2007.
15. Still, S. Information bottleneck approach to predictive inference. *Entropy* **2014**, *16*, 968–989. [CrossRef]
16. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **2001**, *27*, 379–423. [CrossRef]
17. Shannon, C.E.; Weaver, W. *The Mathematical Theory of Communication*; Univ of Illinois Press: Champaign, IL, USA, 1949.
18. Shalizi, C.R.; Moore, C. What is a macrostate? Subjective observations and objective dynamics. *arXiv* **2003**, arXiv:cond-mat/0303625.
19. Israeli, N.; Goldenfeld, N. Coarse-graining of cellular automata, emergence, and the predictability of complex systems. *Phys. Rev. E* **2006**, *73*, 026203. [CrossRef]
20. Görnerup, O.; Jacobi, M.N. A method for finding aggregated representations of linear dynamical systems. *Adv. Complex Syst.* **2010**, *13*, 199–215. [CrossRef]

21. Pfante, O.; Bertschinger, N.; Olbrich, E.; Ay, N.; Jost, J. Comparison between different methods of level identification. *Adv. Complex Syst.* **2014**, *17*, 1450007. [CrossRef]
22. Wolpert, D.H.; Grochow, J.A.; Libby, E.; DeDeo, S. *Optimal High-Level Descriptions of Dynamical Systems*; Santa Fe Institute working paper #2015-06-017; Santa Fe Institute: Santa Fe, NM, USA, 2014.
23. Coello, C. Twenty years of evolutionary multi-objective optimization: A historical view of the field. *IEEE Comput. Intell. Mag.* **2006**, *1*, 28–36. [CrossRef]
24. Schuster, P. Optimization of multiple criteria: Pareto efficiency and fast heuristics should be more popular than they are. *Complexity* **2012**, *18*, 5–7. [CrossRef]
25. Seoane, L.F. Multiobjetive Optimization in Models of Synthetic and Natural Living Systems. PhD Thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2016.
26. Seoane, L.F.; Solé, R. A multiobjective optimization approach to statistical mechanics. *arXiv* **2013**, arXiv:1310.6372.
27. Seoane, L.F.; Solé, R. Phase transitions in Pareto optimal complex networks. *Phys. Rev. E* **2015**, *92*, 032807. [CrossRef]
28. Seoane, L.F.; Solé, R. Systems poised to criticality through Pareto selective forces. *arXiv* **2015**, arXiv:1510.08697.
29. Seoane, L.F.; Solé, R. Multiobjective optimization and phase transitions. In *Proceedings of ECCS*; Springer: Cham, Switzerland, 2014; pp. 259–270.
30. Wolfram, S. Universality and complexity in cellular automata. *Phys. D* **1984**, *10*, 1–35. [CrossRef]
31. Langton, C.G. Computation at the edge of chaos: Phase transitions and emergent computation. *Phys. D* **1990**, *42*, 12–37. [CrossRef]
32. Mitchell, M.; Hraber, P.; Crutchfield, J.P. Revisiting the edge of chaos: Evolving cellular automata to perform computations. *arXiv* **1993**, arXiv:adap-org/9303003.
33. Bak, P. *How Nature Works: The Science of Self-Organized Criticality*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1996.
34. Kauffman, S. *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*; Oxford University Press: Oxford, UK, 1996.
35. Legenstein, R.; Maass, W. What makes a dynamical system computationally powerful. In *New Directions in Statistical Signal Processing: From Systems to Brain*; MIT Press: Cambridge, MA, USA, 2007; pp.127–154.
36. Solé, R. *Phase Transitions*; Princeton U. Press.: Princeton, NJ, USA, 2011.
37. Mora, T.; Bialek, W. Are biological systems poised at criticality? *J. Stat. Phys.* **2011**, *144*, 268–302. [CrossRef]
38. Muñoz, M.A. Colloquium: Criticality and dynamical scaling in living systems. *Rev. Mod. Phys.* **2018**, *90*, 031001. [CrossRef]
39. Corpus of Contemporary American English. Available online: http://corpus.byu.edu/coca/ (accessed on 28 January 2020).
40. NLTK 3.4.5 documentation. Available online: http://www.nltk.org/ (accessed on 28 January 2020).
41. Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620. [CrossRef]
42. Jaynes, E.T. Information theory and statistical mechanics. II. *Phys. Rev.* **1957**, *108*, 171. [CrossRef]
43. Mora, T.; Walczak, A.M.; Bialek, W.; Callan, C.G. Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 5405–5410. [CrossRef]
44. Stephens, G.J.; Bialek, W. Statistical mechanics of letters in words. *Phys. Rev. E* **2010**, *81*, 066119. [CrossRef]
45. Harte, J. *Maximum Entropy and Ecology: A Theory of Abundance, Distribution, and Energetics*; Oxford University Press: Oxford, UK, 2011.
46. Tkačik, G.; Marre, O.; Mora, T.; Amodei, D.; Berry, II, M.J.; Bialek, W. The simplest maximum entropy model for collective behavior in a neural network. *J. Stat. Mech.* **2013**, *2013*, P03011. [CrossRef]
47. Stephens, G.J.; Mora, T.; Tkačik, G.; Bialek, W. Statistical thermodynamics of natural images. *Phys. Rev. Lett.* **2013**, *110*, 018701. [CrossRef]
48. Tkačik, G.; Mora, T.; Marre, O.; Amodei, D.; Palmer, S.E.; Berry, M.J.; Bialek, W. Thermodynamics and signatures of criticality in a network of neurons. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 11508–11513. [CrossRef]
49. Lee, E.D.; Broedersz, C.P.; Bialek, W. Statistical mechanics of the US Supreme Court. *J. Stat. Phys.* **2015**, *160*, 275–301. [CrossRef]

50. Sohl-Dickstein, J.; Battaglino, P.B.; DeWeese, M.R. New method for parameter estimation in probabilistic models: minimum probability flow. *Phys. Rev. Lett.* **2011**, *107*, 220601. [CrossRef]
51. Chomsky, N.; Chomsky, N. An interview on minimalism. In *On Nature and Language*; Cambridge University Press: Cambridge, UK, 2002; pp. 92–161.
52. Hauser, M.D.; Chomsky, N.; Fitch, W.T. The faculty of language: what is it, who has it, and how did it evolve? *Science* **2002**, *298*, 1569–1579. [CrossRef]
53. Berwick, R.C.; Chomsky, N. *Why only Us: Language and Evolution*; MIT Press: Cambridge, MA, USA, 2016.
54. Zipf, G.K., 1949. Human behavior and the principle of least effort. Available online: https://psycnet.apa.org/record/1950-00412-000 (accessed on 28 January 2020).
55. Altmann, E.G.; Gerlach, M. Statistical laws in linguistics. In *Creativity and Universality in Language*; Springer: Cham, Switzerland, 2016; pp. 7–26.
56. Ferrer, I.; Cancho, R.; Solé, R.V. Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 788–791. [CrossRef]
57. Corominas-Murtra, B.; Fortuny, J.; Solé, R.V. Emergence of Zipf's law in the evolution of communication. *Phys. Rev. E* **2011**, *83*, 036115. [CrossRef]
58. Corominas-Murtra, B.; Seoane, L.F.; Solé, R. Zipf's law, unbounded complexity and open-ended evolution. *J. R. Soc. Interface* **2018**, *15*, 20180395. [CrossRef]
59. Bickerton, D. *Language and Species*; University of Chicago Press: Chicago, IL, USA, 1992.
60. Deacon, T.W. *The Symbolic Species: The Co-Evolution of Language and the Brain*; WW Norton & Company: New York, NY, USA, 1998.
61. Crutchfield, J.P.; Young, K. Inferring statistical complexity. *Phys. Rev. Let.* **1989**, *63*, 105. [CrossRef]
62. Crutchfield, J.P. The calculi of emergence: computation, dynamics and induction. *Physica D* **1994**, *75*, 11–54. [CrossRef]
63. Crutchfield, J.P.; Shalizi, C.R. Thermodynamic depth of causal states: Objective complexity via minimal representations. *Phys. Rev. E* **1999**, *59*, 275. [CrossRef]

entropy

Article

A Standardized Project Gutenberg Corpus for Statistical Analysis of Natural Language and Quantitative Linguistics

Martin Gerlach [1] and Francesc Font-Clos [2,*]

[1] Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208, USA; martin.gerlach@northwestern.edu

[2] Center for Complexity and Biosystems, Department of Physics, University of Milan, 20133 Milano, Italy

* Correspondence: francesc.font@unimi.it

Received: 29 November 2019; Accepted: 16 January 2020; Published: 20 January 2020

Abstract: The use of Project Gutenberg (PG) as a text corpus has been extremely popular in statistical analysis of language for more than 25 years. However, in contrast to other major linguistic datasets of similar importance, no consensual full version of PG exists to date. In fact, most PG studies so far either consider only a small number of manually selected books, leading to potential biased subsets, or employ vastly different pre-processing strategies (often specified in insufficient details), raising concerns regarding the reproducibility of published results. In order to address these shortcomings, here we present the Standardized Project Gutenberg Corpus (SPGC), an open science approach to a curated version of the complete PG data containing more than 50,000 books and more than 3×10^9 word-tokens. Using different sources of annotated metadata, we not only provide a broad characterization of the content of PG, but also show different examples highlighting the potential of SPGC for investigating language variability across time, subjects, and authors. We publish our methodology in detail, the code to download and process the data, as well as the obtained corpus itself on three different levels of granularity (raw text, timeseries of word tokens, and counts of words). In this way, we provide a reproducible, pre-processed, full-size version of Project Gutenberg as a new scientific resource for corpus linguistics, natural language processing, and information retrieval.

Keywords: Project Gutenberg; Jensen–Shannon divergence; reproducibility; quantitative linguistics; natural language processing

1. Introduction

Analysis of natural language from a complex systems perspective has provided new insights into statistical properties of language, such as statistical laws [1–9], networks [10–14], language change [15–20], quantification of information content [21–24], or the role of syntactic structures [25] or punctuation [26], etc. In particular, the availability of new and large publicly available datasets such as the google-ngram data [27], the full Wikipedia dataset [28,29], or Twitter [30] opened the door for new large-scale quantitative approaches. One of the main drawbacks of these datasets, however, is the lack of "purity" of the samples resulting from the fact that (i) the composition of the dataset is largely unknown (google-ngram data, see [31,32]); (ii) the texts are a mixture of different authors (Wikipedia); or (iii) the texts are extremely short (Twitter). One approach to ensure large homogeneous samples of data is to analyze literary books – the most popular being from Project Gutenberg (PG) [33] due to their free availability.

Data from PG has been used in numerous cases to quantify statistical properties of natural language. In fact, the statistical analysis of texts in the framework of complex systems or quantitative linguistics is not conceivable without the books from PG. Already in the 1990s the seminal works by

Ebeling et al. [34] and Schurman and Grassberger [35] used up to 100 books from PG in the study of long-range correlations and Baayen [36] investigated the growth curve of the vocabulary. Subsequently, PG has become an indispensable resource for the quantitative analysis of language investigating, e.g., universal properties (such as correlations [37] or scale-free nature of the word-frequency distribution [38–40]) or aspects related to genres [41] or emotions [42]. While we acknowledge that PG has so far been of great use to the community, we also find that it has been handled in a careless and unsystematic way. Our criticisms can be summarized in two points. First, the majority of studies only consider a small subset (typically not more than 20 books) from the more than 50,000 books available in PG. More importantly, the subsets often contain the same manually selected books such as the work "Moby Dick" which can be found in virtually any study using PG data. Thus different works potentially employ biased and correlated subsets. While in some occasions a small set of well-selected books can be sufficient to answer particular linguistic questions, we believe that the enduring task of manually downloading and processing individual PG books has been the major bound to the number of books used in many PG studies. In the study of linguistic laws and its associated exponents, for instance, the variability they display [38,43] can be hard to grasp with reduced samples of PG. Even more clear is the case of double-power laws in Zipfs' law [44], which is only observed in very large samples and has been conjectured to be an artefact of text mixing [39]. Second, different studies use different filtering techniques to mine, parse, select, tokenize, and clean the data or do not describe the methodological steps in sufficient detail. As a result, two studies using the supposedly same PG data might end up with somewhat different datasets. Taken together, these limitations raise concerns about the replicability and generalizability of previous and future studies. In order to ensure the latter, it is pertinent to make corpora widely available in a standardized format. While this has been done for many textual datasets in machine learning (e.g., the UCI machine learning repository [45]) and diachronic corpora for studying language change (e.g., The Corpus of Contemporary American English [46]), such efforts have so far been absent for data from PG.

Here, we address these issues by presenting a standardized version of the complete Project Gutenberg data—the Standardized Project Gutenberg Corpus (SPGC)—containing more than 50,000 books and more than 3×10^9 word-tokens. We provide a framework to automatically download, filter, and process the raw data on three different levels of granularity: (i) the raw text; (ii) a filtered timeseries of word-tokens, and (iii) a list of occurrences of words. We further annotate each book with metadata about language, author (name and year of birth/death), and genre as provided by Project Gutenberg records as well as through collaborative tagging (so-called bookshelves), and show that the latter has more desirable properties such as low overlap between categories. We exemplify the potential of the SPGC by studying its variability in terms of Jensen–Shannon divergence across authors, time and genres. In contrast to standard corpora such as the British National Corpus [47] or the Corpus of Contemporary American English [46], the new Standardized Project Gutenberg Corpus is decentralized, dynamic and multi-lingual. The SPGC is decentralized in the sense that anyone can recreate it from scratch in their computer executing a simple python script. The SPGC is dynamic in the sense that, as new books are added to PG, the SPGC incorporates them immediately, and users can update their local copies with ease. This removes the classic centralized dependency problem, where a resource is initially generated by an individual or institution and initially maintained for certain period of time, after which the resource is no longer updated and remains "frozen" in the past. Finally, the SPGC is multi-lingual because it is not restricted to any language, it simply incorporates all content available in PG (see Section 2.3 for details). These characteristics are in contrast with some standard corpus linguistics practices [48,49], where a corpus is seen as a fixed collection of texts, stored in a way that allows for certain queries to be made, usually in an interactive way. While these corpora are well-suited –logically– for computational linguists interested in specific linguistic patterns, they are less useful for researchers of other disciplines such as physicists interested in long-range correlations in texts or computer scientists willing to train their language models on large corpora. In order to be compatible with a standard corpus model, and to ensure reproducibility of our results, we also provide a static time-stamped version of the corpus, SPGC-2018-07-18 (https://doi.org/10.5281/zenodo.2422560). In summary, we hope that the SPGC will lead to an increase in the availability and reliability of the PG data in the statistical and quantitative analysis of language.

2. Results

2.1. Data Acquisition

Project Gutenberg is a digital library founded in 1971 which archives cultural works uploaded by volunteers. The collection primarily consists of copyright-free literary works (books), currently more than 50,000, and is intended as a resource for readers to enjoy literary works that have entered the public domain. Thus the simplest way for a user to interact with PG is through its website, which provides a search interface, category listings, etc. to facilitate locating particular books of interest. Users can then read them online for free, or download them as plain text or ebook format. While such a manual strategy might suffice to download tens or hundreds of books (given the patience), it does not reasonably scale to the complete size of the PG data with more than 50,000 books.

Our approach consists of downloading the full PG data automatically through a local mirror, see Project Gutenberg's Information About Robot Access page for details. We keep most technical details "under the hood" and instead present a simple, well structured solution to acquire all of PG with a single command. In addition to the book's data, our pipeline automatically retrieves two different datasets containing annotations about PG books. The first set of metadata is provided by the person who uploads the book, and contains information about the author (name, year of birth, year of death), language of the text, subject categories, and number of downloads. The second set of metadata, the so-called bookshelves, provide a categorization of books into collections such as "Art" or "Fantasy", in analogy to the process of collaborative tagging [50].

2.2. Data Processing

In this section, we briefly describe all steps we took to obtain the corpus from the raw data (Figure 1), for details see Section 4. The processing (as of 18 July 2018) yields data for 55,905 books on four different levels of granularity:

- *Raw* data: We download all books and save them according to their PG-ID. We eliminate duplicated entries and entries not in UTF-8 encoding.
- *Text* data: We remove all headers and boiler plate text, see Methods for details.
- *Token* data: We tokenize the text data using the tokenizer from NLTK [51]. This yields a time series of tokens without punctuation, etc.
- *Count* data: We count the number of occurrences of each word-type. This yields a list of tuples (w,n_w), where w is word-type w and n_w is the number of occurrences.

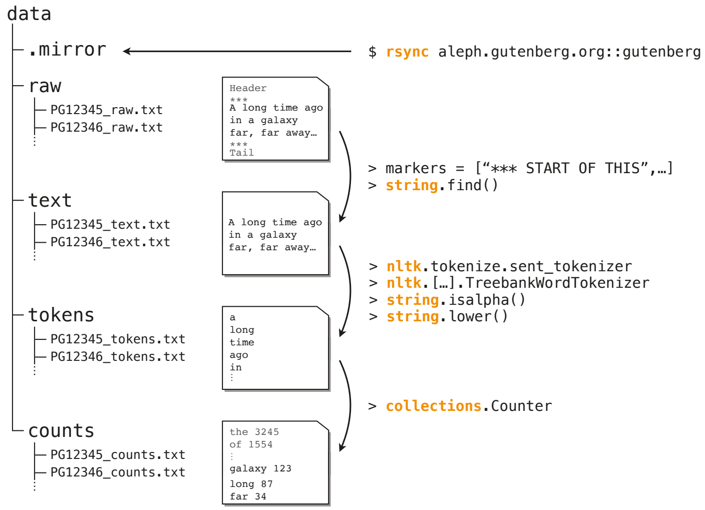


Figure 1. Sketch of the pre-processing pipeline of the Project Gutenberg (PG) data. The folder structure (**left**) organizes each PG book on four different levels of granularity, see example books (**middle**): raw, text, tokens, and counts. On the right we show the basic python commands used in the pre-processing.

2.3. Data Description

We provide a broad characterization of the PG books in terms of their length, language and (when available) inferred date of publication in Figure 2. One of the main reason for the popularity of books from PG is their long text length, which yields large coherent statistical samples without potentially introducing confounding factors originating from, e.g., the mixing of different texts [39]. The length of most PG books exceeds $m = 10^4$ word tokens (Figure 2a) which is larger than typical documents from most web resources. In fact, the distribution shows a heavy-tail for large values of m. Thus we find a substantial fraction of books having more than 10^5 word tokens. Many recent applications in quantitative linguistics aim at tracing diachronic changes. While the metadata does not provide the year of the first publication of each book, we approximate the number of PG books published in year t as the number of PG books for which the author's year of birth is $t_{\text{birth}} + 20 < t$ and the author's year of death is $t < t_{\text{death}}$ (Figure 2b). This reveals that the vast majority of books were first published around the year 1900, however, with a substantial number of books between 1800 and 2000. Part of this is known to be a consequence of the Copyright Term Extension Act of 1998 which, sadly, has prevented books published after 1923 to enter the public domain so far. If no further copyright extensions laws are passed in the future, then this situation will be gradually alleviated year after year, as books published in 1923 will enter the public domain on 1 January 2019, and so on.

While most contemporary textual datasets are in English, the SPGC provides a rich resource to study other languages. Using metadata provided by PG, we find that 81% of the books are tagged as written in English, followed by French (5%, 2864 books), Finnish (3.3%, 1903 books) and German (2.8%, 1644 books). In total, we find books written in 56 different languages, with three (13) languages besides English with more than 1000 (100) books each (Figure 2c). The size of the English corpus is 2.8×10^9 tokens, which is more than one order of magnitude larger than the British National Corpus (10^8 tokens). The second-largest language corpus is made up of French books with $> 10^8$ tokens. Notably, there are six other languages (Finnish, German, Dutch, Italian, Spanish, and Portuguese) that contain $> 10^7$ tokens and still another eight languages (Greek, Swedish, Hungarian, Esperanto, Latin, Danish, Tagalog, and Catalan) that contain $> 10^6$ tokens.

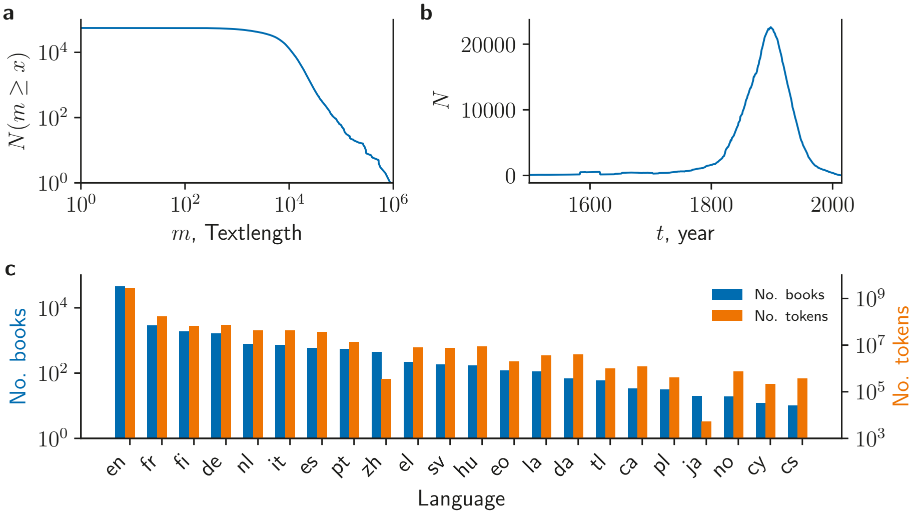


Figure 2. Basic summary statistics from the processed PG data. (**a**) Number of books with a text length larger than m; (**b**) Number of books which are compatible with being published in year t, i.e., year of author's birth is 20 years prior and year of author's death is after t; (**c**) Number of books (left axis) and number of tokens (right axis) which are assigned to a given language based on the metadata. en: English, fr: French, fi: Finnish, de: German, nl: Dutch, it: Italian, es: Spanish, pt: Portuguese, zh: Chinese, el: Greek, sv: Swedish, hu: Hungarian, eo: Esperanto, la: Latin, da: Danish, tl: Tagalog, ca: Catalan, pl: Polish, ja: Japanese, no: Norwegian, cy: Welsh, cs: Czech.

In addition to the "hard-facts" metadata (such as language, time of publication), the SPGC also contains manually annotated topical labels for individual books. These labels allow not only the study of topical variability, but they are also of practical importance for assessing the quality of machine learning applications in Information Retrieval, such as text classification or topic modeling [52]. We consider two sets of topical labels: labels obtained from PG's metadata "subject" field, which we call *subject labels*; and labels obtained by parsing PG's website bookshelf pages, which we call *bookshelf labels*. Table 1 shows that there is certain overlap in the most common labels between the two sets (e.g., Science Fiction or Historical Fiction), but a more detailed analysis of how labels are assigned to books reveals substantial differences (Figure 3). First, subject labels display a very uneven distribution of the number of books per label. That is, most of the subject labels are assigned to very few books (less than 10), with only few subject labels assigned to many books. In comparison, bookshelf labels are more evenly distributed: most of them are assigned to between 10 and 100 books (Figure 3a,c). More importantly, the overlap in the assignment of labels to individual books is much smaller for the bookshelf labels (Figure 3b,d): While roughly 50% of the PG books are tagged with two or more subject labels, up to 85% of books are tagged with a unique bookshelf label. This indicates that the bookshelf labels are more informative because they constitute broader categories and provide a unique assignment of labels to books, and are thus better suited for practical applications such as text classification.

Table 1. Examples for the names of labels and the number of assigned books from bookshelves (left) and subjects (right) metadata.

Rank	Books	Bookshelf	Rank	Books	Subject
1	1341	Science Fiction	1	2006	Fiction
2	509	Children's Book Series	2	1823	Short stories
3	493	Punch	3	1647	Science fiction
4	426	Bestsellers, American, 1895-1923	3	1647	Science fiction
5	383	Historical Fiction	5	746	Historical fiction
6	374	World War I	6	708	Love stories
7	339	Children's Fiction	7	690	Poetry
...	...	...	...	...	...
47	94	Slavery	47	190	Short stories, American
48	92	Western	48	188	Science – Periodicals
49	90	Judaism	49	183	American poetry
50	86	Scientific American	50	180	Drama
51	84	Pirates, Buccaneers, Corsairs, etc.	51	165	Paris (France) – Fiction
52	83	Astounding Stories	52	163	Fantasy literature
53	83	Harper's Young People	53	162	Orphans – Fiction
...	...	...	...	...	...
97	37	Animals-Wild-Reptiles and Amphibians	97	100	Scotland – Periodicals
98	37	Short Stories	98	98	Horror tales
99	36	Continental Monthly	99	97	Canada – Fiction
100	35	Architecture	100	97	France – Court and courtiers
101	35	Bahá'í Faith	101	96	Social classes – Fiction
102	34	Precursors of Science Fiction	102	95	Courtship – Fiction
103	33	Physics	103	95	Seafaring life – Juvenile fiction
...	...	...	...	...	...

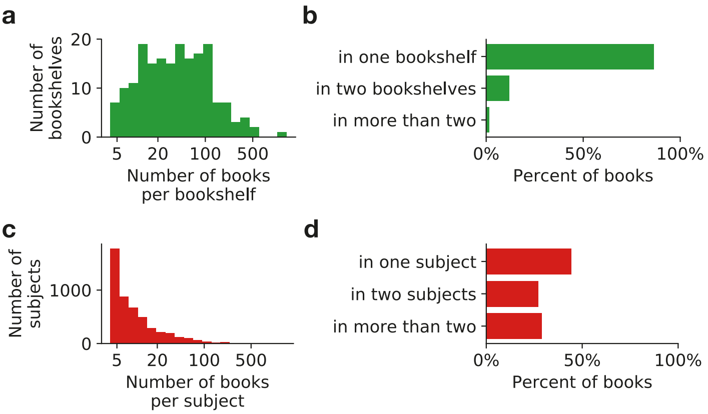


Figure 3. Comparison between bookshelf labels (top, green) and subject labels (bottom, red). (**a**,**c**) Number of labels with a given number of books; (**b**,**d**) Fraction of books with a given number of labels.

2.4. Quantifying Variability in the Corpus

In order to highlight the potential of the SPGC for quantitative analysis of language, we quantify the degree of variability in the statistics of word frequencies across labels, authors, and time. For this, we measure the distance between books i and j using the well-known Jensen–Shannon divergence [53], $D_{i,j}$, with $D_{i,j} = 0$ if the two books are exactly equal in terms of frequencies, and $D_{i,j} = 1$ if they are maximally different, i.e., they do not have a single word in common, see Methods for details.

2.4.1. Labels

We select the 370 books tagged with one of the following bookshelf labels: Art, Biographies, Fantasy, Philosophy and Poetry. After calculating distances $D_{i,j}$ between all pairs of books, in Figure 4 we show an approximate 2-dimensional embedding (Uniform Manifold Approximation and Projection (UMAP), see [54] and Methods for details) in order to visualize which books are more similar to each other. Indeed, we find that books from the same bookshelf tend to cluster together and are well-separated from books belonging to other bookshelves. This example demonstrates the usefulness of the bookshelf labels and that they reflect the topical variability encoded in the statistics of word frequencies.

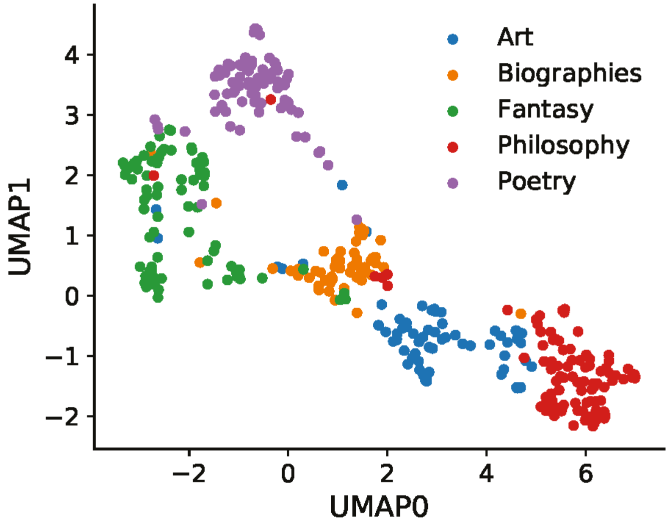


Figure 4. 2-dimensional embedding shows clustering of books from the same bookshelf. Approximate visualization of the pair-wise distances between 370 PG books using Uniform Manifold Approximation and Projection (UMAP) (for details, see Section 4). Each dot corresponds to one book colored according to the bookshelf membership.

2.4.2. Authors

We select all books from the 20 most prolific authors (selected from the authors of the 100 most downloaded books in order to avoid authors such as "Anonymous"). For each author, we draw 1000 pairs of books (i, j) from the same author and compare the distance $D_{i,j}$ with 1000 pairs (i, j') where j' comes from a different author. We observe that the distance between books from the same author is consistently smaller than for two books from different authors – not only in terms of the median, but also in terms of a much smaller spread in the values of $D_{i,j}$ (Figure 5). This consistent variability across authors suggest the potential applicability in the study of stylistic differences, such as in problems of authorship attribution [55,56].

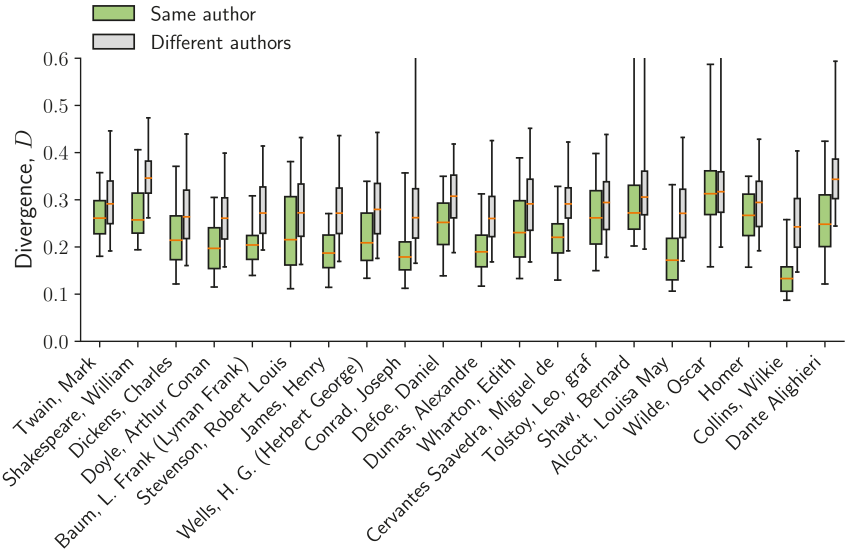


Figure 5. Distance between books from the same author is significantly smaller than distance between books from different authors. For each author, the boxplots shows the $5, 25, 50, 75, 95$-percentile of the distribution of distances from 1000 pairs of books from the same author (**green**) and to a different author (**gray**).

2.4.3. Time

We compare the distance $D_{i,j}$ between pairs of books i, j taken each from a 20-year time period $t_i, t_j \in \{1800 - 1820, \ldots, 1980 - 2000\}$. In Figure 6, we show the distance between two time windows D_{t_i,t_j} by averaging over each 1000 pairs of books. We observe that the average distance increases with increasing separation between the time periods. However, we emphasize that we only observe a substantial increase in D_{t_i,t_j} for large separation between t_i and t_j and later time periods (after 1900). This could be caused by the rough approximation of the publication year and a potential change in composition of the SPGC after 1930 due to copyright laws. In fact, the observed effects are likely a mixture of temporal and topical variability, because the topical composition of PG over time is certainly not uniform. This suggests the limited applicability of PG books for diachronic studies without further filtering (such as subject/bookshelf). Other resources such as the COHA corpus might be more adequate in this case, although potentially a more genre-balanced version of SPGC could be created using the provided metadata.

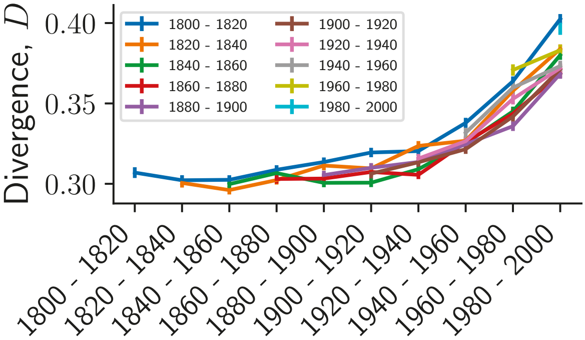


Figure 6. Distance between books increases with their time separation. Average and standard error of the distance between 1000 pairs of books, where the two books in each pair is drawn from two different 20-year time intervals. We fix the first interval and continuously increase the second time interval.

3. Discussion

We have presented the Standardized Project Gutenberg Corpus (SPGC), a decentralized, dynamic multilingual corpus containing more than 50,000 books from more than 20 languages. Combining the textual data with metadata from two different sources we provided not only a characterization of the content of the full PG data but also showed three examples for resolving language variability across subject categories, authors, and time. As part of this work, we provide the code for all pre-processing steps necessary to obtain a full local copy of the PG data. We also provide a static or 'frozen' version of the corpus, SPGC-2018-07-18, which ensures reproducibility of our results and can be downloaded at https://doi.org/10.5281/zenodo.2422560.

We believe that the SPGC will be a first step towards a more rigorous approach for using Project Gutenberg as a scientific resource. A detailed account of each step in the pre-processing, accompanied by the corresponding code, are necessary requirements that will help ensure replicability in the statistical analysis of language and quantitative linguistics, especially in view of the crisis in reproducibility and replicability reported in other fields [57–59]. From a practical point of view, the availability of this resource in terms of the code and the frozen dataset will certainly allow for an easier access to PG data, in turn facilitating the usage of larger and less biased datasets increasing the statistical power of future analysis.

We want to highlight the challenges of the SPGC in particular and PG in general, some of which can hopefully be addressed in the future. First, the PG data only contains copyright-free books. As a result the number of books published after 1930s is comparably small. However, in the future this can be expected to change as copyright for many books will run out and the PG data is continuously growing. This highlights the importance of using a dynamic corpus model that will by default incorporate all new books when the corpus is generated for the first time. Second, the annotation about the books is incomplete, and some books might be duplicated. For example, the metadata lacks the exact date when a book was published, hindering the usage of the PG data for diachronic studies. Different editions of the same book might have been given a different PG identifier, and so they are all included in PG and thus in SPGC. Third, the composition of SPGC is heterogeneous, mixing different genres. However, the availability of document labels from the bookshelf metadata allows for systematic control of corpus composition. For example, it is easy to restrict to or exclude individual genres such as "Poetry".

From a practical perspective, the SPGC has a strong potential to become a complementary resource in applications ranging from computational linguistics to machine learning. A clear limitation of SPGC is that it was designed to fit a wide range use cases, and so the pre-processing and data-selection choices are sub-optimal in many specific cases. However, the modular design of the code allows for researches to modify such choices with ease, and data can always be filtered a posteriori, but not the other way around. Choices are unavoidable, but it is only by providing the full code and data that these choices can later be tailored to specific needs. Overall, we believe the benefits of a standardized version of PG out-weight its potential limitations.

We emphasize that the SPGC contains thousands of annotated books in multiple languages even beyond the Indo-European language family. There is an increasing interest in quantitative linguistics in studies beyond the English language. In the framework of culturomics, texts could be annotated and weighted by additional metadata, e.g., in terms of their 'success' measure as the number of readers [60] or number of PG downloads. For example, it could be expected that the impact of Carroll's "Alice in Wonderland" is larger than that of the "CIA Factbook 1990". Furthermore, with an increase in the quality of the metadata, the identification of the same book in different languages might allow for the construction of high-quality parallel corpora used in, e.g., translation tasks. Finally, in applications of Information Retrieval, metadata labels can be used to evaluate machine learning algorithms for classification and prediction. These and other applications might require additional pre-processing steps (such as stemming) but which could make use of SPGC as a starting point.

In summary, we believe that the SPGC is a first step towards a better usage of PG in scientific studies, and hope that its decentralized, dynamic and multi-lingual nature will lead to further collaborative interdisciplinary approaches to quantitative linguistics.

4. Materials and Methods

4.1. Running the Code

The simplest way to get a local copy of the PG database, with standardized, homogeneous pre-processing, is to clone the git repository

```
$ git clone git@github.com:pgcorpus/gutenberg.git
```

and enter the newly created directory. To get the data, simply run:

```
$ python get_data.py
```

This will download all available PG books in a hidden '.mirror' folder and symlink in the more convenient 'data/raw' folder. To actually process the data, that is, to remove boiler-plate text, tokenize texts, filter and lowercase tokens, and count word type occurrence, it suffices to run

```
$ python process_data.py
```

which will fill in the rest of directories inside 'data' . We use 'rsync' to keep an updated local mirror of aleph.gutenberg.org::gutenberg. Some PG book identifiers are stored in more than one location in PG's server. In these cases, we only keep the latest, most up-to-date version. We do not remove duplicated entries on the basis of book metadata or content. To eliminate boiler-plate text that does not pertain to the books themselves, we use a list of known markers (code adapted from https://github.com/c-w/gutenberg/blob/master/gutenberg/cleanup/strip_headers.py).

4.2. Preprocessing

Texts are tokenized via NLPToolkit [51]. In particular, we set the 'TreebankWordTokenizer' as the default choice, but this can be changed at will. Tokenization works better when the language of the text being analyzed is known. Since the metadata contains a language field for every downloaded book, we pass this information onto the tokenizer. If the language field contains multiple languages (≈ 0.3% of the books), we use the first entry. We only keep tokens composed entirely of alphabetic characters (including accented characters), removing those that contain digits or other symbols. Notice that this is done after tokenization, which correctly handles apostrophes, hyphens, etc. This constitutes a conservative approach to avoid artifacts, for example originating from the digitization process. While one might want to also include tokens with numeric characters in order to keep mentions of, e.g., years, the downside of this approach would be a substantial number of occurrences of undesirable page and chapter numbers. However, we note that the modularized filtering can be easily customized (and extended) to incorporate also other aspects such as stemming as there is no one-size-fits all solution to each individual application. Furthermore, all tokens are lower-cased. While this has undesired consequences in some cases (e.g., some proper nouns can be confounded with unrelated common nouns after being lower-cased), it is a simple and effective way of handling words capitalized after full stop or in dialogues, which would otherwise be (incorrectly) considered different words from their lowercase standard form. We acknowledge that our preprocessing choices might not fit well specific use cases, but they have been designed to favour precision over recall. For instance, we find it preferable to miss mentions of years while ensuring that no page or chapter numbers are included in the final dataset. Notice that the alternative, that of building and additional piece of software that correctly distinguishes these two cases, is in itself complicated, has to handle several edge cases, and ultimately incurs in additional choices.

4.3. Jensen–Shannon Divergence

We use Jensen–Shannon divergence (JSD, [53,61,62]) as a divergence measure between PG books. JSD is an information-theory based divergence measure which, given two symbolic sequences with frequency distributions encoded in vectors p, q can be simply defined as

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

where $H(p)$ denotes the standard Shannon entropy, $H(p) = -\sum_i p_i \log p_i$. In simple and general terms, JSD quantifies how similar two symbolic sequences are on the basis of how frequent or infrequent each symbol is in the two sequences. In our case, this translates to measuring the similarity between two books via the frequencies of their word types. The logarithmic term in the expression of the entropy $H(p)$, however, ensures that the measure is not dominated by high-frequency words, as would happen otherwise, but instead is dependent on differences in frequency along the whole spectrum of usage, from very common to very uncommon words. Therefore, JSD is specially suitable for symbolic sequences that display long tails in the distribution of frequencies, as is the case in natural language. Notice that the distance between one book and a randomly shuffled version of it is exactly 0, from the JSD point of view. This drawback can be alleviated by using JSD on n-grams counts of higher order, that is, taking into account the frequency of pairs of words or bigrams, and so on. However, we do not take this route here since it has the undesirable consequence of exponentially increasing the number of features. For a more technical discussion about JSD and related measures, see [53].

4.4. 2-Dimensional Embedding

We use *Uniform Manifold Approximation and Projection* (UMAP, [54]) for visualization purposes in Figure 4. UMAP is a manifold-learning technique with strong mathematical foundations that aims to preserve both local and global topological structures, see [54] for details. In simple terms, UMAP and other manifold-learning algorithms aim at finding a good low-dimensional representation of a high-dimensional dataset. To do so, UMAP finds the best manifold that preserves topological structures of the data at different scales, again see [54] for details. We used normalized counts data as the input data, with word types playing the role of dimensions (features) and books playing the role of points (samples). Distance between points was computed using the Jensen–Shannon divergence [53]. The end result is the 2-dimensional projection shown in Figure 4. Notice that subject labels were not passed to UMAP, so the observed clustering demonstrates that the statistics of word frequencies encode and reflect the manually-assigned labels.

4.5. Data Availability

The code that we make available as part of this work allows to download and process all available Project Gutenberg books, facilitating the task of keeping an up-to-date and homogeneously processed dataset of a continuously growing resource. In fact, new books are added to Project Gutenberg daily. An unwanted consequence of this feature, however, is that two local versions of the SPGC might differ if they were last updated on different dates. To facilitate and promote reproducibility of our results and possible subsequent analysis, we provide a 'frozen' copy of the SPGC, last updated on 18 July, 2018, containing 55,905 PG books. All statistics and figures reported on this manuscript are based on this version of the data. This data is available at https://doi.org/10.5281/zenodo.2422560.

4.6. Computational Requirements

The 'frozen' dataset of all 55,905 books and all levels of granularity has a size of 65 GB. However, focusing only on the one-gram counts requires only 3.6 GB. Running the pre-processing pipeline of the 'frozen' data took 8 h (without parallelization) on an CPU running at 3.40 GHz.

4.7. Code Availability

Python 3.6 code to download and pre-process all PG books can be obtained at https://github.com/pgcorpus/gutenberg, while python-based jupyter notebooks that reproduce the results of this manuscript can be obtained at https://github.com/pgcorpus/gutenberg-analysis.

Author Contributions: Conceptualization, M.G. and F.F.-C.; Software, M.G. and F.F.-C.; Data Curation, M.G. and F.F.-C.; Writing—Original Draft Preparation, M.G. and F.F.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. F.F.-C. paid the APC.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Altmann, E.G.; Gerlach, M. Statistical laws in linguistics. In *Creativity and Universality in Language*; Degli Esposti, M., Altmann, E.G., Pachet, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2016; pp. 7–26.
2. Ferrer i Cancho, R.; Solé, R.V. Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 788–791. [CrossRef] [PubMed]
3. Petersen, A.M.; Tenenbaum, J.N.; Havlin, S.; Stanley, H.E.; Perc, M. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Sci. Rep.* **2012**, *2*, 943. [CrossRef] [PubMed]
4. Tria, F.; Loreto, V.; Servedio, V.D.P.; Strogatz, S.H. The dynamics of correlated novelties. *Sci. Rep.* **2014**, *4*, 5890. [CrossRef] [PubMed]
5. Corominas-Murtra, B.; Hanel, R.; Thurner, S. Understanding scaling through history-dependent processes with collapsing sample space. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 5348–5353. [CrossRef] [PubMed]
6. Font-Clos, F.; Corral, A. Log-log convexity of type-token growth in Zipf's systems. *Phys. Rev. Lett.* **2015**, *114*, 238701. [CrossRef] [PubMed]
7. Cocho, G.; Flores, J.; Gershenson, C.; Pineda, C.; Sánchez, S. Rank Diversity of Languages: Generic Behavior in Computational Linguistics. *PLoS ONE* **2015**, *10*, e0121898. [CrossRef] [PubMed]
8. Lippi, M.; Montemurro, M.A.; Degli Esposti, M.; Cristadoro, G. Natural Language Statistical Features of LSTM-Generated Texts. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3326–3337. [CrossRef]
9. Mazzolini, A.; Gherardi, M.; Caselle, M.; Cosentino Lagomarsino, M.; Osella, M. Statistics of shared components in complex component systems. *Phys. Rev. X* **2018**, *8*, 021023. [CrossRef]
10. Dorogovtsev, S.N.; Mendes, J.F. Language as an evolving word web. *Proc. R. Soc. B* **2001**, *268*, 2603–2606. [CrossRef]
11. Solé, R.V.; Corominas-Murtra, B.; Valverde, S.; Steels, L. Language networks: Their structure, function, and evolution. *Complexity* **2010**, *15*, 20–26. [CrossRef]
12. Amancio, D.R.; Altmann, E.G.; Oliveira, O.N.; Costa, L.D.F. Comparing intermittency and network measurements of words and their dependence on authorship. *New J. Phys.* **2011**, *13*, 123024. [CrossRef]

13. Choudhury, M.; Chatterjee, D.; Mukherjee, A. Global topology of word co-occurrence networks: Beyond the two-regime power-law. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, China, 23–27 August 2010; pp. 162–170.
14. Cong, J.; Liu, H. Approaching human language with complex networks. *Phys. Life Rev.* **2014**, *11*, 598–618. [CrossRef] [PubMed]
15. Bochkarev, V.; Solovyev, V.; Wichmann, S. Universals versus historical contingencies in lexical evolution. *J. R. Soc. Interface* **2014**, *11*, 20140841. [CrossRef] [PubMed]
16. Ghanbarnejad, F.; Gerlach, M.; Miotto, J.M.; Altmann, E.G. Extracting information from S-curves of language change. *J. R. Soc. Interface* **2014**, *11*, 20141044. [CrossRef]
17. Feltgen, Q.; Fagard, B.; Nadal, J.P. Frequency patterns of semantic change: Corpus-based evidence of a near-critical dynamics in language change. *R. Soc. Open Sci.* **2017**, *4*, 170830. [CrossRef] [PubMed]
18. Gonçalves, B.; Loureiro-Porto, L.; Ramasco, J.J.; Sánchez, D. Mapping the Americanization of English in space and time. *PLoS ONE* **2018**, *13*, e0197741. [CrossRef]
19. Amato, R.; Lacasa, L.; Díaz-Guilera, A.; Baronchelli, A. The dynamics of norm change in the cultural evolution of language. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 8260–8265. [CrossRef]
20. Karjus, A.; Blythe, R.A.; Kirby, S.; Smith, K. Challenges in detecting evolutionary forces in language change using diachronic corpora. *arXiv* **2018**, arXiv:1811.01275.
21. Montemurro, M.A.; Zanette, D.H. Towards the quantification of the semantic information encoded in written language. *Adv. Complex Syst.* **2010**, *13*, 135. [CrossRef]
22. Takahira, R.; Tanaka-Ishii, K.; Debowski, Ł. Entropy rate estimates for natural language—A new extrapolation of compressed large-scale corpora. *Entropy* **2016**, *18*, 364. [CrossRef]
23. Febres, G.; Jaffé, K. Quantifying structure differences in literature using symbolic diversity and entropy criteria. *J. Quant. Linguist.* **2017**, *24*, 16–53. [CrossRef]
24. Bentz, C.; Alikaniotis, D.; Cysouw, M.; Ferrer-i Cancho, R. The entropy of words—Learnability and expressivity across more than 1000 languages. *Entropy* **2017**, *19*, 275. [CrossRef]
25. Ferrer i Cancho, R.; Solé, R.; Köhler, R. Patterns in syntactic dependency networks. *Phys. Rev. E* **2004**, *69*, 51915. [CrossRef] [PubMed]
26. Kulig, A.; Kwapień, J.; Stanisz, T.; Drożdż, S. In narrative texts punctuation marks obey the same statistics as words. *Inf. Sci.* **2016**, *375*, 98–113. [CrossRef]
27. Michel, J.B.; Shen, Y.K.; Aiden, A.P.; Veres, A.; Gray, M.K.; Team, T.G.B.; Pickett, J.P.; Hoiberg, D.; Clancy, D.; Norvig, P.; et al. Quantitative analysis of culture using millions of digitized books. *Science* **2011**, *331*, 176–182. [CrossRef]
28. Masucci, A.P.; Kalampokis, A.; Eguíluz, V.M.; Hernández-García, E. Wikipedia information flow analysis reveals the scale-free architecture of the semantic space. *PLoS ONE* **2011**, *6*, e17333. [CrossRef]
29. Yasseri, T.; Kornai, A.; Kertész, J. A practical approach to language complexity: A Wikipedia case study. *PLoS ONE* **2012**, *7*, e48386. [CrossRef]
30. Dodds, P.S.; Harris, K.D.; Kloumann, I.M.; Bliss, C.A.; Danforth, C.M. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLoS ONE* **2011**, *6*, e26752. [CrossRef]
31. Morse-Gagné, E.E. Culturomics: Statistical traps muddy the data. *Science* **2011**, *332*, 35. [CrossRef]
32. Pechenick, E.A.; Danforth, C.M.; Dodds, P.S. Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE* **2015**, *10*, e0137041. [CrossRef]
33. Hart, M. Project Gutenberg. 1971. Available online: https://www.gutenberg.org (accessed on 18 July 2018)
34. Ebeling, W.; Pöschel, T. Entropy and long-range correlations in literary English. *Europhys. Lett.* **1994**, *26*, 241–246. [CrossRef]
35. Schurmann, T.; Grassberger, P. Entropy estimation of symbol sequences. *Chaos* **1996**, *6*, 414–427. [CrossRef] [PubMed]
36. Baayen, R.H.H. The effects of lexical specialization on the growth curve of the vocabulary. *Comput. Linguist.* **1996**, *22*, 455–480.
37. Altmann, E.G.; Cristadoro, G.; Esposti, M.D. On the origin of long-range correlations in texts. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 11582–11587. [CrossRef]
38. Moreno-Sánchez, I.; Font-Clos, F.; Corral, Á. Large-scale analysis of Zipf's law in English texts. *PLoS ONE* **2016**, *11*, e0147073. [CrossRef]

39. Williams, J.R.; Bagrow, J.P.; Danforth, C.M.; Dodds, P.S. Text mixing shapes the anatomy of rank-frequency distributions. *Phys. Rev. E* **2015**, *91*, 052811. [CrossRef]
40. Tria, F.; Loreto, V.; Servedio, V. Zipf's, Heaps' and Taylor's Laws are determined by the expansion into the adjacent possible. *Entropy* **2018**, *20*, 752. [CrossRef]
41. Hughes, J.M.; Foti, N.J.; Krakauer, D.C.; Rockmore, D.N. Quantitative patterns of stylistic influence in the evolution of literature. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 7682–7686. [CrossRef]
42. Reagan, A.J.; Mitchell, L.; Kiley, D.; Danforth, C.M.; Dodds, P.S. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Sci.* **2016**, *5*, 31. [CrossRef]
43. Ferrer i Cancho, R. The variation of Zipf's law in human language. *Eur. Phys. J. B - Condens. Matter Complex Syst.* **2005**, *44*, 249–257. [CrossRef]
44. Ferrer i Cancho, R.; Solé, R.V. Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited. *J. Quant. Linguist.* **2001**, *8*, 165–173. [CrossRef]
45. Dheeru, D.; Karra Taniskidou, E. UCI Machine Learning Repository, 2017. Available online: https://archive.ics.uci.edu/ml/index.php (accessed on 18 July 2018)
46. Davies, M. The Corpus of Contemporary American English (COCA): 560 Million Words, 1990-Present. 2008. Available online: https://www.english-corpora.org/coca/ (accessed on 18 July 2018)
47. Leech, G. 100 million words of English. *Engl. Today* **1993**, *9*, 9–15. [CrossRef]
48. Biber, D.; Reppen, R. *The Cambridge Handbook of English Corpus Linguistics*; Cambridge University Press: Cambridge, UK, 2015.
49. Jones, C.; Waller, D. *Corpus Linguistics for Grammar: A guide for research*; Routledge: Abingdon, UK, 2015.
50. Cattuto, C.; Loreto, V.; Pietronero, L. Semiotic dynamics and collaborative tagging. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 1461–1464. [CrossRef] [PubMed]
51. Loper, E.; Bird, S. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; Volume 1, pp. 63–70, ETMTNLP '02.
52. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008.
53. Gerlach, M.; Font-Clos, F.; Altmann, E.G. Similarity of symbol frequency distributions with heavy tails. *Phys. Rev. X* **2016**, *6*, 021009. [CrossRef]
54. McInnes, L.; Healy, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv* **2018**, arXiv:1802.03426.
55. Juola, P. Authorship attribution. *Found. Trends® Inf. Retr.* **2008**, *1*, 233–334. [CrossRef]
56. Stamatatos, E. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 538–556. [CrossRef]
57. Ioannidis, J.P.A. Why most published research findings are false. *PLoS Med.* **2005**, *2*, e124. [CrossRef]
58. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **2015**, *349*, aac4716. [CrossRef]
59. Camerer, C.F.; Dreber, A.; Holzmeister, F.; Ho, T.H.; Huber, J.; Johannesson, M.; Kirchler, M.; Nave, G.; Nosek, B.A.; Pfeiffer, T.; et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.* **2018**, *2*, 637–644. [CrossRef]
60. Yucesoy, B.; Wang, X.; Huang, J.; Barabási, A.L. Success in books: A big data approach to bestsellers. *EPJ Data Sci.* **2018**, *7*, 7. [CrossRef]
61. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
62. Grosse, I.; Bernaola-Galván, P.; Carpena, P.; Román-Roldán, R.; Oliver, J.; Stanley, H.E. Analysis of symbolic sequences using the Jensen–Shannon divergence. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **2002**, *65*, 041905. [CrossRef] [PubMed]

entropy

Article

How the Probabilistic Structure of Grammatical Context Shapes Speech

Maja Linke * and Michael Ramscar

Department of Linguistics, University of Tuebingen, Wilhelmstraße 19, 72074 Tuebingen, Germany; michael.ramscar@uni-tuebingen.com
* Correspondence: maja.linke@uni-tuebingen.com

Received: 31 October 2019; Accepted: 7 January 2020; Published: 11 January 2020

Abstract: Does systematic covariation in the usage patterns of forms shape the sublexical variance observed in conversational speech? We address this question in terms of a recently proposed discriminative theory of human communication that argues that the distribution of events in communicative contexts should maintain mutual predictability between language users, present evidence that the distributions of words in the empirical contexts in which they are learned and used are geometric, and thus support this. Here, we extend this analysis to a corpus of conversational English, showing that the distribution of grammatical regularities and the sub-distributions of tokens discriminated by them are also geometric. Further analyses reveal a range of structural differences in the distribution of types in parts of speech categories that further support the suggestion that linguistic distributions (and codes) are subcategorized by context at multiple levels of abstraction. Finally, a series of analyses of the variation in spoken language reveals that quantifiable differences in the structure of lexical subcategories appears in turn to systematically shape sublexical variation in speech signal.

Keywords: speech variance; communicative efficiency; sampling invariance; power laws; communicative distributions

1. Introduction

The words produced in conversational speech often differ substantially from the acoustic signals supposed by canonical dictionary forms [1,2]. The extent to which articulated forms deviate from dictionary models correlates with average word frequency, such that there is a general tendency for shorter and faster articulation in more probable words. This property of speech codes is often taken to suggest that human speech is shaped by the competing requirements of maximizing the success of message transmission while minimizing production effort in ways similar to those described by information coding solutions for electronic message transmission. There are, however, some critical differences between speech and the communication model described by information theory [3]: whereas information theory is concerned with defining the properties of variable length codes optimized for efficient communication in discrete, memoryless systems, human communication codes, at first blush at least, appear neither systematic [3] nor systematically discrete [2,4] or memoryless [5].

In regard to the first point, systematicity, humans learn to communicate by the gradual discrimination of functional (task-relevant) speech dimensions from the samples to which they are exposed, yet because lexical diversity in language samples increases nonlinearly over space and time, the divergence between the samples individuals are exposed to increases as their experience of the linguistic environment grows [5]. A system defined by a probabilistic structure would appear to require that events be distributed in a way that allows the relationships between events probabilities to remain stable independent of the sample size, yet the way that words are distributed across language samples suggests that human languages do not satisfy this requirement.

Considering the second point discreteness, although writing conventions lead to some systematic agreements about what linguistic units are such that words are often thought of as standard discrete linguistic units, speech appears to be different. Human intuition on boundaries in speech diverge as exposure increases. When literate adults, nonliterate adults, and children are asked to divide a speech sequence into units, their intuitions on where any given sequence should be split into multiple units exhibit a systematic lack of agreement [6]; similar effects have been observed when people are asked to discriminate phonetic contrasts [7].

As for memorylessness, which supposes a distribution of events such that an event's probability is independent of the way it is sampled, it has been shown that increased exposure to language leads to a decrease in the informativeness of high-frequency tokens relative to words that they co-occur with such that the informativity relationships between words appear to be unstable across cohorts [5]. For instance, the information that blue provides changes systematically as people successively hear about blue skies, blue eyes, and blue berries, etc. at different rates, an effect that increases nonlinearly with the number of blue covariates that speakers encounter.

To summarize these points, it is clear that adult expectations about events and their probabilities vary with experience. This in turn seems to suggest that the increasing divergence between individual speakers' models will lead to an increase in communication problems between speakers. Nevertheless, sufficiently successful communication between speakers of different experience levels is not only possible but also relatively common. How?

Recent work by Ramscar [3] addresses these apparent communication problems from the perspective of discriminative learning and suggests that, unlike the predefined source codes in artificial communication, human communicative codes are subcategorized by systematic patterns of variation in the way words and arguments are employed. The empirical distributions discriminated by these patterns of variation both serve to minimize communicative uncertainty and to make unattested word forms predictable in context, thereby overcoming some of the problems that arise from the way that linguistic codes are sampled. In support of this argument, Ramscar presents evidence that the empirical distributions shaped by communicative contexts are geometric and suggests that the power laws that commonly characterize word token distributions are not in themselves functional but rather result from the aggregation of multiple functionally distinct communicative distributions [8]. Importantly, unlike power laws, the geometric distribution is sampling invariant and thus directly satisfies many of the constraints defined by information theory [9,10]. Perhaps even more importantly, geometric distributions also appear to maximize the likelihood that, independent of exposure, learning will lead to speakers acquiring similar models of distribution of communicative contrasts in context, thereby enabling a high degree of mutual predictability and helping to explain why human communicative codes actually work as well as they do.

A notable finding in this regard comes from an analysis of names (a universal feature of communicative codes that is almost universally ignored by grammatical theories) and in particular the distributions of English and Sinosphere first names [3,11]. This analysis shows that, historically, prior to the imposition of name laws that distorted their initial distributions, first names across a range of cultures had near identical geometric distributions. Names are a unique aspect of language in that their composition is highly regulated in virtually all modern nation states. Functionally, name sequences serve to discriminate between individuals, and thus, it follows that fixing distributions of name tokens by law in turn fixes the discriminatory power of those distributions. The 20th century is characterized by large global increases in population sizes; that is, the number of individuals that name distributions must serve to discriminate between has increased. In western cultures, this has had two consequences: first, the fixing of last names has caused the increase in information in the distribution of first names in direct proportion to increases in population [3,11]. Second, it has led to an increase in the diversity of regional first name distributions across very large states such as the United States. An interesting consequence of this is that, although the first name distribution in the US as a whole follows a power law, the distribution of names in the individual states still show close fits to the geometric, indicating that the shape of the former distribution may reflect the result of the aggregation of the latter [3,8].

These results suggest that, across space and time, discriminative codes somehow seem to respond to the various communicative pressures imposed by the environment in ways that sustain the sampling invariance that seems to be crucial to efficient, systematic communication, a point that name distributions in particular seem to underline in that individual contributions to the name pool appear, at least at first blush, to be somewhat random. These findings offer a new perspective on the apparent similarities and differences between communication in the human and information theoretical sense and raise some interesting questions in regard to speech. To what extent are speech codes shaped by the competing pressures of providing sufficient contrast to communicate the required distinctions while retaining a sufficiently stable structure to allow mutual predictability over the course of learning? Is the variance in the forms people actually articulate a consequence of the uncertainty in the structure of the immediate context in which they are learned and used, and does this variance have a communicative function?

The following sections briefly review the theoretical background to the present analysis. Section 1.1 reviews some key findings about linguistic distributions that appear to support their communicative function. Section 1.2 describes some of the implications of these findings for speech, and finally, Section 1.3 lays out a set of explicit predictions derived from this theoretical analysis. These are then examined in the rest of the article.

1.1. Grammar as Context—Convention Shapes Learning, Learning Shapes Context

It seems clear that human communication codes are not shared in the predefined way that information theory supposes [3]. Natural languages are learned from exposure to specific, incomplete samples, and these can diverge considerably across cohorts. This in turn suggests that any communicative system operating on global word token probabilities will be inefficient and unsystematic because the bursty/uneven distributions of low-frequency tokens observable in large language samples indicate that a large portion of types will be either over- or underrepresented across the communicative contexts any individual speaker is exposed to. At the same time, the fact that regularities in human languages can be consistently captured and shared through linguistic abstractions at many different levels of description suggests that speech provides speakers (and learners) with probabilistic structures that are sufficiently stable to ensure that most important linguistic conventions will be learnable from samples all speakers are exposed to. For example, Blevins et al. [12] suggest that the existence of grammatical regularities in the distribution of inflectional forms serves to offset many of the problems that arise from the highly skewed distribution of communicative codes, since the neighborhood support provided by morphological distributions makes forms that are otherwise unlikely to be attested to many speakers inferable from a partial sample of a code.

The fact that pseudowords can be interpreted in context [13] (for example, *He drank the dord in one gulp.*) offers another illustration of this point. Here, the lexical context provides sufficient support for the inference that *dord* is likely a drink of some sort, regardless of whether it is familiar to the speaker or correlated to a real life experience. (In the former case, if *dord* were to occur more regularly and in correlation to an actual bottled or cupped substance in the world, it would become a part of the vocabulary, losing its non-word status.) These kinds of context effects appear to rely on the fact that, in the sequences *drink milk*, *drink water*, and *drink beer*, *drink* systematically correlates with words that in turn covary with the consumption of fluids, unlike *eat apple*, *eat banana*, and *eat chicken*.

Given the discriminative nature of learning, it follows that exposure to samples containing this kind of systematic covariance structure will lead to the extraction of clusters (subcategories) of items that are less discriminated from any other items that occur in the same covarying contexts than to unrelated items [3]. Further, there is an abundance of evidence that patterns of systematic covariance of this kind provide a great deal of information, not only at lexical level (where semantically similar words typically share covariance patterns) but also at a grammatical level [3]. For example, in English, different subcategories of verbs can be discriminated from the extent to which they share argument structures with other verbs. The way that verbs co-occur with their arguments appears to provide a level of systematic covariance that nouns appear to lack [14]. For instance, the following sentences would be considered grammatical:

1. John *murdered* Mary's husband.
2. John *ate* Mary's husband.
3. John *chewed* Mary's husband.

However, the following sentence would not be considered grammatical:

4. John *ran* Mary's husband. (*)

One reason for this difference is that *chew*, *eat*, and *murder* share a similar pattern of argument structures (covary systematically) in a way that *run* does not. In contrast, the kinds of grammatical context which predicts a noun (noun phrases) appears to allow any noun—the sentence is grammatical—irrespective of its likelihood (although, obviously, these will vary widely according to context).

5. John *ate*.
6. John *ate* cheese.
7. John *ate* cheese slowly with a toothbrush.

In other words, the systematic covariance of verbs in their argument structures appears to constrain their distribution in context far more than is the case for nouns.

8. Mary *loved*. (*)
9. Mary *loved* cheese.
10. Mary *loved* cheese slowly with a toothbrush. (*)

Accordingly, the distributional patterning of verbs thus appears to reduce uncertainty about not only the lexical properties of upcoming parts of a message but also its structure. In other words, because verbs take arguments, there ought to be less variance in their patterns of covariation and this ought to lead to less overall uncertainty in the context of verb arguments. Consistent with this, Seifart et al. [15] report that slower articulations and more disfluencies precede nouns than verbs across languages, raising further questions about the kind of information that is communicated by variational patterns in speech and, in particular, whether and to what degree, this kind of sublexical variance actually serves a communicative function.

In the next section, we review some evidence that suggests the interactions observed between uncertainty and articulatory variation may indeed be functional.

1.2. Sublexical Variation in Context

It is well established that isolated word snippets extracted from connected speech tend to be suprisingly unintelligible outside of their context. By contrast, when reduced variants are presented to speakers in context, they are able to identify the word without difficulty and to report hearing the full form [16]. Consistent with this, the effect of frequency on speech production has been shown inconsistent across registers, speakers, lexical classes, and utterance positions and there are opaque interactions between context, lexical class, and frequency range.

At first blush, these inconsistencies would appear to limit the scope of functional accounts of speech sound variance, and to date, the effects that are stable enough to be taken as evidence for functional theories are mostly to be found in preselected content words from the mid-frequency range, such that the effects reported rarely align with effects observed in the remaining (by token count, significantly larger) parts of the distribution.

For example, while function words, high frequency discourse markers, and words at utterance boundaries account for the largest portion of variance in speech, their exclusion from the analysis of speech sound variance is such a common practice that it might be considered a de facto standard [17]. Against this background, it is noteworthy that Bell et al. [18] report a divergence in the extent to which the articulation of function and content words across frequency ranges is affected by both frequency and the conditional probability of the collocates. While duration in content words is well predicted by the information provided by the following word but not the preceding word, the effect decreases as the frequency increases and shows a reverse pattern in function words. Similarly, van Son and Pols [19] report a reversal in the correlation between reduction and segmental information in low-information segments and segments at utterance boundaries. The effect of information content is reported to be limited by a *hard floor* in high-frequency segments; that is, most frequent segments fail to support the hypothesis. Standardizing the exclusion of misfits is controversial, especially given that they outnumber the tokens which are typically taken to confirm a hypothesis and account for the largest part of variance in speech [20,21].

The seemingly random and noisy variance in the speech signal appears systematically correlated with uncertainty about the upcoming part of the message. As an example, vowel duration in low- and mid-frequency content words is correlated to the information provided by the upcoming word [18]. Words in less predictable grammatical contexts are on average longer and more disfluent [22]. These fluctuations in duration and sequence structure have been shown to inform listeners' responses. For instance, the duration of common segments in word stems differ between singular and plural forms [23]. Speakers appear to use acoustic differences in word stem as a cue to grammatical context (plural suffix), and incongruence between segmental and durational cues lead to delayed responses in both grammatical number and lexical decision tasks [24]. Similar effects occur at many other levels of description; for example, disfluent instructions (*the . . . uhm . . . camel*) lead to more fixations to objects not predicted by the discourse context [25] and facilitate prediction of unfamiliar objects [26].

The occurrence of silent and filled pauses has been shown to contribute to the perception of fluency [27] and intelligibility [28] as well as improved recall [29]. Importantly however, neither artificially slowed-down speech samples nor samples modified by insertion of pauses are then perceived to be more fluent or intelligible, and indeed, in both cases, these manipulations have been shown to result in impaired performance [30]. Accordingly, the fact that listeners easily interpret reduced sequences from context and reject speech artificially altered to mimic completeness and fluency indicates that hearers are highly sensitive to violations of their expectations about how natural speech should sound and not that they have a preference for completeness and slow and extreme articulation. However, despite the evidence that sublexical variation shapes speaker expectations about the upcoming content, its contribution to successful communication as an informative part of the signal has remained relatively unexplored to date.

However, it is clear that any quantification of the communicative contributions of sublexical variations in context will depend on a consistent definition of context. That is, in order to address the extent to which the quality of articulation and the observed variance in the signal interact with the remaining uncertainty about the message in general terms, it is necessary to first formalize a consistent subset of higher-level abstractions that systematically covary in the degree to which they contribute to uncertainty reduction. The contrast between these subsets can then allow these effects to be analyzed independent of the specific context of any given utterance.

1.3. The Present Study

In comparison to written language, speech often appears to be messy. Instead of the well-formed word sequences that characterize text, spontaneous speech sequences are typically interrupted by silent and filled pauses, left unfinished, depart from word-order conventions, frequently miss word segments or whole words, and rely on clarifying feedback which tends to be short and grammatically incomplete. In consequence, the token distributions that underlie the information structure of written and spoken language differ substantially.

For instance, nouns are less lexically diverse in spoken English then in writing (based on measures derived from the Corpus of Contemporary American English (COCA)), whereas English adjectives tend to be more lexically diverse in speech. While reading and writing are self-paced, speech gives both speakers and hearers less control over timing. This suggests that the moment-to-moment uncertainty experienced in communication may differ in speech as compared to written language, and it may be that more effort is invested in uncertainty reduction in spoken than in written language. From this perspective, the increase in the lexical variety in prenominal adjectives, which in English reduce uncertainty about upcoming nouns [31], might be functional in that it may help manage the extra uncertainty in spoken communication. This raises the question of the degree to which these and other variational changes in spoken English are indeed informative and systematic.

These considerations also suggest that the results of previous analyses of the distributional structure of lexical variety in communicative contexts conducted on text corpora can only offer indirect support when it comes to answering questions about the communicative properties of speech. To address this shortcoming, we conducted a corpus analysis of conversational English [32] to explore the extent to which the distribution and the underlying structure of the grammatical context in which words are embedded interacts with speech signal variation observed across lexical categories. The goal of this analysis was to explore the structural properties of grammatical regularities in speech and their effect on the distributions of the lexical and sublexical contrasts that they discriminate between.

The analysis was conducted in two stages. Part one, presented in Section 3, addresses the distribution of grammatical and lexical contrast in speech and aims to answer the following questions:

- Are distributions of grammatical regularities in speech sampling invariant?
- How do recurrence patterns of grammatical categories and speech sequences inform learning?
- Are distributions of subcategorization frames and types they distinguish between geometric?

Part two of our analysis, presented in Section 4, assesses the concrete consequences of the sublexical variation observed in the speech signal and relates these to the results presented in Section 3, addressing the following questions:

- Are the inconsistent effects of frequency on speech sound variation across categories correlated with structural and distributional aspects of the grammatical and lexical contexts they populate?
- Finally and perhaps most importantly, is the resulting sublexical variance systematic?

2. Materials and Methods

2.1. Data

The Buckeye Corpus [32] contains phonetically transcribed speech from informal interviews with 40 speakers from Columbus, Ohio. The 286, 982 words are annotated with a set of 41 standard aligner phone labels expanded by a set of markers for manner of articulation (nazalization, flaps, glottal stops, and retroflex vocalization). The corpus version we used was extended by Dilts [33] with measures of segmental deletion; dictionary form alignment; and deviation rate normalized by word length, speech rate, and backward and forward conditional probabilities of word ngrams. For the analysis reported here, we excluded from the corpus 8426 words with missing or incomplete duration variables. The data set and the code for the analysis can be found at https://osf.io/bqepj/.

In each of the 1-h interviews, the 40 speakers (who are balanced by age and gender) showed an enormous amount of variability (as assessed by phonetic transcription) in the speech signal. Overall, only 40% of the words are produced in their citation form. Only 38% of word types tend to appear in their non-citation variants more often and the propensity of individual speakers to pronounce word types in their citation form varies widely (between 36% word types and 67% word types). The word *that* appears in 313 variants, including *d ah tq*, *m ah t*, *z eh tq*, and *ng ah*.

We extracted for each citation form and parts-of-speech combination the number of variants observed in the corpus by citation form. The relative frequency counts for each form by parts-of-speech label were taken from the spoken part of COCA, an 80 million token subcorpus of Contemporary American English from transcripts of unscripted conversation on TV and radio programs.

2.2. Probability Distributions Analysis

Plotting a frequency distribution on a log-log plane, with log frequency on the y axis and log of rank order on the x axis, is a common method in the analysis of probabilistic structure. A linear plot indicates that the data conforms to Zipf's law because Zipf's law assumes an exponential increase in the time rate (rank). That is, a linear plot confirms a power law, while distributions we observe here and other variants of aggregate distributions (e.g., Zipf-Mandelbrot) are reported as an anomaly. The latter usually entails the introduction of additional parameters to fit the distribution back to power law.

Because linguists have so far only searched for power laws, the distributions we observe here are, when found, reported as an exception [34]. Ramscar [3] argues that empirical linguistic distributions ought not to be expected to follow power laws. Rather, because learning and mutual predictability require a regular distribution of events over time, human communicative codes ought to be expected to have distributions that retain their structures over time. Accordingly, following Ramscar [3], we employed log-linear plots in these analyses. That is, the linear decrease in probability over discrete time defines a time invariant communicative distribution while the exponential decrease in probability does not. To asses the extent to which the method captures this property, we apply it to a set of subsamples drawn from the original data.

Figure 1a,b shows results from a simulation study capturing fits of the analyzed categories to a geometric distribution and a power-law distribution, respectively, over the first 2500 words from each of the 40 speaker subsamples. The two bottom row panels show the fits to geometric (Figure 1c) and power law (Figure 1d) across 40 random subsamples varying in size between 652 and 19, 363 tokens. As we can see in Figure 1, fits to power law vary with sample size and source across all categories. In contrast, fits to geometric remain relatively stable in empirical distributions independent of sample source and size. This is not the case for aggregate distributions. Accordingly, this method appears to capture the critical property of communicative distributions addressed in this paper.

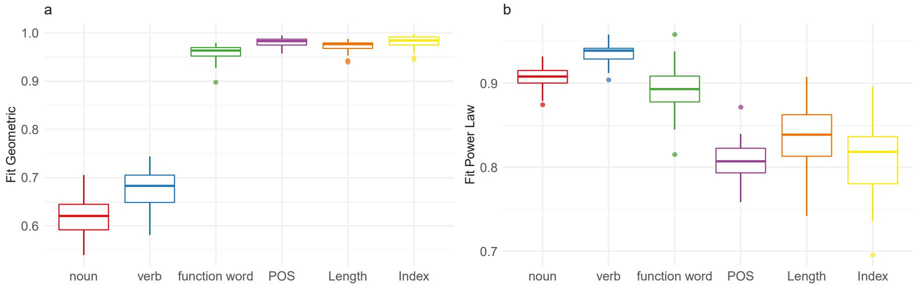


Figure 1. *Cont.*

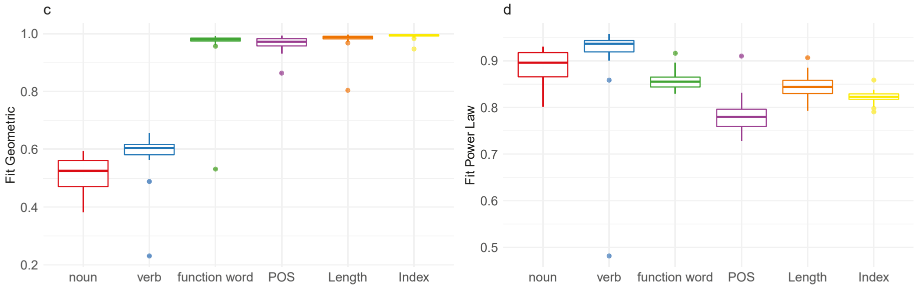


Figure 1. Boxplots of fits to geometric distribution (**a**,**c**) and power law distribution (**b**,**d**) for categories analyzed in Sections 3.1 and 3.2 for the first 2500 words by 40 speakers (**a**,**b**) for 40 random samples ranging in sizes between 652–19,363 (**c**,**d**).

2.3. Statistical Analysis

The results presented in Section 4.1 were analyzed with a generalized additive mixed-effects model (GAMM) [35,36], working with the *mgcv* package for *R*. GAMMs are used for the analysis of complex, often nonlinear patterns involving the interaction of two or more numeric and factorial predictors. Instead of using polynomial functions, GAMMs introduce smoothing splines. A smoothing spline with one predictor fits a curve over multiple basis functions. Smoothing splines with multiple predictors fit multidimensional surfaces. These features allow us to explore interactions between frequency range, context, and lexical category and to reduce the model complexity by identifying relevant predictors which eventually result in linear effects.

3. The Structure of Lexical and Grammatical Variety in Speech: A Corpus Analysis

3.1. Part-of-Speech Token Distributions

3.1.1. Why Parts of Speech?

It is clear that many important regularities in human languages are consistently captured by high-level linguistic abstractions such as, for example, parts-of-speech categories, indicating that languages may be sufficiently structured to allow the discrimination of various functional parts of codes at various levels of abstraction. Ramscar [3] suggests that the probabilistic co-occurrence patterns of words and phrases serve to discriminate subcategories of signals (and hence codes) and that, as well as serving different communicative purposes, these subcategories form distributions that facilitate speaker alignment at various levels of analysis. This raises an obvious question: do the distributional properties of structural regularities in conversational speech actually support this hypothesis?

Parts-of-speech tags are often used to label the various categories that can be extracted from the abstract structure of languages. Different tag sets are used for languages which differ in structure, and the extent to which tags capture detail varies with the particular context in which tagging is employed. These tags are assigned automatically by statistical tools, typically assuming a Markov process, which employs regularities in word co-occurrence patterns over word sequences of varying sizes [37,38]. The fact that taggers achieve high levels of accuracy suggests in turn that high levels of systematicity must be present in distributional patterns. That is, the fact that structural properties of the training set will translate to novel and larger samples implies that the captured properties are sampling invariant. Previous work on text corpora implies that, in text at least, the empirical distributions discriminated by communicative contexts are geometric [3]. This raises a question: do the patterns that emerge during part-of-speech tagging also discriminate distributions with similar empirical properties?

Further, the finding that the probability of types that are subcategorized by these context decreases at a constant rate [3] suggests in turn that different empirical subcategories might serve similar communicative purposes at different levels of specificity. In English, message length in words has been shown to increase as the content of messages increases as a consequence of learning and specialization [39–41]. The apparent systematicity revealed by analyses of covariance patterns in text suggests that communicative codes may be adapted to support the transmission of an unbounded set of messages at multiple levels of description, including length. That is, in speech at least, the considerations reviewed above would seem to suggest that word sequence length (at least in English) may be related to the relative probability of the message with respect to all messages all speakers might want to communicate. This raises a further question: is the distribution of n-grams in speech geometric?

To answer this, we analyzed the distributions of part-of-speech labels, utterance length, and utterance position in the Buckeye Corpus of conversational English [32].

3.1.2. Results

Figure 2a–c shows frequency rank distribution plots of log counts for part-of-speech labels, phrase lengths, and the phrase positions, respectively. The blue line indicates the best fit to log-log scale (power law), while the red line shows the best fit to geometric (geometric is linear; the probability decreases at a constant rate). As we can see, the empirical distribution (represented by the grey points) of part-of-speech labels, utterance length, and utterance position, with R^2 of 0.9725, 0.9957, and 0.9981, respectively, show a close fit to geometric, whereas fits to power law are 0.7798, 0.8109, and 0.8035, respectively.

These results thus suggest that sampling from the functional distributions that can be discriminated by context at this level may indeed result in probability estimates that are similar across speakers, irrespective of discourse context and length.

In addition, they provide some evidence to support the suggestion that hearing a one-word utterance such as *yes*, *okay*, *correct*, or *exactly* or a longer utterance such as *um sort of let them make their own decision when they got older what they wanted to do* is sufficiently stable irrespective of size, again indicating that the distribution of communicative types at different levels of description in conversational speech may be systematic.

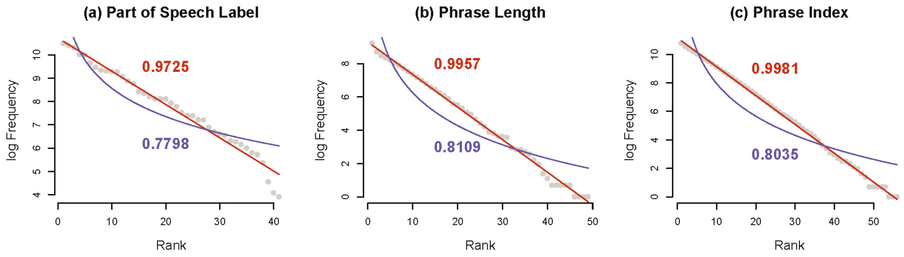


Figure 2. The frequency distributions for the part of speech label (**a**), utterance length (**b**) and utterance position (**c**) categories in the Buckeye Corpus [32]: Grey points show the observed distribution, with fits to a power law distribution (blue line) and a geometric distribution (red line). All three distributions show a close fit to a geometric distribution.

3.1.3. Discussion

Our results show that grammatical subcategories captured by part-of-speech tags have distributions that are likely to lead to an alignment in the probabilistic expectations of speakers regardless of any differences in their exposure to these distributions. They also provide further support for the suggestion that, unlike aggregate word token distributions (which have power law distributions [42]), the empirical distributions that are discriminated by communicative context are geometric [3].

The abstract model of communication defined by Shannon [9] is at heart a deductive process of uncertainty reduction [3]. The model assumes that communicative codes will be distributed so as to ensure that every sequence produced has the same statistical properties. A consequence of this is that any mixture of code samples will have the same statistical properties as any other sample. By contrast, it would appear that, in speech at least, natural languages gradually reduce message uncertainty via a series of sequential subcategorization frames of increasing degree of specificity. Evidence for this suggestion is provided by the differences in type/token ratio of part-of-speech categories, which vary systematically. Further, the shape of the distribution of utterance lengths suggests that the expectations about the distribution of messages of different lengths that speakers learn will likely align, helping the overall system to deal flexibly with the ever-growing number of specific messages that humans are likely to wish to communicate.

In the introduction, we described how constraints on the structure of name sequences have lead to qualitatively different patterns of distribution in English and Sinosphere first names. Legal constraints on last names in English have lead to differentiation between (geometric) local first name distributions which, when aggregated over, fit power laws [3]. Thus, the differences in the extent to which word categories are subjected to grammatical and lexical constraints (Section 1.1) seem to predict differences in the productivity of lexical categories over time, leading to more aggregation in verbs. The analysis presented in the next section aims to explore whether the shape of word frequency distributions of different lexical categories reflect the differences in the way they are constrained by the grammar.

3.2. Covariance, Systematicity, and Subcategorization

3.2.1. Token Distributions across Lexical Categories

As we noted above, high-level descriptions (e.g., parts-of-speech) clearly capture many abstract communicative properties such as animacy, agency and number in nouns or tense, and aspect or argument structure in verbs. However, it seems that the functionality of these categories is further subcategorized by patterns of co-occurrence which encode more specific distinctions between agents, objects, actions, and relationships. This implies that verb and noun frequency distributions are aggregates of functionally distinct subcategories. Consistent with this, Bentz et al. [43] show that aggregates over verbs and nouns are power law distributed while Ramscar [3] confirmed this finding and then showed that the subcategorical distributions of verb and nouns discriminated by communicative context are geometric.

Importantly, previous studies have shown that token distributions in closed class categories (function words and modal verbs) do not follow power laws [43,44]. These departures from the trend to power law in other categories are assumed to be related to the communicative function of high-frequency words. Linguistic theories typically assume that closed class tokens serve a qualitatively different modifying or grammatical function while open classes are considered to contain and transport meanings; that is, they provide lexical contrast.

These previous results thus predict that, when context is not used to subcategorize them, nouns and verbs in English will be distributed differently to function words. To explore these patterns of distribution, we analyzed the word token distributions of these separate parts of speech across the speech samples.

3.2.2. Results

There are 44,722 noun and 45,159 verb tokens in the analyzed sample. With 5817 unique types, nouns are a far more lexically diverse category than verbs with 2574 types. By contrast, the 116,960 function word tokens are represented by 144 unique types.

Figure 3 shows the token distribution of the three largest grammatical categories. We can see that both verbs and nouns have a closer fit to power law compared to geometric distribution: $R^2_{pl} = 0.976$ and $R^2_{geom} = 0.701$ for verbs; $R^2_{pl} = 0.971$ and $R^2_{geom} = 0.772$ for nouns.

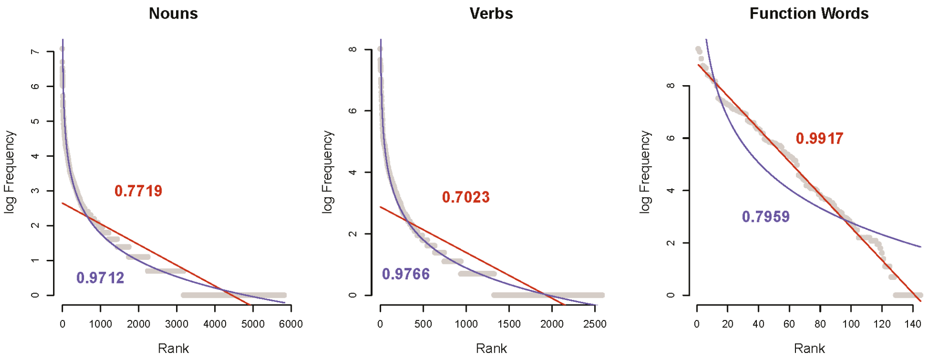


Figure 3. Word frequency distributions of nouns, verbs, and function words in the Buckeye Corpus [32] show that the substantially smaller (compared to nouns) set of verbs has a closer fit to power law distribution, indicating more aggregation. The shape of the distribution in function words suggests that function words form a natural empirical distribution.

By contrast, the 144 unique function words ($n_{tokens} = 11,696$) show an almost perfect fit to geometric $R^2_{pl} = 0.796$, $R^2_{geom} = 0.992$. A separate analysis shows a better fit to geometric over power law in distributions of determiners ($n = 16$, $R^2_{geom} = 0.953$, $R^2_{pl} = 0.830$), pronouns ($n = 28$, $R^2_{geom} = 0.957$, $R^2_{pl} = 0.741$), and prepositions/subordinating conjunctions ($n = 78$, $R^2_{geom} = 0.983$, $R^2_{pl} = 0.863$). The aggregated set of function words, however, improves the fit to geometric.

3.2.3. Discussion

When taken in conjunction with earlier findings [3,43,44], the distribution of function words we observed here supports the suggestion that they form a natural communicative distribution. This in turn suggests that, despite the fact that prepositions (in contrast to determiners and pronouns) distinguish between spatial and temporal relations, prepositions, determiners, and pronouns are part of the same functional subsystem and, at some level, serve the same communicative function.

By contrast, we find that the lexically more diverse categories fit power laws. As previously discussed, these distributions could be the product of aggregating over multiple communicative distributions serving distinct communicative functions. This suggestion is further supported by the observed distributions of verbs and nouns, which suggest that a smaller number of unique verb types appears across a larger number of distinct communicative contexts than is the case for nouns. This observation is supported by the fast growing head in the verb distributions, which appears to result from aggregating over high-frequency verbs, whereas the fast growing tail in the noun distributions appears to reflect the greater volume of low-frequency nouns.

In other words, the results imply that the differences observed between lexical categories do not necessarily warrant categorial distinctions. Rather, the observable differences appear to reflect the extent to which word co-occurrence clusters are shaped by the opposing communicative pressures of prediction and discrimination over the course of learning.

In other words, these results confirm the idea that lexical categories are not equally distributed across utterance positions. The next part of our analysis explores these relationships further.

3.3. Lexical Category, Word Order, and Recurrence Patterns

3.3.1. What Makes a Lexical Category?

The distribution of function words suggests that function words will form the grammatical subcategory that is first discriminated systematically from the speech signal. As a consequence, it seems likely that, as both intuition and many linguistic theories would predict, function words provide a first contextual frame to aid in the learning of other grammatical and contextual categories. Once these basic contextual frames are learned, they will provide context, assisting in the learning of other words. The idea that context will provide information that aids learning in turn suggests lexical diversity will increase with utterance length.

Consistent with this suggestion, Genzel and Charniak [45] have shown that, although caching local probability estimates of a words' occurrence in written samples (to account for the variance in recurrence patterns over time) stabilizes relative entropy over lexical sample size in nouns significantly, the effect is far smaller in verbs and absent in function words. In the light of the foregoing discussion, this might be taken to suggest that patterns of co-occurrence in verbs are less variant than those in nouns and that these patterns are still less variant (and may even be regular) in function words. These considerations suggest in turn that the different subcategories of words systematically reduce uncertainty in communication at different levels of abstraction. To explore whether the different communicative contributions of words from different lexical categories are quantifiable in speech signal, we analyzed the patterns of occurrence of nouns, verbs, and function words (the three largest categories by token count) over utterance length.

3.3.2. Results

The probability density of token occurrence over log normalized utterance length was analyzed by category. As can be seen in the left panel of Figure 4, while the larger parts of tokens in all three categories follow a normal distribution across utterance position (presumably as a consequence of utterance length), there is also evidence of distinct bursts of occurrence which align with the word order typology of English. That is, less specific pronouns are more likely in utterance initial positions, verbs are more likely in utterance medial positions, and nouns are more likely in utterance final positions.

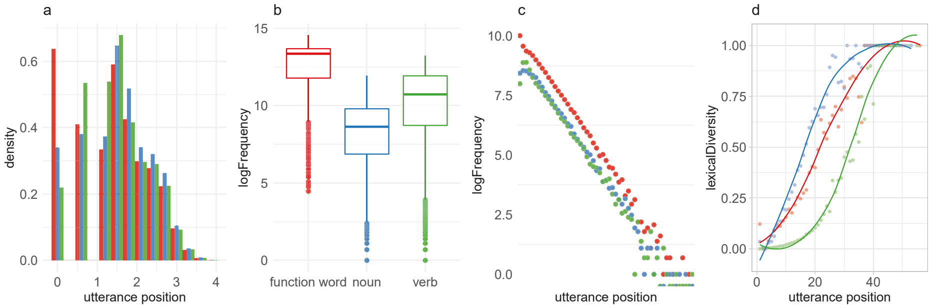


Figure 4. Distributional properties of the three largest (by token count) categories analyzed by utterance position and frequency range: We find that the overall probability of occurrence varies with type and utterance position (**a**), that frequency distributions of lexical classes are not evenly distributed across probability space (**b**), that part-of-speech token probability decreases linearly as a function of utterance position (**c**), and that lexical diversity increases nonlinearly as a function of utterance position (**d**).

The extent to which lexical categories are represented across the probability space (Figure 4b) is correlated to the average utterance position. We find 85% of all function word types in the top 50 tokens, which makes up 51% of the probability mass, and 93% of function word types in the top 100 words, which makes up 64.6% of the probability mass. In other words, function words are high-frequency words.

Further, we observe that, while token probability across lexical categories decreases linearly (Figure 4b) over utterance position, the increase in lexical diversity across all three categories is nonlinear. The right panel of Figure 4 shows smoothness of the normalized type/token ratio as a function of utterance position. We observe significant differences in the patterns of increase between the three lexical classes. The increase in the lexical variety of function words is limited to a small number of tokens in the latter positions of long utterances. The diversity in nouns increases earlier than in verbs.

Figure 5 shows that when words at utterance boundaries are excluded from the analysis, the normalized type/token ratio of nouns and verbs show similar increase patterns while the growth in function words remains unaffected. In contrast, the wide confidence interval in utterance final verbs indicates that the relationship between lexical diversity and utterance length (which can be taken to signify context) is less consistent in verbs than it is in nouns (and pronouns).

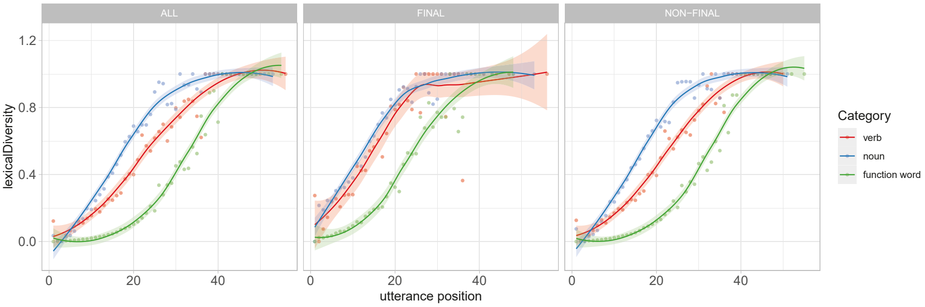


Figure 5. Increase in local lexical diversity (type/token ratio) across utterance position is not linear. The increase rates differ substantially between lexical classes. The differences in the increase rate between verbs and nouns in utterance final position are restricted to utterance initial tokens. The confidence interval in verbs is larger. The differences in the increase rate between verbs and nouns are constituted by the extent to which context affects lexical variety in nonfinal tokens.

3.3.3. Discussion

In sum, we observe significant differences in distributional properties between word categories. Word categories differ with respect to the frequency range they populate, the average utterance position, and lexical diversity. From the perspective of learning, this implies that the properties which distinguish word categories interact with the order in which they are learned while the order in which they are learned appears to be a consequence of the regularity with which they are represented across samples.

We suggest that the aggregation effects in token distributions across lexical classes is correlated to the degree in which category types are regularly distributed across language samples, reflecting the extent to which their communicative function is mediated by the contextual frames they appear within. Our analysis shows that lexical classes differ both in the average utterance position and in the rate at which lexical diversity increases as a function of utterance position and that the increase rate is inversely related to the average utterance position of the class.

In the next part of the analysis, we explore the extent to which the variety of abstract grammatical constructions in which words are embedded can serve to capture the differences in distributional structure and recurrence patterns that we observe across lexical categories.

3.4. Distribution of Grammatical Context

3.4.1. How Do Different Parts of Speech Carry Out Their Communicative Function?

Words often occur in multiple grammatical contexts. The word *claim*, for instance, appears 5989 times in the spoken section of the Corpus of Contemporary American English [46]: 2719 times as a verb and 3270 times as a noun, 3016 times as noun singular, 1994 times as base form verb (1), and 1276 times as an infinitive (2), so that the three instances of *claim* in the three following examples serve distinct communicative functions which are not equally probable.

11. The girls *claim* to have seen the fairies.
12. You may be able to *claim* compensation.
13. The court found no evidence to support her *claim*.

The particular uses of *claim* that speakers intend to communicate will thus be determined by the lexical and grammatical context in which it is used. If one were to count the word *claim* as one type across all the contexts it occurs within without taking into consideration its lexical status, one would run the risk of aggregating over the multiple communicative functions it serves, and this problem will clearly increase as a word's frequency increases.

To explore the extent to which lexical subcategories receive support from these kinds of contextual frames, we next analyzed the distributional properties of the frames that words are embedded within by lexical category.

3.4.2. Results

In the first part of this analysis, we explored the distributions of grammatical context (defined as part-of-speech bigram) that words are embedded in. This was then followed up with an analysis of the word token distributions that these part-of-speech constructions distinguish between.

The left panel of Figure 6 presents the log frequency rank plots of part-of-speech bigram distributions over the three analyzed lexical categories. It shows that all three distributions are geometric and that the slopes differ substantially. The slopes which reflect the extent to which words are subcategorized by grammatical context are inversely correlated to the rate at which lexical diversity increases as a function utterance position in Figure 5.

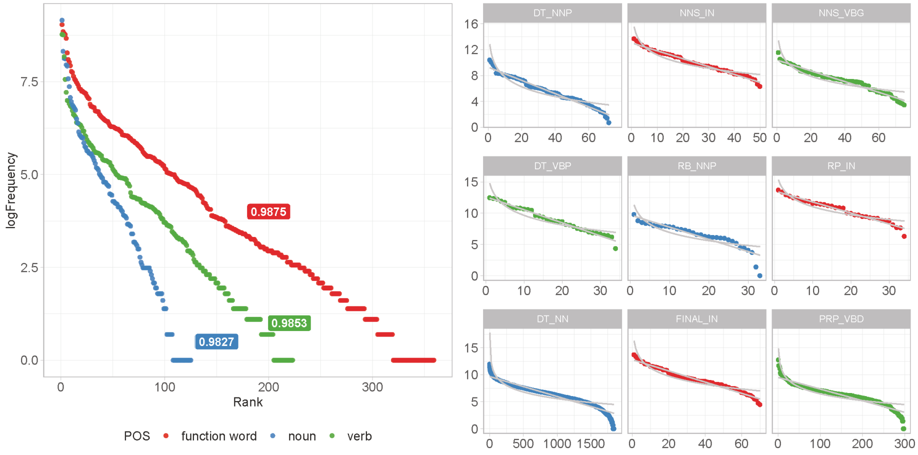


Figure 6. Distribution of contextual distinctions (part of speech bigrams) by lexical class: Nouns appear in a far smaller number of contextual frames; the size of the contextual frame is on average larger. The frequency distribution of verbs within the contextual frame is exponential. In the larger set of nouns, we see effects of aggregation in the low- and high-frequency tails.

We observe a more diverse set of grammatical distinctions between categories of smaller frame size in verbs as compared to nouns. The distribution of the parts-of-speech bigram of the preceding word and the word itself is more diverse, with 124, 223, and 359 parts-of-speech constructions for 5869, 3124, and 144 unique word forms, respectively. The parts-of-speech context on average comprises 37 types of verbs (ranging from 1 to 460) and 119 types of nouns (ranging from 1 to 1862). In contrast, function word contexts on average host 2 distinct function words.

This suggests that the extent to which words are subcategorized by grammatical context is correlated with both lexical diversity and the average utterance position of a category. In consequence, the lexical distributions we find embedded in grammatical frames differ in size and structure. The center and the right panels of Figure 6 show the frequency distributions of the unique words found in two of the smaller subcategorization frames. The smaller (by unique type count) frames show a close fit to geometric irrespective of the word frequency range. In general, we observe more aggregation in noun frames. That is, the extent to which the subsamples extracted from grammatical subcategorization frames show the effects of aggregation appears to be independent of the frequency range of lexical contrast they distinguish between. Instead, aggregation appears correlated to the size of the subsample and, by implication, the extent to which lexical frames serve further subcategorization within the more abstract grammatical frames.

3.4.3. Discussion

The distribution of grammatical constructions suggests that nouns which on average appear in a smaller number of more lexically diverse constructions will receive more support from lexical frames, resulting in less variance in the conditional probability between nouns and the words on which they are conditioned. That is, the more high-frequency nouns tend to appear in larger, high frequency contexts and thus tend to be further subcategorized by smaller lexical subcategorization frames. In contrast, the extent to which the variety of grammatical contexts serves to reduce uncertainty across a smaller (by type count) set of verbs will lead to more variance in the conditional probability between verbs and verb arguments.

In the next section, we explore the effects that the distinct patterns of covariance between high-frequency verbs and high-frequency nouns and their collocates have on the variety of articulated variants we find in the speech corpus.

4. From Information Structure to Speech

The results we have described so far suggest that the structure of speech serves to facilitate efficient message transmission over multiple nested levels of description. The distribution of lexical and grammatical contrasts indicates that information structure *depth* increases over message sequences, supporting gradual increases in the degree to which low-level sublexical contrasts contribute to resolving uncertainty about a message. Consistent with this, it has been shown experimentally that speech rates are perceived as being faster and target words as being longer when cognitive load is increased [47], a response pattern that suggests that speakers adapt their response to the relative uncertainty resulting from utterance context.

The notion that the timescale variance captured in speaker and listener performances reflects adaptation to uncertainty is further supported by evidence showing that the sublexical variation in speech sequences increases with sequence length, a phenomenon characterized by the strengthening of word initial consonants and the lengthening of final vowels. While both effects increase cumulatively as a function of utterance length [48], the interaction between lengthening and strengthening is weak, indicating that hyperarticulation and vowel space expansion are not equally affected by context. Moreover, while low-probability and word initial segments are more likely to be stressed and while segment deletion is more likely in high-frequency phonemes and in latter positions, the frequency effects actually observed in very frequent segments depart from this pattern. Also, the correlation between duration and extreme articulation, and duration and frequency declines as a function of utterance position [49].

The analyses presented in Section 3.4 indicate that average grammatical uncertainty peaks in words that are more likely to occur in utterance initial positions and that average lexical uncertainty peaks in categories which are more likely at utterance final positions. It has also been shown that slow-downs in articulation are associated with uncertainty and that uncertainty leads to an increase in articulatory variance. These effects have been observed both within [23] and across word boundaries [18] as a consequence of syntactic irregularities [22] and appear functional in lexical decision [24] and discourse [26]. Since our analyses show substantial differences across parts of speech in both the extent to which words are predicted by the previous context and the extent to which they serve to predict the upcoming part of the message across the frequency range, this seems to imply that the apparently inconsistent effects of frequency that have been previously observed are both predictable and systematic with respect to the structure of the grammatical context.

This in turn can be taken to suggest that sublexical variance follows as a consequence of an increase in lexical and grammatical variety in which words are embedded and that the variants we observe aim to increase the efficiency in transmission of informative contrast at multiple levels of description. In the next section, we conduct a statistical analysis of the effects of variation in the collocations of words on the number of distinct forms found in the speech corpus.

4.1. Effects of Frequency and Collocate Diversity on Variation

4.1.1. The Distinct Effects of Collocate Diversity And Frequency

Wedel and colleagues have shown that the number of competing minimal pairs in lexical context predict likelihood of vowel merger [50] and voice onset time duration [17], suggesting that what drives speech contrast loss is the extent to which minimal pair competition is resolved in context. In line with this, Piantadosi et al. [51] observe that the relative probability of a word in a lexical context (defined as word sequences ranging between 2–4 words) is a far better predictor of word length than word frequency.

This raises questions: Does this hold for variance too? Is the diversity of collocate contexts across which a word appears a better predictor of the extent to which a type will vary across a speech sample than frequency?

The probability of a known word appearing in a previously unattested context increases with the average word count so that word frequency and collocate diversity are strongly correlated ($r(9190) = 0.70$, $p < 0.0001$). High-frequency words are more likely to be preceded by a larger number of different words and thus tend to appear across a larger number of communicative contexts that vary in size. By implication, there is more variance in the conditional probability between high-frequency words and their collocates. In contrast, words from the mid-frequency range will appear in a smaller number of distinct communicative contexts, leading to less variance in the conditional probability between mid-frequency words and their collocates. In line with this, an analysis by Arnon and Priva [52] shows that, in contrast to results reported by Bell et al. [18], duration in content words is affected by both word and multiword frequency as well as the transitional probability of both the preceding and following collocates when high- and low-frequency trigrams; sequences interrupted by pauses and word final sequences are excluded from the analysis. Finally, the increase in lexical diversity over utterance length (Section 3.3) suggests that low-frequency words tend to appear in a larger number of distinct message contexts, again leading to more variance in the conditional probabilities of low-frequency words at different positions within the sequence with respect to the likelihood of the message.

The discriminative nature of learning predicts that this variance will increase within-context competition over exposure time and that this will minimize the informativeness of contextual cues which predict a large number of lexical contrasts. This in turn predicts more sublexical variation in words that serves as cues to a larger number of collocates, reflecting the uncertainty of the relative context. Taken together, these factors predict distinct patterns of variance across frequency ranges.

4.1.2. Results

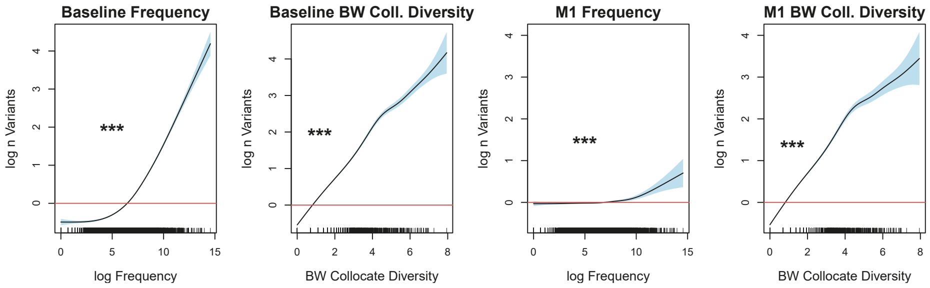
To explore the nonlinear effects of frequency and collocate diversity on observed variance, we fitted generalized additive mixed models (GAMM) [35] using the *mgcv* package for R. In baseline model 1, we model the normalized number of observed corpus variants as a function of the smooth over log frequency. In baseline model 2, we model the number of variants as a function of a smooth over collocate diversity, the log normalized number of preceding words we observe in the corpus.

Model 1 counts show a strong, nonlinear effect of frequency ($p < 0.0001$). It yields an R^2 of 0.435 and explains 43.5% of the deviance in the data ($edf = 5.05$, $AIC = 19852.04$). Model 2 shows a strong, nonlinear effect of diversity of collocates in the preceding position ($p < 0.0001$), explaining 74.6% of the variance in the data ($edf = 6.922$, $R^2 = 0.746$, $AIC = 11777.66$).

We assessed the goodness of fit of both models by the Aikake Information Criterion (AIC). Model 2 improved the score by 8074.38. To contrast the contribution of both predictors, we modeled word variance as a function of smooth over log normalized word frequency and log normalized number of variants observed in the corpus in a combined model 3. Model 3 ($R^2 = 0.746$, $AIC = 11548.74$) reduced the AIC by 228. Both predictors are highly significant ($p < 0.0001$).

Interestingly, the plots show that the frequency effects predicted by the baseline model 1 and the combined model diverge substantially across frequency ranges (see Figure 7a,c), suggesting that the effect of frequency is largely overestimated in the low-frequency and mid-frequency ranges by the baseline model. It further appears that a large part of the frequency effect is confounded by the correlation between word frequency and the number of collocate contexts a word appears within. There remains, however, a strong effect of frequency observable in high-frequency words. The high-frequency part of the data behind the effect comprises 82 function words, 57 nouns, and 47 verbs, representing 69%, 1%, and 2% of unique types, respectively.

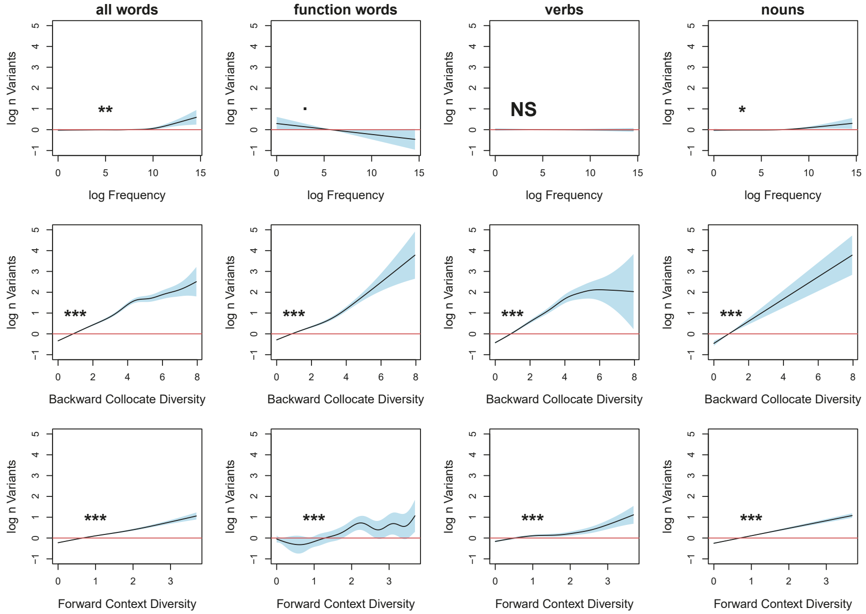
Word frequency appears to influence the extent to which a word varies in form only in high-frequency words and thus holds for type variation across lexical categories to the degree with which the category is represented in the high-frequency tail of the word distribution. We further observe a stronger correlation between collocate diversity and word frequency in function words ($r(143) = 0.882$, $p < 0.0001$), than in verbs ($r(2549) = 0.665$, $p < 0.0001$) and nouns ($r(5626) = 0.593$, $p < 0.0001$).



(**a**) (**b**) (**c**) (**d**)

Figure 7. Baseline word variance model comparison: (**a**) the log normalized number of observed variants as a a function of smooth over log frequency (derived from the spoken part of COCA); (**b**) the log normalized number of observed variants as a function of collocate diversity, the log number of preceding words; and (**c**,**d**) Figure 7a,b in a combined model.

Finally, we fitted a set of combined models, adding in the log number of distinct parts of speech following each word for all words (model 4) and adding in lexical category as a covariate factor (model 5). In model 4, we observe a fairly weak effect of frequency ($p < 0.006$ (see Figure 8a), while the effects of the context predictors were strong. The AIC score is reduced by 531.



(**a**) (**b**) (**c**) (**d**)

Figure 8. Log normalized number of observed variants as a function of smooth over log frequency (row 1) and adjacent token diversity (rows 2 and 3) for all words (**a**), function words (**b**), verbs (**c**), and nouns (**d**): when collocate diversity is taken into account, frequency effects on variation only hold in a minimal proportion of high-frequency nouns and appear to have no effect at all on verbs and function words.

In model 5, the introduction of lexical category as a covariate further reduces the AIC score by 254 points and explains 76.8% ($R^2 = 0.766$) of deviance observed in the data. The effect of frequency is not significant in verbs ($p < 0.816$) and function words ($p < 0.062$) and is statistically significant but weak in nouns ($p < 0.018$). Again, all contextual predictors are highly significant in function words, nouns, and verbs. The same pattern was observed for all of the other analyzed categories apart from the following exceptions: filled pauses and numbers show an effect of frequency ($p < 0.002$); contractions are unaffected by the collocate diversity ($p = 0.21$); and there is no interaction between modal verbs and the upcoming collocate context ($p = 0.1$). Modals, numbers, and contractions comprise 0.008% of the analyzed data set. We observe differences in the effect of preceding collocate diversity between verbs and nouns in that the effect and the confidence interval both increase linearly in nouns while the effect levels off in high-frequency verbs, showing an increase in variance.

A closer examination of the data reveals that the relationship between word frequency and collocate diversity differ significantly across frequency ranges for verbs and nouns. Collocate diversity is much higher in high-frequency verbs and function words than it is in high-frequency nouns. Also, there is far more variance in the effect in high-frequency nouns.

4.1.3. Discussion

The results of this analysis align with the finding that word counts outside of their communicative context contribute little when it comes to explaining variation in articulated forms. Rather, we observe that the largest part of this variance is explained by the diversity of the lexical contexts in which words appear. The remaining effects of frequency are limited to a relatively small number of high-frequency nouns and words from closed categories (numbers, contractions, and filled pauses).

These results are thus consistent with the differences we find in the distributional patterns of lexical categories in that, unlike high-frequency nouns, it would seem that high-frequency verbs are far less likely to be encountered outside of their argument frames (supporting the idea that verbs are encountered as arguments rather than lexical items per se).

Given that our results show that the variance in the observed forms is largely explained by the covariance in the collocate structure and that patterns of covariance are systematic, this finally leads us to the question of the systematicity in the sublexical variance: Is the distribution of the observed contrast geometric?

4.2. Distribution of Word Initial Contrast

4.2.1. Why Word Initial Contrast?

Previous work on sublexical variation shows that the structure of speech sound sequences is such that the probability of speech segments at segment transitions is not independent [49]. Gating paradigm studies have shown that the informativeness of word medial contrast is mediated by the extent to which both the preceding sentence context and word initial phonetic contrasts have minimized uncertainty about the word [53,54]. Accordingly, the entropy in sublexical contrast peaks at word initial boundaries [49]. This suggests that word initial speech contrasts may serve a distinct communicative function in context.

An initial analysis of word initial phonetic label distributions over both observed and citation forms in the corpus revealed poor fits to both power law and exponential distributions, suggesting that the aggregated distribution of the phonetic labels observed in our corpus may result from mixing the underlying communicative distributions. To examine this, we used parts-of-speech classes to provide a simple, objective method for contextually disaggregating individual communicative distributions from the mixed distribution of phonetic labels in our corpus.

4.2.2. Results

The frequency distributions of word initial phone labels were analyzed by parts-of-speech category considering the observed forms as the empirical distribution and the citation form as its model counterpart. Overall, both empirical and model distributions of phone labels show a better fit to geometric than to power-law distribution (Figure 9). However, while the fits to geometric in the model distribution show a larger departure from linearity and large differences in slope between different parts of speech, the observed phones across categories converge on nearly identical distributions with close fits to geometric (Table 1).

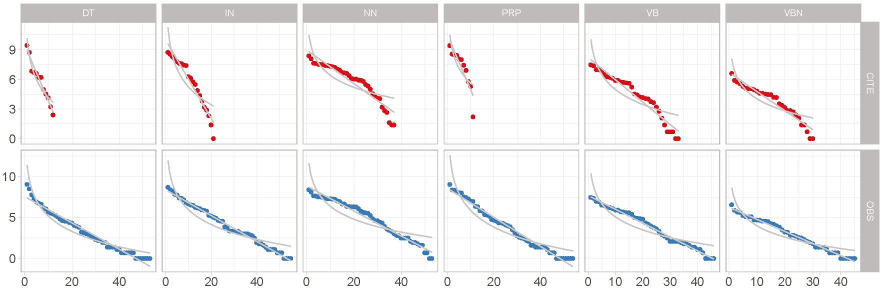


Figure 9. The distribution of word initial phonetic labels in 6 selected parts-of-speech categories: Row 1 shows the distribution presupposed by the dictionary forms, and row 2 shows the distribution of phonetic variants which were actually observed.

Table 1. The distribution of word initial phonetic labels by part-of-speech category (Penn Treebank classification) from the Buckeye Corpus of conversational speech [32]: The first two columns contain slopes from the log frequency - rank model for observed and theoretical distributions, followed by the linear model fit to log frequency - rank (R^2, geometric), model fit to log frequency - log rank (R^2, power law) and the total number of assigned phonetic labels (n_{phon}). The model distribution represents the distribution of labels presupposed by the dictionary forms, while the empirical distribution shows phonetic contrast produced by the speakers.

Part of Speech	**Slope**		R^2_{geom}		R^2_{pl}		n_{phon}	
	emp	*model*	*emp*	*model*	*emp*	*model*	*emp*	*model*
determiner	−0.16	−0.565	0.969	0.953	0.922	0.891	53	12
preposition	−0.157	−0.399	0.993	0.941	0.85	0.697	55	21
personal pronoun	−0.176	−0.554	0.984	0.802	0.87	0.593	54	11
noun, sg. or mass	−0.152	−0.171	0.975	0.895	0.726	0.621	53	37
proper noun, sg.	−0.14	−0.145	0.931	0.889	0.699	0.68	39	31
noun, plural	−0.167	−0.175	0.961	0.873	0.72	0.636	44	35
verb, base form	−0.168	−0.226	0.989	0.941	0.798	0.698	46	33
verb, past tense	−0.172	−0.244	0.972	0.944	0.784	0.768	44	31
verb, gerund/pres.part.	−0.168	−0.21	0.969	0.96	0.812	0.735	46	32
verb, past part.	−0.147	−0.193	0.991	0.934	0.828	0.716	45	30
verb, non-3rd pers.sg.pres.	−0.155	−0.249	0.991	0.977	0.777	0.751	51	31
verb, 3rd pers.sg.pres.	−0.157	−0.218	0.975	0.976	0.866	0.847	42	31
proper noun, pl.	−0.269	−0.342	0.847	0.908	0.938	0.958	12	10
function word	−0.15	−0.229	0.9731	0.8604	0.7317	0.6276	61	26
noun	−0.159	−0.199	0.9689	0.8713	0.7112	0.5797	54	40
verb	−0.164	−0.225	0.9856	0.9234	0.7406	0.6473	55	33

In both function and content words, the empirical distributions significantly improve the fit to a geometric. Importantly, despite substantial differences in the type/token ratio of the lexical classes analyzed, all of the categories have nearly identical empirical distributions with minimal differences in slopes. The exception is plural proper nouns where the data is extremely sparse (this category comprises a mere 50 tokens). Further, while we find that initial phones from several small categories (particles, modals, and filled pauses) have poor fits to either a geometric or a power law, in a similar vein, it is debatable whether these small sets of items constitute separate categories in terms of the covariate structures they populate.

Finally, we extracted time bins of initial phone duration centered by phone category to simulate an artificial set of discrete contrasts such that the simulation assumes a low-level subcategorization of phonetic contrast by duration. Again, across the parts-of-speech categories, the cumulative probability distributions of time bins show close fits to the geometric ($R^2 > 0.9662$) and poor fits to power law ($R^2 < 0.8333$).

4.2.3. Discussion

Our analysis of word initial phonetic labels across different parts-of-speech categories confirms that they are geometrically distributed. The distribution of duration time bins is also geometric. These results thus suggest that what might appear to be random variance in the production of speech sounds may actually reflect a highly systematic distribution of sublexical contrasts.

While word initial variance is observable in all part of speech categories, we find that the extent to which tokens vary is closely correlated to uncertainty that is modulated by the underlying structure of the category. Importantly, despite large differences in the extent to which initial tokens deviate from the citation form, the probability distributions of tokens arising from this variance converge on nearly identical distributional properties across parts of speech.

Finally, we observe that the distribution of word initial phones assumed by the dictionary models show poor fits to geometric and power law, illustrating that, unlike the aggregate lexical contrasts, mixtures over closed sets of items similar in structure do not result in power laws. Instead, the distributions we observe are characterized by a fast growth in the mid-frequency range.

5. General Discussion

We analyzed distributions of the grammatical, lexical, and sublexical varieties in spontaneous conversational speech produced by 40 speakers of American English [32] to assess the effects of the statistical structure of speech on the sublexical variance observed in the signal. Our results show that distributions of regularities in co-occurrence patterns, the lexical contrasts they discriminate between, and the sublexical variety observed in the articulated forms result in distributions which are consistent with previous, similar analyses of written English that satisfy many of the communicative constraints described by information theory [3].

Accordingly, these results also provide further evidence that power law distributions seen in aggregate word frequency distributions are product of mixing functionally relevant distributions that are in themselves geometric [3,8].

The distributions in the analyzed sample suggest that, unlike the codes in artificial communication systems, human speech is a highly structured system of nested communicative distributions shaped by learning. In line with the predictions of learning theory, this suggests that speech variation at positions of high uncertainty is driven by the interaction of regular structures at multiple levels of description and that this variance serves to increases the efficiency of communication by increasing the amount of contrast in signals.

Taken together, our results indicate that the variance in the pronounced forms systematically structures the uncertainty discriminated by communicative contexts, supporting the suggestion that empirical distributions of phonetic contrasts in speech are components of a larger, highly structured communication system.

Author Contributions: conceptualization, M.L. and M.R.; methodology, M.R. and M.L.; verification, M.R.; formal analysis, M.L.; writing—original draft preparation, M.L. and M.R.; writing—review and editing, M.R. and M.L.; visualization, M.L.; supervision, M.R. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge support by Open Access Publishing Fund of University of Tübingen. M.L. was funded by an European Research Council (ERC) advanced grant to Harald Baayen under grant agreement No. 742545.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Port, R.F.; Leary, A.P. Against Formal Phonology. *Linguist. Soc. Am.* **2016**, *81*, 927–964. [CrossRef]
2. Ramscar, M.; Port, R.F. How spoken languages work in the absence of an inventory of discrete units. *Lang. Commun.* **2016**, *53*, 58–74. [CrossRef]
3. Ramscar, M. Source codes in human communication. *arXiv* **2019**, arXiv:1904.03991.
4. Arnon, I.; Ramscar, M. Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition* **2012**, *122*, 292–305. [CrossRef]
5. Ramscar, M.; Hendrix, P.; Shaoul, C.; Milin, P.; Baayen, H. The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Top. Cogn. Sci.* **2014**, *6*, 5–42. [CrossRef]
6. Scribner, Sylvia and Cole, M. *The Psychology of Literacy*; Harvard University Press: Cambridge, MA, USA, 1981.
7. Munson, B.; Edwards, J.; Schellinger, S.K.; Beckman, M.E.; Meyer, M.K. Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of Vox Humana. *Clin. Linguist. Phonics* **2010**, *24*, 245–260. [CrossRef]
8. Newman, M.E.J. Power Laws, Pareto Distributions and Zipf's Law. *Contemp. Phys.* **2005**, *46*, 323–351. [CrossRef]
9. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
10. Hartley, R.V.L. Transmission of Information. *Bell Syst. Tech. J.* **1928**, *7*, 535–563. [CrossRef]
11. Ramscar, M. The empirical structure of word frequency distributions. *arXiv* **2019**, in press.

12. Blevins, J.P.; Ackerman, F.; Malouf, R.; Ramscar, M. Morphology as an adaptive discriminative system. In *Morphological Metatheory*; John Benjamins Publishing: Philadelphia, PA, USA, 2016; pp. 271–302.
13. McDonald, S.; Ramscar, M. Testing the distributioanl hypothesis: The influence of context on judgements of semantic similarity. In Proceedings of the Annual Meeting of the Cognitive Science Society, Edinburgh, UK, 1–4 August 2001.
14. Levin, B. *English Verb Classes and Alternations: A Preliminary Investigation*; University of Chicago Press: Chicago, IL, USA, 1995; Volume 1.
15. Seifart, F.; Strunk, J.; Danielsen, S.; Hartmann, I.; Pakendorf, B.; Wichmann, S.; Witzlack-Makarevich, A.; de Jong, N.H.; Bickel, B. Nouns slow down speech across structurally and culturally diverse languages. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 5720–5725. [CrossRef]
16. Ernestus, M.; Baayen, H.; Schreuder, R. The Recognition of Reduced Word Forms. *Brain Lang.* **2002**, *81*, 162–173. [CrossRef]
17. Wedel, A.; Nelson, N.; Sharp, R. The phonetic specificity of contrastive hyperarticulation in natural speech. *J. Mem. Lang.* **2018**, *100*, 61–88. [CrossRef]
18. Bell, A.; Brenier, J.M.; Gregory, M.; Girand, C.; Jurafsky, D. Predictability effects on durations of content and function words in conversational English. *J. Mem. Lang.* **2009**, *60*, 92–111. [CrossRef]
19. van Son, R.J.J.H.; Pols, L.C.W. An Acoustic Model of Communicative Efficiency in Consonants and Vowels taking into Account Context Distinctiveness. In Proceedings of the International Congress of Phonetic Sciences (ICPhS), Barcelona, Spain, 3–9 August 2003; pp. 2141–2144.
20. Sampson, G. The redundancy of self-organization as an explanation of english spelling. *Language* **2018**, *94*, e43–e47. [CrossRef]
21. Popper, K. *The Logic and Evolution of Scientific Discovery*; Routledge: Abingdon, UK, 2013.
22. Tily, H.; Gahl, S.; Arnon, I.; Snider, N.; Kothari, A.; Bresnan, J. Syntactic probabilities affect pronunciation variation in spontaneous speech. *Lang. Cogn.* **2009**, *1*, 147–165. [CrossRef]
23. Salverda, A.P.; Dahan, D.; McQueen, J.M. The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition* **2003**, *90*, 51–89. [CrossRef]
24. Kemps, R.J.J.K.; Ernestus, M.; Schreuder, R.; Harald Baayen, R. Prosodic cues for morphological complexity: The case of Dutch plural nouns. *Mem. Cogn.* **2005**, *33*, 430–446. [CrossRef]
25. Arnold, J.E.; Fagnano, M.; Tanenhaus, M.K. Disfluencies signal theee, um, new information. *J. Psycholinguist. Res.* **2003**, *32*, 25–36.:1021980931292. [CrossRef]
26. Arnold, J.E.; Kam, C.L.H.; Tanenhaus, M.K. If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* **2007**, *33*, 914–930. [CrossRef]
27. Bosker, H.R.; Pinget, A.f.; Sanders, T.; Jong, N.H.D. What makes speech sound fluent? The contributions of pauses, speed and repairs. *Lang. Test.* **2013**, *30*, 159–175. [CrossRef]
28. Bosker, H.R.; Quené, H.; Sanders, T.; De Jong, N.H. Native 'um's elicit prediction of low-frequency referents, but non-native 'um's do not. *J. Mem. Lang.* **2014**, *75*, 104–116. [CrossRef]
29. Fraundorf, S.H.; Watson, D.G. The disfluent discourse: Effects of filled pauses on recall. *J. Mem. Lang.* **2011**, *65*, 161–175. [CrossRef]
30. Cooke, M.; King, S.; Garnier, M.; Aubanel, V. The listening talker: A review of human and algorithmic context-induced modifications of speech. *Comput. Speech Lang.* **2014**, *28*, 543–571. [CrossRef]
31. Dye, M.; Milin, P.; Futrell, R.; Ramscar, M. Alternative solutions to a language design problem: The role of adjectives and gender marking in efficient communication *Top. Cogn. Sci.* **2018**, *10*, 209–224. [CrossRef]
32. Pitt, M.A.; Johnson, K.; Hume, E.; Kiesling, S.; Raymond, W. The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Commun.* **2005**, 45, 89–95. [CrossRef]
33. Dilts, P.C. Modelling Phonetic Reduction in a Corpus of Spoken English Using Random Forests and Mixed-Effects Regression. Ph.D. Thesis, University of Alberta, Edmonton, AB, USA, 2013.
34. Arbesman, S.; Strogatz, S.H.; Vitevitch, M.S. The structure of phonological networks across multiple languages. *Int. J. Bifurc. Chaos* **2010**, *20*, 679–685. [CrossRef]
35. Wood, S.N. *Generalized Additive Models*; Chapman & Hall/CRC: New York, NY, USA, 2017.
36. Hastie, T.; Tibshirani, R. Exploring the nature of covariate effects in the proportional hazards model. *Biometrics* **1990**, *46*, 1005–1016. [CrossRef]

37. Collins, M. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing—Volume 10*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; pp. 1–8.
38. DeRose, S.J. Grammatical category disambiguation by statistical optimization. *Comput. Linguist.* **1988**, *14*, 31–39.
39. Borensztajn, G.; Zuidema, W.; Bod, R. Children's grammars grow more abstract with age—Evidence from an automatic procedure for identifying the productive units of language. *Top. Cogn. Sci.* **2009**, *1*, 175–188. [CrossRef]
40. Gil, D. How much grammar does it take to sail a boat? In *Language Complexity as an Evolving Variable*; Sampson, G., Gil, D., Trudgill, P., Eds.; Oxford University Press: Oxford, UK, 2009.
41. Klingenstein, S.; Hitchcock, T.; Dedeo, S. The civilizing process in London' s Old Bailey. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 9419–9424. [CrossRef]
42. Baayen, R.H. *Word Frequency Distributions*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2001.
43. Bentz, C.; Kiela, D.; Hill, F.; Buttery, P. Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. *Corpus Linguist. Linguist. Theory* **2014**, *10*, 175–211. [CrossRef]
44. Piantadosi, S.T. Zipf's word frequency law in natural language: A critical review and future directions. *Psychon. Bull. Rev.* **2014**, *21*, 1112–1130. [CrossRef]
45. Genzel, D.; Charniak, E. Entropy rate constancy in text. In Proceedings of the 40th Annual Meeting on Association For Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 199–206.
46. Davies, M. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Lit. Linguist. Comput.* **2010**, *25*, 447–464. [CrossRef]
47. Bosker, H.R.; Reinisch, E.; Sjerps, M.J. Cognitive load makes speech sound fast, but does not modulate acoustic context effects. *J. Mem. Lang.* **2017**, *94*, 166–176. [CrossRef]
48. Fougeron, C.; Keating, P. Articulatory strengthening at the edges of prosodic domains. *J. Acoust. Soc. Am.* **1997**, *101*, 3728–3740. [CrossRef]
49. van Son, R.J.J.H.; Pols, L.C.W. How efficient is speech? *Proc. Inst. Phon. Sci.* **2003**, *25*, 171–184. doi:10.1177/1745691612459060. [CrossRef]
50. Wedel, A.; Jackson, S.; Kaplan, A. Functional Load and the Lexicon: Evidence that Syntactic Category and Frequency Relationships in Minimal Lemma Pairs Predict the Loss of Phoneme contrasts in Language Change. *Lang. Speech* **2013**, *56*, 395–417. [CrossRef]
51. Piantadosi, S.T.; Tily, H.; Gibson, E. Word lengths are optimized for efficient communication. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 3526–3529. [CrossRef]
52. Arnon, I.; Priva, U.C. Time and again: The changing effect of word and multiword frequency on phonetic duration for highly frequent sequences. *Ment. Lex.* **2015**, *9*, 377–400. [CrossRef]
53. Grosjean, F. Spoken word recognition processes and the gating paradigm. *Percept. Psychophys.* **1980**, *28*, 267–283. [CrossRef]
54. Grosjean, F.; Itzler, J. Can semantic constraint reduce the role of word frequency during spoken-word recognition? *Bull. Psychon. Soc.* **1984**, *22*, 180–182. [CrossRef]

entropy

MDPI

Article

Approximating Information Measures for Fields

Łukasz Dębowski

Institute of Computer Science, Polish Academy of Sciences, ul. Jana Kazimierza 5, 01-248 Warszawa, Poland; ldebowsk@ipipan.waw.pl; Tel.: +48-22-3800-553

Received: 20 November 2019; Accepted: 8 January 2020; Published: 9 January 2020

Abstract: We supply corrected proofs of the invariance of completion and the chain rule for the Shannon information measures of arbitrary fields, as stated by Dębowski in 2009. Our corrected proofs rest on a number of auxiliary approximation results for Shannon information measures, which may be of an independent interest. As also discussed briefly in this article, the generalized calculus of Shannon information measures for fields, including the invariance of completion and the chain rule, is useful in particular for studying the ergodic decomposition of stationary processes and its links with statistical modeling of natural language.

Keywords: Shannon information measures; fields; invariance of completion; chain rule

MSC: 94A17

1. Introduction

As it was noticed by Dębowski [1–3], a generalized calculus of Shannon information measures for arbitrary fields—initiated by Gelfand et al. [4] and later developed by Dobrushin [5], Pinsker [6], and Wyner [7]—is useful in particular for studying the ergodic decomposition of stationary processes and its links with statistical modeling of natural language. Fulfilling this need, Dębowski [1] has developed the calculus of Shannon information measures for arbitrary fields, relaxing the requirement of regular conditional probability, assumed implicitly by Dobrushin [5] and Pinsker [6]. He has done it unaware of the classical paper by Wyner [7], which pursued exactly the same idea, with some differences due to an independent interest.

Compared to exposition [7], the added value of the paper [1] was considering continuity and invariance of Shannon information measures with respect to completion of fields. Unfortunately, the proof of Theorem 2 in [1] establishing this invariance and the generalized chain rule contains some mistakes and gaps, which we have discovered recently. For this reason, in this article, we would like to provide a correction and a few new auxiliary results which may be of an independent interest. In this way, we will complete the full generalization of Shannon information measures and their properties, which was developed step-by-step by Gelfand et al. [4], Dobrushin [5], Pinsker [6], Wyner [7], and Dębowski [1]. By the way, we will also rediscuss the linguistic motivations of our results.

The preliminaries are as follows. Fix a probability space $(\Omega, \mathcal{J}, P)$. Fields are set algebras closed under finite Boolean operations, whereas σ-fields are assumed to be closed also under countable unions and products. A field is called finite if it has finitely many elements. A finite partition is a finite collection of events $\{B_j\}_{j=1}^{J} \subset \mathcal{J}$ which are disjoint and whose union equals Ω. The definition proposed by Wyner [7] and Dębowski [1] independently reads as follows:

Definition 1. *For finite partitions $\alpha = \{A_i\}_{i=1}^{I}$ and $\beta = \{B_j\}_{j=1}^{J}$ and a probability measure P, the entropy and mutual information are defined as*

$$H_P(\alpha) := \sum_{i=1}^{I} P(A_i) \log \frac{1}{P(A_i)}, \qquad I_P(\alpha;\beta) := \sum_{i=1}^{I}\sum_{j=1}^{J} P(A_i \cap B_j) \log \frac{P(A_i \cap B_j)}{P(A_i)P(B_j)}. \tag{1}$$

Subsequently, for an arbitrary field $\mathcal{C}$ and finite partitions α and β, we define the pointwise conditional entropy and mutual information as

$$H_P(\alpha||\mathcal{C}) := H_{P(\cdot|\mathcal{C})}(\alpha), \qquad I_P(\alpha;\beta||\mathcal{C}) := I_{P(\cdot|\mathcal{C})}(\alpha;\beta), \tag{2}$$

where $P(E|\mathcal{C})$ is the conditional probability of event $E \in \mathcal{J}$ with respect to the smallest complete σ-field containing $\mathcal{C}$. Subsequently, for arbitrary fields $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$, the (average) conditional entropy and mutual information are defined as

$$H_P(\mathcal{A}|\mathcal{C}) := \sup_{\alpha \subset \mathcal{A}} \boldsymbol{E}_P H_P(\alpha||\mathcal{C}), \qquad I_P(\mathcal{A};\mathcal{B}|\mathcal{C}) := \sup_{\alpha \subset \mathcal{A}, \beta \subset \mathcal{B}} \boldsymbol{E}_P I(\alpha;\beta||\mathcal{C}), \tag{3}$$

where the supremum is taken over all finite subpartitions and $\boldsymbol{E}_P X := \int X dP$ is the expectation. Finally, we define the unconditional entropy $H_P(\mathcal{A}) := H_P(\mathcal{A}|\{\emptyset,\Omega\})$ and mutual information $I_P(\mathcal{A};\mathcal{B}) := I_P(\mathcal{A};\mathcal{B}|\{\emptyset,\Omega\})$, as it is generally done in information theory. When the probability measure P is clear from the context, we omit subscript P from all above notations.

Although the above measures, called Shannon information measures, have usually been discussed for σ-fields, the defining equations (3) also make sense for fields. We observe a number of identities, such as $H(\mathcal{A}) = I(\mathcal{A};\mathcal{A})$ and $H(\mathcal{A}|\mathcal{C}) = I(\mathcal{A};\mathcal{A}|\mathcal{C})$. It is important to stress that Definition 1, in contrast to the earlier expositions by Dobrushin [5] and Pinsker [6], is simpler—as it applies one Radon–Nikodym derivative less—and does not require regular conditional probability, i.e., it does not demand that conditional distribution $(P(E|\mathcal{C}))_{E\in\mathcal{J}}$ be a probability measure almost surely. In fact, the expressions on the right-hand sides of the equations in (3) are defined for all $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. No problems arise when conditional probability is not regular since conditional distribution $(P(E|\mathcal{C}))_{E\in\mathcal{E}}$ restricted to a finite field $\mathcal{E}$ is a probability measure almost surely [8] (Theorem 33.2).

We should admit that in the context of statistical language modeling, the respective probability space is countably generated so regular conditional probability is guaranteed to exist. Thus, for linguistic applications, one might think that expositions [5,6] are sufficient, although for a didactic reason, the approaches proposed by Wyner [7] and Dębowski [1] lead to a simpler and more general calculus of Shannon information measures. Yet, there is a more important reason for Definition 1. Namely, to discuss the ergodic decomposition of entropy rate and excess entropy—some highly relevant results for statistical language modeling, developed in [1] and to be briefly recalled in Section 3—we need the invariance of Shannon information measures with respect to completion of fields. But within the framework of Dobrushin [5] and Pinsker [6], such invariance of completion does not hold for strongly nonergodic processes, which seem to arise quite naturally in statistical modeling of natural language [1–3]. Thus, the approach proposed by Wyner [7] and Dębowski [1] is in fact indispensable.

Thus, let us inspect the problem of invariance of Shannon information measures with respect to completion of fields. A σ-field is called complete, with respect to a given probability measure P, if it contains all sets of outer P-measure 0. Let $\sigma(\mathcal{A})$ denote the intersection of all complete σ-fields containing class $\mathcal{A}$, i.e., $\sigma(\mathcal{A})$ is the completion of the generated σ-field. Let $\mathcal{A} \wedge \mathcal{B}$ denote the intersection of all fields that contain $\mathcal{A}$ and $\mathcal{B}$. Assuming Definition 1, the following statement has been claimed true by Dębowski [1] (Theorem 2):

Theorem 1. *Let $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ be subfields of $\mathcal{J}$.*

1. $I(\mathcal{A};\mathcal{B}|\mathcal{C}) = I(\mathcal{A};\sigma(\mathcal{B})|\mathcal{C}) = I(\mathcal{A};\mathcal{B}|\sigma(\mathcal{C}))$ *(invariance of completion);*
2. $I(\mathcal{A};\mathcal{B}\wedge\mathcal{C}|\mathcal{D}) = I(\mathcal{A};\mathcal{B}|\mathcal{D}) + I(\mathcal{A};\mathcal{C}|\mathcal{B}\wedge\mathcal{D})$ *(chain rule).*

The property stated in Theorem 1. 1 will be referred to as the invariance of completion. It was not discussed by Wyner [7]. The property stated in Theorem 1. 2 is usually referred to as the chain rule or the polymatroid identity. It was proved independently by Wyner [7].

As we have mentioned, the invariance of completion is crucial to prove the ergodic decomposition of the entropy rate and excess entropy of stationary processes. But the proof of the invariance of completion given by Dębowski [1] contains a mistake in the order of quantifiers, and the respective proof of the chain rule is too laconic and contains a gap. For this reason, we would like to supplement the corrected proofs in this article. As we have mentioned, the chain rule was proved by Wyner [7], using an approximation result by Dobrushin [5] and Pinsker [6]. For completeness, we would like to provide a different proof of this approximation result—which follows easily from the invariance of completion—and to supply proofs of both parts of Theorem 1.

The corrected proofs of Theorem 1, to be presented in Section 2, are much longer than the original proofs by Dębowski [1]. In particular, for the sake of proving Theorem 1, we will discuss a few other approximation results, which seem to be of an independent interest. To provide more context for our statements, in Section 3, we will also recall the ergodic decomposition of excess entropy and its application to statistical language modeling.

2. Proofs

Let us write $\mathcal{B}_n \uparrow \mathcal{B}$ for a sequence $(\mathcal{B}_n)_{n\in\mathbb{N}}$ of fields such that $\mathcal{B}_1 \subset \mathcal{B}_2 \subset \cdots \subset \mathcal{B} = \bigcup_{n\in\mathbb{N}} \mathcal{B}_n$. ($\mathcal{B}$ need not be a σ-field.) Our proof of Theorem 1 will rest on a few approximation results and this statement by Dębowski [1] (Theorem 1):

Theorem 2. *Let $\mathcal{A}$, $\mathcal{B}$, $\mathcal{B}_n$, and $\mathcal{C}$ be subfields of $\mathcal{J}$.*

1. $I(\mathcal{A};\mathcal{B}|\mathcal{C}) = I(\mathcal{B};\mathcal{A}|\mathcal{C})$*;*
2. $I(\mathcal{A};\mathcal{B}|\mathcal{C}) \geq 0$ *with the equality if and only if* $P(A\cap B|\mathcal{C}) = P(A|\mathcal{C})P(B|\mathcal{C})$ *almost surely for all* $A\in\mathcal{A}$ *and* $B\in\mathcal{B}$*;*
3. $I(\mathcal{A};\mathcal{B}|\mathcal{C}) \leq \min(H(\mathcal{A}|\mathcal{C}), H(\mathcal{B}|\mathcal{C}))$*;*
4. $I(\mathcal{A};\mathcal{B}_1|\mathcal{C}) \leq I(\mathcal{A};\mathcal{B}_2|\mathcal{C})$ *if* $\mathcal{B}_1 \subset \mathcal{B}_2$*;*
5. $I(\mathcal{A};\mathcal{B}_n|\mathcal{C}) \uparrow I(\mathcal{A};\mathcal{B}|\mathcal{C})$ *for* $\mathcal{B}_n \uparrow \mathcal{B}$*.*

Let $A^c = \Omega \setminus A$. Subsequently, let us denote the symmetric difference

$$A\triangle B := (A\setminus B)\cup(B\setminus A) = (A\cup B)\setminus(A\cap B). \tag{4}$$

Symmetric difference satisfies the following identities, which will be used:

$$A^c\triangle B^c = A\triangle B, \tag{5}$$

$$A\triangle B \subset (A\triangle C)\cup(C\triangle B), \tag{6}$$

$$(A\setminus C)\triangle B \subset (A\triangle B)\cup(C\cap B), \tag{7}$$

$$\left(\bigcup_{i\in C} A_i\right) \triangle \left(\bigcup_{i\in C} B_i\right) \subset \bigcup_{i\in C}(A_i\triangle B_i). \tag{8}$$

Moreover, we will apply the Bonferroni inequalities

$$0 \leq \sum_{1\leq i\leq n} P(A_i) - P\left(\bigcup_{1\leq i\leq n} A_i\right) \leq \sum_{1\leq i<j\leq n} P(A_i\cap A_j) \tag{9}$$

and inequality $P(A) \leq P(B) + P(A \triangle B)$.

In the following, we will derive the necessary approximation results. Our point of departure is the following folklore fact.

Theorem 3 (approximation of σ-fields)**.** *For any field $\mathcal{K}$ and any event $G \in \sigma(\mathcal{K})$, there is a sequence of events $K_1, K_2, \cdots \in \mathcal{K}$ such that*

$$\lim_{n \to \infty} P(G \triangle K_n) = 0. \tag{10}$$

Proof. Denote the class of sets G that satisfy (10) as $\mathcal{G}$. It is sufficient to show that $\mathcal{G}$ is a complete σ-field that contains the field $\mathcal{K}$. Clearly, all $G \in \mathcal{K}$ satisfy (10) so $\mathcal{G} \supset \mathcal{K}$. Now, we verify the conditions for $\mathcal{G}$ to be a σ-field.

1. We have $\Omega \in \mathcal{K}$. Hence, $\Omega \in \mathcal{G}$.
2. For $A \in \mathcal{G}$, consider $K_1, K_2, \cdots \in \mathcal{K}$ such that $\lim_{n \to \infty} P(A \triangle K_n) = 0$. Then, $A \triangle K_n = A^c \triangle K_n^c$, where $K_1^c, K_2^c, \cdots \in \mathcal{K}$. Hence, $A^c \in \mathcal{G}$.
3. For $A_1, A_2, \cdots \in \mathcal{G}$, consider events $K_i^n \in \mathcal{K}$ such that $P(A_i \triangle K_i^n) \leq 2^{-n}$. Then,

$$P\left(\left(\bigcap_{i=1}^{n} A_i\right) \triangle \left(\bigcap_{i=1}^{n} K_i^{i+n}\right)\right) \leq \sum_{i=1}^{n} P(A_i \triangle K_i^{i+n}) \leq 2^{-n}. \tag{11}$$

Moreover,

$$P\left(\left(\bigcap_{i=1}^{\infty} A_i\right) \triangle \left(\bigcap_{i=1}^{n} A_i\right)\right) = P\left(\bigcap_{i=1}^{n} A_i\right) - P\left(\bigcap_{i=1}^{\infty} A_i\right). \tag{12}$$

Hence,

$$\begin{aligned} & P\left(\left(\bigcap_{i=1}^{\infty} A_i\right) \triangle \left(\bigcap_{i=1}^{n} K_i^{i+n}\right)\right) \\ & \leq P\left(\left(\bigcap_{i=1}^{\infty} A_i\right) \triangle \left(\bigcap_{i=1}^{n} A_i\right)\right) + P\left(\left(\bigcap_{i=1}^{n} A_i\right) \triangle \left(\bigcap_{i=1}^{n} K_i^{i+n}\right)\right) \\ & \leq P\left(\bigcap_{i=1}^{n} A_i\right) - P\left(\bigcap_{i=1}^{\infty} A_i\right) - 2^{-n}, \end{aligned} \tag{13}$$

which tends to 0 for n going to infinity. Since $\bigcap_{i=1}^{n} K_i^{i+n} \in \mathcal{K}$, we thus obtain that $\bigcap_{i=1}^{\infty} A_i \in \mathcal{G}$.

Completeness of σ-field $\mathcal{G}$ is straightforward since, for any $A \in \mathcal{G}$ and $P(A \triangle A') = 0$, we obtain $A' \in \mathcal{G}$ using the same sequence of approximating events in field $\mathcal{K}$ as for event A. □

The second approximation result is the following bound:

Theorem 4 (continuity of entropy)**.** *Fix an $\epsilon \in (0, e^{-1}]$ and a field $\mathcal{C}$. For finite partitions $\alpha = \{A_i\}_{i=1}^{I}$ and $\alpha' = \{A_i'\}_{i=1}^{I}$ such that $P(A_i \triangle A_i') \leq \epsilon$ for all $i \in \{1, \ldots, I\}$, we have*

$$\left|H(\alpha|\mathcal{C}) - H(\alpha'|\mathcal{C})\right| \leq I\sqrt{\epsilon} \log \frac{I}{\sqrt{\epsilon}}. \tag{14}$$

Proof. We have the expectation $\int P(A_i \triangle A_i' | \mathcal{C}) dP = P(A_i \triangle A_i') \leq \epsilon$. Hence, by the Markov inequality we obtain

$$P(P(A_i \triangle A_i' | \mathcal{C}) \geq \sqrt{\epsilon}) \leq \sqrt{\epsilon}. \tag{15}$$

Denote

$$B = \left(P(A_i \triangle A_i'|\mathcal{C}) < \sqrt{\epsilon}\right) \text{ for all } i \in \{1, \dots, I\}\right) . \tag{16}$$

From the Bonferroni inequality, we obtain $P(B^c) \leq I\sqrt{\epsilon}$. Subsequently, we observe that $|H(\alpha||\mathcal{C}) - H(\alpha'||\mathcal{C})| \leq \log I$ holds almost surely. Hence,

$$\begin{aligned}\left|H(\alpha|\mathcal{C}) - H(\alpha'|\mathcal{C})\right| &= \left|\int \left[H(\alpha|\mathcal{C}) - H(\alpha'|\mathcal{C})\right] dP\right| \\ &\leq P(B^c)\log I + \int_B \left|H(\alpha||\mathcal{C}) - H(\alpha'||\mathcal{C})\right| dP \\ &\leq I\sqrt{\epsilon}\log I + \int_B \left|H(\alpha||\mathcal{C}) - H(\alpha'||\mathcal{C})\right| dP. \end{aligned} \tag{17}$$

Function $-x\log x$ is subadditive and increasing for $x \in (0, e^{-1}]$. In particular, we have $|(x+y)\log(x+y) - x\log x| \leq -y\log y$ for $x, y \geq 0$. Thus, on the event B we obtain

$$\begin{aligned}\left|H(\alpha||\mathcal{C}) - H(\alpha'||\mathcal{C})\right| &= \left|\sum_{i=1}^{I} P(A_i'|\mathcal{C})\log P(A_i'|\mathcal{C}) - \sum_{i=1}^{I} P(A_i|\mathcal{C})\log P(A_i|\mathcal{C})\right| \\ &\leq -\sum_{i=1}^{I} \left|P(A_i|\mathcal{C}) - P(A_i'|\mathcal{C})\right| \log\left|P(A_i|\mathcal{C}) - P(A_i'|\mathcal{C})\right| \\ &\leq -\sum_{i=1}^{I} P(A_i \triangle A_i'|\mathcal{C})\log P(A_i \triangle A_i'|\mathcal{C}) \\ &\leq -I\sqrt{\epsilon}\log\sqrt{\epsilon} \end{aligned} \tag{18}$$

Plugging (18) into (17) yields the claim. □

Now, we can prove the invariance of completion. Note that

$$I(\alpha;\beta|\mathcal{C}) = H(\alpha|\mathcal{C}) + H(\beta|\mathcal{C}) - H(\alpha \wedge \beta|\mathcal{C}). \tag{19}$$

Proof of Theorem 1. 1 (invariance of completion): Consider some measurable fields $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. We are going to demonstrate

$$I(\mathcal{A};\mathcal{B}|\mathcal{C}) = I(\mathcal{A};\sigma(\mathcal{B})|\mathcal{C}) = I(\mathcal{A};\mathcal{B}|\sigma(\mathcal{C})). \tag{20}$$

Equality $I(\mathcal{A};\mathcal{B}|\mathcal{C}) = I(\mathcal{A};\mathcal{B}|\sigma(\mathcal{C}))$ is straightforward since $P(A|\mathcal{C}) = P(A|\sigma(\mathcal{C}))$ almost surely for all $A \in \mathcal{J}$. It remains to prove $I(\mathcal{A};\mathcal{B}|\mathcal{C}) = I(\mathcal{A};\sigma(\mathcal{B})|\mathcal{C})$. For this goal, it suffices to show that for any $\epsilon > 0$ and any finite partitions $\alpha \subset \mathcal{A}$ and $\beta' \subset \sigma(\mathcal{B})$ there exists a finite partition $\beta \subset \mathcal{B}$ such that

$$\left|I(\alpha;\beta|\mathcal{C}) - I(\alpha;\beta'|\mathcal{C})\right| < \epsilon. \tag{21}$$

Fix then some $\epsilon > 0$ and finite partitions $\alpha := \{A_i\}_{i=1}^{I} \subset \mathcal{A}$ and $\beta' := \left\{B_j'\right\}_{j=1}^{J} \subset \sigma(\mathcal{B})$. Invoking Theorem 3, we know that for each $\eta > 0$ there exists a class of sets $\{C_j\}_{j=1}^{J} \subset \mathcal{B}$ which need not be a partition, such that

$$P(C_j \triangle B_j') \leq \eta \tag{22}$$

for all $j \in \{1, \dots, J\}$. Let us put $B_{J+1}' := \emptyset$ and let us construct sets $D_0 := \emptyset$ and $D_j := \bigcup_{k=1}^{j} C_k$ for $j \in \{1, \dots, J\}$. Subsequently, we put $B_j := C_j \setminus D_{j-1}$ for $j \in \{1, \dots, J\}$ and $B_{J+1} := \Omega \setminus D_J$. In this way, we obtain a partition $\beta := \{B_j\}_{j=1}^{J+1} \subset \mathcal{B}$.

The next step of the proof is showing an analogue of bound (22) for partitions β and β'. To begin, for $j \in \{1, \dots, J\}$, we have

$$\begin{aligned} P(B_j \triangle B'_j) &= P((C_j \setminus D_{j-1}) \triangle B'_j) \le P(C_j \triangle B'_j) + P(D_{j-1} \cap B'_j) \\ &\le \eta + \sum_{k=1}^{j-1} P(C_k \cap B'_j) \\ &\le \eta + \sum_{k=1}^{j-1} \left[P(B'_k \cap B'_j) + P((C_k \cap B'_j) \triangle (B'_k \cap B'_j)) \right] \\ &\le \eta + \sum_{k=1}^{j-1} \left[0 + P(C_k \triangle B'_k) \right] \le j\eta. \end{aligned} \tag{23}$$

Now, we observe for $j, k \in \{1, \dots, J\}$ and $j \ne k$ that

$$P(C_j) \ge P(B'_j) - P(C_j \triangle B'_j) \ge P(B'_j) - \eta \tag{24}$$

$$\begin{aligned} P(C_j \cap C_k) &\le P(B'_j \cap B'_k) + P((C_j \cap C_k) \triangle (B'_j \cap B'_k)) \\ &\le 0 + P(C_j \triangle B'_j) + P(C_k \triangle B'_k) \le 2\eta. \end{aligned} \tag{25}$$

Hence, by the Bonferroni inequality we derive

$$\begin{aligned} P(B_{J+1} \triangle B'_{J+1}) &= P((\Omega \setminus D_J) \triangle \emptyset) = P(\Omega \setminus D_J) = 1 - P(D_J) \\ &\le 1 - \sum_{1 \le j \le J} P(C_j) + \sum_{1 \le j < k \le J} P(C_j \cap C_k) \\ &\le 1 - \sum_{1 \le j \le J} P(B'_j) + J\eta + \sum_{1 \le j < k \le J} 2\eta = J^2\eta. \end{aligned} \tag{26}$$

Resuming our bounds, we obtain

$$P((A_i \cap B_j) \triangle (A_i \cap B'_j)) \le P(B_j \triangle B'_j) \le J^2\eta \tag{27}$$

for all $i \in \{1, \dots, I\}$ and $j \in \{1, \dots, J+1\}$. Then, invoking Theorem 4 yields

$$\begin{aligned} \left| I(\alpha; \beta | \mathcal{C}) - I(\alpha; \beta' | \mathcal{C}) \right| &\le \left| H(\alpha \wedge \beta | \mathcal{C}) - H(\alpha \wedge \beta' | \mathcal{C}) \right| + \left| H(\beta | \mathcal{C}) - H(\beta' | \mathcal{C}) \right| \\ &\le I(J+1)\sqrt{J^2\eta} \log \frac{I(J+1)}{\sqrt{J^2\eta}} + (J+1)\sqrt{J^2\eta} \log \frac{J+1}{\sqrt{J^2\eta}}. \end{aligned} \tag{28}$$

Taking η sufficiently small, we obtain (21), which is the desired claim. □

Some consequence of the above result is this approximation result proved by Dobrushin [5] and Pinsker [6] and used by Wyner [7] to demonstrate the chain rule. Applying the invariance of completion, we supply a different proof than Dobrushin [5] and Pinsker [6].

Theorem 5 (split of join)**.** *Let $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ be subfields of $\mathcal{J}$. We have*

$$I(\mathcal{A}; \mathcal{B} \wedge \mathcal{C} | \mathcal{D}) = \sup_{\alpha \subset \mathcal{A}, \beta \subset \mathcal{B}, \gamma \subset \mathcal{C}} \boldsymbol{E}\, I(\alpha; \beta \wedge \gamma || \mathcal{D}), \tag{29}$$

where the supremum is taken over all finite subpartitions.

Proof. Define class

$$\mathcal{E} := \bigcup_{\beta \subset \mathcal{B}, \gamma \subset \mathcal{C}} \sigma(\beta \wedge \gamma). \tag{30}$$

It can be easily verified that $\mathcal{E}$ is a field such that $\sigma(\mathcal{E}) = \sigma(\mathcal{B} \wedge \mathcal{C})$. Thus, for all finite partitions $\beta \subset \mathcal{B}$ and $\gamma \subset \mathcal{C}$ we have $\beta \wedge \gamma \subset \mathcal{E}$. Moreover, by definition of $\mathcal{E}$, for each finite partition $\varepsilon \subset \mathcal{E}$ there exists finite partitions $\beta \subset \mathcal{B}$ and $\gamma \subset \mathcal{C}$ such that partition $\beta \wedge \gamma$ is finer than ε. Hence, by Theorem 2.4, we obtain in this case,

$$\mathbf{E}\, I(\alpha; \varepsilon || \mathcal{D}) \le \mathbf{E}\, I(\alpha; \beta \wedge \gamma || \mathcal{D}) \le I(\alpha; \mathcal{E} | \mathcal{D}). \tag{31}$$

In consequence, by Theorem 1. 1, we obtain the claim

$$\begin{aligned} I(\mathcal{A}; \mathcal{B} \wedge \mathcal{C} | \mathcal{D}) = I(\mathcal{A}; \mathcal{E} | \mathcal{D}) &= \sup_{\alpha \subset \mathcal{A}, \varepsilon \subset \mathcal{E}} \mathbf{E}\, I(\alpha; \varepsilon || \mathcal{D}) \\ &= \sup_{\alpha \subset \mathcal{A}, \beta \subset \mathcal{B}, \gamma \subset \mathcal{C}} \mathbf{E}\, I(\alpha; \beta \wedge \gamma || \mathcal{D}). \end{aligned} \tag{32}$$

□

The final approximation result which we need to prove the chain rule is as follows:

Theorem 6 (convergence of conditioning)**.** *Let $\alpha = \{A_i\}_{i=1}^I$ be a finite partition and let $\mathcal{C}$ be a field. For each $\epsilon > 0$, there exists a finite partition $\gamma' \subset \sigma(\mathcal{C})$ such that for any partition $\gamma \subset \sigma(\mathcal{C})$ finer than γ' we have*

$$|H(\alpha|\mathcal{C}) - H(\alpha|\gamma)| \le \epsilon. \tag{33}$$

Proof. Fix an $\epsilon > 0$. For each $n \in \mathbb{N}$ and $A \in \mathcal{J}$, partition

$$\gamma_A := \{((k-1)/n < P(A|\mathcal{C}) \le k/n) : k \in \{0, 1, \dots, n\}\} \tag{34}$$

is finite and belongs to $\sigma(\mathcal{C})$. If we consider partition $\gamma' := \bigwedge_{i=1}^I \gamma_{A_i}$, it remains finite and still satisfies $\gamma' \subset \sigma(\mathcal{C})$. Let a partition $\gamma \subset \sigma(\mathcal{C})$ be finer than γ'. Then,

$$|P(A_i|\mathcal{C}) - P(A_i|\gamma)| \le 1/n \tag{35}$$

almost surely for all $i \in \{1, \dots, I\}$. We also observe

$$|H(\alpha|\mathcal{C}) - H(\alpha|\gamma)| \le \int |H(\alpha||\mathcal{C}) - H(\alpha||\gamma)|\, dP. \tag{36}$$

We recall that function $-x \log x$ is subadditive and increasing for $x \in (0, e^{-1}]$. In particular, we have $|(x+y)\log(x+y) - x \log x| \le -y \log y$ for $x, y \ge 0$. Hence, for $n \ge e$ we obtain almost surely

$$\begin{aligned} |H(\alpha||\mathcal{C}) - H(\alpha||\gamma)| &= \left| \sum_{i=1}^I P(A_i|\mathcal{C}) \log P(A_i|\mathcal{C}) - \sum_{i=1}^I P(A_i|\gamma) \log P(A_i|\gamma) \right| \\ &\le - \sum_{i=1}^I |P(A_i|\mathcal{C}) - P(A_i|\gamma)| \log |P(A_i|\mathcal{C}) - P(A_i|\gamma)| \\ &\le \frac{I \log n}{n}. \end{aligned} \tag{37}$$

Taking n so large that $n^{-1} I \log n \le \epsilon$ yields the claim. □

Taking the above into account, we can demonstrate the chain rule. Our proof essentially follows the ideas of Wyner [7], except for invoking Theorem 6.

Proof of Theorem 1. 2 (chain rule): Let $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ be arbitrary fields, and let α, β, γ, and δ be finite partitions. The point of our departure is the chain rule for finite partitions [9] (Equation 2.60)

$$I(\alpha;\beta\wedge\gamma)=I(\alpha;\beta)+I(\alpha;\gamma|\beta). \tag{38}$$

By Definition 1 and Theorems 1. 1, 5, and 6, conditional mutual information $I(\mathcal{A};\mathcal{B}|\mathcal{C})$ can be approximated by $I(\alpha;\beta|\gamma)$, where we take appropriate limits of refined finite partitions with a certain care.

In particular, by Theorems 1. 1, 5, and 6, taking sufficiently fine finite partitions of arbitrary fields $\mathcal{B}$ and $\mathcal{C}$, the chain rule (38) for finite partitions implies

$$I(\alpha;\mathcal{B}\wedge\mathcal{C})=I(\alpha;\mathcal{B})+I(\alpha;\mathcal{C}|\mathcal{B}), \tag{39}$$

where all expressions are finite. Hence, we also obtain

$$\begin{aligned}0&=[I(\alpha;\mathcal{B}\wedge\mathcal{C}\wedge\mathcal{D})-I(\alpha;\mathcal{D})-I(\alpha;\mathcal{B}\wedge\mathcal{C}|\mathcal{D})]\\&\quad-[I(\alpha;\mathcal{B}\wedge\mathcal{D})-I(\alpha;\mathcal{D})-I(\alpha;\mathcal{B}|\mathcal{D})]\\&\quad-[I(\alpha;\mathcal{B}\wedge\mathcal{C}\wedge\mathcal{D})-I(\alpha;\mathcal{B}\wedge\mathcal{D})-I(\alpha;\mathcal{C}|\mathcal{B}\wedge\mathcal{D})]\\&=I(\alpha;\mathcal{B}|\mathcal{D})+I(\alpha;\mathcal{C}|\mathcal{B}\wedge\mathcal{D})-I(\alpha;\mathcal{B}\wedge\mathcal{C}|\mathcal{D}),\end{aligned}$$

where all expressions are finite. Having established the above claim for a finite partition α, we generalize it to

$$I(\mathcal{A};\mathcal{B}\wedge\mathcal{C}|\mathcal{D})=I(\mathcal{A};\mathcal{B}|\mathcal{D})+I(\mathcal{A};\mathcal{C}|\mathcal{B}\wedge\mathcal{D}) \tag{40}$$

for an arbitrary field $\mathcal{A}$, taking its appropriately fine finite partitions. □

3. Applications

This section borrows its statements largely from Dębowski [1–3] and is provided only to sketch some context for our research and justify its applicability to statistical language modeling. Let $(X_i)_{i\in\mathbb{Z}}$ be a two-sided infinite stationary process over a countable alphabet $\mathbb{X}$ on a probability space $(\mathbb{X}^{\mathbb{Z}},\mathcal{X}^{\mathbb{Z}},P)$, where $X_k((\omega_i)_{i\in\mathbb{Z}}):=\omega_k$. We denote random blocks $X_j^k:=(X_i)_{j\le i\le k}$ and complete σ-fields $\mathcal{G}_j^k:=\sigma(X_j^k)$ generated by them. By the generalized calculus of Shannon information measures, i.e., Theorems 1 and 2, we can define the entropy rate h_P and the excess entropy E_P of process $(X_i)_{i\in\mathbb{Z}}$ as

$$h_P:=\lim_{n\to\infty}H_P(\mathcal{G}_0|\mathcal{G}_{-n}^{-1})=H_P(\mathcal{G}_0|\mathcal{G}_{-\infty}^{-1})\text{ if }\mathbb{X}\text{ is finite,} \tag{41}$$

$$E_P:=\lim_{n\to\infty}I_P(\mathcal{G}_{-n}^{-1};\mathcal{G}_0^{n-1})=I_P(\mathcal{G}_{-\infty}^{-1};\mathcal{G}_0^{\infty}), \tag{42}$$

see [10] for more background.

Let $T((\omega_i)_{i\in\mathbb{Z}}):=(\omega_{i+1})_{i\in\mathbb{Z}}$ be the shift operation and let $\mathcal{I}:=\left\{A\in\mathcal{X}^{\mathbb{Z}}:T^{-1}(A)=A\right\}$ be the invariant σ-field. By the Birkhoff ergodic theorem [11], we have $\sigma(\mathcal{I})\subset\sigma(\mathcal{G}_{-\infty})\cap\sigma(\mathcal{G}_{\infty})$ for the tail σ-fields $\mathcal{G}_{-\infty}:=\bigcap_{n=1}^{\infty}\mathcal{G}_{-\infty}^{-n}$ and $\mathcal{G}_{\infty}:=\bigcap_{n=1}^{\infty}\mathcal{G}_{n}^{\infty}$. Hence, by Theorems 1 and 2 we further obtain expressions

$$h_P=H_P(\mathcal{G}_0|\mathcal{G}_{-\infty}^{-1})=H_P(\mathcal{G}_0|\mathcal{G}_{-\infty}^{-1}\wedge\mathcal{I})\text{ if }\mathbb{X}\text{ is finite,} \tag{43}$$

$$E_P=I_P(\mathcal{G}_{-\infty}^{-1};\mathcal{G}_0^{\infty})=H_P(\mathcal{I})+I_P(\mathcal{G}_{-\infty}^{-1};\mathcal{G}_0^{\infty}|\mathcal{I}). \tag{44}$$

Denoting the conditional probability $F(A) := P(A|\mathcal{I})$, which is a random stationary ergodic measure by the ergodic decomposition theorem [12], we notice that $H_P(\mathcal{G}_0|\mathcal{G}_{-\infty}^{-1} \wedge \mathcal{I}) = \mathbf{E}_P H_F(\mathcal{G}_0|\mathcal{G}_{-\infty}^{-1})$ and $I_P(\mathcal{G}_{-\infty}^{-1};\mathcal{G}_0^{\infty}|\mathcal{I}) = \mathbf{E}_P I_F(\mathcal{G}_{-\infty}^{-1};\mathcal{G}_0^{\infty})$, and consequently we obtain the ergodic decomposition of the entropy rate and excess entropy, which reads

$$h_P = \mathbf{E}_P h_F \text{ if } \mathbb{X} \text{ is finite,} \tag{45}$$

$$E_P = H_P(\mathcal{I}) + \mathbf{E}_P E_F. \tag{46}$$

Formulae (45) and (46) were derived by Gray and Davisson [13] and Dębowski [1] respectively. The ergodic decomposition of the entropy rate (45) states that a stationary process is asymptotically deterministic, i.e., $h_P = 0$, if and only if almost all its ergodic components are asymptotically deterministic, i.e., $h_F = 0$ almost surely. In contrast, the ergodic decomposition of the excess entropy (46) states that a stationary process is infinitary, i.e., $E_P = \infty$, if some of its ergodic components are infinitary, i.e., $E_F = \infty$ with a nonzero probability, or if $H_P(\mathcal{I}) = \infty$, i.e., if the process is strongly nonergodic in particular, see [14,15].

The linguistic interpretation of the above results is as follows. There is a hypothesis by Hilberg [16] that the excess entropy of natural language is infinite. This hypothesis can be partly confirmed by the original estimates of conditional entropy by Shannon [17], by the power-law decay of the estimates of the entropy rate given by the PPM compression algorithm [18], by the approximately power-law growth of vocabulary called Heaps' or Herdan's law [2,3,19,20], and by some other experiments applying neural statistical language models [21,22]. In parallel, Dębowski [1–3] supposed that the very large excess entropy in natural language may be caused by the fact that texts in natural language describe some relatively slowly evolving and very complex reality. Indeed, it can be mathematically proved that if the abstract reality described by random texts is unchangeable and infinitely complex, then the resulting stochastic process is strongly nonergodic, i.e., $H_P(\mathcal{I}) = \infty$ in particular [1–3]. Consequently, its excess entropy is infinite by formula (46). We suppose that a similar mechanism may work for natural language, see [23–26] for further examples of abstract stochastic mechanisms leading to infinitary processes.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Dębowski, Ł. A general definition of conditional information and its application to ergodic decomposition. *Stat. Probab. Lett.* **2009**, *79*, 1260–1268. [CrossRef]
2. Dębowski, Ł. On the Vocabulary of Grammar-Based Codes and the Logical Consistency of Texts. *IEEE Trans. Inf. Theory* **2011**, *57*, 4589–4599. [CrossRef]
3. Dębowski, Ł. Is Natural Language a Perigraphic Process? The Theorem about Facts and Words Revisited. *Entropy* **2018**, *20*, 85. [CrossRef]
4. Gelfand, I.M.; Kolmogorov, A.N.; Yaglom, A.M. Towards the general definition of the amount of information. *Dokl. Akad. Nauk. SSSR* **1956**, *111*, 745–748. (In Russian)
5. Dobrushin, R.L. A general formulation of the fundamental Shannon theorems in information theory. *Uspekhi Mat. Nauk.* **1959**, *14*, 3–104. (In Russian)
6. Pinsker, M.S. *Information and Information Stability of Random Variables and Processes*; Holden-Day: San Francisco, CA, USA, 1964.
7. Wyner, A.D. A definition of conditional mutual information for arbitrary ensembles. *Inf. Control.* **1978**, *38*, 51–59. [CrossRef]
8. Billingsley, P. *Probability and Measure*; John Wiley: New York, NY, USA, 1979.
9. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley: New York, NY, USA, 1991.
10. Crutchfield, J.P.; Feldman, D.P. Regularities unseen, randomness observed: The entropy convergence hierarchy. *Chaos* **2003**, *15*, 25–54. [CrossRef]

11. Birkhoff, G.D. Proof of the ergodic theorem. *Proc. Natl. Acad. Sci. USA* **1932**, *17*, 656–660. [CrossRef]
12. Rokhlin, V.A. On the fundamental ideas of measure theory. *Am. Math. Soc. Transl. Ser. 1* **1962**, *10*, 1–54.
13. Gray, R.M.; Davisson, L.D. The ergodic decomposition of stationary discrete random processses. *IEEE Trans. Inf. Theory* **1974**, *20*, 625–636. [CrossRef]
14. Löhr, W. Properties of the Statistical Complexity Functional and Partially Deterministic HMMs. *Entropy* **2009**, *11*, 385–401. [CrossRef]
15. Crutchfield, J.P.; Marzen, S. Signatures of infinity: Nonergodicity and resource scaling in prediction, complexity, and learning. *Phys. Rev. E* **2015**, *91*, 050106. [CrossRef]
16. Hilberg, W. Der bekannte Grenzwert der redundanzfreien Information in Texten—eine Fehlinterpretation der Shannonschen Experimente? *Frequenz* **1990**, *44*, 243–248. [CrossRef]
17. Shannon, C. Prediction and entropy of printed English. *Bell Syst. Tech. J.* **1951**, *30*, 50–64. [CrossRef]
18. Takahira, R.; Tanaka-Ishii, K.; Dębowski, Ł. Entropy Rate Estimates for Natural Language—A New Extrapolation of Compressed Large-Scale Corpora. *Entropy* **2016**, *18*, 364. [CrossRef]
19. Herdan, G. *Quantitative Linguistics*; Butterworths: London, UK, 1964.
20. Heaps, H.S. *Information Retrieval—Computational and Theoretical Aspects*; Academic Press: New York, NY, USA, 1978.
21. Hahn, M.; Futrell, R. Estimating Predictive Rate-Distortion Curves via Neural Variational Inference. *Entropy* **2019**, *21*, 640. [CrossRef]
22. Braverman, M.; Chen, X.; Kakade, S.M.; Narasimhan, K.; Zhang, C.; Zhang, Y. Calibration, Entropy Rates, and Memory in Language Models. *arXiv* **2019**, arXiv:1906.05664.
23. Dębowski, Ł. Mixing, Ergodic, and Nonergodic Processes with Rapidly Growing Information between Blocks. *IEEE Trans. Inf. Theory* **2012**, *58*, 3392–3401. [CrossRef]
24. Dębowski, Ł. On Hidden Markov Processes with Infinite Excess Entropy. *J. Theor. Probab.* **2014**, *27*, 539–551. [CrossRef]
25. Travers, N.F.; Crutchfield, J.P. Infinite Excess Entropy Processes with Countable-State Generators. *Entropy* **2014**, *16*, 1396–1413. [CrossRef]
26. Dębowski, Ł. Maximal Repetition and Zero Entropy Rate. *IEEE Trans. Inf. Theory* **2018**, *64*, 2212–2219. [CrossRef]

entropy

MDPI

Article

Productivity and Predictability for Measuring Morphological Complexity

Ximena Gutierrez-Vasques [1,*,†] and Victor Mijangos [2,†]

[1] Language and Space Lab, URPP Language and Space, University of Zurich, 8006 Zurich, Switzerland
[2] Institute of Philological Research, National Autonomous University of Mexico, 04510 Mexico City, Mexico; vmijangosc@ciencias.unam.mx
* Correspondence: ximena.gutierrezv@spur.uzh.ch
† These authors contributed equally to this work.

Received: 31 October 2019; Accepted: 23 December 2019; Published: 30 December 2019

Abstract: We propose a quantitative approach for quantifying morphological complexity of a language based on text. Several corpus-based methods have focused on measuring the different word forms that a language can produce. We take into account not only the productivity of morphological processes but also the predictability of those morphological processes. We use a language model that predicts the probability of sub-word sequences within a word; we calculate the entropy rate of this model and use it as a measure of predictability of the internal structure of words. Our results show that it is important to integrate these two dimensions when measuring morphological complexity, since languages can be complex under one measure but simpler under another one. We calculated the complexity measures in two different parallel corpora for a typologically diverse set of languages. Our approach is corpus-based and it does not require the use of linguistic annotated data.

Keywords: language complexity; morphology; TTR; language model; entropy rate

1. Introduction

Languages of the world differ from each other in unpredictable ways [1,2]. Language complexity focuses on determine how these variations occurs in terms of complexity (size of grammar elements, internal structure of the grammar).

Conceptualizing and quantifying linguistic complexity is not an easy task, many quantitative and qualitative dimensions must be taken into account [3]. In general terms, the complexity of a system could be related to the number and variety of elements, but also to the elaborateness of their interrelational structure [4,5].

In recent years, morphological complexity has attracted the attention of the research community [1,6]. Morphology deals with the internal structure of words [7]. Several corpus-based methods are successful in capturing the number and variety of the morphological elements of a language by measuring the distribution of words over a corpus. However, they may not capture other complexity dimensions such as the predictability of the internal structure of words. There can be cases where a language is considered complex because it has a rich morphological productivity, i.e., great number of morphs can be encoded into a single word. However, the combinatorial structure of these morphs in the word formation process can have less uncertainty than other languages, i.e., more predictable.

We would like to quantify the morphological complexity by measuring the type and token distributions over a corpus, but also by taking into account the predictability of the sub-word sequences within a word [8].

We assume that the predictability of the internal structure of words reflects the difficulty of producing novel words given a set of lexical items (stems, suffixes or morphs). We take as our method the statistical language models used in natural language processing (NLP), which are a useful tool for estimating a probability distribution over sequences of words within a language. However, we adapt this notion to the sub-word level. Information theory-based measures (entropy) can be used to estimate the predictiveness of these models.

Previous Work

Despite the different approaches and definitions of linguistic complexity, there are some main distinctions between the absolute and the relative complexity [3]. The former is defined in terms of the number of parts of a linguistic system; and the latter (more subjective) is related to the cost and difficulty faced by language users. Another important distinction includes global complexity that characterizes entire languages, e.g., as easy or difficult to learn. In contrast, particular complexity focuses only in a specific language level, e.g., phonological, morphological, syntactic.

In the case of morphology, languages of the world have different word production processes. Therefore, the amount of semantic and grammatical information encoded at the word level, may vary significantly from language to language. In this sense, it is important to quantify the morphological richness of languages and how it varies depending on their linguistic typology. Ackerman and Malouf [9] highlight two different dimensions that must be taken into account: the enumerative (e-complexity) that focuses on delimiting the inventories of language elements (number of morphosyntactic categories in a language and how they are encoded in a word); and the integrative complexity (i-complexity) that focuses on examining the systematic organization underlying the surface patterns of a language (difficulty of the paradigmatic system).

Coterell et al. [10] investigate a trade-off between the e-complexity and i-complexity of morphological systems. The authors propose a measure based on the size of a paradigm but also on how hard is to jointly predict all the word forms in a paradigm from the lemma. They conclude that "a morphological system can mark a large number of morphosyntactic distinctions [. . .] or it may have a high-level of unpredictability (irregularity); or neither. However, it cannot do both".

Moreover, Bentz et al. [11] distinguishes between paradigm-based approaches that use typological linguistic databases for quantifying the number of paradigmatic distinctions of languages as an indicator of complexity; and corpus-based approaches that estimate the morphological complexity directly from the production of morphological instances over a corpus.

Corpus-based approaches represent a relatively easy and reproducible way to quantify complexity without the strict need for linguistic annotated data. Several corpus-based methods share the underlying intuition that morphological complexity depends on the morphological system of a language, such as its inflectional and derivational processes; therefore, a very productive system will produce a lot of different word forms. This morphological richness can be captured using information theory measures [12,13] or type-token relationships [14], just to mention a few.

It is important to mention that enumerative complexity has been approached using a paradigm-based or a corpus-based perspective. However, the methods that target the integrative complexity seem to be more paradigm-based oriented (which can restrict the number of languages covered). With that in mind, the measures that we present in this work are corpus-based and they do not require access to external linguistic databases.

2. Methodology

In this work, we quantify morphological complexity by combining two different measures over parallel corpora: (a) the type-token relationship (TTR); and (b) the entropy rate of a sub-word language model as a measure of predictability. In this sense, our approach could be catalogued as a corpus-based method for measuring absolute complexity of a specific language level (morphology).

2.1. The Corpora

Parallel corpora are a valuable resource for many NLP tasks and for linguistics studies. Translation documents preserve the same meaning and functions, to a certain extent, across languages. This allows analysis/comparison of the morphological and typological features of languages.

We used two different parallel corpora that are available for a wide set of languages. On one hand, we used a portion of the Parallel Bible Corpus [15]; in particular, we used a subset of 1150 parallel verses that overlapped across 47 languages (the selection of languages and pre-processing of this dataset was part of the Interactive Workshop on Measuring Language Complexity (IWMLC 2019) http://www.christianbentz.de/MLC2019_index.html). These languages are part of the WALS 100-language sample, a selection of languages that are typologically diverse [16] (https://wals.info/languoid/samples/100).

On the other hand, we used the JW300 parallel corpus that compiles magazine articles for many languages [17] (these articles were originally obtained from the Jehovah's Witnesses website https://www.jw.org). In this case, we extracted a subset of 68 parallel magazine articles that overlapped across 133 languages. Table 1 summarizes information about the corpora.

Table 1. General information about the parallel corpora.

Corpus	Languages Covered	Total Tokens	Avg. Tokens Per Language
Bibles	47	1.1 M	24.8 K
JW300	133	22.4 M	168.9 K

We ran the experiments in both corpora independently. The intersection of languages covered by the two parallel corpora is 25. This shared set of languages was useful to compare the complexity rankings obtained with our measures, i.e., test if our complexity measures are consistent across different corpora.

It is important to mention that no sentence alignment was applied to the corpora. The Bibles corpus was already aligned at the verse level while the JW300 corpus was only aligned at the document level. However, for the aim of our experiments, alignment annotation (at the sentence or verse level) was not required.

2.2. Type-Token Relationship (TTR)

The type-token relationship (TTR) has proven to be a simple, yet effective, way to quantify the morphological complexity of a language using relatively small corpora [14]. It has also shown a high correlation with other types of complexity measures such as paradigm-based approaches that are based on typological information databases [11].

Morphologically rich languages will produce many different word forms (types) in a text, this is captured by measures such as TTR. From a linguistic perspective, Joan Bybee [18] affirms that "the token frequency of certain items in constructions [i.e., words] as well as the range of types [. . .] determines representation of the construction as well as its productivity".

TTR can be influenced by the size of a text (Heaps' law) or even by the domain of a corpus [19,20]. Some alternatives to make TTR more comparable include normalizing the text size or using logarithm, however, Covington and McFall [19] argue that these strategies are not fully successful, and they propose the moving-Average Type-Token Ratio. On the other hand, using parallel corpora has shown to be a simple way to make TTR more comparable across languages [21,22]. In principle, translations preserve the same meaning in two languages, therefore, there is no need for the texts to have the exact same length in tokens.

We calculated the TTR for a corpus by simply using Equation (1). Where $\#types$ are the different word types in the corpus (vocabulary size), and $\#tokens$ is the total number of word tokens in the corpus. Values closer to 1 would represent greater complexity. This simple way of measuring TTR, without any normalization, has been used in similar works [11,22,23].

$$\text{TTR} = \frac{\#types}{\#tokens} \tag{1}$$

We use this measure as an easy way to approach the e-complexity dimension; i.e., different morphosyntactic distinctions, and their productivity, could be reflected in the type and token distribution over a corpus.

2.3. Entropy Rate of a Sub-Word Language Model

Entropy as a measure of unpredictability represents a useful tool to quantify different linguistic phenomena, in particular, the complexity of morphological systems [9,12,24].

Our method aims to reflect the predictability of the internal structure of words in a language. We conjecture that morphological processes that are irregular/suppletive, unproductive, etc., will increase the entropy of a model that predicts the probability of sequences of morphs/sub-word units within a word.

To do this, we estimate a stochastic matrix P, where each cell contains the transition probability between two sub-word units in that language (see example Table 2). These probabilities are estimated using the corpus and a neural language model that we will describe below.

Table 2. Toy example of a stochastic matrix using the trigrams contained in the word 'cats'. The symbols #, $ indicate beginning/end of a word.

	#ca	**cat**	**ats**	**ts$**
#ca	0.01	0.06	0.07	0.33
cat	0.9	0.04	0.05	0.22
ats	0.06	0.78	0.05	0.23
ts$	0.03	0.12	0.83	0.22

We calculate the stochastic matrix P as follows (2):

$$P = p_{ij} = p(w_j|w_i) \tag{2}$$

where w_i and w_j are sub-word units. We used a neural probabilistic language model to estimate a probability function.

2.3.1. Sub-Word Units

Regarding to sub-word units, one initial thought would be to use character sequences that correspond to the linguistic notion of morphemes/morphs. However, it could be difficult to perform morphological segmentation to all the languages in the corpora. There are unsupervised morphological segmentation approaches, e.g., Morfessor [25], BPE encoding [26], but they still require parameter tuning to control over-segmentation/under-segmentation (making these approaches not completely language independent).

Instead of this, we focused on fixed-length sequences of characters (n-grams), which is more easily applicable to all the languages in the corpora. This decision is also driven by the evidence that trigrams encode morphological properties of the word [27]. Moreover, in some tasks such as language modeling, the use of character trigrams seems to lead to better word vector representations than unsupervised morphological segmentations [28].

Therefore, we trained the language models using character trigrams. We also took into account unigrams (characters) sequences, since there are languages with syllabic writing systems in the datasets and in these cases a single character can encode a whole syllable.

2.3.2. Neural Language Model

Our model was estimated using a feedforward neural network; this network gets trained with pairs of consecutive n-grams that appear in the same word. Once the network is trained we can retrieve from the output layer the probability p_{ij} for any pair of n-grams. This architecture is based on [29]; however, we used character n-grams instead of words. The network comprises the following layers: (1) an input layer of one-hot vectors representing the n-grams; (2) an embedding layer; (3) a hyperbolic tangent hidden layer; (4) and finally, an output layer that contains the conditional probabilities obtained by a SoftMax function defined by Equation (3).

$$p_{ij} = \frac{e^{a_{ij}}}{\sum_k e^{a_{ik}}} \tag{3}$$

The factor a_{ij} in Equation (3) is the jth output of the network when the n-gram w_i is the input. The architecture of the network is presented in Figure 1.

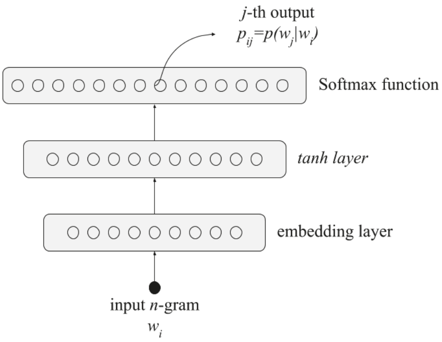


Figure 1. Neural probabilistic language model architecture, w_i, w_j are n-grams.

Once the neural network is trained, we can build the stochastic matrix P using the probabilities obtained for all the pairs of n-grams. We determine the entropy rate of the matrix (P) by using Equation (4) [30]:

$$H(P) = -\sum_{i=1}^{N} \mu_i \sum_{j=1}^{N} p_{ij} log_N p_{ij} \tag{4}$$

where p_{ij} are the entries of the matrix P, N is the size of the n-grams vocabulary, and μ represents the stationary distribution. This stationary distribution can be obtained using Equation (5), for each $i = 1, \ldots, N$:

$$\mu_i = \frac{1}{N} \sum_{k=1}^{N} p_{ik} \tag{5}$$

This equation defines a uniform distribution (we selected a uniform distribution since we observed that the stationary distribution, commonly defined by $P\mu = \mu$, was uniform for several small test corpora. Due to the neural probabilistic function, we can guarantee that the matrix P is irreducible; we assume that the irreducibility of the matrices is what determines the uniform stationary distribution. See [31]). To normalize the entropy, we use the logarithm base N. Thus, $H(P)$ can take values from 0 to 1. A value close to 1 would represent higher uncertainty in the sequence of n-grams within the words in a certain language, i.e., less predictability in the word formation processes.

The overall procedure can be summarized in the following steps: (the code is available at http://github.com/elotlmx/complexity-model)

1. For a given corpus, divide every word into its character n-grams. A vocabulary of size N (the number of n-grams) is obtained.

2. Calculate the probability of transitions between n-grams, $p_{ij} = p(w_j|w_i)$. This is done using the neural network described before.
3. A stochastic matrix $P = p_{ij}$ is obtained.
4. Calculate the entropy rate of the stochastic matrix $H(P)$.

3. Results

We applied the measures to each language contained in the JW300 and Bibles corpora. We use the notations H_1, H_3 for the entropy rate calculated with unigrams and trigrams respectively; TTR is the type-token relationship.

To combine the different complexity dimensions, we ranked the languages according to each measure, then we averaged the obtained ranks for each language (since we ranked the languages from the most complex to the less complex, we used the inverse of the average in order to be consistent with the complexity measures (0 for least complex, 1 for the most complex)). The notation for these combined rankings are the following: TTR+H_1 (TTR rank averaged with H_1 rank); TTR+H_2 (TTR rank averaged with H_2 rank); TTR+H_1+H_3 (TTR rank averaged with H_1 and H_3 ranks). In all the cases the scales go from 0 to 1 (0 for the least complex and 1 for the most complex).

Tables 3 and 4 contain the measures described above for each corpus. These tables only show the set of 25 languages that are shared between the two corpora. In Figures 2 and 3 we plot these different complexities, and their combinations. The complete list of languages and results are included in Appendices A and B.

Table 3. Complexity measures on the Bibles corpus (H_1: unigrams entropy; H_3: trigrams entropy; TTR: Type-token relationship); bold numbers indicate the highest and the lowest values for each measure, the rank is in brackets.

Language	H_1	H_3	TTR	TTR+H_1	TTR+H_3	TTR+H_1+H_3
Arabic	0.726 (3)	0.748 (4)	0.31 (3)	0.333 (2)	0.333 (3)	0.333 (2)
Burmese	0.74 (2)	0.823 (2)	**0.791** (1)	**0.667** (1)	**0.667** (1)	**0.6** (1)
Eastern Oromo	0.652 (10)	0.573 (22)	0.196 (9)	0.105 (7)	0.065 (18)	0.073 (12)
English	0.703 (5)	0.667 (10)	0.082 (19)	0.083 (11)	0.069 (16)	0.088 (10)
Fijian	0.569 (19)	0.519 (24)	0.048 (24)	0.047 (21)	**0.042** (24)	0.045 (24)
Finnish	0.696 (6)	0.59 (20)	0.266 (5)	0.182 (5)	0.08 (9)	0.097 (8)
French	0.607 (17)	0.609 (18)	0.139 (12)	0.069 (16)	0.067 (17)	0.064 (18)
Georgian	0.632 (12)	0.67 (9)	0.238 (6)	0.105 (7)	0.133 (5)	0.107 (5)
German	0.588 (18)	0.664 (12)	0.136 (13)	0.065 (17)	0.08 (9)	0.07 (13)
Hausa	0.61 (16)	0.614 (17)	0.098 (18)	0.059 (19)	0.057 (21)	0.059 (21)
Hindi	0.54 (22)	0.729 (6)	0.057 (22)	0.045 (23)	0.071 (13)	0.06 (20)
Indonesian	0.662 (9)	0.599 (19)	0.115 (17)	0.077 (12)	0.056 (22)	0.067 (16)
Korean	**0.394** (25)	**0.861** (1)	0.348 (2)	0.074 (14)	0.667 (1)	0.107 (5)
Modern Greek	0.683 (7)	0.655 (14)	0.181 (10)	0.118 (6)	0.083 (8)	0.097 (8)
Malagasy (Plateau)	0.568 (20)	**0.519** (24)	0.14 (11)	0.065 (17)	0.056 (22)	0.054 (23)
Russian	**0.751** (1)	0.732 (5)	0.225 (8)	0.222 (4)	0.154 (4)	0.214 (3)
Sango	0.538 (23)	0.56 (23)	**0.025** (25)	**0.042** (25)	**0.042** (24)	**0.042** (25)
Spanish	0.647 (11)	0.656 (13)	0.133 (15)	0.077 (12)	0.071 (13)	0.077 (11)
Swahili	0.613 (14)	0.576 (21)	0.233 (7)	0.091 (9)	0.071 (13)	0.07 (13)
Tagalog	0.632 (12)	0.629 (16)	0.121 (16)	0.071 (15)	0.063 (19)	0.068 (15)
Thai	0.554 (21)	0.752 (3)	0.055 (23)	0.045 (23)	0.074 (11)	0.063 (19)
Turkish	0.705 (4)	0.63 (15)	0.297 (4)	0.25 (3)	0.105 (6)	0.13 (4)
Vietnamese	0.406 (24)	0.684 (8)	0.066 (20)	0.045 (23)	0.071 (13)	0.058 (22)
Western Farsi	0.67 (8)	0.705 (7)	0.135 (14)	0.091 (9)	0.095 (7)	0.103 (7)
Yoruba	0.613 (14)	0.666 (11)	0.064 (21)	0.057 (20)	0.062 (20)	0.065 (17)

Table 4. Complexity measures on the JW300 corpus (H_1: unigrams entropy; H_3: trigrams entropy; TTR: Type-token relationship); bold numbers indicate the highest and the lowest values for each measure, the rank is in brackets.

Language	H_1	H_3	TTR	TTR+H_1	TTR+H_3	TTR+H_1+H_3
Arabic	0.586 (8)	0.826 (2)	0.171 (4)	0.166 (4)	**0.333** (1)	**0.214** (1)
Burmese	0.514 (19)	0.75 (5)	0.016 (22)	0.048 (23)	0.074 (12)	0.065 (17)
Eastern Oromo	0.552 (14)	0.568 (23)	0.111 (6)	0.1 (10)	0.068 (16)	0.069 (15)
English	0.682 (2)	0.712 (12)	0.053 (16)	0.111 (9)	0.071 (14)	0.1 (7)
Fijian	0.517 (18)	0.66 (17)	0.022 (21)	0.051 (21)	0.052 (23)	0.053 (21)
Finnish	0.563 (10)	0.628 (20)	**0.184** (1)	0.181 (2)	0.095 (6)	0.096 (9)
French	0.522 (17)	0.673 (16)	0.072 (11)	0.071 (14)	0.074 (12)	0.068 (16)
Georgian	0.563 (10)	0.728 (9)	0.175 (2)	0.153 (6)	0.181 (2)	0.136 (3)
German	0.636 (3)	0.686 (14)	0.084 (9)	0.166 (4)	0.086 (9)	0.115 (5)
Hausa	0.527 (16)	0.619 (22)	0.035 (18)	0.058 (17)	**0.05** (25)	0.053 (21)
Hindi	0.591 (6)	0.783 (3)	0.023 (19)	0.076 (12)	0.086 (9)	0.103 (6)
Indonesian	0.556 (12)	0.624 (21)	0.051 (17)	0.068 (15)	0.052 (23)	0.06 (19)
Korean	0.349 (24)	**0.907** (1)	0.057 (14)	0.052 (20)	0.133 (4)	0.076 (14)
Modern Greek	0.594 (5)	0.753 (4)	0.09 (8)	0.153 (6)	0.166 (3)	0.176 (2)
Malagasy (Plateau)	0.499 (22)	**0.537** (25)	0.062 (12)	0.058 (17)	0.054 (22)	0.05 (24)
Russian	0.5 (21)	0.722 (11)	0.137 (5)	0.076 (12)	0.125 (5)	0.081 (12)
Sango	0.385 (23)	0.724 (10)	**0.01** (25)	**0.041** (24)	0.057 (20)	0.051 (23)
Spanish	0.59 (7)	0.65 (18)	0.079 (10)	0.117 (8)	0.071 (14)	0.085 (10)
Swahili	0.598 (4)	0.565 (24)	0.098 (7)	0.181 (2)	0.064 (18)	0.085 (10)
Tagalog	0.514 (19)	0.676 (15)	0.054 (15)	0.057 (19)	0.066 (17)	0.06 (19)
Thai	0.552 (14)	0.74 (7)	0.013 (24)	0.051 (21)	0.064 (18)	0.065 (17)
Turkish	**0.684** (1)	0.65 (18)	0.175 (2)	**0.5** (1)	0.09 (8)	0.13 (4)
Vietnamese	**0.344** (25)	0.692 (13)	0.014 (23)	**0.041** (24)	0.055 (21)	**0.049** (25)
Western Farsi	0.569 (9)	0.738 (8)	0.061 (13)	0.09 (11)	0.095 (6)	0.1 (7)
Yoruba	0.553 (13)	0.748 (6)	0.023 (19)	0.062 (16)	0.08 (11)	0.078 (13)

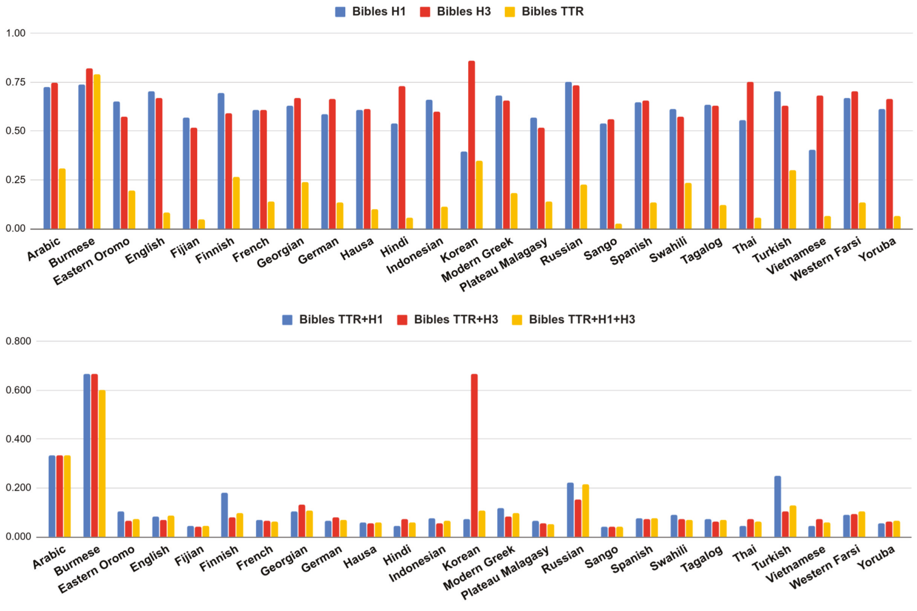


Figure 2. Different complexity measures (**above**) and their combinations (**below**) from Bibles corpus.

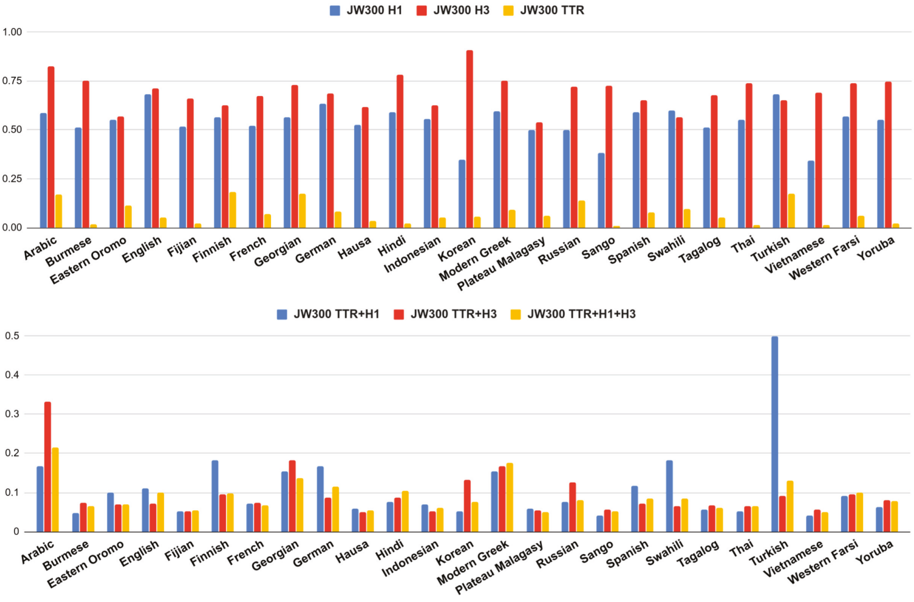


Figure 3. Different complexity measures (**above**) and their combinations (**below**) from JW300 corpus.

We can see that languages can be complex under one measure but simpler under another one. For instance, in Figures 2 and 3, we can easily notice that Korean is the most complex language if we only take into account the entropy rate using trigrams (H_3). However, this entropy dramatically drops using unigrams (H_1); therefore, when we combine the different measures, Korean is not the most complex language anymore.

There are cases such as English where its TTR is one of the lowest. This is expected since English is a language with poor inflectional morphology. However, its entropy is high. This suggests that a language such as English, usually not considered morphologically complex, may have many irregular forms that are not so easy to predict for our model.

We can also find the opposite case, where a language has a high TTR but low entropy, suggesting that it may produce many different word forms, but the inner structure of the words was "easy" to predict. This trend can be observed in languages such as Finnish (high TTR, low H_3), Korean (high TTR, low H_1) or Swahili (high TTR, low H_3).

The fact that a language has a low value of TTR does not necessarily imply that its entropy rate should be high (or vice versa). For instance, languages such as Vietnamese or Malagasy (Plateau), have some of the lowest values of entropy (H_1, H_3); however, their TTR values are not among the highest in the shared subset. In this sense, these languages seem to have low complexity in both dimensions.

Burmese language constitutes a peculiar case, it behaves differently among the two corpora. Burmese complexity seems very high in all dimensions (TTR and entropies) just in the Bibles corpora. We conjecture that TTR is oddly high due to tokenization issues [32]: this is a language without explicit word boundary delimiters, if the words are not well segmented then the text will have many different long words without repetitions (high TTR). The tokenization pre-processing of the Bibles was based only on whitespaces and punctuation marks, while the JW300 had a more sophisticated tokenization. In the latter, Burmese obtained a very low TTR and H_1 entropy.

Cases with high complexity in both dimensions were less common. Arabic was perhaps the language that tends to be highly complex under both criteria (TTR and entropy) and this behavior remained the same for the two corpora. We conjecture that this is related to the root-and-pattern morphology of the language, i.e., these types of patterns were difficult to predict for our sequential character n-grams language model. We will discuss more about this in Section 4.

3.1. Correlation across Corpora

Since our set of measures was applied to two different parallel corpora, we wanted to check if the complexities measures were, more or less, independent from the type of corpora used, i.e., languages should get similar complexity ranks in the two corpora.

We used Spearman's correlation [33] for the subset of shared languages across corpora. Table 5 shows the correlation coefficient for each complexity measure between the two corpora. Burmese language was excluded from the correlations due to the tokenization problems.

Table 5. Correlation of complexities between the JW300 and Bibles corpora (H_1: unigrams entropy; H_3: trigrams entropy; TTR: Type-token relationship).

	H_1	H_3	TTR	TTR+H_1	TTR+H_3	TTR+H_1+H_3
Correlation	0.520	0.782	0.890	0.776	0.858	0.765

Although the Bibles and the JW300 corpora belong to the same domain (religion), they greatly differ in size and in the topics covered (they are also parallel at different levels). Despite this, all the measures were positively correlated. The weaker correlation was obtained with H_1, while complexity measures such as TTR or TTR+H_3 were strongly correlated across corpora.

The fact that the complexity measures are correlated among the two corpora suggest that they are not very dependent of the corpus size, topics and other types of variations.

3.2. Correlation between Complexity Measures

In addition to the correlation across different corpora, we were interested in how the different complexity measures correlate between them (in the same corpus). Tables 6 and 7 show the Spearman's correlation between measures in each corpus.

Table 6. Spearman's correlations between measures in the corpus JW300 (all languages considered) (H_1: unigrams entropy; H_3: trigrams entropy; TTR: Type-token relationship).

	H_1	H_3	TTR	TTR+H_1	TTR+H_3	TTR+H_1+H_3
H_1	1.0	0.271	0.423	0.839	0.471	0.788
H_3	-	1.0	0.112	0.238	0.746	0.64
TTR	-	-	1.0	0.843	0.732	0.709
TTR+H_1	-	-	-	1.0	0.72	0.892
TTR+H_3	-	-	-	-	1.0	0.909
TTR+H_1+H_3	-	-	-	-	-	1.0

Table 7. Spearman's correlations between measures in the Bibles corpus (all languages considered) (H_1: unigrams entropy; H_3: trigrams entropy; TTR: Type-token relationship).

	H_1	H_3	TTR	TTR+H_1	TTR+H_3	TTR+H_1+H_3
H_1	1.0	0.276	0.384	0.828	0.464	0.810
H_3	-	1.0	0.006	0.152	0.693	0.585
TTR	-	-	1.0	0.815	0.654	0.637
TTR+H_1	-	-	-	1.0	0.668	0.866
TTR+H_3	-	-	-	-	1.0	0.862
TTR+H_1+H_3	-	-	-	-	-	1.0

In both corpora, the entropy-based measures (specially H_3) were poorly correlated (or not correlated) with the type-token relationship TTR. If these two types of measures are capturing, in fact, two different dimensions of the morphological complexity then it should be expected that they are not correlated.

The combined measures (TTR+H_1, TTR+H_3 and TTR+H_1+H_3) tend to be strongly correlated between them. It seems that all of them can combine, to some extent, the two dimensions of complexity (productivity and predictability).

Surprisingly, the entropy-based measures (H_1 and H_3) are weakly correlated between them, despite both trying to capture predictability. We conjecture that this could be related to the fact that for some languages, is more suitable to apply a trigram model and for some others the unigram model. For instance, in the case of Korean, one character is equivalent to a whole syllable (syllabic writing system). When we took combinations of three characters (trigrams) the model became very complex (high H_3), this does not necessarily reflect the real complexity. On the other hand, languages such as Turkish, Finnish or Yaqui (see Appendix B) obtained a very high value of H_1 (difficult to predict using only unigrams, very long words), but if we use the trigrams the entropy H_3 decreasse, trigram models may be more appropriate for these type of languages.

3.3. Correlation with Paradigm-Based Approaches

Finally, we compared our corpus-based morphological complexity measures against two paradigm-based measures. First, we used the C_{WALS} measure proposed by [11], it is based on 28 morphological features/chapters extracted from the linguistic database WALS [16]. This measure maps each morphological feature to a numerical value, the complexity of a language is the average of the values of the morphological features.

The measure C_{WALS} was originally applied to 34 typologically diverse languages. However, we only took 19 languages (the shared set of languages with our Bibles corpus). We calculated the correlation between our complexity measures and C_{WALS} (Table 8).

In addition, we included the morphological counting complexity (*MCC*) as implemented by [34]. Their metric counts the number of inflectional categories for each language, the categories are obtained from the annotated lexicon UniMorph [35].

This measure was originally applied to 21 languages (mainly Indo-European), we calculated the correlation between *MCC* and our complexity measures using the JW300 corpus (which contained all of those 21 languages) Table 8.

Appendices C and D contain the list of languages used for each measure and the complexities.

Table 8. Spearman's correlation between C_{WALS}, *MCC* and our complexity measures (H_1: unigrams entropy; H_3: trigrams entropy; TTR: Type-token relationship).

	H_1	H_3	TTR	TTR+H_1	TTR+H_3	TTR+H_1+H_3
C_{WALS}	0.322	−0.392	0.882	0.730	0.395	0.406
MCC	0.064	0.024	0.851	0.442	0.585	0.366

C_{WALS} and TTR are strongly correlated, this was already pointed out by [11]. However, our entropy-based measures are weakly correlated with C_{WALS}, it seems that they are capturing different things. *MCC* metric shows a similar behavior, it is highly correlated with TTR but not with H_1 (unigrams entropy) or H_3 (trigrams entropy).

It has been suggested that databases such as WALS, which provide paradigmatic distinctions of languages, reflect mainly the e-complexity dimension [2]. This could explain the high correlation between C_{WALS}, *MCC*, and measures such as TTR. However, the i-complexity may be better captured by other types of approaches, e.g., the entropy rate measure that we have proposed.

The weak correlation between our entropy-based measures and C_{WALS} (even negative correlation in the case of H_3) could be a hint of the possible trade-off between the i-complexity and e-complexity. However, further investigation is needed.

4. Discussion

Our corpus-based measures tried to capture different dimensions that play a role in the morphological complexity of a language. H_1 and H_3 are focused on the predictability of the internal structure of words, while TTR is focused on how many different word forms can a language produce. Our results show that these two approaches poorly correlate, especially H_3 and TTR (0.112 for JW300 and 0.006 for the Bibles), which give us a lead that these quantitative measures are capturing different aspects of the morphological complexity.

This is interesting since, in fields such as NLP, languages are usually considered complex when their morphology allows them to encode many morphological elements within a word (producing many different word forms in a text). However, a language that is complex in this dimension can also be quite regular (low entropy) in its morphological processes, e.g., a predictable/regular process can be applied to a large number of roots, producing many different types; this is a common phenomenon in natural languages [36].

We can also think in the opposite case, a language with poor inflectional morphology may have low TTR; however, it may have suppletive/irregular patterns that will not be fully reflected in TTR but they will increase the entropy of a model that tries to predict these word forms.

The aim of calculating the entropy rate of our language models was to reflect the predictability of the internal structure of words (how predictable sequences of n-grams are in a given language). We think this notion is closer to the concept of morphological integrative complexity (i-complexity); however, there are probably many other additional criteria that play a role in this type of complexity. In any case, it is not common to find works that try to conceptualize this complexity dimension based only on raw corpora, our work could be an initial step towards that direction.

Measures such as H_3, TTR (and all the combined versions) were consistent across the two parallel corpora. This is important since these corpora had different sizes and characteristics (texts from the JW300 corpus were significantly bigger than the Bibles one). These corpus-based measures may not necessarily require big amounts of text to grasp some typological differences and quantify the morphological complexity across languages.

The fact that measures such as C_{WALS} highly correlated with TTR but negative correlated with H_3, suggests that C_{WALS} and TTR are capturing the same type of complexity, closer to the e-complexity criteria. This type of complexity may be easier to capture by several methods, contrary to the i-complexity dimension, which is related to the predictability of forms, among other morphological phenomena.

Adding typological information of the languages could help to improve the complexity analysis. As a preliminary analysis, in Appendix E we classified a subset of languages as concatenative vs isolating morphology using WALS. As expected, there is a negative (weak) correlation between the TTR and H_3. However, this sign of possible trade-off is more evident in isolating languages compared to the ones that are classified as concatenative. This may be related to the fact that languages with isolating tendency do not produce many different word forms (low TTR); however, their derivative processes were difficult to predict for our sub-word language model (high entropy). More languages and exhaustive linguistic analysis are required.

One general advantage of our proposed measures for approaching morphological complexity is that they do not require linguistic annotated data such as morphological paradigms or grammars. The only requirement is to use parallel corpora, even if the texts are not fully parallel at the sentence level.

There are some drawbacks that are worth to discuss. We think that our approach of entropy rate of a sub-word language model may be especially suitable for concatenative morphology. For instance, languages with root-and-pattern morphology may not be sequentially predictable, making the entropy of our models go higher (Arabic is an example); however, these patterns may be predictable using a different type of model.

Furthermore, morphological phenomena such as stem reduplication may seem quite intuitive from a language user perspective; however, if the stem is not frequent in the corpus, it could be difficult for our language model to capture these patterns. In general, derivational processes could be less predictable by our model than the inflectional ones (more frequent and systematic).

On the other hand, these measures are dealing with written language, therefore, they can be influenced by factors such as the orthography, the writing systems, etc. The corpus-based measures that we used, especially TTR, are sensitive to tokenization and word boundaries.

The lack of a "gold-standard" makes it difficult to assess the dimensions of morphological complexity that we are successfully capturing. The type-token relationship of a language seems to agree more easily with other complexity measures (Section 3.3). On the other hand, our entropy rate is based on sub-word units, this measure did not correlate with the type-token relationship, nor with the degree of paradigmatic distinctions obtained from certain linguistic databases. We also tested an additional characteristic, the average word length per language (see Appendix F), and this does not strongly correlate either with H_3 or H_1.

Perhaps the question of whether this latter measure can be classified as i-complexity remains open. However, we think our entropy-based measure is reflecting to some extent the difficulty of predicting a word form in a language, since the entropy rate would increase with phenomena like: (a) unproductive processes; (b) allomorphy; (c) complex system of inflectional classes; and (d) suppletive patterns [37], just to mention a few.

Both approaches, TTR and the entropy rate of a sub-word language model, are valid and complementary, we used a very simple way to combine them (average of the ranks). In the future, a finer methodology can be used to integrate these two corpus-based quantitative approximations.

5. Future Work

In this section, we discuss some of the limitations that could be addressed as future work. The use of parallel corpora offers many advantages for comparing characteristics across languages. However, it is very difficult to find parallel corpora that cover a great amount of languages and that is freely available. Usually, the only available resources belong to specific domains, moreover, the parallel texts tend to be translations from one single language, e.g., English. It would be interesting to explore how these conditions affect the measurement of morphological complexity.

The character n-grams that we used for training the language models could be easily replaced by other types of sub-word units in our system. A promising direction could be testing different morphological segmentation models. Nevertheless, character trigrams seem to be a good initial point, at least for many languages, since these units may be capturing syllable information and this is related to morphological complexity [38,39].

Our way to control the influence of a language script system in the complexity measures was to consider two different character n-gram sizes. We noticed that trigrams (H_3) could be more suitable for languages with Latin script, while unigrams (H_1) may be better for other script systems (like Korean or Japanese). Automatic transliteration and other types of text pre-processing could be beneficial for this task.

There are still many open questions, as a future work we would like to make a more fine-grained typological analysis of the languages and complexity trends that resulted from these measures. Another promising research direction would be to quantify other processes that also play a role in the morphological complexity. For example, adding a tone in tonal languages is considered to add morphological complexity [3].

6. Conclusions

In this work we tried to capture two dimensions of morphological complexity. Languages that have a high TTR have the potential of encoding many different functions at the word level, therefore, they produce many different word forms. On the other hand, we proposed that the entropy rate of a sub-word language model could reflect how uncertain are the sequences of morphological elements within a word, languages with high entropy may have many irregular phenomena that are harder to predict than other languages. We were particularly interested in this latter dimension, since there are less quantitative methods, based on raw corpora, for measuring it.

The measures were consistent across two different parallel corpora. Moreover, the correlation between the different complexity measures suggest that our entropy rate approach is capturing a different complexity dimension than measures such as TTR or C_{WALS}.

Deeper linguistic analysis is needed; however, corpus-based quantitative measures can complement and deepen the study of morphological complexity.

Author Contributions: Conceptualization, V.M. and X.G.-V.; Investigation, X.G.-V. and V.M.; Methodology, X.G.-V. and V.M.; Writing—original draft, X.G.-V. and V.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Swiss Government Excellence Scholarship and the Mexican Council of Science and Technology (CONACYT). Fellowships 2019.0022 and 442471

Acknowledgments: We thank the reviewers, and Tanja Samardzic, for their valuable comments.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Complexity Measures for JW300 Corpus

Table A1. Complexity measures on the JW300 corpus (for all languages).

Language	H_1	H_3	TTR	TTR+H_1	TTR+H_3	TTR+H_1+H_3
Afrikaans	0.566 (73)	0.674 (69)	0.047 (82)	0.013 (79)	0.013 (76)	0.013 (79)
Amharic	0.582 (56)	0.875 (4)	0.2 (8)	0.031 (22)	0.167 (1)	0.044 (8)
Arabic	0.586 (53)	0.827 (6)	0.171 (15)	0.029 (25)	0.095 (4)	0.041 (12)
Azerbaijani	0.661 (6)	0.728 (32)	0.151 (21)	0.074 (5)	0.038 (15)	0.051 (5)
Bicol	0.622 (18)	0.69 (57)	0.049 (79)	0.021 (44)	0.015 (63)	0.019 (40)
Cibemba	0.527 (107)	0.581 (115)	0.108 (39)	0.014 (73)	0.013 (79)	0.011 (99)
Bulgarian	0.56 (80)	0.68 (66)	0.091 (45)	0.016 (63)	0.018 (46)	0.016 (62)
Bislama	0.548 (88)	0.662 (75)	0.009 (132)	0.009 (120)	0.01 (117)	0.01 (112)
Bengali	0.546 (90)	0.801 (9)	0.06 (69)	0.013 (83)	0.026 (26)	0.018 (46)
Cebuano	0.543 (93)	0.708 (42)	0.051 (75)	0.012 (87)	0.017 (54)	0.014 (71)
Chuukese	0.579 (58)	0.618 (104)	0.037 (90)	0.014 (75)	0.01 (107)	0.012 (91)
Seychelles Creole	0.593 (46)	0.645 (87)	0.024 (107)	0.013 (77)	0.01 (107)	0.012 (85)
Czech	0.668 (4)	0.777 (13)	0.125 (29)	0.061 (10)	0.048 (9)	0.065 (4)
Danish	0.617 (22)	0.695 (53)	0.063 (65)	0.023 (36)	0.017 (55)	0.021 (34)
German	0.636 (14)	0.686 (62)	0.084 (50)	0.031 (22)	0.018 (47)	0.024 (29)
Ewe	0.488 (124)	0.717 (39)	0.05 (77)	0.01 (109)	0.017 (53)	0.012 (85)
Efik	0.61 (30)	0.657 (80)	0.043 (85)	0.017 (56)	0.012 (94)	0.015 (64)
Modern Greek	0.594 (44)	0.753 (19)	0.09 (47)	0.022 (40)	0.03 (21)	0.027 (22)
English	0.682 (3)	0.713 (41)	0.053 (74)	0.026 (29)	0.017 (51)	0.025 (26)
Spanish	0.59 (48)	0.65 (82)	0.079 (54)	0.02 (51)	0.015 (63)	0.016 (57)
Estonian	0.623 (17)	0.663 (74)	0.155 (19)	0.056 (12)	0.022 (32)	0.027 (22)
Western Farsi	0.569 (71)	0.739 (27)	0.061 (68)	0.014 (71)	0.021 (34)	0.018 (43)
Finnish	0.563 (75)	0.628 (96)	0.184 (9)	0.024 (34)	0.019 (40)	0.017 (52)
Fijian	0.517 (111)	0.66 (77)	0.022 (115)	0.009 (124)	0.01 (105)	0.01 (115)
French	0.522 (110)	0.674 (68)	0.072 (62)	0.012 (90)	0.015 (59)	0.012 (85)
Ga	0.547 (89)	0.664 (73)	0.046 (83)	0.012 (90)	0.013 (83)	0.012 (89)
Kiribati	0.506 (118)	0.592 (113)	0.031 (101)	0.009 (118)	0.009 (125)	0.009 (128)
Gujarati	0.542 (95)	0.835 (5)	0.048 (81)	0.011 (93)	0.023 (28)	0.017 (53)
Gun	0.575 (65)	0.691 (55)	0.024 (108)	0.012 (92)	0.012 (91)	0.013 (81)
Hausa	0.527 (106)	0.619 (102)	0.035 (94)	0.01 (107)	0.01 (109)	0.01 (114)
Hebrew	0.595 (43)	0.763 (17)	0.17 (16)	0.034 (19)	0.061 (6)	0.039 (13)
Hindi	0.591 (47)	0.783 (10)	0.022 (111)	0.013 (81)	0.017 (57)	0.018 (46)

Table A1. *Cont.*

Language	H_1	H_3	TTR	TTR+H_1	TTR+H_3	TTR+H_1+H_3
Hiligaynon	0.564 (74)	0.699 (48)	0.045 (84)	0.013 (81)	0.015 (62)	0.015 (70)
Hiri Motu	0.543 (94)	0.604 (111)	0.012 (128)	0.009 (122)	0.008 (131)	0.009 (129)
Croatian	0.63 (16)	0.735 (30)	0.109 (38)	0.037 (15)	0.029 (22)	0.036 (16)
Haitian Creole	0.552 (85)	0.662 (76)	0.022 (114)	0.01 (106)	0.011 (104)	0.011 (105)
Hungarian	0.694 (1)	0.747 (22)	0.172 (14)	0.133 (2)	0.056 (7)	0.081 (3)
Armenian	0.575 (64)	0.736 (29)	0.117 (33)	0.021 (44)	0.032 (19)	0.024 (29)
Indonesian	0.556 (82)	0.624 (97)	0.051 (76)	0.013 (81)	0.012 (98)	0.012 (94)
Igbo	0.576 (60)	0.613 (107)	0.032 (99)	0.013 (83)	0.01 (115)	0.011 (101)
Iloko	0.611 (29)	0.64 (89)	0.08 (53)	0.024 (32)	0.014 (71)	0.018 (48)
Icelandic	0.637 (11)	0.704 (45)	0.09 (46)	0.035 (18)	0.022 (30)	0.029 (19)
Isoko	0.569 (70)	0.656 (81)	0.02 (116)	0.011 (99)	0.01 (111)	0.011 (102)
Italian	0.595 (40)	0.614 (106)	0.082 (52)	0.022 (41)	0.013 (87)	0.015 (65)
Japanese	0.302 (133)	0.914 (1)	0.024 (106)	0.008 (128)	0.019 (42)	0.012 (85)
Georgian	0.563 (77)	0.729 (31)	0.175 (12)	0.022 (38)	0.047 (10)	0.025 (27)
Kongo	0.534 (100)	0.619 (103)	0.022 (112)	0.009 (114)	0.009 (127)	0.01 (122)
Greenlandic	0.538 (98)	0.623 (99)	0.335 (1)	0.02 (47)	0.02 (38)	0.015 (65)
Cambodian	0.509 (117)	0.779 (12)	0.011 (129)	0.008 (129)	0.014 (69)	0.012 (96)
Kannada	0.587 (52)	0.754 (18)	0.239 (3)	0.036 (16)	0.095 (4)	0.041 (10)
Korean	0.349 (131)	0.907 (2)	0.057 (71)	0.01 (110)	0.027 (24)	0.015 (69)
Kikaonde	0.553 (83)	0.541 (127)	0.087 (48)	0.015 (68)	0.011 (99)	0.012 (96)
Kikongo	0.486 (126)	0.541 (128)	0.079 (55)	0.011 (94)	0.011 (103)	0.01 (118)
Kirghiz	0.563 (76)	0.695 (51)	0.144 (24)	0.02 (49)	0.027 (25)	0.02 (39)
Luganda	0.601 (36)	0.539 (129)	0.14 (25)	0.033 (20)	0.013 (79)	0.016 (61)
Lingala	0.526 (108)	0.633 (93)	0.04 (88)	0.01 (105)	0.011 (101)	0.01 (109)
Silozi	0.539 (97)	0.598 (112)	0.033 (97)	0.01 (103)	0.01 (119)	0.01 (116)
Lithuanian	0.637 (13)	0.706 (43)	0.167 (17)	0.067 (9)	0.033 (18)	0.041 (10)
Kiluba	0.544 (92)	0.56 (125)	0.112 (35)	0.016 (64)	0.012 (89)	0.012 (91)
Tshiluba	0.489 (123)	0.617 (105)	0.074 (60)	0.011 (96)	0.012 (94)	0.01 (107)
Luvale	0.545 (91)	0.525 (133)	0.145 (23)	0.018 (55)	0.013 (83)	0.012 (90)
Mizo	0.595 (42)	0.681 (65)	0.04 (87)	0.016 (67)	0.013 (78)	0.015 (63)
Latvian	0.582 (57)	0.745 (24)	0.123 (32)	0.022 (38)	0.036 (16)	0.027 (24)
Mauritian Creole	0.583 (55)	0.624 (98)	0.019 (117)	0.012 (90)	0.009 (127)	0.011 (103)
Plateau Malagasy	0.499 (122)	0.538 (131)	0.062 (66)	0.011 (100)	0.01 (111)	0.009 (124)
Marshallese	0.587 (51)	0.718 (38)	0.022 (113)	0.012 (86)	0.013 (76)	0.015 (67)
Macedonian	0.571 (68)	0.698 (49)	0.083 (51)	0.017 (58)	0.02 (38)	0.018 (46)
Malayalam	0.607 (32)	0.701 (47)	0.272 (2)	0.059 (11)	0.041 (14)	0.037 (15)
Moore	0.561 (79)	0.724 (34)	0.027 (104)	0.011 (96)	0.014 (66)	0.014 (75)
Marathi	0.612 (27)	0.738 (28)	0.095 (44)	0.028 (27)	0.028 (23)	0.03 (18)
Maltese	0.616 (24)	0.683 (63)	0.075 (59)	0.024 (33)	0.016 (58)	0.021 (36)
Burmese	0.514 (113)	0.75 (20)	0.016 (121)	0.009 (126)	0.014 (69)	0.012 (93)
Nepali	0.524 (109)	0.768 (15)	0.096 (43)	0.013 (76)	0.034 (17)	0.018 (44)
Niuean	0.389 (129)	0.646 (86)	0.013 (125)	0.008 (131)	0.009 (122)	0.009 (131)
Dutch	0.604 (34)	0.683 (64)	0.061 (67)	0.02 (50)	0.015 (61)	0.018 (42)
Norwegian	0.605 (33)	0.723 (35)	0.056 (72)	0.019 (53)	0.019 (42)	0.021 (34)
Sepedi	0.514 (114)	0.637 (90)	0.037 (91)	0.01 (111)	0.011 (101)	0.01 (112)
Chichewa	0.567 (72)	0.562 (124)	0.124 (31)	0.019 (52)	0.013 (81)	0.013 (80)
Eastern Oromo	0.552 (86)	0.568 (121)	0.111 (36)	0.016 (61)	0.013 (85)	0.012 (88)
Ossetian	0.575 (63)	0.688 (61)	0.077 (57)	0.017 (60)	0.017 (55)	0.017 (53)
Punjabi	0.572 (66)	0.816 (7)	0.025 (105)	0.012 (88)	0.018 (47)	0.017 (51)
Pangasinan	0.612 (28)	0.66 (78)	0.058 (70)	0.02 (46)	0.014 (73)	0.017 (49)
Papiamento (Curaçao)	0.603 (35)	0.704 (46)	0.031 (102)	0.015 (70)	0.014 (73)	0.016 (55)
Solomon Islands Pidgin	0.642 (9)	0.64 (88)	0.013 (123)	0.015 (69)	0.009 (122)	0.014 (76)
Polish	0.617 (23)	0.745 (23)	0.152 (20)	0.047 (13)	0.047 (10)	0.045 (6)
Ponapean	0.533 (102)	0.576 (118)	0.032 (98)	0.01 (107)	0.009 (129)	0.009 (123)
Portuguese	0.595 (41)	0.697 (50)	0.075 (58)	0.02 (47)	0.019 (44)	0.02 (38)
Romanian	0.609 (31)	0.695 (52)	0.071 (63)	0.021 (43)	0.017 (51)	0.021 (36)
Russian	0.5 (121)	0.722 (37)	0.137 (26)	0.014 (74)	0.032 (20)	0.016 (57)
Kirundi	0.534 (101)	0.636 (91)	0.15 (22)	0.016 (62)	0.018 (49)	0.014 (73)

Table A1. *Cont*.

Language	H_1	H_3	TTR	TTR+H_1	TTR+H_3	TTR+H_1+H_3
Kinyarwanda	0.599 (38)	0.57 (120)	0.134 (28)	0.03 (24)	0.014 (73)	0.016 (59)
Sango	0.385 (130)	0.725 (33)	0.01 (130)	0.008 (133)	0.012 (91)	0.01 (111)
Sinhala	0.578 (59)	0.742 (25)	0.079 (56)	0.017 (56)	0.025 (27)	0.021 (34)
Slovak	0.614 (26)	0.767 (16)	0.124 (30)	0.036 (17)	0.043 (12)	0.042 (9)
Slovenian	0.637 (12)	0.69 (56)	0.111 (37)	0.041 (14)	0.022 (32)	0.029 (20)
Samoan	0.536 (99)	0.629 (95)	0.017 (119)	0.009 (117)	0.009 (125)	0.01 (121)
Shona	0.622 (19)	0.538 (130)	0.18 (10)	0.069 (7)	0.014 (67)	0.019 (41)
Albanian	0.648 (8)	0.723 (36)	0.073 (61)	0.029 (26)	0.021 (37)	0.029 (20)
Sranantongo	0.54 (96)	0.562 (123)	0.01 (131)	0.009 (125)	0.008 (133)	0.009 (132)
Sesotho (Lesotho)	0.465 (128)	0.58 (116)	0.033 (95)	0.009 (123)	0.009 (122)	0.009 (130)
Swedish	0.621 (20)	0.706 (44)	0.066 (64)	0.024 (34)	0.019 (44)	0.023 (31)
Swahili	0.598 (39)	0.566 (122)	0.098 (41)	0.025 (31)	0.012 (93)	0.015 (68)
Swahili (Congo)	0.562 (78)	0.586 (114)	0.098 (41)	0.017 (59)	0.013 (82)	0.013 (82)
Tamil	0.618 (21)	0.715 (40)	0.234 (6)	0.074 (5)	0.043 (12)	0.045 (7)
Telugu	0.66 (7)	0.811 (8)	0.211 (7)	0.143 (1)	0.133 (2)	0.136 (1)
Thai	0.552 (87)	0.74 (26)	0.013 (124)	0.009 (112)	0.013 (75)	0.013 (83)
Tigrinya	0.666 (5)	0.891 (3)	0.162 (18)	0.087 (4)	0.095 (4)	0.115 (2)
Tiv	0.576 (61)	0.659 (79)	0.017 (120)	0.011 (94)	0.01 (113)	0.012 (98)
Tagalog	0.514 (115)	0.676 (67)	0.054 (73)	0.011 (100)	0.014 (67)	0.012 (94)
Otetela	0.529 (105)	0.605 (110)	0.085 (49)	0.013 (78)	0.013 (88)	0.011 (100)
Setswana	0.503 (120)	0.612 (108)	0.031 (100)	0.009 (120)	0.01 (118)	0.009 (127)
Tongan	0.532 (103)	0.688 (60)	0.023 (110)	0.009 (115)	0.012 (96)	0.011 (104)
Chitonga	0.558 (81)	0.647 (85)	0.177 (11)	0.022 (41)	0.021 (35)	0.017 (50)
Tok Pisin	0.575 (62)	0.632 (94)	0.008 (133)	0.01 (104)	0.009 (130)	0.01 (109)
Turkish	0.684 (2)	0.65 (83)	0.175 (13)	0.133 (2)	0.021 (35)	0.031 (17)
Tsonga	0.572 (67)	0.571 (119)	0.036 (93)	0.012 (85)	0.009 (124)	0.011 (106)
Tatar	0.593 (45)	0.689 (58)	0.116 (34)	0.025 (30)	0.022 (31)	0.022 (32)
Chitumbuka	0.588 (49)	0.534 (132)	0.108 (40)	0.022 (38)	0.012 (97)	0.014 (78)
Twi	0.469 (127)	0.664 (72)	0.039 (89)	0.009 (116)	0.012 (90)	0.01 (107)
Tahitian	0.487 (125)	0.669 (70)	0.012 (127)	0.008 (130)	0.01 (111)	0.009 (126)
Ukrainian	0.601 (37)	0.775 (14)	0.136 (27)	0.031 (22)	0.049 (8)	0.038 (14)
Umbundu	0.531 (104)	0.56 (126)	0.048 (80)	0.011 (98)	0.01 (115)	0.01 (119)
Urdu	0.631 (15)	0.781 (11)	0.033 (96)	0.018 (54)	0.019 (42)	0.025 (28)
Venda	0.512 (116)	0.619 (101)	0.031 (103)	0.009 (118)	0.01 (114)	0.009 (125)
Vietnamese	0.344 (132)	0.692 (54)	0.014 (122)	0.008 (131)	0.011 (100)	0.01 (117)
Waray-Waray	0.586 (54)	0.665 (71)	0.042 (86)	0.014 (72)	0.013 (85)	0.014 (72)
Wallisian	0.517 (112)	0.577 (117)	0.013 (126)	0.008 (127)	0.008 (132)	0.008 (133)
Xhosa	0.615 (25)	0.647 (84)	0.237 (4)	0.069 (7)	0.023 (29)	0.027 (24)
Yapese	0.639 (10)	0.635 (92)	0.018 (118)	0.016 (65)	0.01 (120)	0.014 (76)
Yoruba	0.553 (84)	0.749 (21)	0.023 (109)	0.01 (102)	0.015 (59)	0.014 (73)
Maya	0.587 (50)	0.688 (59)	0.05 (78)	0.016 (65)	0.015 (65)	0.016 (60)
Zande	0.505 (119)	0.62 (100)	0.037 (92)	0.009 (112)	0.01 (105)	0.01 (120)
Zulu	0.57 (69)	0.609 (109)	0.235 (5)	0.027 (28)	0.018 (50)	0.016 (55)

Appendix B. Complexity Measures Bibles Corpus

Table A2. Complexity measures on the Bibles corpus (for all languages).

Language	H_1	H_3	TTR	TTR+H_1	TTR+H_3	TTR+H_1+H_3
Amele	0.568 (37)	0.59 (29)	0.134 (26)	0.031 (36)	0.036 (34)	0.032 (36)
Alamblak	0.673 (11)	0.643 (18)	0.203 (15)	0.076 (8)	0.06 (8)	0.068 (9)
Bukiyip	0.651 (16)	0.591 (28)	0.119 (32)	0.041 (25)	0.033 (37)	0.039 (28)
Apurinã	0.592 (29)	0.523 (43)	0.205 (14)	0.046 (19)	0.035 (36)	0.034 (33)
Mapudungun	0.598 (27)	0.596 (27)	0.145 (20)	0.041 (25)	0.042 (20)	0.04 (26)
Egyptian Arabic	0.725 (5)	0.748 (4)	0.31 (4)	0.222 (2)	0.25 (3)	0.23 (2)
Barasana-Eduria	0.526 (45)	0.577 (35)	0.146 (19)	0.031 (36)	0.037 (31)	0.03 (40)
Chamorro	0.678 (10)	0.663 (13)	0.13 (29)	0.051 (17)	0.046 (16)	0.056 (12)
German	0.588 (30)	0.663 (13)	0.136 (24)	0.037 (29)	0.054 (12)	0.044 (18)
Daga	0.585 (32)	0.545 (41)	0.095 (39)	0.028 (40)	0.025 (44)	0.026 (44)
Modern Greek	0.683 (9)	0.655 (16)	0.181 (17)	0.076 (8)	0.06 (8)	0.071 (7)
English	0.703 (7)	0.667 (10)	0.082 (40)	0.042 (22)	0.04 (24)	0.052 (13)
Basque	0.655 (14)	0.588 (31)	0.224 (13)	0.074 (10)	0.045 (17)	0.051 (15)

Table A2. *Cont*.

Language	H_1	H_3	TTR	TTR+H_1	TTR+H_3	TTR+H_1+H_3
Fijian	0.568 (37)	0.519 (44)	0.048 (46)	0.024 (42)	0.022 (47)	0.023 (46)
Finnish	0.696 (8)	0.589 (30)	0.266 (6)	0.142 (5)	0.055 (10)	0.068 (9)
French	0.606 (25)	0.609 (24)	0.139 (23)	0.041 (25)	0.042 (20)	0.041 (23)
Paraguayan Guaraní	0.613 (21)	0.642 (19)	0.174 (18)	0.051 (17)	0.054 (12)	0.051 (15)
Eastern Oromo	0.652 (15)	0.573 (38)	0.196 (16)	0.064 (12)	0.037 (31)	0.043 (19)
Hausa	0.609 (24)	0.613 (23)	0.098 (38)	0.032 (32)	0.032 (39)	0.035 (30)
Hindi	0.54 (43)	0.729 (6)	0.057 (43)	0.023 (43)	0.04 (24)	0.032 (36)
Indonesian	0.661 (13)	0.598 (26)	0.115 (34)	0.042 (22)	0.033 (37)	0.041 (23)
Popti’	0.624 (20)	0.646 (17)	0.108 (37)	0.035 (30)	0.037 (31)	0.04 (26)
Kalaallisut	0.572 (35)	0.455 (47)	0.542 (2)	0.054 (15)	0.04 (24)	0.035 (30)
Georgian	0.632 (18)	0.67 (9)	0.238 (9)	0.071 (11)	0.111 (5)	0.081 (5)
West Kewa	0.573 (34)	0.583 (33)	0.113 (35)	0.028 (40)	0.029 (41)	0.029 (41)
Halh Mongolian	0.745 (3)	0.601 (25)	0.228 (11)	0.142 (5)	0.055 (10)	0.076 (6)
Korean	0.393 (47)	0.861 (1)	0.348 (3)	0.04 (27)	0.5 (2)	0.058 (11)
Lango (Uganda)	0.602 (26)	0.558 (40)	0.112 (36)	0.032 (32)	0.026 (43)	0.029 (41)
San Miguel El Grande Mixtec	0.57 (36)	0.614 (22)	0.125 (30)	0.03 (39)	0.038 (27)	0.034 (33)
Burmese	0.739 (4)	0.822 (2)	0.791 (1)	0.4 (1)	0.666 (1)	0.428 (1)
Wichí Lhamtés Güisnay	0.586 (31)	0.585 (32)	0.117 (33)	0.031 (36)	0.03 (40)	0.031 (38)
Nama (Namibia)	0.576 (33)	0.665 (11)	0.131 (28)	0.032 (32)	0.05 (14)	0.041 (23)
Western Farsi	0.67 (12)	0.705 (7)	0.135 (25)	0.054 (15)	0.062 (7)	0.068 (9)
Plateau Malagasy	0.567 (39)	0.518 (45)	0.14 (21)	0.032 (32)	0.029 (41)	0.028 (43)
Imbabura Highland Quichua	0.598 (27)	0.492 (46)	0.249 (8)	0.057 (14)	0.037 (31)	0.037 (29)
Russian	0.75 (1)	0.732 (5)	0.225 (12)	0.153 (4)	0.117 (4)	0.166 (3)
Sango	0.537 (44)	0.56 (39)	0.024 (47)	0.021 (47)	0.023 (46)	0.023 (46)
Spanish	0.647 (17)	0.656 (15)	0.133 (27)	0.045 (20)	0.047 (15)	0.05 (17)
Swahili	0.612 (22)	0.575 (36)	0.233 (10)	0.06 (13)	0.043 (18)	0.043 (19)
Tagalog	0.632 (18)	0.629 (20)	0.121 (31)	0.04 (27)	0.038 (27)	0.042 (21)
Thai	0.554 (41)	0.752 (3)	0.055 (44)	0.023 (43)	0.042 (20)	0.034 (33)
Turkish	0.705 (6)	0.629 (20)	0.297 (5)	0.181 (3)	0.08 (6)	0.096 (4)
Vietnamese	0.406 (46)	0.684 (8)	0.066 (41)	0.022 (45)	0.04 (24)	0.031 (38)
Sanumá	0.546 (42)	0.574 (37)	0.05 (45)	0.022 (45)	0.024 (45)	0.024 (45)
Yagua	0.563 (40)	0.524 (42)	0.266 (6)	0.042 (22)	0.04 (24)	0.033 (35)
Yaqui	0.748 (2)	0.579 (34)	0.14 (21)	0.086 (7)	0.036 (34)	0.052 (13)
Yoruba	0.612 (22)	0.665 (11)	0.064 (42)	0.031 (36)	0.037 (31)	0.04 (26)

Appendix C. Complexity Measures Using C_{WALS}

Table A3. C_{WALS} complexity for the subset of languages shared with the Bibles corpus.

Language	C_{WALS}	H_1	H_3	TTR	TTR+H_1	TTR+H_3	TTR+H_1+H_3
Amele	0.456 (9)	0.568 (17)	0.59 (13)	0.134 (13)	0.066 (17)	0.076 (16)	0.069 (18)
Apurinã	0.573 (5)	0.592 (15)	0.523 (17)	0.205 (8)	0.087 (12)	0.08 (14)	0.075 (16)
Basque	0.647 (4)	0.655 (8)	0.588 (14)	0.224 (7)	0.133 (5)	0.095 (9)	0.103 (7)
Eastern Oromo	0.487 (8)	0.652 (9)	0.573 (16)	0.196 (9)	0.111 (9)	0.08 (14)	0.088 (11)
Egyptian Arabic	0.563 (6)	0.725 (3)	0.748 (1)	0.31 (1)	0.5 (1)	1.0 (1)	0.6 (1)
English	0.329 (15)	0.703 (5)	0.667 (4)	0.082 (17)	0.09 (10)	0.095 (9)	0.115 (6)
German	0.397 (13)	0.588 (16)	0.663 (6)	0.136 (12)	0.071 (14)	0.111 (5)	0.088 (11)
Halh Mongolian	0.516 (7)	0.745 (2)	0.601 (11)	0.228 (5)	0.285 (3)	0.125 (4)	0.166 (4)
Hausa	0.322 (16)	0.609 (13)	0.613 (10)	0.098 (16)	0.069 (15)	0.076 (16)	0.076 (15)
Imbabura Quichua	0.662 (3)	0.599 (14)	0.492 (19)	0.25 (3)	0.117 (8)	0.09 (12)	0.083 (14)
Indonesian	0.336 (14)	0.661 (7)	0.598 (12)	0.115 (15)	0.09 (10)	0.074 (18)	0.088 (11)
Modern Greek	0.452 (11)	0.683 (6)	0.655 (8)	0.181 (10)	0.125 (6)	0.111 (5)	0.125 (5)
Plateau Malagasy	0.309 (17)	0.567 (18)	0.518 (18)	0.14 (11)	0.069 (15)	0.069 (19)	0.063 (19)
Russian	0.453 (10)	0.751 (1)	0.732 (2)	0.225 (6)	0.285 (3)	0.25 (2)	0.333 (2)
Spanish	0.44 (12)	0.647 (10)	0.656 (7)	0.133 (14)	0.083 (13)	0.095 (9)	0.096 (8)
Swahili	0.675 (2)	0.612 (11)	0.575 (15)	0.233 (4)	0.125 (6)	0.105 (7)	0.096 (8)
Turkish	0.775 (1)	0.705 (4)	0.629 (9)	0.297 (2)	0.333 (2)	0.181 (3)	0.2 (3)
Vietnamese	0.141 (19)	0.406 (19)	0.684 (3)	0.066 (18)	0.054 (19)	0.095 (9)	0.075 (16)
Yoruba	0.178 (18)	0.612 (11)	0.665 (5)	0.064 (19)	0.066 (17)	0.083 (13)	0.085 (13)

Appendix D. Complexity Measures Using *MCC*

Table A4. *MCC* complexity for the subset of languages shared with the JW300 corpus.

Language	*MCC*	H_1	H_3	TTR	TTR+H_1	TTR+H_3	TTR+H_1+H_3
Bulgarian	96.0 (7)	0.56 (20)	0.68 (16)	0.091 (10)	0.016 (20)	0.018 (13)	0.016 (18)
Czech	195.0 (2)	0.668 (3)	0.777 (1)	0.125 (6)	0.061 (3)	0.048 (2)	0.065 (2)
Danish	15.0 (20)	0.617 (9)	0.695 (11)	0.063 (19)	0.023 (12)	0.017 (16)	0.021 (13)
Dutch	26.0 (19)	0.604 (13)	0.683 (15)	0.061 (20)	0.02 (18)	0.015 (19)	0.018 (16)
English	6.0 (21)	0.682 (2)	0.713 (7)	0.053 (21)	0.026 (9)	0.017 (16)	0.025 (10)
Estonian	110.0 (5)	0.623 (7)	0.663 (18)	0.155 (4)	0.056 (4)	0.022 (8)	0.027 (8)
Finnish	198.0 (1)	0.563 (19)	0.628 (20)	0.184 (1)	0.024 (10)	0.019 (11)	0.017 (17)
French	30.0 (18)	0.522 (21)	0.674 (17)	0.072 (16)	0.012 (21)	0.015 (19)	0.012 (21)
German	38.0 (16)	0.636 (6)	0.686 (14)	0.084 (12)	0.031 (8)	0.018 (13)	0.024 (11)
Hungarian	94.0 (8)	0.694 (1)	0.747 (4)	0.172 (2)	0.133 (1)	0.056 (1)	0.081 (1)
Italian	52.0 (13)	0.595 (14)	0.614 (21)	0.082 (13)	0.022 (14)	0.013 (21)	0.015 (20)
Latvian	81.0 (9)	0.582 (18)	0.745 (5)	0.123 (8)	0.022 (14)	0.036 (5)	0.027 (8)
Lithuanian	152.0 (3)	0.637 (4)	0.706 (8)	0.167 (3)	0.067 (2)	0.033 (6)	0.041 (5)
Modern Greek	50.0 (14)	0.594 (16)	0.753 (3)	0.09 (11)	0.022 (14)	0.03 (7)	0.027 (8)
Polish	112.0 (4)	0.617 (9)	0.745 (5)	0.152 (5)	0.047 (5)	0.047 (3)	0.045 (3)
Portuguese	77.0 (10)	0.595 (14)	0.697 (10)	0.075 (15)	0.02 (18)	0.019 (11)	0.02 (15)
Romanian	60.0 (12)	0.609 (12)	0.695 (11)	0.071 (17)	0.021 (16)	0.017 (16)	0.021 (13)
Slovak	40.0 (15)	0.614 (11)	0.767 (2)	0.124 (7)	0.036 (7)	0.043 (4)	0.042 (4)
Slovenian	100.0 (6)	0.637 (4)	0.69 (13)	0.111 (9)	0.041 (6)	0.022 (8)	0.029 (6)
Spanish	71.0 (11)	0.59 (17)	0.65 (19)	0.079 (14)	0.02 (18)	0.015 (19)	0.016 (18)
Swedish	35.0 (17)	0.621 (8)	0.706 (8)	0.066 (18)	0.024 (10)	0.019 (11)	0.023 (12)

Appendix E. Correlation Using Typological Classifications

For each language in the intersection set between the Bibles and JW300 corpora, we extracted its information about the feature 20A: "Fusion of Selected Inflectional Formatives" (WALS database). We focused on the languages classified as "concatenative" or "isolating". For each corpus, we calculated the correlations within complexity measures for concatenative languages and the correlations within the isolating ones (Tables A5 and A6).

Table A5. Spearman's correlation between complexity measures in concatenative and isolating languages (Bibles corpus).

		H_1	H_3	TTR
Concatenative	H_1	1.0	0.233	0.618
	H_3	-	1.0	−0.121
	TTR	-	-	1.0
Isolating	H_1	1.0	−0.355	0.513
	H_3	-	1.0	−0.178
	TTR	-	-	1.0

Table A6. Spearman's correlation between complexity measures in concatenative and isolating languages (JW300 corpus).

		H_1	H_3	TTR
Concatenative	H_1	1.0	−0.12	0.296
	H_3	−12	1.0	−0.369
	TTR	-	-	1.0
Isolating	H_1	1.0	−0.011	0.438
	H_3	-	1.0	−0.741
	TTR	-	-	1.0

Appendix F. Correlation Using Average Word Length

We calculate the average word length per language in both corpora. This is formulated as the average of the number of characters per word. Tables A7 and A8 show the correlations of the average word length with the other measures for the Bibles and JW300 corpora, respectively.

Table A7. Spearman's correlation between complexity measures and the average length per word in the Bibles corpus.

	H_1	H_3	TTR	TTR+H_1	TTR+H_3	TTR+H_1+H_3
Average Word Length	0.354	−0.421	0.697	0.628	0.141	0.278

Table A8. Spearman's correlation between complexity measures and the average length per word in the JW300 corpus.

	H_1	H_3	TTR	TTR+H_1	TTR+H_3	TTR+H_1+H_3
Average Word Length	0.296	−0.359	0.735	0.606	0.265	0.315

References

1. Sampson, G.; Gil, D.; Trudgill, P. *Language Complexity as an Evolving Variable*; Oxford University Press: Oxford, UK, 2009; Volume 13.
2. Meinhardt, E.; Malouf, R.; Ackerman, F. *Morphology Gets More and More Enumeratively Complex, Unless It Doesn't*; LSA Summer Institute: Lexington, KY, USA, 2017.
3. Miestamo, M.; Sinnemäki, K.; Karlsson, F. Grammatical complexity in a cross-linguistic perspective. *Lang. Complexity Typol. Contact Chang.* **2008**, 23–41. [CrossRef]
4. Simon, H.A. *The Architecture of Complexity*; MIT Press: Cambridge, MA, USA, 1996.
5. Sinnemäki, K. *Language Universals and Linguistic Complexity: Three Case Studies in Core Argument Marking*; University of Helsinki: Helsinki, Finland, 2011.
6. Baerman, M.; Brown, D.; Corbett, G.G. *Understanding and Measuring Morphological Complexity*; Oxford University Press: New York, NY, USA, 2015.
7. Haspelmath, M.; Sims, A.D. *Understanding Morphology*; Hodder Education: London, UK, 2010.
8. Montermini, F.; Bonami, O. Stem spaces and predictability in verbal inflection. *Lingue e Linguaggio* **2013**, *12*, 171–190.
9. Ackerman, F.; Malouf, R. Morphological organization: The low conditional entropy conjecture. *Language* **2013**, *89*, 429–464. [CrossRef]
10. Cotterell, R.; Kirov, C.; Hulden, M.; Eisner, J. On the complexity and typology of inflectional morphological systems. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 327–342. [CrossRef]
11. Bentz, C.; Ruzsics, T.; Koplenig, A.; Samardzic, T. A comparison between morphological complexity measures: Typological data vs. language corpora. In Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (cl4lc), Osaka, Japan, 11 December 2016; pp. 142–153.
12. Blevins, J.P. The information-theoretic turn. *Psihologija* **2013**, *46*, 355–375. [CrossRef]
13. Bonami, O.; Beniamine, S. Joint predictiveness in inflectional paradigms. *Word Struct.* **2016**, *9*, 156–182. [CrossRef]
14. Kettunen, K. Can type-token ratio be used to show morphological complexity of languages? *J. Quant. Linguist.* **2014**, *21*, 223–245. [CrossRef]
15. Mayer, T.; Cysouw, M. Creating a massively parallel bible corpus. *Oceania* **2014**, *135*, 40.
16. Dryer, M.S.; Haspelmath, M. The World Atlas of Language Structures Online. 2013. Available online: https://wals.info/ (accessed on 8 December 2019).
17. Agić, Ž.; Vulić, I. JW300: A wide-coverage parallel corpus for low-resource languages. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguist, Florence, Italy, 28 July–2 August 2019.
18. Bybee, J. *Language, Usage and Cognition*; Cambridge University Press: Cambridge, UK, 2010.

19. Covington, M.A.; McFall, J.D. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *J. Quant. Linguist.* **2010**, *17*, 94–100. [CrossRef]
20. Tweedie, F.J.; Baayen, R.H. How variable may a constant be? Measures of lexical richness in perspective. *Comput. Humanit.* **1998**, *32*, 323–352. [CrossRef]
21. Kelih, E. The type-token relationship in Slavic parallel texts. *Glottometrics* **2010**, *20*, 1–11.
22. Mayer, T.; Wälchli, B.; Rohrdantz, C.; Hund, M. From the extraction of continuous features in parallel texts to visual analytics of heterogeneous areal-typological datasets. *Lang. Process. Grammars. Role Funct. Oriented Comput. Model.* **2014**, 13–38. [CrossRef]
23. Gerz, D.; Vulić, I.; Ponti, E.M.; Reichart, R.; Korhonen, A. On the relation between linguistic typology and (limitations of) multilingual language modeling. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 316–327.
24. Baerman, M. Paradigmatic chaos in Nuer. *Language* **2012**, *88*, 467–494. [CrossRef]
25. Smit, P.; Virpioja, S.; Grönroos, S.A.; Kurimo, M. Morfessor 2.0: Toolkit for statistical morphological segmentation. In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, 26–30 April 2014; pp. 21–24.
26. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. *arXiv* **2015**, arXiv:1508.07909.
27. Baayen, R.H.; Chuang, Y.Y.; Blevins, J.P. Inflectional morphology with linear mappings. *Ment. Lex.* **2018**, *13*, 230–268. [CrossRef]
28. Vania, C.; Lopez, A. From characters to words to in between: Do we capture morphology? *arXiv* **2017**, arXiv:1704.08352.
29. Bengio, Y.; Ducharme, R.; Vincent, P.; Jauvin, C. A neural probabilistic language model. *J. Mach. Learn. Res.* **2003**, *3*, 1137–1155.
30. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
31. Freedman, A. Convergence Theorem for Finite Markov Chains. *Proc. REU* **2017**. Available online: http://math.uchicago.edu/~may/REU2017/REUPapers/Freedman.pdf (accessed on 24 December 2019).
32. Ding, C.; Aye, H.T.Z.; Pa, W.P.; Nwet, K.T.; Soe, K.M.; Utiyama, M.; Sumita, E. Towards Burmese (Myanmar) Morphological Analysis: Syllable-based Tokenization and Part-of-speech Tagging. *ACM Trans. Asian Low-Resour. Lang. Inf. Process. (TALLIP)* **2019**, *19*, 5. [CrossRef]
33. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **1987**, *100*, 441–471. [CrossRef]
34. Cotterell, R.; Mielke, S.J.; Eisner, J.; Roark, B. Are all languages equally hard to language-model? *arXiv* **2018**, arXiv:1806.03743.
35. Kirov, C.; Sylak-Glassman, J.; Knowles, R.; Cotterell, R.; Post, M. A Rich Morphological Tagger for English: Exploring the Cross-Linguistic Tradeoff Between Morphology and Syntax. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3–7 April 2017; Short Papers; Association for Computational Linguistics: Valencia, Spain, 2017; Volume 2, pp. 112–117.
36. Blevins, J.P.; Milin, P.; Ramscar, M. The Zipfian paradigm cell filling problem. In *Perspectives on Morphological Organization*; Brill: Leiden, The Netherlands, 2017; pp. 139–158.
37. Mel'čuk, I.A. On suppletion. *Linguistics* **1976**, *14*, 45–90. [CrossRef]
38. Peters, A.M.; Menn, L. False Starts and Filler Syllables: Ways to Learn Grammatical Morphemes. *Language* **1993**, *69*, 742–777. [CrossRef]
39. Coupé, C.; Oh, Y.M.; Dediu, D.; Pellegrino, F. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Sci. Adv.* **2019**, *5*, eaaw2594. [CrossRef]

entropy

MDPI

Article

Entropy Rate Estimation for English via a Large Cognitive Experiment Using Mechanical Turk

Geng Ren [1], Shuntaro Takahashi [2] and Kumiko Tanaka-Ishii [3,*]

[1] Sorbonne Université, École Polytechnique Universitaire, 75005 Paris, France; geng.ren@etu.upmc.fr
[2] Graduate School of Engineering, The University of Tokyo, Tokyo 113-8654, Japan; takahashi@cl.rcast.u-tokyo.ac.jp
[3] Research Center for Advanced Science and Technology, The University of Tokyo, Tokyo 153-8904, Japan
* Correspondence: kumiko@cl.rcast.u-tokyo.ac.jp

Received: 30 October 2019; Accepted: 3 December 2019; Published: 6 December 2019

Abstract: The entropy rate h of a natural language quantifies the complexity underlying the language. While recent studies have used computational approaches to estimate this rate, their results rely fundamentally on the performance of the language model used for prediction. On the other hand, in 1951, Shannon conducted a cognitive experiment to estimate the rate without the use of any such artifact. Shannon's experiment, however, used only one subject, bringing into question the statistical validity of his value of $h = 1.3$ bits per character for the English language entropy rate. In this study, we conducted Shannon's experiment on a much larger scale to reevaluate the entropy rate h via Amazon's Mechanical Turk, a crowd-sourcing service. The online subjects recruited through Mechanical Turk were each asked to guess the succeeding character after being given the preceding characters until obtaining the correct answer. We collected 172,954 character predictions and analyzed these predictions with a bootstrap technique. The analysis suggests that a large number of character predictions per context length, perhaps as many as 10^3, would be necessary to obtain a convergent estimate of the entropy rate, and if fewer predictions are used, the resulting h value may be underestimated. Our final entropy estimate was $h \approx 1.22$ bits per character.

Keywords: entropy rate; natural language; crowd source; Amazon Mechanical Turk; Shannon entropy

1. Introduction

Entropy rates h of natural languages have been used to investigate the complexity underlying these languages. The entropy rate of a sequence measures the amount of information per character [1] and indicates that the number of possible sequences is 2^{hn} for a sequence of length n.

Following the development of information theory and an abundance of data resources, recent studies have used computational approaches for finding the entropy rates of natural languages. Starting from the first attempt made by [2], which used a three-gram, word-level language model, various compression algorithms have been utilized [3,4]. The most recent study makes use of a state-of the art neural language model [5]. However, such computational attempts have a drawback; i.e., the computation of h requires a computational language model with which to predict the probability distribution of every character. As a result, the value of h reflects not only the complexity of the language but also the performance of the model. Indeed, in natural language processing, such an estimate of h is used as an indicator of the goodness-of-fit of a language model [6]. Recently reported decreases in the upper bound of h, for which the current minimum for English is 1.08 bpc [7] are simply highlighting improvements in the computational model.

Originally, Shannon's study [1] and some work that followed [8–11] used cognitive methods to estimate the entropy rate h. The original scientific interest in h had to do with the complexity of human language. Given this perspective, the performance of a computational model should not be involved in obtaining a value of h.

The studies using cognitive approaches can be reconsidered from two perspectives. First, they were all based on limited-scale experiments. In all of these studies, a subject was asked to predict the n-th character given the preceding $n-1$ characters. According to [11], Shannon's spouse was his only subject. Even the most recent cognitive study [11] relied on just eight subjects. Experimenting on such a small scale raises the question of the statistical validity of the acquired estimate.

Second, none of the cognitive approaches considered the limit with respect to the context length n. While the estimated values should be evaluated at infinite n by the definition of the entropy rate, the reported values are obtained at some finite n. In Shannon [1], the value $h = 1.3$ bits per character (bpc) for English was obtained at $n = 100$, and Moradi et al. [11] concluded that the estimated value does not decrease beyond $n \geq 32$ and reported a rate of $h \approx 1.6$ bpc. For extrapolation, however, a large number of observations becomes necessary in order to capture the dependence of the entropy rate on n well.

To that end, we conducted a large-scale cognitive test to acquire the English language entropy rate h through Amazon Mechanical Turk (AMT). AMT is a crowd-sourcing service offered by Amazon that allowed us to gather a large number of participants in a short time and at a reasonable cost. We focused on the entropy rate in English to make a fair comparison with Shannon [1] and other works. Other languages possibly have different values of the entropy rate, as can be seen in the comparison made in [4]. We collected a total of 172,954 character predictions from 683 different subjects. To the best of our knowledge, the scale used in this experiment was more than two times larger than any used in previous studies. At such a scale, the effects of factors that may influence the estimation of the entropy rate can be examined. Our analysis implies that Shannon's original experiment had an insufficient sample size with which to find a convergent estimate. We finally obtained $h \approx 1.22$ bpc for English, which is smaller than Shannon's original result of $h = 1.3$ bpc.

2. Entropy Rate Estimation

2.1. Entropy Rate and n-Gram Entropy

Definition 1. Shannon entropy

Let X be a stochastic process $\{X_t\}_{t=1}^{\infty}$, where each element belongs to a finite character set $\mathcal{X}$. Let $X_i^j = X_i, X_{i+1}, \ldots, X_{j-1}, X_j$ for $i < j$ and $P(X_i^j)$ be the probability of X_i^j. The Shannon entropy of a stochastic process $H(X_1^n)$ is defined as

$$H(X_1^n) = -\sum_{X_1^n} P(X_1^n) \log P(X_1^n). \tag{1}$$

Definition 2. Entropy rate

The entropy rate h of a stochastic process X is defined as

$$h = \lim_{n\to\infty} \frac{1}{n} H(X_1^n), \tag{2}$$

if such a value exists [12]. The entropy rate h is the average amount of information per element in a sequence of infinite length.

In the following, let F_n be the prediction complexity of X_n given X_1^{n-1}, as follows:

$$F_n \equiv H(X_n | X_1^{n-1}). \tag{3}$$

In other words, F_n quantifies the average uncertainty of the n-th character given a character string with length $n-1$. If the stochastic process X is stationary, F_n reaches the entropy rate h as n tends to infinity, as follows [12]:

$$h = \lim_{n\to\infty} F_n. \tag{4}$$

In this work, h was estimated via F_n. A human *subject* was given X_1^{n-1} characters and asked to predict the next character X_n. We aimed to collect a large number of predictions from many subjects. For a subject and a phrase, let a *sample* indicate the *prediction* of a X_n given a particular X_1^{n-1}.

An *experimental session* is defined as a subject and phrase pair. For every experimental session, a subject first predicts X_1, then X_2 given X_1, then X_3 given X_1^2, then X_4 given $X_1^3, \ldots, X_n$ given X_1^{n-1}, and so on. Therefore, in an experimental session, a number of observations are acquired for a given phrase, with the maximum number of observations being the character length of the phrase.

2.2. Shannon's Method

If a subject guesses a character given a string of length n, the answer will be correct or incorrect. In Shannon's setting and ours, the *prediction* of X_n *by a subject is accomplished by making multiple guesses*, one character at a time, until he/she reaches the correct answer. In other words, a prediction for character X_n in this setting consists of a series of guesses.

The number of guesses required to reach the correct answer reflects the predictability of that character and should relate to the probability of that character X_n appearing after X_1^{n-1}. Let q_i^n denote the probability that a subject requires *i guesses* in a prediction to find the correct letter following a block of length $n-1$.

Shannon deduced the following inequality [1]:

$$\sum_{i=1}^{K} i(q_i^n - q_{i+1}^n)\log i \leq F_n \leq -\sum_{i=1}^{K} q_i^n \log q_i^n. \tag{5}$$

Here, K is the number of characters in the set; in this work, $K = 27$, since the English alphabet consists of 26 letters and the space symbol. This setting corresponds to the settings used in previous works [9,11] using the cognitive approach to acquire the entropy rate in order for our results to be comparable with those reported in these works. Note that this lower bound is the lower bound of the upper bound of h and not the direct lower bound of h. For each context length n, the probability q_i^n can be calculated for a set of samples.

In Shannon's original experiment, 100 phrases of length 100 were taken from *Jefferson the Virginian*, a biography of ex-US President Thomas Jefferson authored by Dumas Malone. In each experimental session, the subject (i.e., only his spouse, according to [11]) was asked to predict the next character given a block of length $n-1$. She continued in this manner for $n = 1, 2, \ldots, 15$, and 100 for each phrase; consequently, Shannon acquired 16 observations for each phrase. He used 100 different phrases; therefore, he collected $16 \times 100 = 1600$ observations from his spouse in total. He then calculated q_i^n for $n = 1, 2, \ldots, 15$, and 100, each based on 100 observations, and the upper and lower bounds of h were computed based on the leftmost and rightmost terms of the inequality (5), respectively. Shannon observed a decrease in the bounds with respect to n and obtained an upper bound of $h = 1.3$ bpc for $n = 100$.

Moradi et al. [11] conducted Shannon's experiment under two different settings. In the first experiment, they used 100 phrases of length $n = 64$ from *Scruples II*, a romance novel authored by Judith Krantz. In the first setting, a single subject participated, and they calculated the upper bounds from $n = 1$ to $n = 64$ based on 100 observations. They reported that the entropy rate reached $h \approx 1.6$ bpc at $n = 32$ and that larger values of n did not contribute to decreasing the upper bound. In the second setting, the eight participants were given phrases extracted from four different books, and the values of the upper bound at $n = 32$ were reported, which ranged between $h = 1.62$ and $h = 3.00$ bpc.

Jamison and Jamison [9] used 50 and 40 phrases, both taken from some unspecified source, for each of two subjects, respectively. They conducted the experiment for $n = 4, 8, 12$, and 100 and obtained $h = 1.63$ and $h = 1.67$ bpc for the two subjects at $n = 100$ based on 50 and 40 phrase samples, respectively.

Note how the reported values deviate greatly from Shannon's $h = 1.3$ bpc. In all these experiments, since the number of subjects was small, the number of observations was limited, making the statistical validity questionable.

2.3. Cover King's Method

While Shannon's method only considers the likelihood of the correct answer for each X_n, Cover and King wanted to collect the distribution for each X_n. Hence, instead of counting the number of guesses required, a subject was asked to assign a probability distribution to the nth character given the preceding string of length $n-1$. Precisely, in Cover and King [10], a *prediction by a subject is the character distribution* of X_n.

They designed this experiment using a *gambling* framework, following their theory of information in gambling [13,14]. A subject assigned odds to every character which could be used for X_n; i.e., a probability distribution.

Cover and King [10] conducted two experiments separately. In the first experiment, phrases were extracted from *Jefferson the Virginian* for 12 subjects. The maximum length of a phrase was set as $n = 75$. The estimated value of the upper bound of h for the 12 subjects ranged between $h = 1.29$ bpc and $h = 1.90$ bpc. In the second experiment, phrases were taken from *Contact: The First Four Minutes* (a science book on psychology authored by Leonard M. Zunin); lengths of $n = 220$ were used, and two subjects participated. The estimated values of h produced by the two subjects were $h = 1.26$ bpc and $h = 1.30$ bpc.

We conducted Cover and King's experiment using the similar framework, as explained in detail in the following section. Compared with the experiment proposed by Shannon, however, their experiment demanded too much from each subject since he/she had to set the odds for all 27 characters every time. The majority of the subjects abandoned the experiment before completing the assignment, and it was difficult to collect a large number of reliable observations. Therefore, we could not utilize this method effectively and focused on Shannon's framework instead.

2.4. Summary of the Scales Used in Previous Studies

Table 1 summarizes the experimental settings of the previous reports [1,9–11]. We refer to the total number of observations as the sum of the count of the predictions made by the subjects for different phrases and context lengths. For example, in Shannon's case, the total number of observations was 1600, as one subject was asked to make predictions for 16 different context lengths (i.e., $n = 1, 2, \ldots, 15$, and 100) for each of 100 different phrases. The third and fourth columns in the table list the numbers of distinct subjects and phrases used in each study, respectively. Note that a phrase could be tested by multiple subjects or a subject could test multiple phrases, depending on the experimental setting.

Table 1. Comparison of the scales of cognitive experiments undertaken in previous works for the entropy rate estimation in English [1,9–11] and that of the present work.

	Total Number of Samples	**Number of Subjects**	**Number of Phrases**	**Max n for a Session**	**Number of Sample Per n**
Shannon [1]	1600	1	100	100	100
Jamison and Jamison [9]	360	2	50 and 40	100	50 and 40
Cover and King [10] No.1	440	2	1	220	2
Cover and King [10] No.2	900	12	1	75	12
Moradi et al. [11] No.1	6400	1	100	64	100
Moradi et al. [11] No.2	3200	8	400	32	100
Our Experiment	172,954	683	225	87.51	1954.86

The fifth and sixth columns present the average maximum value of n obtained in one session and the mean number of observations per n, respectively, where n represents the offset of a character from the beginning of a phrase. Both of these values were fixed in the previous works.

3. Cognitive Experiment Using Mechanical Turk

3.1. The Mechanical Turk Framework

Our experimental framework was implemented through Amazon Mechanical Turk, a workplace service offered by *Amazon*. AMT puts up tasks called HITs (human intelligence tasks) and *workers* do them. AMT has been used previously as a research tool for conducting large-scale investigations that require human judgment, ranging from annotating image data [15,16], to collecting text and speech data [17,18], behavioral research [19], judging music and documents [20,21], and identifying complex patterns in brain activity [22].

With AMT, the experimenter is able to collect a large number of observations on a wide range of topics. Compared with standard in-laboratory studies, however, such an experiment is open to anonymous subjects, and thus, control is limited. For example, in our case, a subject could use any external information to predict the next character. In particular, we were unable to prohibit subjects from conducting a search for the $n-1$ characters to obtain the answer for the next character. Furthermore, the English fluency of the subjects was unknown. Thus, the results should be examined from this perspective as well; see Section 5.2.

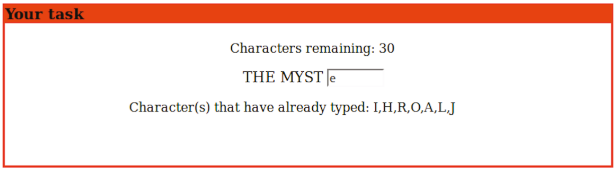
An experimental user interface based on Shannon's original proposal was developed. The most important requirement of the design was the adequacy of the task load since a subject could easily lose their concentration and abandon a prediction during the experiment. We designed the user interface to be as simple as possible so as to lessen the psychological demand on the subjects.

3.2. Experimental Design

In this HIT, a subject was asked to start from the beginning fragment of a sentence, and then guess character after character of the remainder of the sentence. Figure 1 shows the interface used in the experiment. As shown, a subject received three types of information:

1. The number of characters still available for use.
2. The preceding $n-1$ characters.
3. The set of incorrect characters already used.

In this framework, once a subject decides on their guess, they input it and press enter to submit it. If the guess is correct, the context is updated to length n, and the task continues with the prediction of the $n+1$-th character. If the answer is incorrect, the subject must guess what the n-th character is until obtaining the correct answer. Subjects were informed in advance of the number of characters in the remaining phrase to avoid anyone abandoning the task.



Your task

Characters remaining: 30

THE MYST e

Character(s) that have already typed: I,H,R,O,A,L,J

Figure 1. Our user interface for our cognitive experiment on Amazon Mechanical Turk. It provides: (**i**) the number of characters still available for use, (**ii**) the preceding $n-1$ characters, and (**iii**) the set of incorrect characters already used.

If a phrase is too long, subjects become easily distracted. Therefore, it was necessary to adjust the length of time provided for an experimental session. Too short a time raises the cognitive load, whereas too long a time decreases a subject's interest. After multiple trials across multiple options, such as putting a constant cap on the time allowed for each guess, we chose to allow a maximum number of guesses for every phrase. After some preliminary tests, this number was fixed to the character length of the phrase. Therefore, a subject was able to complete the task only if they always guessed all of the characters correctly. Most of the time, then, a subject was unable to finish a phrase.

The phrases were taken from the *Wall Street Journal*. In particular, 225 sentences were randomly extracted for this experiment and used as the experimental phrases. Their average length was 150.97. All characters were capitalized, and non-alphabetical symbols other than spaces were removed, duplicating the settings in previous works [1,9–11]. Hence, the characters were limited to the 26 letters of the alphabet, all in capital letters, and the space symbol. Table 2 lists the top ten most frequently used words and two successive words used in the experiment. As shown, they are relatively simple words that do not require specialized knowledge to predict correctly.

Table 2. The top ten most frequently used words along with two subsequent words appearing in the phrases used in our experiment.

Rank	Word	Frequency	Two Subsequent Words	Frequency
1	market	15	interest rates	4
2	company	13	future contracts	3
3	investment	11	program trading	3
4	price	11	stock market	3
5	people	11	money managers	3
6	companies	10	same time	2
7	stock	9	wide variety	2
8	buy	9	time around	2
9	officials	7	higher dividends	2
10	growth	7	some firms	2

We considered multiple variations of Shannon's experiment. The experiment could have consisted of guessing a character of a different phrase every time; thus, increasing the cognitive load for the subject by having them read through a different phrase every time. Another possibility was to proceed even if the character guess was incorrect. Since multiple subjects participated, it would then still be possible to acquire the probability of a correct guess. Such a method would decrease the task load substantially. However, this idea was not adopted since some subjects could choose random characters for all predictions. Finally, we reached the conclusion that Shannon's framework was well designed and utilized it in this work.

3.3. Experimental Outcomes

The last row of Table 1 provides the summary for the cognitive experiment. We collected 172,954 observations from 683 different subjects, whose residences were limited to the United States, Canada, Great Britain, and Australia. The mean of the maximum values of n for each experimental session was 87.51. The mean number of observations collected for $n \leq 70$ was 1954.86.

These numbers are by far the largest collected for this type of experiment [1,9–11], in terms of both the total number of observations and the number of subjects. While these values were fixed in the previous works, they varied in our experiment due to the use of Mechanical Turk.

Figure 2 shows the number of samples acquired for different context lengths $n - 1$. As the context length $n - 1$ increased, the number of observations decreased because, in our experiment, the number of guesses could reach the maximum number of guesses allowed for a phrase, as mentioned in the previous section. For up to $n = 70$, over 85% of the subjects made guesses. Beyond $n = 70$, however, the number of subjects making guesses decreased quickly. As we discuss later, having a large number of observations is crucial for acquiring a good estimate of the entropy rate within a statistically reasonable margin.

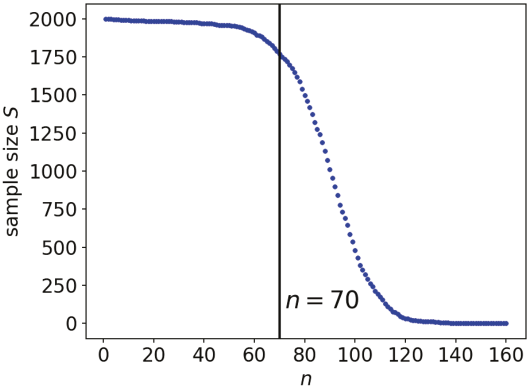


Figure 2. The number of observations collected for the predictions made for the n-th character. The vertical line indicates $n = 70$, which provided the minimum direct estimate of $h_{expmin} = 1.407$ in our experiment.

3.4. Human Prediction Accuracy with Respect to Context Length

Shannon [1] originally reported that the upper bound decreases with respect to the context length for up to $n = 100$. This result implies that a human is able to improve their prediction performance with more context. However, the later experiment by [11] disagreed with Shannon's [1], as they reported that the upper bound did not decrease for $n \geq 32$. Therefore, the question remains as to whether longer contextual phrases help humans to predict future characters more accurately. Hence, we examined whether the prediction performance of subjects improved with a longer contextual phrase length, based on all observations collected.

Figure 3 shows the probability that a subject provided the correct n-th character with their first guess. At $n = 1$ (i.e., the subject was asked to predict the first character of a phrase with no context given), the probability was below 20%. The probability improved greatly from $n = 1$ to $n = 2$, as it reached above 50% for $n = 2$. As n increased to $n = 100$, the probability roughly monotonically increased to nearly 80%. Based on this result, a subject improves their accuracy in predicting the next character as the context length n increases, at least up to $n = 100$, which supports Shannon's claim.

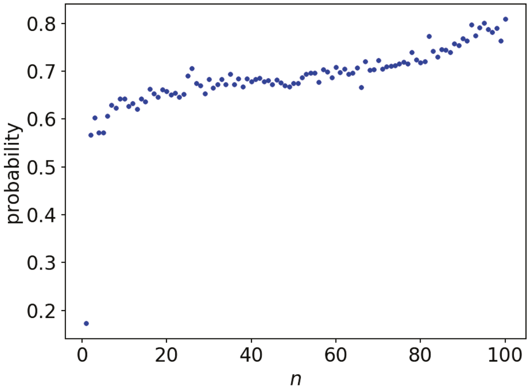


Figure 3. The probability that the subject needed only one guess to make the correct prediction of n-th character.

This result also implies that the subjects of our experiment exhibited reasonable performances since it was a major concern that the collected observations might be of low quality due to the online experimental setting.

3.5. The Datapoints of the Bounds for n

Using all of the observations, the upper and lower bounds can be estimated with Equation (5) for every n. The number of collected observations varies with respect to n, as shown in Figure 2. Figure 4 shows the plots of the upper and lower bounds computed for $n = 1, 2, \ldots, 70$ using all of the collected observations. The blue plot indicates the upper bound, whereas the red plot shows the lower bound. For the upper bound, the blue plot exhibits a decreasing tendency, although the values fluctuate along with n. Our main interest lies in the upper bound.

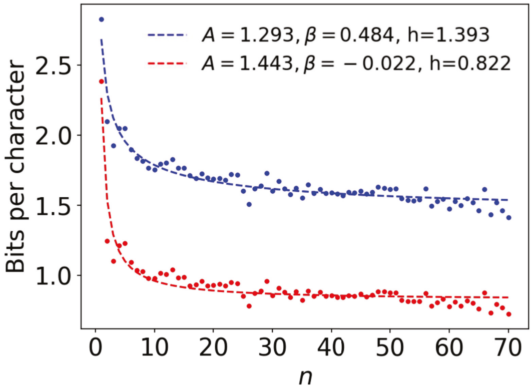


Figure 4. The plots of the upper bound (**blue**) and the lower bound (**red**) acquired from all observations and their extrapolations via ansatz functions of f_1 (dashed lines).

Plots of both bounds have large fluctuations for $n > 70$ due to the decrease in the sample size for large n, which will be examined later in Section 5.1. The minimum experimental value of the upper bound was $h_{expmin} \equiv 1.407$ bpc, which was located at $n = 70$. Since this is the minimum of the direct experimental values, any computed entropy rate larger than this would appear to be invalid. In the remainder of this paper, the observations collected up to $n = 70$ are utilized.

4. Extrapolation of the Bounds with an Ansatz Function

As mentioned in the Introduction, the other drawback of the previous studies utilizing the cognitive approach to the entropy rate lies in not extrapolating the experimental values. Precisely, in the previous cognitive experiments [1,10,11], the reported entropy rate values were the direct upper bounds at the largest n used, such as $n = 100$ in [1].

As the entropy rate, by definition, is the value of F_n with n tending to infinity, its upper and lower bounds, as n tends to infinity, must be considered and can be examined via some extrapolation functions.

4.1. Ansatz Functions

As the mathematical nature of a natural language time series is unknown, such a function can only be an ansatz function. The first ansatz function was proposed by Hilberg [23], who hypothesized that the entropy rate decreases according to the power function with respect to n based on the experimental results of Shannon [1]. This function is as follows:

$$f_1(n) = An^{\beta-1} + h, \qquad\qquad \beta < 1. \tag{6}$$

Originally, this function was proposed without the h term. There have been theoretical arguments as to whether $h = 0$ [2–5,7,24,25]; therefore, a function with the h term was considered in this work.

Takahira et al. [4] suggested another possibility that modifies the function $f_1(n)$ slightly, which is as follows:

$$f_2(n) = \exp\left(An^{\beta-1} + h\right), \qquad\qquad \beta < 1. \tag{7}$$

They observed that the stretched exponential function $f_2(n)$ leads to a smaller value of h by roughly 0.2 bpc in a compression experiment for English characters.

Schümann and Grassberger [3] introduced another function $f_3(n)$ based on their experimental result:

$$f_3(n) = An^{\beta-1} \log n + h, \qquad\qquad \beta < 1. \tag{8}$$

These three ansatz functions f_1, f_2, and f_3 will be evaluated based on their fit to the data points discussed in the previous section. For f_1 and f_3, h is the estimated value at infinite n, whereas in the case of f_2, the estimated value of the upper and lower bounds at infinity is e^h.

4.2. Comparison among Ansatz Functions Using All Estimates

Every ansatz function was fitted to the plots of the upper and lower bounds via the Levenberg–Marquardt algorithm for minimizing the square error. The ansatz functions' fits to the data points mentioned in Section 3.5, are shown in Figure 4 for f_1 and in Figure A1 in the Appendix A for f_2 and f_3.

For f_1 and f_2, the fits converged well and the errors were also moderate. The mean-root-square error of f_1 was 0.044, quite close to the error of f_2, which was 0.043. Both the entropy rate estimates also converged to similar values of h; namely, $h = 1.393$ and $h = 1.353$ bpc, respectively, for the upper bounds. The values of β, were 0.484 and 0.603 for f_1 and f_2, respectively, suggesting monotonic decay in both cases.

On the other hand, f_3 presented some problems. The function did not fit well, and the error was 0.069. Above all, f_3's extrapolated upper bound was $h = 1.573$ bpc. The value is larger than the minimum experimental value $h_{expmin} = 1.407$ bpc considered in Section 3.5.

This tendency of f_3 to overestimate the value h may be the result of $f_3(n)$ having been designed based on the convergence of the entropy rate of some random sequence. Therefore, a suitable ansatz function would be either f_1 or f_2. As seen, they provide similar results, which is consistent with the original observation provided in [4]. Consequently, we focus on f_1, the most conventional ansatz, in the following section.

5. Analysis via the Bootstrap Technique

Section 2.3 mentioned that the scale of our experiment was significantly larger than the scales used in previous experiments [1,9,11]. The large number of observations allowed us to investigate the effect of the number of observations via the bootstrap technique, which uses subsets of the experimental samples.

5.1. The Effect of the Sample Size

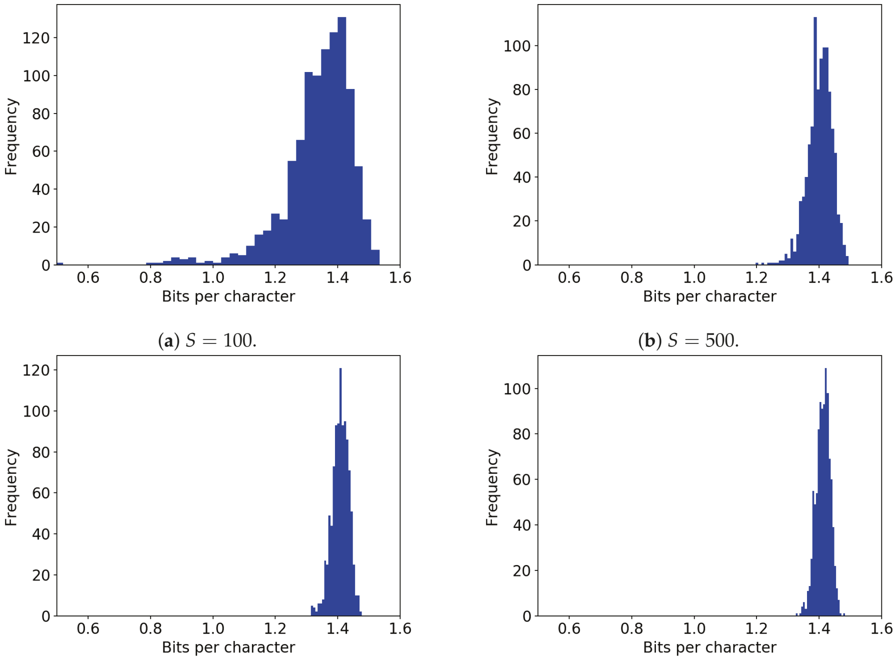
B sets of observations, each of which include S records of the experimental sessions, were sampled without redundancy. Let S be referred to as the *sample size* in the following discussion. As defined in Section 2.1, a record of an experimental session consists of a series of the number of guesses for each context of length $n-1$ produced by the same subject for a phase.

For each set, the upper bound of every n is the rightmost term in Equation (5), and an acquired set of points is extrapolated with the ansatz function f_1. We obtain B different values of h. In addition to their mean value, it would be reasonable to examine the interval between some bounds for the entropy rate estimate. We consider these bounds based on the fixed percentile of B values of h. We set $B = 1000$ and acquired the means and both bounds at 5% upper/lower percentiles for different values of S.

Figure 5 shows the histograms of h values for $S = 100, 500, 1000$, and 1500. At $S = 100$, the estimated values vary widely, and the 5% percentile bounds are $h = 1.124$ bpc and $h = 1.467$ bpc, as shown in Table 3. The previous experiments, including Shannon's study [1,9,11], used a maximum of $S = 100$ observations for certain values of n. Our results suggest that the values reported by these works have large intervals around them and should not be considered to be general results.

Furthermore, for small S, the estimated values also tend to be biased towards smaller values. The mean value at $S = 100$ was $h = 1.340$ bpc, which is about 0.07 bpc smaller than the value $h = 1.412$ bpc obtained for $S = 1000$. This underestimation occurred due to the fact that an event with small probability cannot be sampled when the sample size is small. Such events with small probabilities then contribute to increasing the entropy. When their contributions are ignored, the estimate tends to be smaller than its true value. Consequently, Shannon's original experiment could have underestimated the upper bound.

These observations suggest that a large sample size is necessary to obtain convergence of the upper bound. As observed in the values reported in Table 3, the histograms Figure 5, the red data points, and the shaded area in Figure 6, the differences between the 5% upper/lower percentile bounds decrease with larger sample size S. At $S = 1000$, the difference between the bounds is smaller than 0.1 bpc, which is a reasonably acceptable margin of error.



(**a**) $S = 100$.

(**b**) $S = 500$.

(**c**) $S = 1000$.

(**d**) $S = 1500$.

Figure 5. Histograms for the estimated values of the upper bound of the entropy rate h for different sample sizes. (**a**) $S = 100$; (**b**) $S = 500$; (**c**) $S = 1000$; (**d**)$S = 1500$.

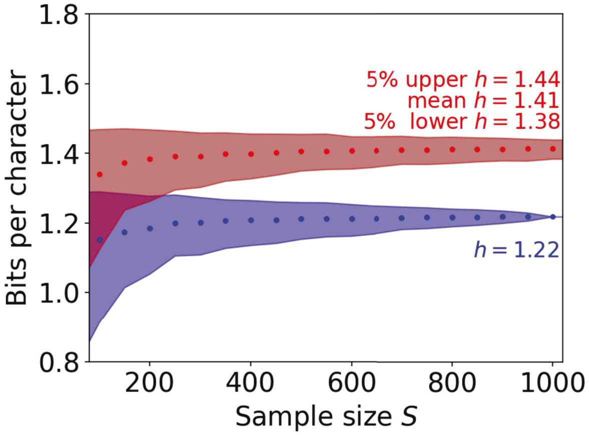


Figure 6. The estimated upper bounds with ansatz function f_1 using: (1) 1000 experimental sessions with the best prediction performances (**blue**), and (2) all experimental sessions (**red**), with the values reported in Table 3. The blue and red points indicate the mean values for the $B = 1000$ sets, and the shaded areas indicate the 5% percentile bounds.

Table 3. The means and the 5% percentile-bound-intervals for the upper bound of h found by using the ansatz function f_1 for $S = 100, 500, 1000$, and 1500. The number of sets is $B = 1000$. The error is large for a small sample sizes, such as $S = 100$, as the difference between the 5% percentile upper and lower bounds is larger than 0.3 bpc. This difference decreases with increasing S and eventually becomes smaller than ±0.1 bpc for $S \geq 1000$.

Sample Size S	**Mean**	**5% Upper**	**5% Lower**
100	1.340	1.467	1.124
200	1.383	1.468	1.263
300	1.391	1.459	1.302
400	1.398	1.456	1.327
500	1.405	1.455	1.349
1000	1.412	1.438	1.383
1500	1.411	1.444	1.374

5.2. The Effect of Variation on Subjects' Estimation Performances

Our experiment was conducted with anonymous subjects, and therefore, was less controlled than an in-laboratory experiment. Such factors could influence the entropy rate estimate; therefore, the bias is examined in this section.

Although the residences of the participants were limited to native English speaking countries, as mentioned in Section 3.3, we could not control the native tongues of our participants. Although our phrases were extracted from the *Wall Street Journal* and the terms and expressions were easy to understand, even for non-natives (see Table 2), the results might be biased. In addition, the experiment was not supervised on site; therefore, subject conditions could have varied.

In principle, the entropy rate measures the maximal predictability of the text. Therefore, each estimated value should be obtained based on the maximal performance of the subject. Here, we consider estimating the entropy rate with only the best-performed experimental sessions. We first defined the performance of an experimental session as the average number of guesses required to predict the succeeding character X_n. The experimental sessions for which the maximal n was less than 70 were filtered out in order to keep the sample size the same for all $n = 1 \ldots 70$.

Next, the experimental sessions were sorted by performance, and the $S = 1000$ best sessions are selected. Note that this S was necessary for obtaining convergence, as seen in the previous section.

We evaluated the mean and 5% percentile bounds of the best-performing set by measuring the upper bound h from $B = 1000$ sets of $S = 100, 150, 200, \ldots, 1000$ sub-samples. At $S = 1000$, there is only one possible set; therefore, h can have just one value. The results are shown in Figure 6. The blue data points in the middle show the means, and the blue-colored areas around them shows the intervals contained within the 5% percentile bounds. Similar to the results for all experiment sessions (shown as red data points and a red-shaded area), the widths of the intervals are quite large for small sample sizes, such as $S = 100$, and decrease towards $S = 1000$. The mean value of the upper bound increased with respect to S, which is also similar to the result for all experiment sessions.

Using just the selected experimental sessions, the final estimated value converged to $h \approx 1.22$ bpc, which is smaller than the value estimated when using all experimental sessions h_{expmin} and those acquired by previous cognitive experiments.

6. Discussion

6.1. Computational versus Cognitive Methods

In parallel with the cognitive approach, computational approaches have also attempted to estimate the entropy rate's upper bound for natural language. Such an approach requires that some language model be used, and previous estimates have been found with, for example, the n-gram language model [2], compression algorithm model [3,4], and neural language model [5,7]. In particular, Brown et al. [2] constructed the word-level n-gram language model and obtained $h = 1.63$ bpc, whereas Takahira et al. [4] conducted a compression experiment using giga byte-scale newspaper corpora and obtained an estimate of $h = 1.32$.

In addition to the compression algorithms and n-gram language models, recent works have also employed neural language models, which potentially have higher capacities for accurately predicting future characters. Recently, Dai et al. [7] reported $h = 1.08$ bpc when using Transformer XL on text8. This dataset is a collection of natural language text taken from Wikipedia and cleaned to the point of having only 26 alphabet characters and space corresponding to the setting of the Shannon's experiment. That h value was smaller than our estimated value, suggesting that humans may not be able to outperform computational models in character guessing games. Nevertheless, it is worth considering the differences in the conditions of the experiments.

The primary factor is the context length. Dai et al. [7]'s model utilized several hundred context lengths to acquire their results. The high performance of the neural language models can be explained, at least partially, by their ability to utilize long contexts. However, humans *are* also able to utilize long contexts, at least as long as $n \approx 10^2$, to improve their prediction performances, whereas our experiment used the context lengths of up to $n = 70$ to obtain the upper bound for h.

Furthermore, while a cognitive experiment obtains the upper bound of the entropy rate from the number of guesses, when using the computational model, the estimate is calculated based on the probability assigned to the correct character. With a distribution at hand, the upper bound of the computational model can be evaluated more tightly and precisely. The design of an experiment that incorporates a longer context length and character probability distributions is a direction of research that may be pursued in future work.

6.2. Application to Other Languages and Words

This work focused on English, which is the most studied language within the context of entropy rate estimation. Shannon's experiment is applicable to other languages if the alphabet size of the writing system is comparable with that of English.

In contrast, for ideographic languages such as Chinese and Japanese, which have much larger alphabet sizes, it is practically impossible to conduct Shannon's experiment. A prediction could involve thousands of trials until a subject reaches the correct character. Therefore, a new experimental design is required to estimate the entropy rate for these languages with large alphabet sizes.

Such an experimental setting would be also applicable to the estimation of the entropy rate at the word level, which could be interesting to investigate via a cognitive approach. Humans partly generate text word by word and character by character (sound by sound). Thus, any analysis could reveal new information about linguistic communication channels, including their distortions, as studied in [26,27].

6.3. Nature of h Revealed by Cognitive Experimentation

Provided with some previous work and the good fit of an ansatz extrapolation function while assuming that $h \geq 0$ and using what we consider reliable data points, we arrived at h = 1.22.

There is more than one way, however, to investigate the true value of h. Figure 4 shows how data points for larger n become lower than the estimated ansatz, perhaps suggesting that the values tend to zero even for larger n. It could be the case that h goes to zero. Indeed, a function without an h term (i.e., $h = 0$) would fit reasonably well if the upper bound is evaluated only with relatively small datapoints of n such as $n \leq 70$. Overall, our analysis does not rule out the possibility of the zero entropy rate.

One observation gained from this work that highlighted the sample size is that data points are distributed and statistical margins must be considered. Hence, h should be considered as having a distribution and not as a single value. One such way of analysis was described in Section 5.

7. Conclusions

This paper presented a large-scale cognitive experiment for estimating the entropy rate for English. Using AMT, we conducted Shannon's experiment online and collected 172,954 character predictions in total across 683 subjects. It was by far the largest cognitive experiment conducted thus far, and the scale enabled us to analyze the factors that influence the estimation.

While Shannon implied that subjects' prediction performances improved with increasing context length, others disagreed with his implication. Our experiment showed that subjects' prediction performances improved consistently with increasing context length, at least up to 100 characters.

Further, we investigated the influence of the number of observations on the estimation via the bootstrap technique. One of the most important insights gained is that the number of prediction observations must be at least 1000 in order to produce an estimate with a reasonable margin of error. In the case of small samples, the value of h could be potentially underestimated. Hence, Shannon's original experiment and other previous experiments provided estimates that could have been underestimated. We believe that this present work reports a statistically reliable estimate with a reasonable margin of error.

Due to the online environment, the performances of the subjects varied, and the upper bound should be evaluated based on filtered results. With a sufficient number of well-performing samples, we obtained an upper bound of $h \approx 1.22$ bpc, which is slightly smaller than Shannon's reported value of $h = 1.3$ bpc.

Future work could include finding a new experimental design, one in which the participants use longer contexts to predict the next character; thus, reducing the cognitive load. Such an experiment would contribute to the tighter evaluation of the upper bound of the entropy rate. It would be also interesting to examine the entropy rates of other languages and at the word level while still utilizing a cognitive experiment.

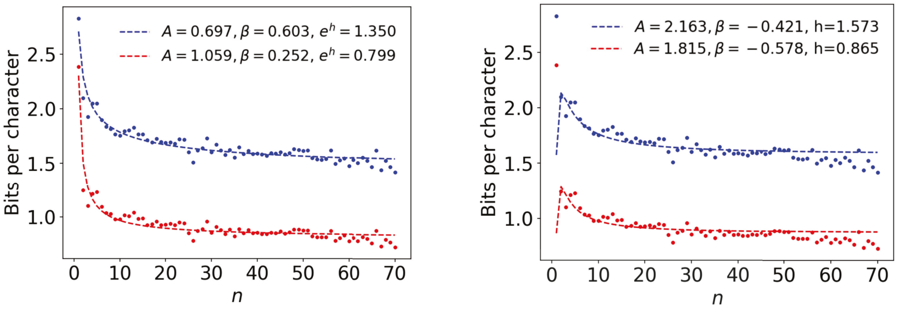
Author Contributions: Conceptualization, K.T.-I.; Data curation, G.R.; Funding acquisition, K.T.-I.; Investigation, G.R., K.T.-I. and S.T.; Methodology, G.R., K.T.-I. and S.T.; Project administration, K.T.-I.; Resources, K.T.-I.; Software, G.R. and S.T. ; Supervision, K.T.-I.; Validation, S.T. and K.T.-I.; Visualization, G.R., S.T. and K.T.-I.; Writing—original draft, S.T. and K.T.-I.; Writing—review & editing, K.T.-I. and S.T.

Funding: This research was funded by RISTEX-HITE 176100000214 of Japan Science Technology Agency, Japan.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The fits of f_2 and f_3 to the same data points (as opposed to f_1, shown in Figure 4) are shown in Figure A1.



(**a**) $f_2(n)$.

(**b**) $f_3(n)$.

Figure A1. The plots of the upper bounds (**blue**) and lower bounds (**red**) acquired from all observations and their extrapolations via the ansatz functions f_2 and f_3 (shown as the dashed lines).

References

1. Shannon, C.E. Prediction and entropy of printed English. *Bell Syst. Tech. J.* **1951**, *30*, 50–64. [CrossRef]
2. Brown, P.F.; Pietra, S.A.D.; Pietra, V.J.D.; Lai, J.C.; Mercer, R.L. An Estimate of an Upper Bound for the Entropy of English. *Comput. Linguist.* **1992**, *18*, 31–40.
3. Schümann, T.; Grassberger, P. Entropy estimation of symbol sequences. *Chaos* **1996**, *6*, 414–427. [CrossRef] [PubMed]
4. Takahira, R.; Tanaka-Ishii, K.; Dębowski, Ł. Entropy Rate Estimates for Natural Language—A New Extrapolation of Compressed Large-Scale Corpora. *Entropy* **2016**, *18*, 364. [CrossRef]
5. Takahashi, S.; Tanaka-Ishii, K. Cross Entropy of Neural Language Models at Infinity—A New Bound of the Entropy Rate. *Entropy* **2018**, *20*, 839. [CrossRef]
6. Manning, C.; Schutze, H. *Foundations of Statistical Natural Language Processing*; MIT Press: Cambridge, MA, USA, 1999.
7. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 2978–2988.
8. Burton, N.G.; Licklider, J.C.R. Long-Range Constraints in the Statistical Structure of Printed English. *Am. J. Psychol.* **1955**, *68*, 650–653. [CrossRef] [PubMed]
9. Jamison, D.; Jamison, K. A note on the entropy of partially-known languages. *Inf. Control* **1968**, *12*, 164–167. [CrossRef]
10. Cover, T.M.; King, R.C. A Convergent Gambling Estimate of the Entropy of English. *IEEE Trans. Inf. Theory* **1978**, *24*, 413–421. [CrossRef]
11. Moradi, H.; Grzymala-Busse, J.; Roberts, J. Entropy of English Text: Experiments with Humans and a Machine Learning System Based on Rough Sets. *Inf. Sci.* **1998**, *104*, 31–47. [CrossRef]
12. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley-Interscience: Hoboken, NJ, USA, 2006.
13. Kelly, J.L., Jr. A New Interpretation of Information Rate. *Bell Syst. Tech. J.* **1956**, *35*, 917–926. [CrossRef]
14. Breiman, L. Optimal gambling systems for favorable games. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Berkeley, CA, USA, 1961; Volume 1, pp. 65–78.
15. Sorokin, A.; Forsyth, D. Utility data annotation with Amazon Mechanical Turk. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Anchorage, AK, USA, 23–28 June 2008.

16. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Berg, M.B.A.C.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
17. Callison-Burch, C.; Dredze, M. Creating speech and language data with Amazon's Mechanical Turk. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, CA, USA, 6 June 2010; pp. 1–12.
18. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–4 November 2016; pp. 2383–2392.
19. Mason, W.; Suri, S. Conducting behavioral research on Amazon's Mechanical Turk. *Behav. Res. Methods* **2012**, *44*, 1–23. [CrossRef] [PubMed]
20. Urbano, J. Morato, M.M.; Martín, D. Crowdsourcing Preference Judgments for Evaluation of Music Similarity Tasks. In *ACM SIGIR Workshop on Crowdsourcing for Search Evaluation*; ACM: New York, NY, USA, 2010.
21. Alonso, O.; Mizzaro, S. Using crowdsourcing for TREC relevance assessment. *Inf. Process. Manag. Int. J.* **2012**, *48*, 1053–1066. [CrossRef]
22. Warby, S.C.; Wendt, S.L.; Welinder, P.; Munk, E.G.S.; Carrillo, O.; Sorensen, H.B.D.; Jennum, P.; Peppard, P.E.; Perona, P.; Mignot, E. Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. *Nat. Methods* **2014**, *11*, 385–392. [CrossRef] [PubMed]
23. Hilberg, W. Der bekannte Grenzwert der redundanzfreien Information in Texten-eine Fehlinterpretation der Shannonschen Experimente? *Frequenz* **1990**, *44*, 243–248. [CrossRef]
24. Genzel, D.; Charniak, E. Entropy Rate Constancy in Text. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 June 2002; pp. 199–206.
25. Levy, R.; Jaeger, T.F. Speakers optimize information density through information density through syntactic reduction. In Proceedings of the 19th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 4–7 December 2006; pp. 849–856.
26. Berger, T. Distortion Theory for Sources with Abstract Alphabets and Memory. *Inf. Control* **1968**, *13*, 254–273. [CrossRef]
27. Shannon, C.E. Coding theorems for a discrete source with a fidelity criterion. *Int. Conv. Rec.* **1959**, *7*, 142–163.

entropy

Article

Semantic Entropy in Language Comprehension

Noortje J. Venhuizen *, Matthew W. Crocker and Harm Brouwer

Department of Language Science & Technology, Saarland University, 66123 Saarbrücken, Germany; crocker@coli.uni-saarland.de (M.W.C.); brouwer@coli.uni-saarland.de (H.B.)

* Correspondence: noortjev@coli.uni-saarland.de

Received: 30 October 2019; Accepted: 25 November 2019; Published: 27 November 2019

Abstract: Language is processed on a more or less word-by-word basis, and the processing difficulty induced by each word is affected by our prior linguistic experience as well as our general knowledge about the world. Surprisal and entropy reduction have been independently proposed as linking theories between word processing difficulty and probabilistic language models. Extant models, however, are typically limited to capturing linguistic experience and hence cannot account for the influence of world knowledge. A recent comprehension model by Venhuizen, Crocker, and Brouwer (2019, *Discourse Processes*) improves upon this situation by instantiating a comprehension-centric metric of surprisal that integrates linguistic experience and world knowledge at the level of interpretation and combines them in determining online expectations. Here, we extend this work by deriving a comprehension-centric metric of entropy reduction from this model. In contrast to previous work, which has found that surprisal and entropy reduction are not easily dissociated, we do find a clear dissociation in our model. While both surprisal and entropy reduction derive from the same cognitive process—the word-by-word updating of the unfolding interpretation—they reflect different aspects of this process: state-by-state expectation (surprisal) versus end-state confirmation (entropy reduction).

Keywords: natural language; entropy; neural networks

1. Introduction

Language is processed on a more or less word-by-word basis, and certain words induce more processing effort (as reflected in higher reading times; RTs) than others. Inspired by Shannon's [1] theory of communication, it has been proposed that the informativity of a word is proportional to the processing effort that it induces. One way to quantify word informativity is using the notion of *surprisal*, which is a metric that quantifies the expectancy of a word [2,3]; the less expected a word is in a given context, the higher its surprisal (also called *self-information*). A second metric for word informativity is the *entropy reduction* induced by a word, which quantifies the extent to which the word decreases the amount of uncertainty about what is being communicated [4]. Surprisal and entropy reduction have been independently proposed as relevant linking hypotheses between probabilistic language models and processing difficulty [5–15]. That is, instantiations of these metrics provide a computational-level explanation (in terms of Marr [16]) of how the probability of a word in a linguistic context (estimated using language models) affects processing difficulty. There exists, however, a range of experimental findings that show that the processing difficulty of individual words is not only affected by their probability as part of the (local) linguistic context but is also affected by the larger discourse and visual context as well as by general knowledge about the world (see, e.g., [17–32]). Hence, in order to explain these findings in terms of word informativity, the information-theoretic metrics of surprisal and entropy reduction should take into account the probabilistic structure of the world, above and beyond that of the linguistic signal alone. This means that existing instantiations of these information-theoretic metrics, which are generally based on language models, should either be augmented with a probabilistic notion of extra-linguistic knowledge or be redefined in terms of the underlying cognitive processes.

In this paper, we take the latter approach by building upon previous work by Venhuizen et al. [33] (henceforth, VCB), who put forward a model of language comprehension in which surprisal estimates are derived from the probabilistic, distributed meaning representations that the model constructs on a word-by-word basis. By systematically manipulating the model's linguistic experience (the linguistic input history of the model) and world knowledge (the probabilistic knowledge captured within the representations), VCB show that, like human comprehenders, the model's comprehension-centric surprisal estimates are sensitive to both of these information sources. Since surprisal in this model directly derives from the process of incremental linguistic comprehension, the model offers an explanation at Marr's representational and algorithmic level of how linguistic experience and world knowledge can affect processing difficulty as quantified by surprisal. Given that entropy reduction has been argued to be a relevant predictor of processing difficulty independent of surprisal [15], we here extend these results by deriving a comprehension-centric metric of entropy from the meaning representations that the model constructs. Whereas previous instantiations of entropy in language are defined over linguistic structures (e.g., Probabilistic Context-Free Grammar, PCFG, states [4,14], parts-of-speech [8], or individual words [15]), we here define entropy as the amount of uncertainty relative to the state of affairs of the world. That is, the entropy reduction of a word w_t quantifies how much uncertainty regarding the current state of affairs is taken away by processing word w_t. Empirical support for such an approach comes from a recent study of situated language comprehension, which manipulated only the visual context, thus keeping (linguistic) surprisal constant [34]. Words that reduce referential entropy to a greater extent—with respect to a visual context—led to increased processing effort for otherwise identical utterances.

We investigate whether the comprehension-centric notions of surprisal and entropy reduction make differential predictions within the model and how these metrics relate to the underlying cognitive process of comprehension. Based on the results, we conclude that surprisal and entropy reduction derive from a single cognitive process—comprehension as navigation through meaning space—and that they reflect different aspects of this process: state-by-state expectation (surprisal) versus end-state confirmation (entropy reduction). Critically, while previous language model-based instantiations have found that surprisal and entropy reduction are not easily dissociated [15], the comprehension-centric perspective on word informativity predicts that surprisal and entropy reduction differentially reflect effects of linguistic experience and world knowledge during online comprehension.

In what follows, we first introduce the probabilistic, distributed meaning representations used by VCB [33], from a novel, formal semantic perspective (cf. [35]) (Section 2.1). Next, we describe the comprehension model (Section 2.2.1) as well as how processing in this model gives rise to a comprehension-centric notion of surprisal (Section 2.2.2). From here, a comprehension-centric notion of entropy is derived (Section 2.3). The remainder of the paper, then, explores how and why comprehension-centric entropy reduction differs from comprehension-centric surprisal (Section 3). Finally, we discuss the implications of our findings and outline directions for further study (Section 4).

2. Comprehension-Centric Surprisal and Entropy

VCB [33] present a computational model of language comprehension that explicates how world knowledge and linguistic experience are integrated at the level of interpretation and combine in determining online expectations. To this end, they present a neural network model that constructs a representation of utterance meaning on an incremental, word-by-word basis. It is shown that word surprisal naturally derives from the incremental construction of these meaning representations, and that it is affected by both linguistic experience (the linguistic input history of the model) and world knowledge (the probabilistic knowledge captured within the representations). Here, we will show that in addition to this comprehension-centric notion of surprisal, the meaning representations also allow for the definition of a comprehension-centric notion of entropy.

2.1. Meaning in a Distributional Formal Meaning Space

The notion of surprisal presented in [33] exploits the rich, probabilistic meaning representations that are constructed by the comprehension model on a word-by-word basis. These representations, which are based on the Distributed Situation-state Space framework [36,37], are argued to instantiate situation models that allow for world knowledge-driven inference. Following [35], we here reconceptualize this approach in terms of model-theoretic semantics, thereby emphasizing the generalizability of the framework.

Based on a set of propositions $\mathcal{P}$, and a set of models formal models $\mathcal{M}$ (which can be defined as combinations of the propositions in $\mathcal{P}$), we can define a meaning space: $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ (see Figure 1). Importantly, the set of models $\mathcal{M}$ is assumed to reflect the state of the world truth-conditionally and probabilistically (i.e., reflecting the probabilistic structure of the world). The meaning of a proposition $p \in \mathcal{P}$ is defined as the vector $\vec{v}(p)$ that, for each $M \in \mathcal{M}$, assigns a 1 iff M satisfies p ($M \vDash p$) and a 0 otherwise. The resulting meaning vector captures the truth conditions of individual propositions indirectly by identifying the models that satisfy the proposition. Because the meaning vectors of all propositions are defined with respect to the same set of models, the distributional meaning of any $p \in \mathcal{P}$ is defined in relation to all other $p' \in \mathcal{P}$; that is, propositions that have related meanings will be true in many of the same models and hence have similar meaning vectors.

	p_1	p_2	p_3	$\ldots$	p_n
M_1	1	0	0	$\ldots$	1
M_2	0	1	1	$\ldots$	1
M_3	1	1	0	$\ldots$	0
$\ldots$	.	.	.	$\ldots$	.
M_m	0	1	0	$\ldots$	0

Figure 1. Example of a meaning space $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$, where $\mathcal{M} = \{M_1, \ldots, M_m\}$ defines the set of models and $\mathcal{P} = \{p_1, \ldots, p_n\}$ the set of propositions. Rows represent models as combinations of propositions and columns represent meaning vectors that derive from this space, such that: $\vec{v}_i(p_j) = 1$ *iff* $M_i \vDash p_j$.

Given well-defined sets of models $\mathcal{M}$ and propositions $\mathcal{P}$ (i.e., $\mathcal{P}$ fully describes the set of propositions that can be captured in $\mathcal{M}$), the resulting vector space $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ offers distributed representations that are compositional and probabilistic. To start with the former, the meaning space $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ not only allows for deriving the meaning vectors of individual propositions in $\mathcal{P}$ but also combinations thereof. That is, given a definition of negation and conjunction over meaning vectors, the meaning of any logical combination of propositions in the semantic space can be defined. The meaning vector $\vec{v}(p)$ of a proposition $p \in \mathcal{P}$ defines its truth values relative to $\mathcal{M}$, which means that we can define its negation $\vec{v}(\neg p)$ as the vector that assigns 0 to all $M \in \mathcal{M}$ such that p is satisfied in M and 1 otherwise:

$$\vec{v}_i(\neg p) = 1 \text{ iff } M_i \nvDash p \text{ for } 1 \leq i \leq |\mathcal{M}|. \tag{1}$$

The meaning of the conjunction $p \wedge q$, given $p, q \in \mathcal{P}$, then, is defined as the vector $\vec{v}(p \wedge q)$ that assigns 1 to all $M \in \mathcal{M}$ such that M satisfies both p and q and 0 otherwise:

$$\vec{v}_i(p \wedge q) = 1 \text{ iff } M_i \vDash p \text{ and } M_i \vDash q \text{ for } 1 \leq i \leq |\mathcal{M}|. \tag{2}$$

The probabilistic nature of the meaning space $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ derives from the fact that the meaning vectors for individual propositions in $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ inherently encode their probability. Given a set of models $\mathcal{M}$ that reflects the probabilistic nature of the world, the probability of any formula φ can be defined by the number of models that satisfy φ, divided by the total number of models:

$$P(\varphi) = |\{M \in \mathcal{M} \mid M \vDash \varphi\}|/|\mathcal{M}|. \tag{3}$$

Thus, (logical combinations of) propositions that are true in a large set of models will obtain a high probability and vice versa. Given that this directly allows for the definition of the conjunctive probability of two formulas, we can also define the conditional probability of any formula ψ given φ:

$$P(\psi|\varphi) = P(\varphi \wedge \psi)/P(\varphi). \tag{4}$$

In order to obtain sensible probability estimations about propositional co-occurrence in $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$, the set of models $\mathcal{M}$ needs to reflect the probabilistic structure of the actual world regarding the truth-conditions and co-occurrence of each proposition $p \in \mathcal{P}$. Arriving at such a set of models $\mathcal{M}$ is a non-trivial exercise. One possible strategy would be to deduce the meaning space from annotated corpora or knowledge bases with world knowledge-driven inferences (e.g., [38]), or from crowd-sourced human data on propositional co-occurrence (e.g., [39]). However, in order to empirically evaluate how the information-theoretic notion of entropy (reduction) is affected by the structure of the world, the co-occurrence between propositions needs to be defined in a controlled manner. Therefore, the meaning representations used here (following VCB [33]) are induced from a high-level description of the structure of the world, using an incremental, inference-driven construction procedure [35].

2.2. A Model of Surprisal Beyond the Words Given

2.2.1. The Comprehension Model

The model presented by VCB [33] is a simple recurrent neural network (SRN) [40] consisting of three groups of artificial logistic dot-product neurons: an INPUT layer (21 units), HIDDEN layer (100 units), and OUTPUT layer (150 units) (see Figure 2). Time in the model is discrete, and at each processing time-step t, activation flows from the INPUT through the HIDDEN layer to the OUTPUT layer. In addition to the activation pattern at the INPUT layer, the HIDDEN layer also receives its own activation pattern at time-step $t-1$ as input (effectuated through an additional CONTEXT layer, which receives a copy of the activation pattern at the HIDDEN layer prior to feed-forward propagation). The HIDDEN and the OUTPUT layers both receive input from a bias unit (omitted in Figure 2). The model was trained using bounded gradient descent [41] to map sequences of localist word representations constituting the words of a sentence onto a meaning vector from $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$, representing the meaning of that sentence.

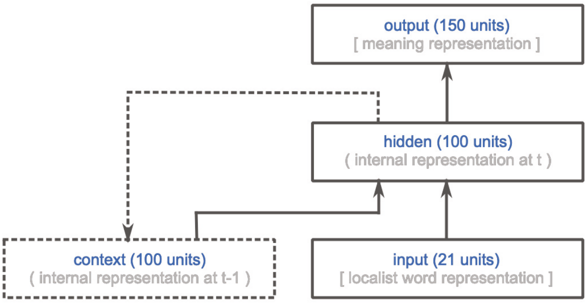

output (150 units)
[meaning representation]
hidden (100 units)
(internal representation at t)
context (100 units)
(internal representation at t-1)
input (21 units)
[localist word representation]

Figure 2. Graphic depiction of the simple recurrent neural network from [33]. Boxes represent groups of artificial neurons, and solid arrows between boxes represent full projections between the neurons in a projecting and a receiving group. The dashed lines indicate that the CONTEXT layer receives a copy of the activation pattern at the HIDDEN layer at the previous time-step. See text for details.

The sentences on which the model is trained describe situations in a world that is defined in terms of three persons ($p \in \{beth, dave, thom\}$), two places ($x \in \{cinema, restaurant\}$), two types of food ($f \in \{dinner, popcorn\}$), and three drinks ($d \in \{champagne, cola, water\}$), which can be combined using the following seven predicates: *enter(p,x)*, *ask_menu(p)*, *order(p,f/d)*, *eat(p,f)*, *drink(p,d)*, *pay(p)*, and *leave(p)*. The resulting set of propositions $\mathcal{P}$ ($|\mathcal{P}| = 45$) fully describes the world. A meaning space was constructed from these atomic propositions by sampling a set of 10K models $\mathcal{M}$, while taking into account world knowledge in terms of hard and probabilistic constraints on propositional co-occurrence; for instance, a person can only enter a single place (hard), ordering water is more common than ordering champagne (probabilistic), and eating popcorn is more likely in the cinema than in the restaurant (probabilistic) (see [33] for details). In order to employ meaning vectors derived from this meaning space in the SRN, a subset $\mathcal{M}'$ consisting of 150 models was algorithmically selected from $\mathcal{M}$, such that $\mathcal{M}'$ adequately reflected the structure of the world ([33], Appendix B). Situations in the world were described using sentences from a language consisting of 21 words. The grammar described in [33] generates a total of 117 different (active) sentences, consisting of simple noun phrase-verb phrase (NP VP) sentences and coordinated (NP VP and VP) sentences. Sentence-initial NPs identify persons, and VPs directly map onto the aforementioned propositions. The semantics assigned to the sentences were meaning vectors from $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ reflecting propositional (simple sentences) or conjunctive meanings (coordinated sentences). In order to induce differential linguistic experience in the model, some of these sentences were encountered more often than others during training; in particular, the sentences "*NP$_{person}$ ordered dinner/champagne*" occurred nine times more often than "*NP$_{person}$ ordered popcorn/water*" (whereas the frequency of the different NPs was held constant throughout the training set, see [33] for details). The resulting training set consisted of 237 sentences, which the model encountered 5000 times during training (see [33] for details on other training parameters).

After training, the model successfully learned to map sequences of word representations (representing sentences) onto meaning vectors from $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ that describe the semantics of these sentences. Since the aim is to investigate how information-theoretic metrics can be derived from the processing behavior of the model, the effects need to be tightly controlled, which is why the model is not tested using a separate set of unseen test sentences (note, however, that other models employing similar meaning representations have shown generalization to unseen sentences and semantics, in both comprehension [37] and production [42]). Instead, the performance of the model was evaluated using a comprehension score *comprehension(a,b)* [37] that indicates how well meaning vector a is understood to be the case from meaning vector b, resulting in a score that ranges from -1 (perfectly understood not to be the case) to $+1$ (perfectly understood to be the case). The average comprehension score of the intended target given the model's output vector over the entire training set was 0.89, which means that after processing a sentence, the model almost perfectly infers the intended meaning of the sentence. This shows that, due to the structured nature of the meaning representations, the (rather simple) SRN architecture suffices to obtain the desired comprehension behavior. It should be noted, however, that the meaning representations could also be employed in a more cognitively plausible architecture, in order to gain more insight into the cognitive processes underlying incremental language comprehension, for instance, by linking model behavior to electrophysiological correlates [43].

2.2.2. A Comprehension-Centric Notion of Surprisal

On the basis of its linguistic input, the comprehension model incrementally constructs a meaning vector at its OUTPUT layer that captures the meaning of the sentence so far; in other words, the model effectively navigates the meaning space $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ on a word-by-word basis. That is, each incoming word w_t induces a transition from a point in meaning space $\vec{v}_{t-1}$ to the next $\vec{v}_t$. Figure 3 provides a visualization of this navigation process. This figure is a three-dimensional representation of the 150-dimensional meaning space (for a subset of the atomic propositions), derived using multidimensional scaling (MDS). The grey points in this space correspond to propositional meaning vectors. As this figure illustrates, meaning in $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ is defined in terms of co-occurrence; propositions that co-occur frequently in $\mathcal{M}$ (e.g., *order(beth,cola)*, and *drink(beth,cola)*) are positioned close to each other in space. Note that multidimensional scaling from 150 into three dimensions necessarily results in a significant loss of information; therefore, distances between points in the meaning space shown in Figure 3 should be interpreted with care. The coloured points show the model's word-by-word output for the sentences "*beth entered the cinema and ordered [popcorn/dinner]*" (as the function words "*the*" and "*and*" trigger minimal transitions in meaning space, they are left out in Figure 3 to enhance readability). The navigational trajectory (indicated by the arrows) illustrates how the model assigns intermediate points in meaning space to each (sub-sentential) sequence of words. For instance, at the word "*beth*", the model navigates to a point in meaning space that is in between the meanings of the propositions pertaining to *beth*. The prior probability of propositions in $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ ("world knowledge"), as well as the sentences on which the model was trained ("linguistic experience") together determine the model's trajectory through meaning space. For instance, while the model was exposed to the sentences "*beth entered the restaurant and ordered popcorn*" and "*beth entered the restaurant and ordered dinner*" equally often, the meaning vector at the word "*ordered*" is closer to *order(beth,popcorn)* ($cos(\theta) = 0.70$) than to *order(beth,dinner)* ($cos(\theta) = 0.16$), because the former is more probable in the model's knowledge of the world (see [33] for details).

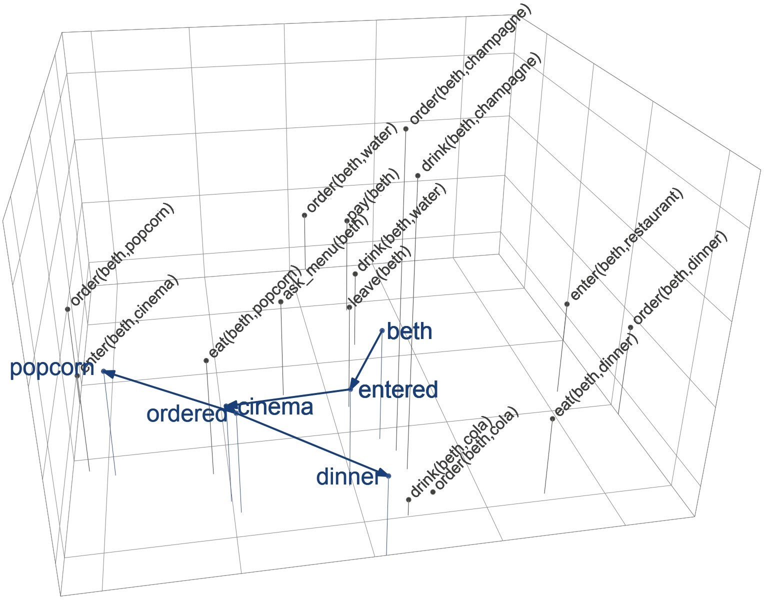


Figure 3. Three-dimensional visualization of the meaning space (by means of multidimensional scaling) for a subset of the atomic propositions (those pertaining to *beth*). Coloured points and arrows show the word-by-word navigational trajectory of the model from [33] for the sentence *beth entered the cinema and ordered [popcorn/dinner]* (function words are omitted; see text for details).

Based on the view of comprehension as meaning space navigation, VCB [33] define surprisal in terms of the points in meaning space that the model incrementally constructs. As a result, surprisal in the model essentially reflects the distance of transitions in meaning space: in case the meaning vector after processing word w_t (i.e., $\vec{v}_t$) is close to the previous point in meaning space $\vec{v}_{t-1}$, the transition induced by word w_t is small, indicating that this word is unsurprising. If, on the other hand, $\vec{v}_t$ is far away from $\vec{v}_{t-1}$, the transition induced by word w_t is big, and thus, this word is highly surprising. Because of the probabilistic nature of the meaning representations that are derived from $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$, the conditional probability $P(\vec{v}_t|\vec{v}_{t-1})$ can be calculated directly from the meaning vectors (see Equation (4)). This, then, results in the following definition of surprisal:

$$S(w_t) = -\log P(\vec{v}_t|\vec{v}_{t-1}). \tag{5}$$

That is, the surprisal induced by word w_t is inversely proportional to the conditional probability of the meaning vector constructed after processing word w_t, given the meaning vector constructed after processing words $w_1, \ldots, w_{t-1}$. VCB [33] show that this comprehension-centric notion of surprisal is sensitive to both the *world knowledge* represented in the meaning representations as well as to the *linguistic experience* of the model. World knowledge derives from the probabilistic structure of the meaning space, as determined by the truth-conditional and probabilistic co-occurrences between propositions in $\mathcal{M}$. As a result, (sub-)propositional meaning vectors that are likely to co-occur in the world will be close to each other in meaning space. Hence, the surprisal of word w_t will be affected by how likely its resultant meaning vector $\vec{v}_t$ is to co-occur with the previous meaning vector $\vec{v}_{t-1}$ in $\mathcal{M}$. The linguistic experience of the model, in turn, is determined by frequency differences within the set of training items. When certain sentence-final meaning vectors occur more frequently in the training data, this will affect the word-by-word navigation of the model through meaning space; that is, the meaning vector constructed at word w_{t-1} will be closer to the more frequent sentence-final meanings than to the less frequent ones. As a result, surprisal of word w_t will be lower if w_t moves the model towards a point in space that is closer to a more frequent sentence-final meaning vector. Crucially, since world knowledge and linguistic experience in the model derive from different probability distributions (i.e., over models versus training items), they need not be in unison. VCB [33] show that their notion of surprisal reflects a weighted average predictability derived from both of these sources.

2.3. Deriving a Comprehension-Centric Notion of Entropy

Entropy is a metric that quantifies the amount of uncertainty in a given state. In the context of language processing, entropy reduction defines the extent to which a word decreases the amount of uncertainty about what is being communicated, which is hypothesized to affect cognitive processing [4,14,44]. In terms of the model presented above, language comprehension can be viewed as navigating the meaning space $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ on a word-by-word basis (see again Figure 3). At each point in time t, the model finds itself at a point in meaning space, defined by the meaning vector $\vec{v}_t$, that reflects the meaning of words $w_1, \ldots, w_t$ (as derived from the linguistic experience and world knowledge available to the model). This navigational process effectively aims to recover which combinations of propositions satisfy the meaning at time t. That is, at each point in space the model tries to determine the current state of affairs in terms of the propositions in $\mathcal{P}$ (i.e., each $p \in \mathcal{P}$ is either true or false). In other words, the meaning vector $\vec{v}_t$ inherently reflects uncertainty about which fully specified state of affairs corresponds to the current point in space. The notion of entropy can be used to quantify this uncertainty.

Given $|\mathcal{P}| = n$, there are 2^n fully specified states of affairs. In order to calculate entropy, we need a probability distribution over this entire set. This will, however, quickly become infeasible (in the current model, $|\mathcal{P}| = 45$, resulting in $2^{45} > 10^{13}$ probabilities) [45]. Critically, however, not all combinations of propositions are licensed by world knowledge; only those states of affairs that correspond to one of the models that constitutes the meaning space will obtain a probability $P > 0$ (since all other combinations will yield the zero vector $\vec{0}$). That is, the models in $\mathcal{M}$ themselves represent fully specified states of affairs, which, like any other combination of propositions, can be represented as a meaning vector $\vec{v}_{M_i}$ (e.g., $\vec{v}_{M_1} = \vec{v}(p_1 \wedge \neg p_2 \wedge \ldots \wedge p_n)$; see Figure 1), which will have a 1 for exactly that unit that corresponds to model M_i. By definition these model vectors inherently carry a probability, which can be used to define entropy. To this end, we define a probability distribution over the set of meaning vectors that identify unique models in $\mathcal{M}$, i.e., $\mathcal{V}_{\mathcal{M}} = \{\vec{v}_M \mid \vec{v}_M(i) = 1$ *iff* $M_i = M$ and M is a unique model in $\mathcal{M}\}$. The probabilities of the unique models in $\mathcal{V}_{\mathcal{M}}$ form a proper probability distribution since they are by definition mutually exclusive ($P(\vec{v}_1 \wedge \vec{v}_2) = 0$ for each $\vec{v}_1, \vec{v}_2 \in \mathcal{V}_{\mathcal{M}}$ such that $\vec{v}_1 \neq \vec{v}_2$), and their probabilities sum to 1 since $\mathcal{V}_{\mathcal{M}}$ covers the entire meaning space: $\bigvee_{\vec{v}\in\mathcal{V}_{\mathcal{M}}} = \vec{1}$. At time step t, entropy can then be defined as follows:

$$H(t) = -\sum_{\vec{v}_M \in \mathcal{V}_{\mathcal{M}}} P(\vec{v}_M|\vec{v}_t) \log P(\vec{v}_M|\vec{v}_t). \tag{6}$$

Following this definition, entropy will be zero if the current meaning vector $\vec{v}_t$ singles out a unique model. If, on the other hand, all models are equally likely at t (i.e., the probability distribution over all possible models is uniform), entropy will be maximal with respect to t.

In the psycholinguistic literature, entropy has been linked to processing difficulty via the entropy reduction hypothesis (ERH), which states that the reduction of entropy "is positively related to human sentence processing difficulty" ([4], p. 650). The entropy reduction between two states, as triggered by word w_t, is defined as the difference between the entropy at state $t-1$ and the entropy at state t:

$$\Delta H(w_t) = H(t-1) - H(t). \tag{7}$$

In terms of the comprehension-centric notion of entropy, this means that an increase in processing effort is predicted for words that more greatly reduce uncertainty about fully specified states of affairs. Crucially, however, the difference in entropy between time step $t-1$ and t is not necessarily positive; that is, as the model navigates the meaning space on a word-by-word basis, individual words may in principle result in either an increase or a decrease in the uncertainty about which state of affairs is being communicated. While there is no negative entropy reduction in the current model due to the structure of the training data in which the coordinated sentences describe increasingly specific states of affairs, training the model to achieve broader empirical coverage may lead it to exhibit behavior in which it finds itself in a state of relative certainty about the communicated state of affairs, which is then challenged by additional input that moves the model toward a state of increased uncertainty (note that any slightly negative entropy reduction values shown in the plots below result from noise due to the processing behavior of the model). Hence, just as a decrease in entropy reflects the transition from a state of uncertainty to a state of greater certainty, an increase in entropy reflects the transition from a state of certainty to a state of uncertainty. As a result, both positive and negative changes in uncertainty are predicted to increase processing effort. Thus, in contrast to the—syntactically defined—ERH from [4], according to which entropy reduction (and hence, processing effort) is zero in case the current state is more uncertain than the previous state, the comprehension-centric perspective on entropy predicts that both a reduction and an increase in entropy result in an increase in processing effort. That is, the processing difficulty indexed by entropy reduction is a direct reflection of the absolute degree of change in (un)certainty ($|\Delta H(w_t)|$) about the communicated state of affairs, as induced by word w_t: the larger the change in (un)certainty between state $\vec{v}_{t-1}$ prior to processing word w_t and state $\vec{v}_t$ after processing w_t, the higher the processing difficulty.

In what follows, we will investigate how these comprehension-centric notions of entropy and entropy reduction behave in the online comprehension model from [33] and how they relate to the notion of surprisal described in Section 2.2.2.

3. Entropy Reduction in Online Comprehension

3.1. Comprehension-Centric Entropy Reduction versus Surprisal

Surprisal and entropy reduction have been independently proposed as a linking theory between probabilistic language models and human processing difficulty in online word-by-word comprehension [2–4]. Moreover, it has been shown that these information theoretic metrics also independently account for variability in word processing difficulty [15]. A first step, therefore, is to examine the degree to which the predictions of the comprehension-centric instantiations of surprisal and entropy reduction align in the model.

Figure 4 (left) plots the online surprisal estimates for each training sentence of the VCB model [33] against the corresponding online entropy reduction estimates (we here use the term "online" in order to differentiate the model-derived surprisal and entropy reduction metrics from the "offline" metrics derived from the model's training data; see Section 3.2 below). Overall, there is no significant relationship between the two metrics ($r = 0.0261$, $p = 0.408$). However, given that sentence-initial words minimally reduce uncertainty about sentence-final meaning (all sentences start with a proper name, and all models in $\mathcal{M}$ satisfy at least one proposition concerning each person), they induce a rather uniform surprisal (mean = 1.24, sd = 0.01) and entropy reduction (mean = 0.107, sd = 0.01) profile, which may cloud the relationship between these metrics. To account for this, Figure 4 (middle) shows the estimates for all but the sentence-initial word. This now reveals a significant relationship between surprisal and entropy reduction ($r = 0.177$, $p < 0.01$), albeit a weak one ($R^2 = 0.0315$), leaving the majority of variance unaccounted for. Finally, as the last word of an utterance maximally disambiguates (or confirms anticipated) utterance meaning, it is also of interest to look at these separately. Figure 4 (right) plots the estimates for all sentence-final words. At this position, there is no significant relationship between surprisal and entropy reduction ($r = 0.0778$, $p = 0.233$).

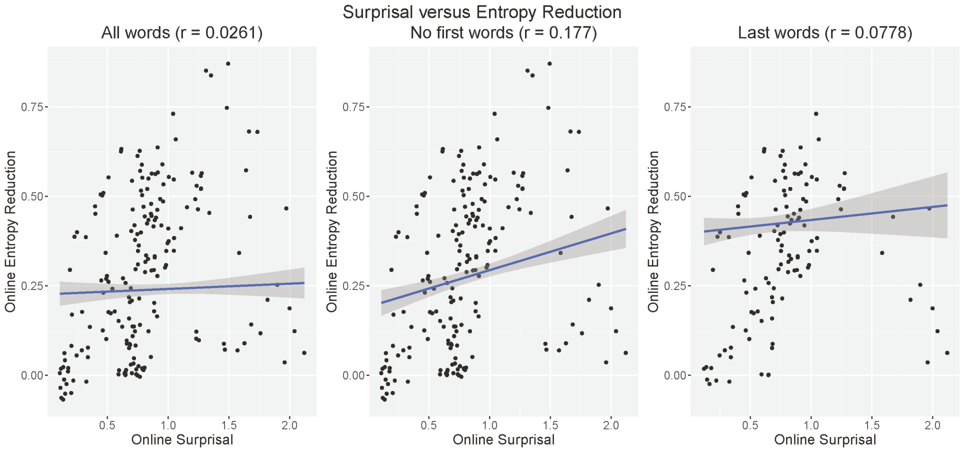


Figure 4. Comparison between online entropy reduction and online surprisal estimates. The scatter plots show the correlation between the surprisal and entropy reduction estimates for all words (**left**), all but the first words (**middle**), and for the last words only (**right**). The solid blue lines depict the corresponding linear regressions with their 95% confidence intervals. The Pearson correlation efficient (r) is shown at the top of each plot.

In summary, when we ignore the rather uniform surprisal and entropy reduction profiles at the sentence-initial words, we observe a weak positive correlation between the two metrics. This relationship, which does not appear to be driven by disambiguation or confirmation at sentence-final words, explains about 3% of the variance and hence leaves the majority of variability unaccounted for. This raises the question of where and how the comprehension-centric instantiations of the metrics diverge. VCB [33] explored the online surprisal metric by investigating its sensitivity to different degrees of linguistic experience and probabilistic world knowledge. Hence, one way forward is to examine the sensitivity of online entropy reduction under these constellations and to identify where and how it differs from online surprisal.

3.2. *Effects of Linguistic Experience versus World Knowledge*

The comprehension model of VCB [33] maps sentences onto their their corresponding probabilistic meaning vectors on an incremental, word-by-word basis. Crucially, the model is exposed to certain sentence-semantic pairs more frequently than others during training, thereby shaping its linguistic experience. In addition, as each meaning vector inherently carries its own probability in the meaning space, certain sentences can map onto meanings that are more likely than others, which provides the model with world knowledge. These individual sources of knowledge, which influence the behavior of the model, can be independently quantified in the training data using surprisal.

The linguistic experience that the model is exposed to can be quantified using the offline *linguistic* surprisal, which is straightforwardly estimated from the sentences that the model is trained on [2,3]:

$$\mathrm{S}_{ling}(w_t) = -\log P(w_t \mid w_{1,\ldots,t-1}). \tag{8}$$

If a word w_t frequently occurs after the prefix $w_1, \ldots, w_{t-1}$, its conditional probability will be high and its linguistic surprisal low (and vice versa). Crucially, this linguistic surprisal metric is not influenced by the world knowledge contained within meaning vectors; it solely derives from the distribution of word sequences in the set of training sentences.

World knowledge, in turn, can be quantified using offline *situation* surprisal, which is derived from the meaning vectors corresponding to the training sentences, rather than the sentences themselves. That is, given a sequence of words $w_1, \ldots, w_t$, a situation vector $\mathrm{sit}(w_{1,\ldots,t})$ can be derived by taking the disjunction of the semantics of all sentences that are consistent with this prefix. For instance, the situation vector of the prefix 'Dave drank' is defined as $\mathrm{sit}(\text{Dave drank}) = \vec{v}(drink(dave, water) \vee drink(dave, cola) \vee drink(dave, champagne))$, the disjunction of all meaning vectors consistent with the word sequence "Dave drank". The offline situation surprisal induced by a next word is then defined as follows:

$$\mathrm{S}_{sit}(w_t) = -\log P(\mathrm{sit}(w_{1,\ldots,t}) \mid \mathrm{sit}(w_{1,\ldots,t-1})). \tag{9}$$

If an incoming word w_t leads to a situation vector that is highly likely given the situation vector for the disjunctive semantics consistent with the words $w_1, \ldots, w_{t-1}$, its conditional probability—which is estimated through its conditional belief—will be high and its situation surprisal low and vice versa. This offline situation surprisal metric is independent of linguistic experience; it is only sensitive to probabilistic world knowledge encoded within the meaning space.

By differentially manipulating linguistic experience and world knowledge, VCB [33] investigate the behavior of their comprehension-centric, online surprisal metric under three constellations:

1. Manipulation of linguistic experience only: the model is presented with sentences that differ in terms of their occurrence frequency in the training data (i.e., differential linguistic surprisal) but that keep the meaning vector probabilities constant (i.e., equal situation surprisal).
2. Manipulation of world knowledge only: the model is presented with sentences that occur equally frequently in the training data (i.e., equal linguistic surprisal) but differ with respect to their probabilities within the meaning space (i.e., differential situation surprisal).
3. Manipulation of both linguistic experience and world knowledge: to investigate the interplay between linguistic experience and world knowledge, the model is presented with sentences in which the linguistic experience and world knowledge are in conflict with each other (i.e., linguistic experience dictates an increase in linguistic surprisal whereas world knowledge dictates a decrease in situation surprisal or vice versa).

Here, we compare the comprehension-centric notion of online entropy reduction ($\Delta\mathrm{H}_{\mathrm{onl}}$; see Equations (6) and (7)) to online surprisal ($\mathrm{S}_{\mathrm{onl}}$; see Equation (5)) under these three constellations, which are constructed by manipulating offline linguistic surprisal ($\mathrm{S}_{\mathrm{ling}}$; see Equation (8)), reflecting the linguistic experience of the model, and offline situation surprisal ($\mathrm{S}_{\mathrm{sit}}$; see Equation (9)), reflecting the world knowledge available to the model. Figure 5 shows the difference in surprisal/entropy reduction for these manipulations.

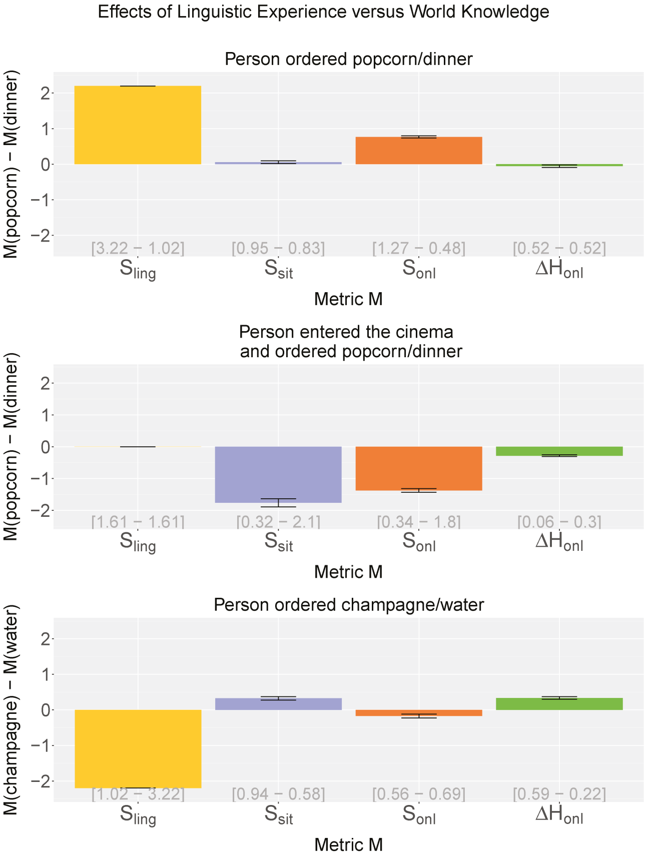


Figure 5. Effects of linguistic experience (LE) versus world knowledge (WK) on linguistic surprisal (S_{ling}), situation surprisal (S_{sit}), online surprisal (S_{onl}), and online entropy reduction (ΔH_{onl}). Bars represent differences between two target words. **Top**: Effects of LE for the contrast "*NP_{person} ordered popcorn* [T]/*dinner* [C]". **Middle**: Effects of WK for '*NP_{person} entered the cinema and ordered popcorn* [T]/*dinner* [C]". **Bottom**: Interplay between LE and WK for the contrast "*NP_{person} ordered champagne* [T]/*water* [C]". Error bars show standard errors (n = 3). Individual means are shown in brackets (T-C).

When only linguistic experience is manipulated (the sentence "*NP_{person} ordered dinner*" is more frequent than "*NP_{person} ordered popcorn*") and world knowledge is held constant ($P(\text{order(dinner)}) = P(\text{order(popcorn)})$ in the meaning space), online surprisal (S_{onl}) pairs with offline linguistic surprisal (S_{ling}) in that "popcorn" is more effortful than "dinner". Online entropy reduction (ΔH_{onl}), in turn, like offline situation surprisal (S_{sit}), shows no effect (the negligible differences between conditions are attributable to noise from the dimension selection procedure [33]); see

Figure 5 (top). By contrast, when only world knowledge is manipulated ($P(\text{order(popcorn)}|\text{cinema}) > P(\text{order(dinner)}|\text{cinema}))$, and linguistic experience is held constant ("*NP_{person} entered the cinema and ordered popcorn/dinner*" are equally frequent), both online surprisal (S_{onl}) and online entropy reduction (ΔH_{onl}) pair with offline situation surprisal (S_{sit}) in that "dinner" is more effortful than "popcorn", while offline linguistic surprisal (S_{ling}) shows no effect (Figure 5, middle). Finally, when there is a mismatch between linguistic experience ("*NP_{person} ordered champagne*" is more frequent than "*NP_{person} ordered water*") and world knowledge ($P(\text{order(champage)} < P(\text{order(water)}))$, online surprisal ($S_{\text{onl}}$) pairs with offline linguistic surprisal (S_{ling}) in that "water" is more effortful than "champagne" (Figure 5, bottom). Online entropy reduction (ΔH_{onl}), in turn, again aligns with offline situation surprisal (S_{sit}) in that "champagne" is more effortful than "water". Indeed, a correlation analysis between online entropy reduction and offline situation surprisal for all words in the training data reveals a strong positive correlation between the two metrics ($r = 0.834$, $p < 0.01$).

In addition to comparing online entropy reduction (ΔH_{onl}) and online surprisal (S_{onl}) to offline linguistic surprisal (S_{ling}) and offline situation surprisal (S_{sit}), we could gain further insight by comparing them to the entropy reduction counterparts of these offline metrics: offline linguistic entropy reduction (ΔH_{ling}) and offline situation entropy reduction (ΔH_{sit}), which can both be straightforwardly estimated from the sentence-semantics pairs in the training data [45]. However, as all contrasts shown in Figure 5 concern sentence-final contrasts, there will in fact be no effects on linguistic entropy reduction: within each contrast, entropy will be the same for the control and target conditions at the penultimate word position, and in both conditions, the sentence-final words will reduce entropy to zero. Hence, there will be no difference in linguistic entropy reduction between the conditions. As for offline situation entropy reduction, this could be estimated by replacing $\vec{v}_t$ in Equation (6) with the situation vector $\text{sit}(w_{1,\ldots,t})$—the disjunction of all meaning vectors consistent with the prefix $w_1, \ldots, w_t$—that is also used for offline situation surprisal (see above). However, it turns out that for the model at hand, this yields the exact same predictions as offline situation surprisal (ΔH_{sit}) and hence will not lead to any further insights. This is a mathematical artefact of using strictly binary meaning vectors that represent disjunctions over propositions, which yield a uniform probability distribution over models. Under such a constellation, offline situation entropy and offline situation surprisal will not diverge, as the former is the weighted average of the latter.

In sum, online entropy reduction makes different predictions than online surprisal; while online surprisal reflects expectancy based on linguistic experience (∼offline linguistic surprisal) and world knowledge (∼offline situation surprisal), online entropy reduction consistently aligns with world knowledge and appears relatively insensitive to linguistic frequency differences.

3.3. Online Entropy Reduction as the Sentence Unfolds

Figure 6 shows the development of the surprisal and entropy reduction metrics across two sets of sentences "NP_{person} entered the NP_{place} and asked [...]" (top) and "NP_{person} entered the NP_{place} and ordered [...]" (bottom). These sets of sentences differ in terms of structural frequency (linguistic experience) and probability of their corresponding semantics (world knowledge): the latter sentences (ordering something after entering a place) are both more frequent and their semantics more probable than the former (asking for something after entering a place). This difference is reflected in the offline linguistic surprisal (S_{ling}) and offline situation surprisal (S_{sit}) metrics at the verb of the coordinated sentence, which is in non-final position: both predict higher surprisal for "asked" relative to "ordered".

Indeed, online entropy reduction (ΔH_{onl}) aligns with the trajectory of offline situation surprisal (S_{sit}). Being consistent with both offline surprisal metrics, it predicts larger entropy reduction at "asked" than at "ordered" ($\Delta H_{\text{onl}}(\textit{asked}) - \Delta H_{\text{onl}}(\textit{ordered}) = 0.38 - 0.02 = 0.36$). By contrast, online surprisal (S_{onl}) follows a completely different path as the sentences unfold: after the second word, it does not align with any of the other metrics. Instead, online surprisal is relatively high at the beginning of the sentence, and at the critical words, it predicts only a slight (negative) difference in surprisal between "asked" and "ordered" ($\mathrm{S}_{\mathrm{onl}}(asked) - \mathrm{S}_{\mathrm{onl}}(ordered) = 0.68 - 0.75 = -0.07$). Hence, online entropy reduction and online surprisal develop differently as the sentences unfold, and they make qualitatively different predictions at the critical verb.

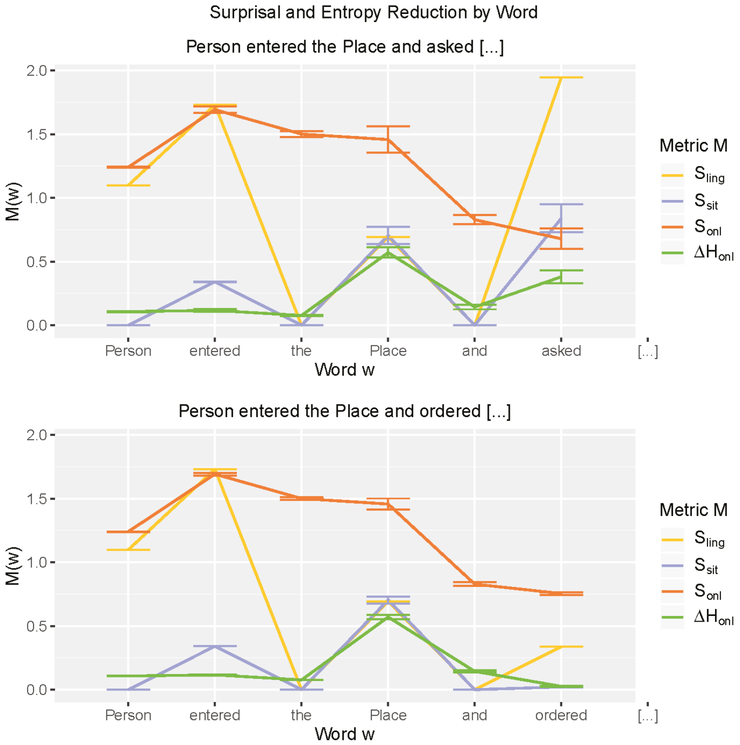


Figure 6. Word-by-word surprisal and entropy reduction metrics for two sets of sentences: "NP_{person} entered the NP_{place} and asked [. . .]" (n = 6, top) and "NP_{person} entered the NP_{place} and ordered [. . .]" (n = 30, bottom). Metrics shown are linguistic surprisal ($\mathrm{S}_{\mathrm{ling}}$), situation surprisal ($\mathrm{S}_{\mathrm{sit}}$), online surprisal ($\mathrm{S}_{\mathrm{onl}}$), and online entropy reduction ($\Delta\mathrm{H}_{\mathrm{onl}}$). Error bars show standard errors.

The trajectory of online entropy reduction is relatively straightforward to understand: entropy reduction stays relatively low throughout the sentence, except for the points at which propositional meanings can be singled out (i.e., at "Place", *enter(Person,Place)* is derived, and at "asked", *ask_menu(Person)*). In turn, online surprisal is relatively high for the sentence-initial words. This is due to the way in which the model navigates through meaning space; it will start out with relatively uniform meaning vectors (∼high entropy, see Figure 7 below) and gradually move toward more polarized vectors with more units approximating 0 and 1 (∼lower entropy). Since surprisal derives from the conditional probability between two model-derived meaning vectors, it will be affected by the amount of polarization of these vectors, i.e., less polarized vectors generally lead to higher surprisal. Indeed, if we quantify the polarization at t in terms of entropy $H(t)$ and the interaction in entropy between time step $t-1$ and t as $H(t-1) * H(t)$, we obtain a significant positive relationship between surprisal and this interaction ($r = 0.532$, $p < 0.01$). In fact, this also explains the differential effect of surprisal and entropy reduction at the critical word: the vector $\vec{v}_{asked}$, constructed after processing "asked", is more polarized than the vector $\vec{v}_{ordered}$, constructed after "ordered" ($|\{i|\vec{v}_{asked}(i) < 0.1 \vee \vec{v}_{asked}(i) > 0.9\}| = 126$ and $|\{i|\vec{v}_{ordered}(i) < 0.1 \vee \vec{v}_{ordered}(i) > 0.9\}| = 91$, for the sentences "thom entered the restaurant and $\underline{\text{asked}}$/$\underline{\text{ordered}}$"), since "asked" directly disambiguates the sentence-final meaning and "ordered" does not ("asked" is necessarily followed by "for the menu", whereas "ordered" has different continuations, such as "cola", "dinner", etc.). As the amount of polarization affects the conditional probability, it may thereby obscure the effect of linguistic experience and world knowledge reflected in the two offline surprisal measures.

Since online entropy is defined relative to fully specified states of affairs, which are themselves represented as meaning vectors identifying unique models in $\mathcal{M}$, entropy effectively quantifies the amount of polarization of the meaning vectors; low entropy states are more polarized. To illustrate how this polarization develops as the sentence unfolds, Figure 7 shows the entropy at each word of each training sentence of the VCB model (note that the distribution of sentence-lengths in the training data is as follows: 2 words (6), 3 words (150), 4 words (3), 5 words (18), 6 words (6), 7 words (33), 8 words (15), and 9 words (6); for a detailed description of the training data, see [33]). A first thing to note is that entropy reduces as sentences unfold ($r = -0.8$; $p < 0.01$). As the model processes sentences on a word-by-word basis, it moves through points in space that render it increasingly clear which propositions are the case and which are not, thereby reducing uncertainty about the state of affairs conveyed by the utterance. Secondly, this figure shows that sentence-final entropy remains relatively high (in comparison to the maximum entropy for the 150 non-duplicate models constituting the meaning space in [33], which is: $-\log(\frac{1}{150}) = 5.01$ nats (=7.23 bits)). Indeed, when entropy is defined relative to fully specified states of the world, individual sentences will not reduce entropy to zero (in contrast to previous instantiations of linguistic entropy that are defined over sentence-final structures [4,8,14,15]): for instance, the sentence "beth ordered cola" is satisfied by all models in which *order(beth,cola)* is the case but is not explicit about all the other propositions that can co-occur with it, thus leaving significant uncertainty with respect to the fully-specified state of affairs.

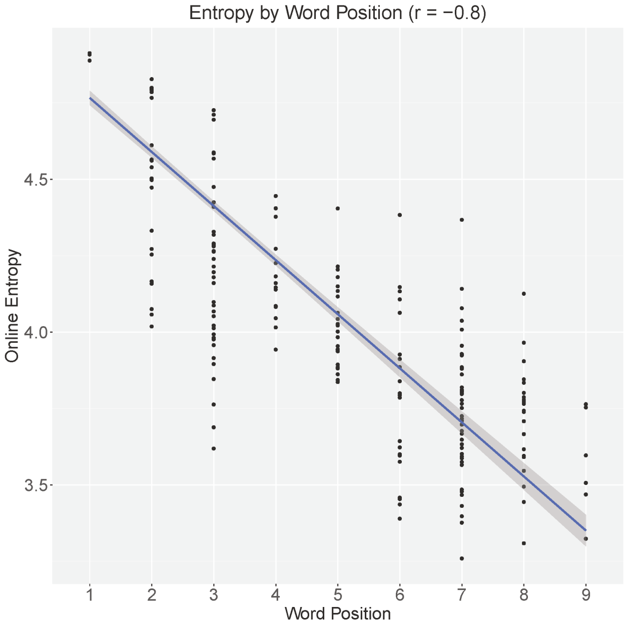


Figure 7. Online entropy by word position. The solid blue line depicts a linear regression and its 95% confidence interval. The Pearson correlation coefficient (r) is shown at the top.

4. Discussion

We have derived a comprehension-centric notion of online semantic entropy, based on a comprehension model that incrementally constructs probabilistic distributed meaning representations. Instead of defining entropy over the probabilistic structure of the language, we here define it in terms of the structure of the world [45]. That is, in line with the comprehension-centric notion of surprisal presented by VCB [33], entropy derives from the model's incremental navigation through meaning space, which is guided by both linguistic experience and world knowledge [33]. More specifically, at time step t, entropy in this model quantifies the amount of uncertainty at t with respect to fully specified states of affairs, i.e., the combinations of propositions that constitute the meaning space.

While surprisal is estimated from the probabilistic properties of previous and current states of processing—and hence naturally falls out of probabilistic language (processing) models—entropy derives from the probabilities of all possible future states (e.g., every possible continuation of the sentence at hand), which makes it typically less straightforward to estimate. Indeed, given that the set of possible sentences that can be produced is non-finite, this quickly becomes infeasible, and some state-limiting mechanism is required in order for entropy to be estimated (e.g., see [15]). In the present model, by contrast, this is mitigated by the fact that entropy, like surprisal, directly derives from the finite dimensions of the utterance meaning representations that the model constructs on a word-by-word basis. That is, at each time step t, the model produces a vector $\vec{v}(t)$ representing the activity pattern over $|\mathcal{M}|$ neuron-like processing units, and entropy directly derives from these $|\mathcal{M}|$ states. While this offers an account of entropy (and surprisal) at the level of representations—and hence at Marr's [16] representational and algorithmic level—it does raise questions about the ecological status of $\mathcal{M}$. We see $\mathcal{M}$ as a set of representative, maximally informative models reflecting the structure of the world. That is, we do not take each $M \in \mathcal{M}$ to instantiate a single observation of a state-of-affairs but rather as an exemplar state-of-affairs, which combines with the other exemplars in $\mathcal{M}$ to represent the probabilistic structure of the world. In this sense, $\mathcal{M}$ can be seen as an abstraction of our accumulated experience with the world around us. Indeed, this gives rise to the question of how $\mathcal{M}$ could be acquired, developed, and altered as children and adults navigate the world over time. While this is a question for language acquisition that is beyond the scope of this article, one speculative approach could be to implement $\mathcal{M}$ as a self-organization map (SOM), which consists of the running average of maximally informative states of affairs (e.g., see [37]) and which interfaces with the comprehension model. Of course, despite this perspective on the set of states of affairs $\mathcal{M}$ that constitutes our meaning space, the number of dimensions needed to capture real human world knowledge will significantly exceed the limited dimensions of the current model. As a result, entropy is predicted to be high in general, and individual sentences are predicted to reduce entropy only marginally. Critically, however, sentences are generally interpreted in context (be it a linguistic or extra-linguistic context), which significantly constrains the set of states of affairs that contribute to the word-derived entropy: for instance, a context in which "beth enters the restaurant" will effectively reduce our meaning space to only those states of affairs that are related to (beth) going to a restaurant. Hence, entropy calculation regarding fully specified states of affairs becomes both feasible and intuitive when taking a context-dependent (or dynamic) perspective on language comprehension.

Using the comprehension model presented in [33], we have investigated how the comprehension-centric notion of entropy reduction behaves during online comprehension and how it relates to online surprisal. We have found that online entropy reduction and surprisal correspond to differential processing metrics, which may be reflected in different behavioral effects (cf. [15]). Critically, entropy reduction and surprisal here are not conceived as reflecting different underlying cognitive processes as both derive from the model's comprehension process as navigation through meaning space. They do, however, describe distinct aspects of this navigation process; whereas surprisal reflects the transition in meaning space from one word to the next, entropy reduction quantifies how much uncertainty is reduced with respect to the state of the world. This explains why entropy reduction seems less sensitive to effects of linguistic experience than surprisal; even though the point in meaning space at which the model arrives at time step t is determined by both linguistic experience and world knowledge (as reflected in the online surprisal estimates [33]), entropy is calculated relative to fully specified states of affairs, which means that it will be more sensitive to probabilities that derive from the structure of the world than to those deriving from linguistic frequency effects. This is especially true in the current setup of the model, where linguistic experience is limited to word frequency effects (sentence structures are relatively invariant across the training data). Hence, to the extent that linguistic experience can restrict which states of affairs are consistent with the current meaning vector, it may affect online entropy reduction. However, the presented set of contrasts illustrates that online surprisal is inherently more sensitive than entropy reduction to effects of linguistic experience. Overall, the observation that entropy reduction is highly sensitive to the probabilistic structure of the world is consistent with recent findings from situated language comprehension [34].

A consequence of deriving entropy from fully specified states of affairs is that entropy stays relatively high after processing sentence-final words. As discussed above, this is because of the structure of the world and the world knowledge-driven inferences that are inherent to the meaning representations: after a sentence is processed, its literal propositional content and any highly likely or necessary propositions that co-occur with it, are inferred to be the case, but there also remains a vast amount of uncertainty regarding other propositions that could co-occur with it. This is consistent with a perspective on language comprehension in which pragmatic inference is an inherent part of incremental, word-by-word processing. In fact, one could argue that the model instantiates a perspective in which comprehension *is* pragmatic inference; the literal propositional content of an utterance has no special status—there is only the probabilistic inferences that derive from processing an utterance (which will typically entail the literal propositional content). This leads to another prediction regarding the difference between surprisal and entropy reduction in our model: surprisal, which derives directly from two subsequent points in meaning space, effectively reflects how the likelihood of inferred propositions changes *locally*, as it only takes into account the inferences contained within these points. Entropy reduction, in turn, looks at the difference in entropy between these points, which explicitly factors in the likelihood of all possible inferences. Entropy reduction thus reflects how the likelihood of inferred propositions changes *globally*, i.e., with respect to the full set of possible inferences that could be drawn. Hence, in the current instantiation of the model, the surprisal of the word "restaurant" in the sentence "beth entered the restaurant" is driven by the change in likelihood between the (probabilistic) inferences made at the word "the" and those made at the word "restaurant", while its entropy reduction is determined by the difference in uncertainty about the full set of inferences available to the model.

In sum, in the comprehension-centric perspective on surprisal and entropy reduction formalized in the current model, the metrics derive from a single process—word-by-word meaning space navigation—but differ in which aspects of this process they elucidate. That is, the processing of an incoming word moves the model from a previous point to a next point in space. The exact coordinates of these points depend on the linguistic experience of the model as well as the world knowledge contained within the meaning space that it navigates. Surprisal quantifies how likely the next point is given the previous one and thereby effectively how expected the input was. Surprisal can thus be thought of as reflecting *state-by-state expectation*, where input that moves the model to unexpected points in space yields high surprisal. Entropy, in turn, quantifies how likely each fully-specified state of affairs constituting the meaning space is, given the current point in space. Entropy reduction, then, is effectively a metric of *end-state confirmation*, where higher reduction of uncertainty about the propositions that are communicated to be the case, i.e., stronger confirmation of the communicated state-of-affairs, leads to higher reduction of entropy. This characterization appears to be in line with recent theories and models from the text comprehension literature, in which the notion of *validation*—the process of evaluating consistency of incoming linguistic information with the previous linguistic context and general knowledge about the world—has a central role [46–48]. The above described conceptualization of entropy reduction in terms of end-state confirmation might indeed turn out to be an index of the degree of, or effort induced by, validating the incoming input against the larger context and knowledge about the world. To the extent that this mapping is correct, one could explore the dissociation between entropy reduction and surprisal even further by turning to experimental designs that pit global knowledge of the world against local textual/discourse coherence—a good point to start this investigation is by turning to the text comprehension literature [17,19,21,27,49,50].

Taken together, the conceptualization of comprehension as meaning-space navigation predicts a dichotomy in which surprisal and entropy reduction—while often correlated—differentially index effort during incremental, expectation-based comprehension: state-by-state expectation (surprisal) versus end-state confirmation (entropy reduction). That is, while both metrics derive from transitions between states in meaning space, surprisal approximates the distance of this transition, whereas entropy reduction reflects a change in the inherent nature of these states: the degree of certainty regarding the state of affairs being communicated.

Author Contributions: Conceptualization, H.B., N.J.V. and M.W.C.; Methodology, H.B., N.J.V. and M.W.C.; Software, H.B.; Validation, N.J.V., H.B. and M.W.C.; Formal analysis, N.J.V.; Investigation, N.J.V. and H.B.; Resources, N.J.V.; Data curation, N.J.V.; Writing–original draft preparation, N.J.V. and H.B.; Writing–review and editing, M.W.C.; Visualization, N.J.V. and H.B.; Supervision, M.W.C.; Project administration, M.W.C.; Funding acquisition, H.B. and M.W.C.

Funding: This research was funded by the Deutsche Forschungsgemeinschaft, SFB/CRC 1102 "Information density and linguistic encoding" (Project A1) awarded to M.W.C. and H.B.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
2. Hale, J.T. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2001; pp. 1–8.
3. Levy, R. Expectation-based syntactic comprehension. *Cognition* **2008**, *106*, 1126–1177. [CrossRef] [PubMed]
4. Hale, J.T. Uncertainty about the rest of the sentence. *Cogn. Sci.* **2006**, *30*, 643–672. [CrossRef] [PubMed]
5. Boston, M.F.; Hale, J.T.; Kliegl, R.; Patil, U.; Vasishth, S. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *J. Eye Mov. Res.* **2008**, *2*, 1–12.
6. Demberg, V.; Keller, F. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* **2008**, *109*, 193–210. [CrossRef] [PubMed]
7. Frank, S.L. Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In *Proceedings of the 31st Annual Conference of the Cognitive Science Societ*y; Cognitive Science Society: Austin, TX, USA, 2009; pp. 1139–1144.
8. Roark, B.; Bachrach, A.; Cardenas, C.; Pallier, C. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2009; pp. 324–333.
9. Smith, N.J.; Levy, R. Optimal Processing Times in Reading: A Formal Model and Empirical Investigation. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*; Cognitive Science Society: Austin, TX, USA, 2008; pp. 595–600.
10. Brouwer, H.; Fitz, H.; Hoeks, J. Modeling the Noun Phrase versus Sentence Coordination Ambiguity in Dutch: Evidence from Surprisal Theory. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*; Association for Computational Linguistics: Uppsala, Sweden, 2010; pp. 72–80.
11. Blache, P.; Rauzy, S. Predicting linguistic difficulty by means of a morpho-syntactic probabilistic model. In Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC-2011), Singapore, 16–18 December 2011; pp. 160–167.
12. Wu, S.; Bachrach, A.; Cardenas, C.; Schuler, W. Complexity metrics in an incremental right-corner parser. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2010; pp. 1189–1198.
13. Frank, S.L. Uncertainty reduction as a measure of cognitive processing effort. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2010; pp. 81–89.
14. Hale, J.T. What a rational parser would do. *Cogn. Sci.* **2011**, *35*, 399–443. [CrossRef]
15. Frank, S.L. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Top. Cogn. Sci.* **2013**, *5*, 475–494. [CrossRef]
16. Marr, D. *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*; W. H. Freeman: San Francisco, CA, USA, 1982.
17. O'Brien, E.J.; Albrecht, J.E. Comprehension strategies in the development of a mental model. *J. Exp. Psychol. Learn. Mem. Cogn.* **1992**, *18*, 777–784. [CrossRef]
18. Albrecht, J.E.; O'Brien, E.J. Updating a mental model: Maintaining both local and global coherence. *J. Exp. Psychol. Learn. Mem. Cogn.* **1993**, *19*, 1061–1070. [CrossRef]
19. Morris, R.K. Lexical and message-level sentence context effects on fixation times in reading. *J. Exp. Psychol. Learn. Mem. Cogn.* **1994**, *20*, 92–102. [CrossRef]
20. Hess, D.J.; Foss, D.J.; Carroll, P. Effects of global and local context on lexical processing during language comprehension. *J. Exp. Psychol. Gen.* **1995**, *124*, 62–82. [CrossRef]

21. Myers, J.L.; O'Brien, E.J. Accessing the discourse representation during reading. *Discourse Process.* **1998**, *26*, 131–157. [CrossRef]
22. Altmann, G.T.; Kamide, Y. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* **1999**, *73*, 247–264. [CrossRef]
23. Van Berkum, J.J.A.; Hagoort, P.; Brown, C.M. Semantic integration in sentences and discourse: Evidence from the N400. *J. Cogn. Neurosci.* **1999**, *11*, 657–671. [CrossRef] [PubMed]
24. Van Berkum, J.J.A.; Zwitserlood, P.; Hagoort, P.; Brown, C.M. When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cogn. Brain Res.* **2003**, *17*, 701–718. [CrossRef]
25. Van Berkum, J.J.A.; Brown, C.M.; Zwitserlood, P.; Kooijman, V.; Hagoort, P. Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *J. Exp. Psychol. Learn. Mem. Cogn.* **2005**, *31*, 443–467. [CrossRef]
26. Garrod, S.; Terras, M. The contribution of lexical and situational knowledge to resolving discourse roles: Bonding and resolution. *J. Mem. Lang.* **2000**, *42*, 526–544. [CrossRef]
27. Cook, A.E.; Myers, J.L. Processing discourse roles in scripted narratives: The influences of context and world knowledge. *J. Mem. Lang.* **2004**, *50*, 268–288. [CrossRef]
28. Knoeferle, P.; Crocker, M.W.; Scheepers, C.; Pickering, M.J. The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition* **2005**, *95*, 95–127. [CrossRef]
29. Knoeferle, P.; Habets, B.; Crocker, M.W.; Münte, T.F. Visual scenes trigger immediate syntactic reanalysis: Evidence from ERPs during situated spoken comprehension. *Cereb. Cortex* **2008**, *18*, 789–795. [CrossRef]
30. Camblin, C.C.; Gordon, P.C.; Swaab, T.Y. The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *J. Mem. Lang.* **2007**, *56*, 103–128. [CrossRef] [PubMed]
31. Otten, M.; Van Berkum, J.J.A. Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Process.* **2008**, *45*, 464–496. [CrossRef]
32. Kuperberg, G.R.; Paczynski, M.; Ditman, T. Establishing causal coherence across sentences: An ERP study. *J. Cogn. Neurosci.* **2011**, *23*, 1230–1246. [CrossRef] [PubMed]
33. Venhuizen, N.J.; Crocker, M.W.; Brouwer, H. Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. *Discourse Process.* **2019**, *56*, 229–255. doi:10.1080/0163853X.2018.1448677. [CrossRef]
34. Tourtouri, E.N.; Delogu, F.; Sikos, L.; Crocker, M.W. Rational over-specification in visually-situated comprehension and production. *J. Cult. Cogn. Sci.* **2019**. doi:10.1007/s41809-019-00032-6. [CrossRef]
35. Venhuizen, N.J.; Hendriks, P.; Crocker, M.W.; Brouwer, H. A Framework for Distributional Formal Semantics. In *Logic, Language, Information, and Computation*; Iemhoff, R., Moortgat, M., de Queiroz, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2019; pp. 633–646, doi:10.1007/978-3-662-59533-6_39. [CrossRef]
36. Frank, S.L.; Koppen, M.; Noordman, L.G.; Vonk, W. Modeling knowledge-based inferences in story comprehension. *Cogn. Sci.* **2003**, *27*, 875–910. [CrossRef]
37. Frank, S.L.; Haselager, W.F.; van Rooij, I. Connectionist semantic systematicity. *Cognition* **2009**, *110*, 358–379. [CrossRef]
38. Bos, J.; Basile, V.; Evang, K.; Venhuizen, N.J.; Bjerva, J. The Groningen Meaning Bank. In *Handbook of Linguistic Annotation*; Ide, N., Pustejovsky, J., Eds.; Springer: Dordrecht, The Netherlands, 2017; pp. 463–496.
39. Wanzare, L.D.A.; Zarcone, A.; Thater, S.; Pinkal, M. DeScript: A Crowdsourced Database of Event Sequence Descriptions for the Acquisition of High-quality Script Knowledge. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*; Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., et al., Eds.; European Language Resources Association (ELRA): Paris, France, 2016.
40. Elman, J.L. Finding structure in time. *Cogn. Sci.* **1990**, *14*, 179–211. [CrossRef]
41. Rohde, D.L.T. A Connectionist Model of Sentence Comprehension and Production. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2002.

42. Calvillo, J.; Brouwer, H.; Crocker, M.W. Connectionist Semantic Systematicity in Language Production. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*; Papafragou, A., Grodner, D., Mirman, D., Trueswell, J.C., Eds.; Cognitive Science Society: Philadelphia, PA, USA, 2016; pp. 2555–3560.
43. Brouwer, H.; Crocker, M.W.; Venhuizen, N.J.; Hoeks, J.C.J. A Neurocomputational Model of the N400 and the P600 in Language Processing. *Cogn. Sci.* **2017**, *41*, 1318–1352. [CrossRef]
44. Hale, J.T. The information conveyed by words in sentences. *J. Psycholinguist. Res.* **2003**, *32*, 101–123. [CrossRef]
45. Frank, S.L.; Vigliocco, G. Sentence comprehension as mental simulation: An information-theoretic perspective. *Information* **2011**, *2*, 672–696. [CrossRef]
46. Singer, M. Validation in reading comprehension. *Curr. Dir. Psychol. Sci.* **2013**, *22*, 361–366. [CrossRef]
47. O'Brien, E.J.; Cook, A.E. Coherence threshold and the continuity of processing: The RI-Val model of comprehension. *Discourse Process.* **2016**, *53*, 326–338. [CrossRef]
48. Richter, T. Validation and comprehension of text information: Two sides of the same coin. *Discourse Process.* **2015**, *52*, 337–355. [CrossRef]
49. Gerrig, R.J.; McKoon, G. The readiness is all: The functionality of memory-based text processing. *Discourse Process.* **1998**, *26*, 67–86. [CrossRef]
50. Cook, A.E.; O'Brien, E.J. Knowledge activation, integration, and validation during narrative text comprehension. *Discourse Process.* **2014**, *51*, 26–49. [CrossRef]

entropy

Article

Linguistic Laws in Speech: The Case of Catalan and Spanish

Antoni Hernández-Fernández [1,2,*,†], Iván G. Torre [3,†], Juan-María Garrido [4] and Lucas Lacasa [5,*]

[1] Societat Catalana de Tecnologia, Secció de Ciències i Tecnologia, Institut d'Estudis Catalans, Carrer del Carme 47, 08001 Barcelona, Catalonia, Spain

[2] Complexity and Quantitative Linguistics Lab, LARCA Research Group, Institut de Ciències de l'Educació, Universitat Politècnica de Catalunya, Av. Doctor Marañón 44-50, 08028 Barcelona, Catalonia, Spain

[3] Departamento de Matemática Aplicada y Estadistíca, Universidad Politécnica de Madrid, Plaza Cardenal Cisneros, 28040 Madrid, Spain; ivan.gonzalez.torre@upm.es

[4] Laboratorio de Fonética Antonio Quilis, Facultad de Filología, Universidad Nacional de Educación a Distancia, 28040 Madrid, Spain; jmgarrido@flog.uned.es

[5] School of Mathematical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, UK

[*] Correspondence: antonio.hernandez@upc.edu (A.H.-F.); l.lacasa@qmul.ac.uk (L.L.)

[†] These authors contributed equally to this work.

Received: 31 October 2019 ; Accepted: 21 November 2019; Published: 26 November 2019

Abstract: In this work we consider Glissando Corpus—an oral corpus of Catalan and Spanish—and empirically analyze the presence of the four classical linguistic laws (Zipf's law, Herdan's law, Brevity law, and Menzerath–Altmann's law) in oral communication, and further complement this with the analysis of two recently formulated laws: lognormality law and size-rank law. By aligning the acoustic signal of speech production with the speech transcriptions, we are able to measure and compare the agreement of each of these laws when measured in both physical and symbolic units. Our results show that these six laws are recovered in both languages but considerably more emphatically so when these are examined in physical units, hence reinforcing the so-called 'physical hypothesis' according to which linguistic laws might indeed have a physical origin and the patterns recovered in written texts would, therefore, be just a byproduct of the regularities already present in the acoustic signals of oral communication.

Keywords: Zipf's law; Brevity law; Menzerath–Altmann's law; Herdan's law; lognormal distribution; size-rank law; quantitative linguistics; Glissando corpus; scaling; speech

1. Introduction

Linguistic laws are statistical regularities and properties of linguistic elements (i.e., phonemes, syllables, words or sentences) which can be formulated mathematically and estimated quantitatively [1]. While linguistic laws have been thoroughly studied over the last century [1–3], the debate on its ultimate origin is still open. In what follows we summarise the classic four linguistic laws [3], recently revised, mathematically substantiated and expanded to two other laws [4] (see Table 1 for specific formulations):

1. Zipf's law. After some notable precursors (as Pareto [5], Estoup [6] or Condon [7] among others), George Kingsley Zipf formulated and explained in [8,9] one of the most popular quantitative linguistic observations known in his honor as Zipf's Law. He observed that the number of occurrences of words with a given rank can be expressed as $f(r) \sim r^{-\alpha}$, when ordering the words of written corpus in decreasing order by their frequency. This is a solid linguistic law proven in many written corpus [10] and in speech [11], even though its variations have been discussed in many contexts [12–14].

Table 1. Main linguistic laws, according to Torre and collaborators [4]. From left to right columns: name of the linguistic law, its mathematical formulation, details on its magnitudes and parameters; and finally some basic references about each law. While Zipf's law is naturally defined and measured in symbolic units (texts or speech transcriptions), Herdan-Heaps, Brevity, Size-Rank and Menzerath-Altmann laws can be measured both in symbolic and physical units. Lognormality law is only defined in physical units (time duration).

	Mathematical Formulation	Details	References
Zipf's law	$f(r) \sim r^{-\alpha}$	f: frequency r: rank α: parameter	[5–9]
Herdan-Heaps' law	$V \sim L^{\beta}$	L: text size / time elapsed V: vocabulary β: parameter	[15–17]
Brevity law	$f \sim \exp(-\lambda\ell), \quad \lambda > 0$	f: frequency ℓ: size λ: parameter	[4,8,9,18,19]
Size-rank law	$\ell \sim \theta \log(r),\ \theta = \frac{\alpha}{\lambda}$	ℓ: size r: rank θ: parameter	[4,9,18]
Menzerath-Altmann's law	$y(n) = an^{b}\exp(-cn)$	n: size of the whole y: size of the parts a, b, c: parameters	[4,20–24]
Lognormality law	$\mathrm{p}(t;\mu,\sigma) = \frac{1}{t\sigma\sqrt{2\pi}}e^{-\frac{(\ln(t)-\mu)^2}{2\sigma^2}}$	t: time duration σ, μ: parameters	[4,25–28]

2. Herdan's law. Although with little-known precedents [15], Herdan's law [16] (also known as Heap's law, because it was also formulated later by Heaps in [17]) describes that the average growth of new different words V in a text of size L follows $V \sim L^{\alpha}, \alpha < 1$ [16]. Thus, Herdan's law shows the evolution of the number V of different words in a text (types) as its size increases, measured in the total number of words (L). L obviously is obtained by the summation of the number of occurrences of each word (tokens), for each different words types that appear in the text.

3. Brevity law. Also known as Zipf's law of abbreviation, its original qualitative statement claims that the more a word is used, the shorter it tends to be [8,9,18]. In texts or transcriptions, usually the way of measuring the word size is using the number of characters that compose the word. In this way, brevity law has been empirically proven in texts from almost a thousand languages of eighty different linguistic families [19], but also holds acoustically when measuring the time duration of words [29,30].
   The leap from the classical qualitative conception of brevity law to a quantitative proposal has recently been made [4,31]. In information-theoretic terms [32], if a certain symbol i has a probability p_i of appearing in a given symbolic code with a $\mathcal{D}$-ary alphabet, then its minimum (optimal) expected description length $\ell_i^* = -\log_{\mathcal{D}}(p_i)$. Deviating from optimality can be effectively modelled by adding a pre-factor, such that the description length of symbol i is $\ell_i \sim -\frac{1}{\lambda_{\mathcal{D}}}\log_{\mathcal{D}}(p_i)$, where $0 < \lambda_{\mathcal{D}} \leq 1$. So, the closer $\lambda_{\mathcal{D}}$ is to one, the closer it is the system to optimal compression. Reordering terms, one finds an exponentially decaying dependence between the frequency of a unit and its size (see [4] for further details on the mathematical formulation).

4. Size-rank law. Zipf's law and brevity law involve frequencies. Taking advantage of the new mathematical formulation of the latter, these can now be combined [4] in such a way the "size" ℓ_i of a unit i is mathematically related to its rank r_i via α (Zipf) and λ (brevity law) exponents. Experimentally, $\theta = \frac{\alpha}{\lambda}$ is therefore an observable parameter which indeed combines Zipf and

Brevity exponents in a size-rank plot, and this law predicts that the larger linguistic units tend to have a higher rank following a specific logarithmic relation [4].

5. Menzerath-Altmann law. Again after some forerunners [20], Paul Menzerath established that there is a negative correlation between the length of a linguistic construct and the length of its constituents [21,22]. Subsequently, a mathematical formulation law was heuristically proposed by Gabriel Altmann [23,24]: if n stands for the size of the linguistic construct and y is the constituent size, then $y(n) = an^b \exp(-cn)$, being a, b and c free parameters of the model, whose interpretation remains controversial [33]. Definitely, Menzerath–Altmann's law could be simplified and generalized qualitatively as "the longer a language construct the shorter its components (constituents)." [23,34]. This law has been revised in different linguistic levels under multiple and polyhedral perspectives [1,33,34], but above all in written texts. Recently some researchers are turning back to the phonetic origins of the law [35] and new mathematical models explaining the actual formulation have been proposed [4].

6. Lognormality law. Previous studies have found consistently lognormal distributions for spoken phonemes in several languages [25–28,36] and in word and breath groups (BGs) duration for English [4,37]. In [4] it was confirmed that the time duration of phonemes, words and breath groups in speech are well described by lognormal distribution for the English language. Moreover, in [4] a general stochastic model was presented to explain and justify such lognormality at all linguistic levels only assuming that the lowest (phonemic) level follows a lognormal distribution, hence claiming the universal validity of the lognormal shape and its proposal as a 'lognormality law'.

A critical review of the literature about linguistic laws shows that the majority of empirical studies have been conducted by analyzing written corpus or transcripts, while research on linguistic laws in speech have been limited to small and fragmented corpora. In this work, we have followed the protocol of [4] to systematically explore linguistic laws in speech in two new languages: Catalan and Spanish. We simultaneously use both physical and symbolic magnitudes to examine the abovementioned laws at three different linguistic levels in physical and symbolic space. We certify that these hold approximately well in both languages, but better so in physical space. As we will see, this new evidence in Catalan and Spanish further support the physical hypothesis as an alternative to the classical approach of the so-called symbolic hypothesis in the study of language. On one hand, the symbolic hypothesis states that these statistical laws emerge in language use as a consequence of its symbolic representation [4]. On the other hand, the physical hypothesis claims that the linguistic laws emerge initially in oral communication—possibly as a consequence of physical, acoustic, physiological mechanisms taking place and driven by communication optimization pressures—and the emergence of similar laws in written texts can thus be regarded as a byproduct of such more fundamental regularities [4,38,39]. We aim to recover in this way a more naturalistic approach to the study of language, somehow cornered after many years of written corpus studies in computational linguistics.

2. Results

Our results are based on an analysis of Glissando corpus [40]. For illustration, Table 2 summarises some general characteristics of this oral corpus (for more details, see Section 4, Materials and Methods). For the phonetic inventory of Spanish and Catalan, note that only phonemes that appear effectively in the Glissando corpus have been taken into account, without considering other phonemes that could appear in other linguistic varieties of both languages [41].

In the next subsections we provide a systematic study of each of the six laws detailed above in Glissando corpus. Table 3 summarizes the fitted exponents and parameters for all the linguistic laws explored in this work at the level of words, whereas Table 4 does the same at the phonemic level, with the exception, as explained in the following section, of the lognormality law for Catalan and

Spanish. As we will see, linguistic laws are again recovered with only slight differences with respect to English [4] and some technical details that are worth detailing for each law.

Table 2. Main characteristics of Glissando. This Table summarises main characteristics of Glissando corpus [40] across linguistic levels for both Catalan and Spanish. For reference, a comparison to Buckeye corpus (English) is provided [4,42,43]. We report the total number of linguistic elements (phonemes, words and breath groups (BG)), specifying the number of different linguistic elements (types) and the total (tokens). Since time duration distribution of linguistic levels are usually heavy-tailed [4], we use median duration (instead of mean) as a reference.

	Number of Elements					**Median Duration (secs.)**		
	Phonemes		**Words**		**BG**	**Phon**	**Word**	**BG**
	Tokens	**Types**	**Tokens**	**Types**	**Tokens**			
Catalan	3×10^5	35	8×10^4	5×10^3	2×10^4	0.05	0.20	0.8
Spanish	2×10^5	32	5×10^4	4×10^3	1×10^4	0.05	0.21	0.9
English	8×10^5	64	3×10^5	9×10^3	5×10^4	0.07	0.20	1.1

Table 3. Summary of exponents and parameters for the case of words. Results are on reasonable good agreement to those found for English in [4]. Note that actual fit of lognormality law in Spanish and Catalan was not carried out due to low-resolution problems of the Glissando corpus, however we certified that this law also holds (see the text).

Words	**Zipf**	**Herdan-Heaps**	**Brevity**	**Size-rank**	**Menzerath-Altmann**			**Lognormality**	
	α	β	λ	θ	a	b	c	μ	σ
Catalan	1.42	0.62	23.8	0.060	0.301	−0.132	−0.004	-	-
Spanish	1.41	0.63	24.1	0.058	0.336	−0.148	−0.003	-	-
English	1.41	0.63	20.6	0.07	0.364	−0.227	−0.0067	−1.62	0.66

Table 4. Summary of exponents and parameters for the case of phonemes. Results are on reasonable good agreement to those found for English in [4]. Note that actual fit of lognormality law in Spanish and Catalan was not carried out due to low-resolution problems of the Glissando corpus, however we certified that this law also holds (see the text).

Phonemes	**Yule**		**Brevity**	**Menzerath-Altmann**			**Lognormality**	
	a	b	λ	a	b	c	μ	σ
Catalan	0.04	0.90	297	0.092	−0.355	−0.037	-	-
Spanish	0.16	0.89	76	0.102	−0.393	−0.032	-	-
English	0.25	0.96	127	0.18	−0.23	−0.007	−2.68	0.59

2.1. Lognormality Law and Low-Resolution Effects

Recently [4], after exploring most common plausible family of probability distributions with the use of maximum likelihood estimation method (MLE) [44], compelling statistical evidence showed that the time duration distribution in speech in an English corpus is lognormally distributed across linguistic scales (phonemes, words, and BG), and such regularity was robust for individual speakers. Moreover, a generative mechanism able to explain the stability of the lognormal law for different linguistic scales was also proposed [4], suggesting that such regularity is indeed universal, hence proposing the so-called lognormality law. Here we explore the fulfillment of such new law in two additional languages, Catalan and Spanish. Empirical results for the time duration distribution $P(t)$ of phonemes, words, and breath-groups (BGs) are depicted in the main plots of the top panels of

Figure 1. Since a lognormal distribution appears normal (Gaussian) in linear-logarithmic axis, we have logarithmically rescaled the time duration variable t as

$$t' = \frac{\log(t) - \langle \log(t) \rangle}{\sigma(\log(t))}.$$

Accordingly, if $P(t)$ is lognormal, then $P(t')$ is a standard Gaussian $\mathcal{N}(0,1)$ regardless of the linguistic scale. This fact is numerically checked in the inset panels of Figure 1.

Overall the data approximately collapse to the Gaussian shape—hence validating the lognormality law—however, there are small deviations, and these are notably stronger for phonemes at short timescales (note that only the right branch of the Gaussian is recovered and deviations are found at $t' < 0$). In what follows we argue that this is indeed an artifact due to finite-precision and lower-bound resolution of the Catalan and Spanish corpus, rather than a genuine, linguistic effect.

First, note that lower-bound segmentation time in the Glissando Corpus is 30 ms, and the corpus has a precision (granularity) of 10 ms. The lower-bound segmentation time precludes us from experimentally observing the left-end of the phoneme time duration distribution. Furthermore, these artifacts can propagate up to a higher scale (i.e., to words), as evidenced by the fact that words with duration of 30, 40 or 50 ms turn out to be always composed by a single phoneme, words with time duration of 60, 70 and 80 ms have one or two phonemes, and so on.

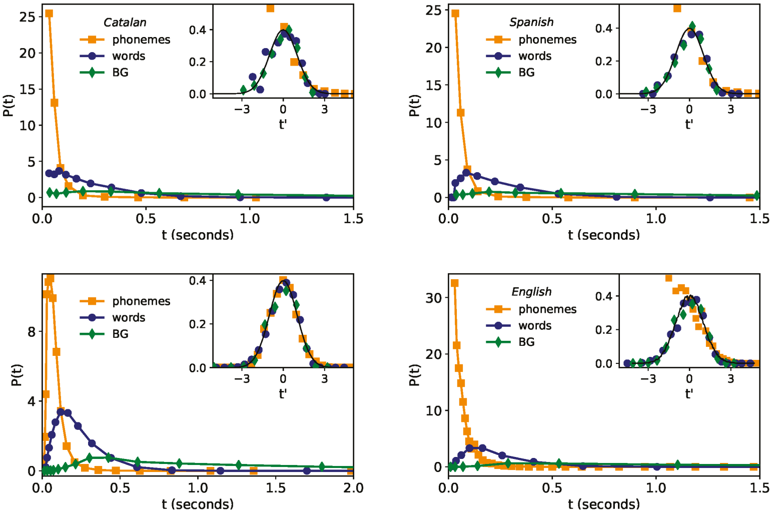


Figure 1. Lognormality law for time duration. (**outer panels**) Time duration distribution of phonemes (orange), words (blue) and BGs (green) for Glissando corpus: Catalan (**top left**) and Spanish (**top right**). For comparison, in the bottom left panel we show the results of English from Buckeye corpus (extracted from [4]), where Buckeye has finer statistics (higher resolution) than Glissando. A coarsened version of the English corpus—developed to be comparable with Glissando's resolution—is plotted in the bottom, right panel (see the text for details). (**inset panels**) Collapse of all distributions after time rescaling $t' = (\log(t) - \langle \log(t) \rangle / std(\log(t)))$ (where $std(\log(t))$ stands for the standard deviation of the random variable $\log t$). If time durations at all levels comply with a lognormal distribution, then the collapsed data should approach a standard Gaussian $\mathcal{N}(0,1)$ (solid line), in good agreement with the results. Small deviations found in Catalan and Spanish are similarly found in the coarsened version of English, thus concluding that such deviations are mainly due to finite-precision and lower-bound detectability effects, and the lognormality law otherwise holds.

In order to certify that these low-resolution issues are indeed underpinning the deviations from the pure lognormality law, we have added the following experiment. The so-called Buckeye corpus (English corpus) has higher resolution than Glissando and precision and is, therefore, free from these issues (also, Buckeye corpus has larger sample sizes than Glissando, see Table 2). Indeed for the Buckeye corpus, compliance to the lognormality law has recently found to be excellent (see bottom left panel of Figure 1). We thus proceed to construct a coarsened, low-resolution version of the Buckeye corpus, comparable to the particularities of the Glissando corpus under study, by rounding up time durations in Buckeye data to a precision of 10 ms and, by further setting the minimum observable time duration (the lower limit segmentation) to 30 ms (we do not deal with further limitations such as that words shorter than 60 ms are always composed of one phoneme in Glissando). The resulting time distribution of phonemes, words, and BGs in this coarsened Buckeyed corpus are plotted in the right panel of Figure 1. Interestingly, similar deviations from the lognormality law to the ones found in the Glissando corpus are now recovered in the low-resolution version of the Buckeye corpus. This evidence supports our hypothesis that the lognormality law indeed holds well in Catalan and Spanish, albeit it might not be fully observable at the phoneme level in the Glissando corpus. Furthermore, this analysis flags an important issue: low-resolution effects such as low precision and a too-large lower limit segmentation time can induce important deviations and hinder the observation of the true, underlying distribution.

To further investigate these effects, it is worth discussing at this point that the origin of the lognormality law has been mathematically discussed recently in terms of a stochastic model [4]. Suppose that phoneme time durations can be modeled by a random variable Y, which is indeed lognormally distributed. Since words can be understood as concatenation of phonemes, then the time duration of words can thus be modeled by a random variable $Z = \sum_{i=1}^{n} Y_i$, where each Y_i is in principle a different lognormal distribution and n is yet another random variable which describes the number of phonemes shaping up a word. Whereas when n is large the central limit theorem predicts Z is asymptotically normal when n is small and under some additional conditions, Z is well approximated by a lognormal distribution [4], thereby explaining why the time duration of words is indeed found to be lognormally distributed in practice. Now, how would Z be distributed if we imposed on its sampling the artifacts detected in Glissando, such as a large lower-bound detectability threshold, finite precision, or a smallish sample size? To illustrate these effects, we have run a numerical test where we initially sample words of duration Z, constructed by concatenating phonemes with time duration Y where $Y = \exp(X)$ and X is a Gaussian random variable. Y is therefore lognormal and if we log-rescale it $\tilde{Y} = [\log Y - \langle \log Y \rangle]/std(\log Y)$ (where $std(\log Y)$ stands for the standard deviation of the random variable $\log Y$) we should recover a standard Gaussian $\mathcal{N}(0,1)$. This distribution is shown (black curve) in the left panel of Figure 2, whereas the case of words is plotted in the right panel of the same figure, approximately recovering again the lognormal shape (standard Gaussian in rescaled units). Then, we have repeated the same experiment and 'lowered its resolution' by imposing the following: (i) the precision of Y is rounded to two decimal digits, imitating the precision of 10 ms found in Glissando, (ii) any synthetic phoneme shorter than a lower bound $Y < 0.03$ s is forced to have the minimal allowed duration, $Y = 0.03$ s. Results for this low-resolution version of the original experiment are then shown as purple curves in the same Figure 2. In particular, we can see how the lognormal shape of the phoneme time duration is significantly affected for shorter timescales, and such issues propagate to the word case at short timescales. The phenomenology is similar to what we found by comparing the results on the Buckeye corpus (English) versus the same results on a low-resolution version of the Buckeye corpus (bottom panels of Figure 1). All in all, this provides yet additional evidence explaining why the lognormality law might not be fully observable across all linguistic scales if the corpus has these kinds of limitations.

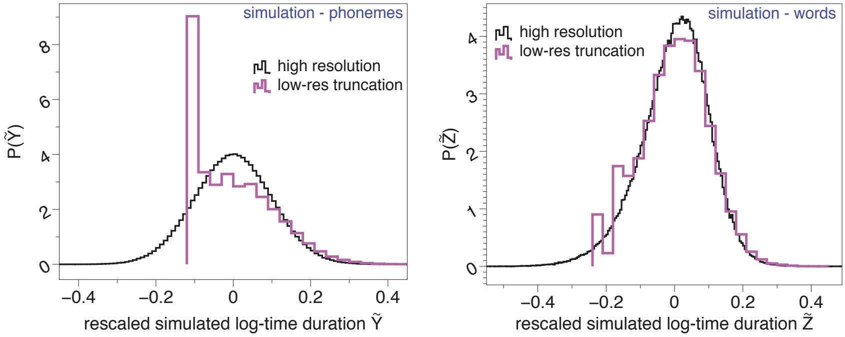


Figure 2. Lognormality law truncation. (**left**) Rescaled log-time duration distribution of synthetic 'phonemes' $P(\tilde{Y})$, estimated by (i) sampling $Y = \exp(X)$ where X is normally distributed $X \sim (\mu, \sigma^2)$ with $\mu = -3, \sigma = 2$, and then (ii) rescaling $\tilde{Y} = [\log Y - \langle \log Y \rangle]/std(\log Y)$ (where $std(\log Y)$ stands for the standard deviation of the random variable $\log Y$). If Y is lognormal, then $\tilde{Y} \sim \mathcal{N}(0,1)$. (**right**) Rescaled log-time duration distribution of synthetic 'words' $P(\tilde{Z})$, obtained using the stochastic model of [4] by concatenating n phonemes where n is another random variable whose distributed is approximated empirically. As the left panel, if Z is lognormal, then $\tilde{Z} \sim \mathcal{N}(0,1)$. In both panels, the black curve is the original, high resolution experiment whereas the purple curve is the result of (i) reducing the precision by rounding off to two decimal digits, (ii) reducing the sampling size to match differences between Buckeye and Glissando, and (iii) impose a lower-bound detectability $\tau = 0.03$ s (akin to the 30 ms of Glissando), such that all synthetically generated phonemes with a duration $Y < \tau$ are rounded to 0.03 s. Whereas lognormality is recovered in the original experiment, this shape is smeared out as soon as the lower-bound detectability threshold and other low-resolution artifacts are imposed, thereby explaining why the lognormality law might not be fully observable in Glissando.

2.2. Zipf's Law for Words and Yule Distribution for Phonemes

Results for Zipf's law are reported in Figure 3. The estimation of exponent α obtained for word frequencies applying the methodology of Clauset et al [45,46] are in agreement with those previously found [4] for the second regime in English (see Table 3), with $\alpha \approx 1.41$ (Spanish and English) and $\alpha \approx 1.42$ (Catalan), pointing to the robustness of the law also in speech. However, in the case of phonemes, whereas a Yule distribution can be fitted following the MLE method [44], fits are not very good—perhaps due to lack of statistics—and there are some slight differences between the distribution parameters of Catalan, Spanish and English (see Table 4). We conclude at this point that the Yule shape might not be universal for phoneme distribution and this hypothesis should be carefully revisited.

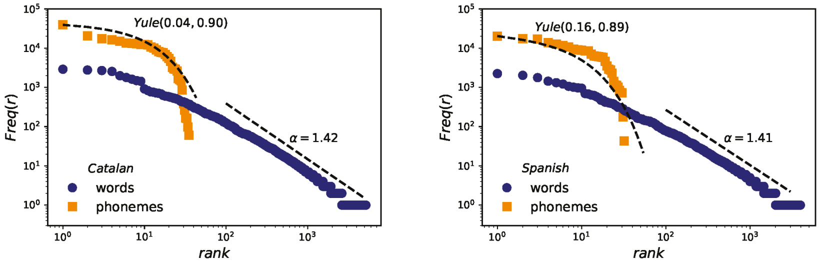


Figure 3. Zipf's law. Log-log frequency-rank of phonemes (orange squares) and words (blue circles) for the case of Catalan (**left**) and Spanish (**right**). Words are fitted to a power law distribution following [45,46] and leading to $x_{min} = 1$ and slopes almost similar for both languages. Phonemes are fitted to a Yule distribution with the help of the maximum likelihood estimation method (MLE).

2.3. Herdan–Heaps's Law

This law accounts for the sublinear increase of the number of different words V, and can be measured in physical units (i.e., as a function of the time elapsed T, $V(T) \sim T^{\gamma}$) or in symbolic units (i.e., as a function of the total number of words spoken L, $V(L) \sim L^{\beta}$). Results are reported in Figure 4, certifying that (i) this law holds both in Catalan and Spanish and (ii) both in symbolic (β) and physical (γ) units, (iii) with a scaling exponent $\beta \approx \gamma$, in good agreement to previous results [4] found for English: $\beta \approx 0.63$ (Spanish and English) and $\beta \approx 0.62$ (Catalan). In fact, this does not come as a surprise, given that a number of works have derived an inverse relationship between Zipf's and Herdan's exponents using different assumptions (see [47] or [48] for a review).

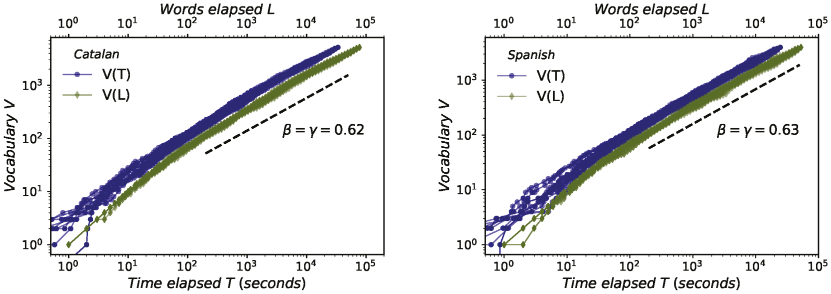


Figure 4. Herdan–Heaps's law. Sublinear increase of number of different words V versus time elapsed T (blue circles) and versus total number of words spoken L (green diamonds) for Catalan (**left**) and Spanish (**right**). As we are leading with a multiauthor corpus, each line represents a different way of permuting the order of concatenating each speaker. In every case we find scaling laws $V(L) \sim L^{\beta}$ and $V(T) \sim T^{\gamma}$ which holds for about three decades. The scaling exponents β and γ are estimated for each permutation using the least-squares method, and the average value of each of them over all permutations is shown in the figure. We find $\beta \approx \gamma$, as previously justified in [4], while its numerical value is on agreement with the one found for English [4].

2.4. Brevity Law

A mathematical formulation of brevity law was developed in [4] based on the information-theoretic principle of compression [32,49], when the size of the units is expressed in symbolic and physical units. In Figure 5 we report the results obtained for brevity law in the Glissando corpus for the case of words (left panel for Catalan, right panel for Spanish). Raw data (light grey) was fitted to the theoretical exponential law (see Table 1), and the best fit is depicted as a red dashed line. Also, a data binning is added (blue dots) to be able to visually compare it with the fit (red dashed line).

When word size is measured in physical units (word duration), the best exponential fit to the raw data (red dashed line) accurately matches the binned data, with similar fitting parameters $\lambda \approx 23.8$ (Catalan) and $\lambda \approx 24.1$ (Spanish) (to be compared with $\lambda \approx 20.6$ for English in Buckeye Corpus [4]), with significant Spearman correlations. Note that deviations of binned data from the red dashed line take place for short timescales: we argue that these are indeed related to the finite-precision and resolution issues discussed before, which propagate into (short) words.

When word size is measured in symbolic units (i.e., in the number of phonemes and number of characters), the law is again recovered (inset panels of Figure 5). Interestingly, the mathematical formulation of this law assigns a specific interpretation of the exponent λ when units are measured in symbolic space (i.e., when a code is available): the exponent in this case is always bounded $0 \leq \lambda \leq 1$ and quantifies the deviation of the language under study from compression optimality, where the closer to 1 the closer to optimality [4]. Out of the three languages, results suggest that Spanish ($\lambda_{\mathcal{D}} = 0.56$ for phonemes and $\lambda_{\mathcal{D}} = 0.60$ for characters) is slightly closest to optimality, followed by English ($\lambda_{\mathcal{D}} = 0.5$

for phonemes and $\lambda_{\mathcal{D}} = 0.6$ for characters) and Catalan ($\lambda_{\mathcal{D}} = 0.49$ for phonemes and $\lambda_{\mathcal{D}} = 0.53$ for characters).

In the case of phoneme duration the statistics—and thus fit—are much poorer, especially for Catalan (we recall again on the finite-precision and resolution issues of Glissando corpus). Nevertheless, Spearman correlations are significant, both for Spanish and Catalan (Figure 6), although Spearman's correlation is better for Spanish (−0.54) than for Catalan (−0.3).

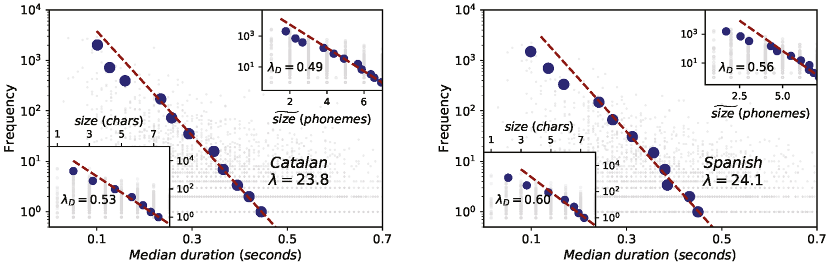


Figure 5. Brevity law: words (Catalan on the left panel and Spanish on the right panel). Red dashed lines are fits to the exponential law $f \sim \exp(-\lambda\ell)$, where ℓ is the word size which can be measured in physical units (mean duration) (**outer panels**) or in symbolical units (number of phonemes or number of characters, inset panels). See the text for and Table 3 for data fits and interpretation. Blue dots are the result of a data binning. Note that the fits are performed to the raw data, but the resulting exponential shape accurately matches the binned data within a range (deviations occur for shorter sizes, when the resolution and finite-precision issues of the Glissando corpus are important). Spearman test shows consistent negative correlations for the three formulations for the case of Catalan of −0.27, while for the case of Spanish the correlation is slightly stronger in physical magnitudes (−0.25) than in symbolic units (−0.22).

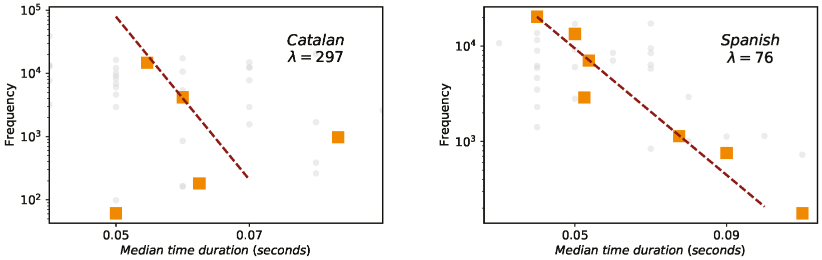


Figure 6. Brevity law: phonemes (Catalan on the left panel and Spanish on the right panel). Red dashed lines are fits to the exponential law $f \sim \exp(-\lambda\ell)$, where ℓ is the phoneme size measured in physical units (mean duration). Orange squares are the result of a data binning. Spearman test always denote negative correlations (−0.3 for Catalan, −0.54 for Spanish) but the data sample is too small to evaluate the agreement to the exponential law.

2.5. Size-Rank Law

The size-rank law mathematically connects the Brevity and Zipf's laws, indicating that the words of larger rank tend to have larger size [4]. Results for the case of words are depicted in Figure 7. Interestingly, despite the precision problems of Glissando for short durations already described previously, the size-rank law holds more robustly than the brevity law for both Catalan and Spanish. The slight variations in the exponent θ of Catalan (0.06) and Spanish (0.058), with respect to English (0.07), are here a consequence of the variations in the λ exponents of brevity law (Table 3).

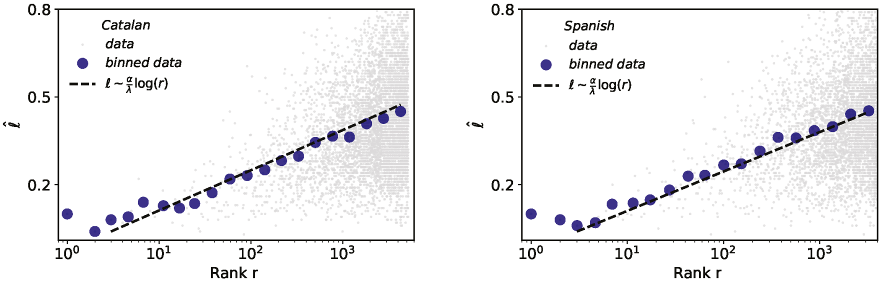


Figure 7. Size-rank law for words. Linear-log representation of word size ℓ versus rank of all words (blue dots denote binned data) in Catalan (**left**) and Spanish (**right**). The black dashed line is a fit of raw data (light grey dots) to the size-rank law (see Table 1), i.e., the fit of this law is not done to the binned data, however its agreement is excellent.

2.6. Menzerath–Altmann's Law (MAL)

The results of fittings of the Catalan and English corpus to MAL for different scales are depicted in Figures 8 and 9. For the scale of BGs vs words (Figure 8), MAL holds well when the size of the constituent is measured in physical units of time duration (outer panels) and it is either poorly or not fulfilled when the size is measured in symbolic units such as number of letters or number of phonemes per word (inset panels). Coefficients of determination $R^2 = 0.47$ for Catalan and $R^2 = 0.84$ for Spanish when size is measured in time duration, to bee compared with Catalan $R^2 = 0.23$ (Catalan, characters), $R^2 = 0.11$ (Catalan, phonemes), $R^2 = 0.04$ (Spanish, characters) and $R^2 = 0.08$ (Spanish, phonemes). These results are in agreement with the case of English [4]. Overall, better agreement to MAL is found for Spanish than for Catalan in time duration. Results and agreement to MAL also hold at the word vs phoneme scale. In fact, these results are new clear evidence in favor of the acoustical origin of the law [21] and the physical model explained in [4]. Note that while the size of the BGs are not large enough to reach to observe the range where MAL is inverted (at $b/c \approx 34$ words [4]), the value of the exponents (see Tables 3 and 4) certifies that such regime inversion indeed exists.

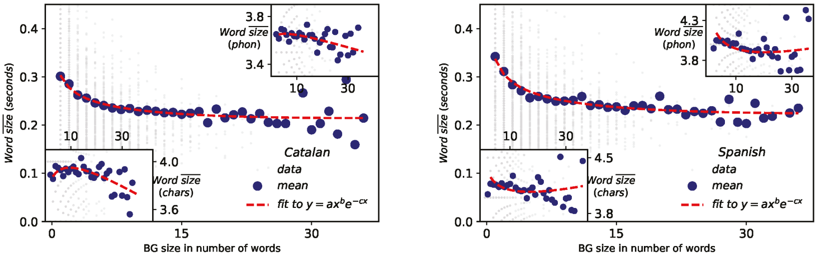


Figure 8. Menzerath–Altmann law: BG vs words Representation of BG size measured in number of words versus the mean size of those words for Catalan (**left**) and Spanish (**right**), where the size of the words can be measured in physical magnitudes (**main panel**) or symbolic units (phonemes or number of characters, inset panels). Each grey point represents one BG, whereas blue circles are the mean duration of BGs. MAL holds in physical magnitudes (with coefficient of determination $R^2 = 0.47$ for Catalan and $R^2 = 0.84$ for Spanish), while it is poorly fulfilled when the size is measured symbolically (Catalan: $R^2 = 0.23$ for character units and $R^2 = 0.11$ for phoneme units; Spanish: $R^2 = 0.04$ for character units and $R^2 = 0.08$ for phoneme units). Fitted parameters a, b, c are reported in Table 3.

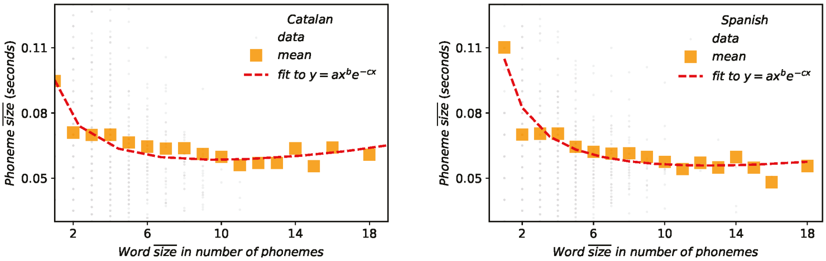


Figure 9. Menzerath–Altmann law: words–phonemes relation between the word size measured in number of phonemes versus the size of those phonemes in physical magnitudes. Orange squares represent the mean size of each word. Fitted parameters are shown in Table 4 coefficient of determination for these are $R^2 = 0.75$ for Catalan and $R^2 = 0.9$ for Spanish.

3. Discussion

The relevance of the study of the frequency of linguistic elements and its relative marginalization by the academy in the twentieth century was eloquently highlighted by Bybee [50] (p.6), stating that:

> *The other major theoretical factor working against an interest in frequency of use in language is the distinction, traditionally traced back to Ferdinand de Sausurre (1916), between the knowledge that speakers have of the signs and structures of their language and the way language is used by actual speakers communicating with one another. American structuralists, including those of the generativist tradition, accept this distinction and assert furthermore that the only worthwhile object of study is the underlying knowledge of language (Chomsky 1965 and subsequent works). In this view, any focus on the frequency of use of the patterns or items of language is considered irrelevant.*

While agreeing with Bybee [50], let us add here that the relevance of the study of the frequency of linguistic elements goes beyond the mere study of texts and, indeed, also concerns orality, which at the end of the day is at the basis of most human languages—with obviously some notable exceptions such as sign languages. In this work, we have studied the statistical properties and the onset of linguistic laws in the structure of speech in Catalan and Spanish. Set aside a precedent in pre-phonemic levels [39], to the best of our knowledge this is the first study of these characteristics in these two languages. Not only the frequencies of phonemes, words or breath groups of Catalan and Spanish have been examined here in speech transcriptions, but, perhaps more importantly, the presence of similar linguistic laws in these languages has been analyzed in inherently acoustic magnitudes (time duration), confirming some results previously found for English [4] and finding new evidence. Let us now summarise some of these findings and discuss some open problems for future research.

First, we have concluded that in order to fully observe the lognormality law at all linguistic scales the corpus under study needs to have sufficiently high resolution, and in particular a sufficiently low segmentation time, as close as possible to the physiological limit given by the glottal pulse (about 10 ms, see [51]). For too large lower-bounds (30 ms in the case of Glissando) and finite-precision, strong deviations from the lognormality law at the phoneme level will necessarily emerge, and some of these effects can mildly percolate at the word level. While it is not a definite proof, our experimental and numerical investigations supports our hypothesis that the lognormality law indeed holds well in Catalan and Spanish, complementing previous results in English [4], albeit it might not be fully observable at the phoneme level in Glissando corpus.

With respect to Zipf's law for the frequency of words, as well as with Herdan's law and the size-rank law, the experimental evidence found here for Catalan and Spanish reinforces the universality of these linguistic laws in speech. The mathematical models that relate these laws to each other are fully verified [4]. We can also highlight here the empirical strength of the size-rank law, very robust despite the variability of the data and the mentioned limitations of the corpus.

In the case of the brevity law for words, in Catalan and Spanish, the mathematical formulation derived from information theory and optimal coding [31] recently developed [4] is fully verified, and the law holds quantitatively. In a recent work [31] it has been established that optimal non-singular coding predicts that the length of linguistic elements should grow approximately as the logarithm of its frequency rank, which is consistent with Zipf's law of abbreviation and our approach [4], but more work is needed to certify any optimality ranking between languages.Future work shall also extend this analysis to other languages (ideally from different linguistic families beyond the Indo–European) one and/or with different writing systems, and would also explore the connection between the brevity exponents and other complexity metrics of language or the so-called orthographic transparency, defined as the more or less direct relationship in the conversion between graphemes and phonemes for each language [52].

Finally, MAL has been fully certified to hold in Catalan and Spanish *only* when measured in physical units, in line with previous evidence for English [4] and providing additional evidence to suggest that this is indeed a fundamentally acoustic law. In a similar vein, the bulk of results provided in this work is yet another empirical support to the validity of the 'physical hypothesis', and we hope it provides encouragement for other researchers to follow-up and address the necessary challenges to fully verify this hypothesis and its implications for theoretical linguistics.

To round off let us discuss some additional open problems, potential objects of future research. In the case of the Spanish written corpus, the frequencies of the words, lemmas and even their punctuation marks have been extensively reviewed before [53], and similarly for the case of Catalan and Spanish—in a bilingual context—the evolution of words and lemmas during childhood and adolescence in written production has even been analyzed [54]. In future work, a similar approach could be explored within speech, by analyzing in detail whether linguistic regularities emerge both in pauses, interruptions and other elements of acoustic variability and also in prosody, which turn out to be fundamental for example in clinical linguistics [55]. Finally, in relation to the ontogeny of language, and similarly to recent works which study the evolution of Zipf's law in language acquisition [13] as well as the law of brevity [56], the evolution of linguistic laws in speech is an open problem which also deserves investigation.

4. Materials and Methods

Glissando is a speech corpus for Spanish and Catalan which has been specially designed for prosodic studies, but that can be used also for other purposes [40]. It includes more than 12 hours of speech in Catalan and Spanish, recorded under optimal acoustic conditions, orthographically transcribed, phonetically aligned and annotated with prosodic information (location of the stressed syllables and prosodic phrasing). They are composed of three subcorpora: the 'news', a corpus of read news texts; the 'Task dialogues', a set of three task-oriented dialogues, covering three different interaction situations; and the 'free dialogues', a subcorpus of informal conversations. In this article, we aimed to research on orality and spontaneous conversation and compare the results with previous studies in English [4], so from those three we focused on 'task dialogues' and 'free dialogues'.

We had simultaneous access to (i) the aligned speech signal, (ii) its symbolic transcription and (iii) its phonetic transcription. In this way, we had control over all linguistic levels of phonemes and words, while the annotations allowed us to define another linguistic level: the breath group (BG). BGs are defined as the sequence of utterances between pauses in speech for breathing or longer [57]. A more detailed description of the Glissando corpus can be found in [40].

Annotation files containing symbolic and phonetic transcription were derived from the orthographic transcription using an automatic phonetic transcription tool, a phonetic aligner and a tool for the automatic annotation of prosodic boundaries [58]. The phonetic transcription tool converted the orthographic text into a chain of phonetic symbols, representing the theoretical pronunciation of the text. This phonetic transcription was time-aligned with the speech signal using the phonetic aligner, which established where to locate the initial and final boundaries for each phoneme. These aligners have some margin of error in the placement of these boundaries, so the output of this automatic process is usually revised by hand [40]. In the case of the 'News' subcorpus, this automatic annotation was manually revised by several experts in Phonetics, to obtain a transcription of the actual pronunciation of the speakers and not a theoretical one, as provided by the automatic tools, and to correct misplaced boundaries. During this process, some phonetic symbols not necessarily related to Spanish or Catalan were included to transcribe anomalous or deviated pronunciations. The 'dialogues' subcorpus, however, was not subject to this revision process [40]. This lack of manual revision might explain some of the specific phenomenology observed in this work, such as the lower-bound time duration resolution segmentation of the corpus, which could indeed be related to the limitations of the phonetic aligner.

Data and Reproducibility

From the whole corpus, we have used data corresponding to 'Free dialogues' and 'Task dialogues' in order to make the results more comparable with a previous work [4] which was based in the analysis of spontaneous speech conversation. Glissando corpus is a freely accessible corpus for non-commercial uses. It can be obtained through ELRA (http://catalog.elra.info/en-us/repository/browse/ELRA-S0406/ for the Spanish subcorpus and http://catalog.elra.info/en-us/repository/browse/ELRA-S0407/ for the Catalan subcorpus.).

Post-processed data of the Glissando corpus created by the authors of this article are available at https://doi.org/10.6071/M3XW9T, while scripts for generating the results are available at https://github.com/ivangtorre/ling-law-speech-spanish-catalan. We used Python 3.7 for the analysis. Levenberg–Marquardt algorithm, Kolmogorov–Smirnov distance, Spearman test and most of MLE fits use Scipy 1.3.0. MLE fits for power laws that are self-coded. Other libraries such as Numpy 1.16.2, Pandas 0.24.2 or Matplotlib 3.1.0 were also used. Fits to Zipf's law are done with R and PowerRlaw [46].

Author Contributions: All authors contributed to conceptualization, methodology, investigation, validation and writing (original draft preparation and review and editing of final version). J.-M.G. provided the database and data curation. I.G.T. performed data postprocessing, data analysis and programming; L.L. and A.H.-F. performed the formal analysis, supervision, project administration, resources and funding acquisition.

Funding: I.G.T. was supported by the grant FIS2017-84151-P (Ministerio de Economia, Industria y Competitividad, Gobierno de España). A.H.-F. was supported by the grant TIN2017-89244-R (MACDA) (Ministerio de Economia, Industria y Competitividad, Gobierno de España). I.G.T. and A.H.-F. were supported by the project PRO2019-S03 (RCO03080 Lingüística Quantitativa) de l'Institut d'Estudis Catalans. The development of the Glissando corpus was supported by projects FFI2008-04982-C03-01 and FFI2011-29559-C02-01 of the Ministerio de Ciencia e Innovación (Gobierno de España). L.L. was supported by EPSRC Early Career Fellowship EP/P01660X/1.

Acknowledgments: The authors especially appreciate the previous work of all collaborators in the Glissando corpus, without which this work would not have been possible.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

BG	Breath group
LND	Lognormal distribution
MAL	Menzerath–Altmann's Law

References

1. Köhler, R.; Altmann, G.; Piotrowski, R.G. *Quantitative Linguistik/Quantitative Linguistics: Ein Internationales Handbuch/an International Handbook*; Walter de Gruyter: Berlin, Germany, 2008; Volume 27.
2. Grzybek, P. History of quantitative linguistics. *Glottometrics* **2012**, *23*, 70–80.
3. Best, K.H.; Rottmann, O. *Quantitative Linguistics, an Invitation*; RAM-Verlag: Ludenscheid, Germany, 2017.

4. Torre, I.G.; Luque, B.; Lacasa, L.; Kello, C.T.; Hernández-Fernández, A. On the physical origin of linguistic laws and lognormality in speech. *R. Soc. Open Sci.* **2019**, *6*. [CrossRef] [PubMed]
5. Pareto, V. *Cours d'économie Politique*; Librairie Droz; Imprime en Suisse: Geneva, Swizerland, 1964; Volume 1. (In French)
6. Estoup, J.B. *Gammes Sténographiques. Recueil de Textes Choisis pour L'acquisition Méthodique de la Vitesse, Précédé d'une Introduction par J.-B. Estoup*; Sténographique: Paris, France, 1912. (In French)
7. Condon, E.U. Statistics of vocabulary. *Science* **1928**, *67*, 300. [CrossRef] [PubMed]
8. Zipf, G.K. *The Psychobiology of Language, an Introduction to Dynamic Philology*; Houghton–Mifflin: Boston, MA, USA, 1935.
9. Zipf, G.K. *Human Behavior and the Principle of Least Effort*; Addison–Wesley: Cambridge, MA, USA, 1949.
10. Altmann, E.G.; Gerlach, M. Statistical laws in linguistics. In *Creativity and Universality in Language*; Springer: Cham, Germany, 2016; pp. 7–26.
11. Bian, C.; Lin, R.; Zhang, X.; Ma, Q.D.Y.; Ivanov, P.C. Scaling laws and model of words organization in spoken and written language. *EPL (Europhysics Letters)* **2016**, *113*, 18002. [CrossRef]
12. Ferrer-i Cancho, R. The variation of Zipf's law in human language. *Eur. Phys. J. B* **2005**, *44*, 249–257. [CrossRef]
13. Baixeries, J.; Elvevag, B.; Ferrer-i Cancho, R. The evolution of the exponent of Zipf's law in language ontogeny. *PLoS ONE* **2013**, *8*. [CrossRef]
14. Neophytou, K.; van Egmond, M.; Avrutin, S. Zipf's Law in Aphasia Across Languages: A Comparison of English, Hungarian and Greek. *J. Quant. Linguist.* **2017**, *24*, 178–196. [CrossRef]
15. Kuraszkiewicz, W.; Łukaszewicz, J. Ilość różnych wyrazów w zależności od długości tekstu. *Pamiętnik Literacki: Czasopismo Kwartalne Poświęcone Historii i Krytyce Literatury Polskiej* **1951**, *42*, 168–182. (In Polish)
16. Herdan, G. *Type-Token Mathematics: A Textbook of Mathematical Linguistics*; De Gruyter Mouton: Berlin, Germany, 1960.
17. Heaps, H.S. *Information Retrieval, Computational and Theoretical Aspects*; Academic Press: Cambridge, MA, USA, 1978.
18. Zipf, G.K. *Selected Studies of the Principle of Relative Frequency in Language*; De Gruyter Mouton: Berlin, Germany, 1932.
19. Bentz, C.; i Cancho, R.F. *Zipf's Law of Abbreviation as a Language Universal*; Universitätsbibliothek Tübingen: Tübingen, The Netherlands, 2016.
20. Grégoire, A. Variation de la dure de la syllabe française suivant sa place dans les groupements phonetiques. *La Parole* **1899**, *1*, 161–176. (In French)
21. Menzerath, P.; Oleza, J. *Spanische Lautdauer: Eine Experimentelle Untersuchung*; De Gruyter Mouton: Berlin, Germany, 1928. (In German)
22. Menzerath, P. *Die Architektonik des Deutschen Wortschatzes*; Dümmler: Berlin, Germany, 1954; Volume 3. (In German)
23. Altmann, G. Prolegomena to Menzerath's law. *Glottometrika* **1980**, *2*, 1–10.
24. Altmann, G.; Schwibbe, M. *Das Menzertahsche Gesetz in Informationsverbarbeitenden Systemen*; Georg Olms: Hildesheim, Germany, 1989. (In German)
25. Herdan, G. The relation between the dictionary distribution and the occurrence distribution of word length and its importance for the study of Quantitative Linguistics.*Biometrika* **1958**, *45*, 222–228. [CrossRef]
26. Rosen, K.M. Analysis of speech segment duration with the lognormal distribution: A basis for unification and comparison. *J. Phon.* **2005**, *33*, 411–426. [CrossRef]
27. Gopinath, D.P.; Veena, S.; Nair, A.S. Modeling of Vowel Duration in Malayalam Speech using Probability Distribution. In Proceedings of the Speech Prosody, Campinas, Brazil, 6–9 May 2008; pp. 6–9.
28. Shaw, J.A.; Kawahara, S. Effects of surprisal and entropy on vowel duration in Japanese. *Language Speech* **2017**, *62*, 80–114. [CrossRef] [PubMed]
29. Gahl, S. Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language* **2008**, *84*, 474–496. [CrossRef]
30. Tomaschek, F.; Wieling, M.; Arnold, D.; Baayen, R.H. Word frequency, Vowel Length and Vowel Quality in Speech Production: An EMA Study of the Importance of Experience. Available online: https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/5957 (accessed on 23 November 2019).

31. Ferrer-i-Cancho, R.; Bentz, C.; Seguin, C. Optimal coding and the origins of Zipfian laws. *arXiv* **2019**, arXiv:1906.01545.
32. Cover, T.M.; Thomas, J.A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*; Wiley-Interscience: New York, NY, USA, 2006.
33. Cramer, I. The Parameters of the Altmann-Menzerath Law. *J. Quant. Linguist.* **2005**, *12*, 41–52. [CrossRef]
34. Grzybek Peter, N.; Stadlober, E.; Kelih Emmerich, N. The Relationship of Word Length and Sentence Length: The Inter-Textual Perspective. In *Advances In Data Analysis*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 611–618.
35. Mačutek, J.; Chromý, J.; Koščová, M. Menzerath-Altmann Law and Prothetic /v/ in Spoken Czech. *J. Quant. Linguist.* **2019**, *26*, 66–80. [CrossRef]
36. Sayli, O. Duration Analysis and Modeling for Turkish Text-to-Speech Synthesis. Master's Thesis, Bogaziei University, Istanbul, Turkey, 2002.
37. Greenberg, S.; Carvey, H.; Hitchcock, L.; Chang, S. Temporal properties of spontaneous speech-a syllable-centric perspective. *J. Phon.* **2003**, *31*, 465–485. [CrossRef]
38. Luque, J.; Luque, B.; Lacasa, L. Scaling and universality in the human voice. *J. R. Soc. Interface* **2015**, *12*, 20141344. [CrossRef]
39. Torre, I.G.; Luque, B.; Lacasa, L.; Luque, J.; Hernández-Fernández, A. Emergence of linguistic laws in human voice. *Sci. Rep.* **2017**, *7*, 43862. [CrossRef]
40. Garrido, J.M.; Escudero, D.; Aguilar, L.; Cardeñoso, V.; Rodero, E.; de-la Mota, C.; González, C.; Rustullet, S.; Larrea, O.; Laplaza, Y.; et al. Glissando: A corpus for multidisciplinary prosodic studies in Spanish and Catalan. *Lang. Resour. Eval.* **2013**, *47*, 945–971. [CrossRef]
41. Fernández Planas, A. *Así se Habla: Nociones Fundamentales de Fonética General y Española.; Apuntes de Catalán, Gallego y Euskara*; Horsori Editorial: Barcelona, Spain, 2005. (In Spanish)
42. Pitt, M.A.; Dilley, L.; Johnson, K.; Kiesling, S.; Raymond, W.; Hume, E.; Fosler-Lussier, E. Buckeye Corpus of Conversational Speech, 2nd release; Columbus, OH: Department of Psychology, Ohio State University, 2007. Available online: http://sldr.org/voir_depot.php?id=776&lang=en&sip=0 (accessed on 23 November 2019).
43. Pitt, M.A.; Johnson, K.; Hume, E.; Kiesling, S.; Raymond, W. The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Commun.* **2005**, *45*, 89–95. [CrossRef]
44. Eliason, S.R. *Maximum Likelihood Estimation: Logic and Practice*; Sage Publications: Tucson, AZ, USA, 1993; Volume 96.
45. Clauset, A.; Shalizi, C.R.; Newman, M.E. Power-law distributions in empirical data. *SIAM Rev.* **2009**, *51*, 661–703. [CrossRef]
46. Gillespie, C.S. Fitting Heavy Tailed Distributions: The poweRlaw Package. *J. Stat. Softw.* **2015**, *64*, 1–16. [CrossRef]
47. Lü, L.; Zhang, Z.K.; Zhou, T. Zipf's law leads to Heaps' law: Analyzing their relation in finite-size systems. *PLoS ONE* **2010**, *5*, e14139. [CrossRef] [PubMed]
48. Font-Clos, F.; Boleda, G.; Corral, A. A scaling law beyond Zipf's law and its relation to Heaps' law. *New J. Phys.***2013**, *15*, 093033. [CrossRef]
49. Ferrer-i Cancho, R. Compression and the origins of Zipf's law for word frequencies. *Complexity* **2016**, *21*, 409–411. [CrossRef]
50. Bybee, J. *Frequency of Use and the Organization of Language*; Oxford University Press: Oxford, UK, 2007.
51. Quatieri, T.F. *Discrete-Time Speech Signal Processing: Principles and Practice*; Prentice Hall PTR: Upper Saddle River, NJ, USA, 2002.
52. Borleffs, E.; Maassen, B.A.M.; Lyytinen, H.; Zwarts, F. Measuring orthographic transparency and morphological-syllabic complexity in alphabetic orthographies: A narrative review. *Read. Writ.* **2017**, *30*, 1617–1638. [CrossRef]
53. Rojo, G. Sobre la configuración estadística de los corpus textuales. *Lingüística* **2017**, *33*, 121–134. (In Spanish) [CrossRef]
54. Tolchinsky, L.; Martí, A.; Llaurado, A. The growth of the written lexicon in Catalan From childhood to adolescence. *Writ. Lang. Lit.* **2010**, *13*, 206–235. [CrossRef]
55. Baken, R.; Orlikoff, R. *Clinical Measurement of Speech and Voice (Speech Science)*; Cengage Learning: Boston, MA, USA, 2000.

56. Casas, B.; Hernández-Fernández, A.; Català, N.; i Cancho, R.F.; Baixeries, J. Polysemy and brevity versus frequency in language. *Comput. Speech Lang.* **2019**, *58*, 1–50. [CrossRef]
57. Tsao, Y.C.; Weismer, G. Interspeaker variation in habitual speaking rate: Evidence for a neuromuscular component. *J. Speech Lang. Hear. Res.* **1997**, *40*, 858–866. [CrossRef] [PubMed]
58. Garrido, J.M. SegProso: A Praat-Based Tool for the Automatic Detection and Annotation of Prosodic Boundaries in Speech Corpora. In Proceedings of the TRASP 2013, Barcelona, Spain, 30 August 2013, pp. 74–77.

entropy

Article

Estimating Predictive Rate–Distortion Curves via Neural Variational Inference

Michael Hahn [1,*] and Richard Futrell [2]

[1] Department of Linguistics, Stanford University, Stanford, CA 94305, USA
[2] Department of Language Science, University of California, Irvine, CA 92697, USA
[*] Correspondence: mhahn2@stanford.edu

Received: 13 May 2019; Accepted: 26 June 2019; Published: 28 June 2019

Abstract: The Predictive Rate–Distortion curve quantifies the trade-off between compressing information about the past of a stochastic process and predicting its future accurately. Existing estimation methods for this curve work by clustering finite sequences of observations or by utilizing analytically known causal states. Neither type of approach scales to processes such as natural languages, which have large alphabets and long dependencies, and where the causal states are not known analytically. We describe Neural Predictive Rate–Distortion (NPRD), an estimation method that scales to such processes, leveraging the universal approximation capabilities of neural networks. Taking only time series data as input, the method computes a variational bound on the Predictive Rate–Distortion curve. We validate the method on processes where Predictive Rate–Distortion is analytically known. As an application, we provide bounds on the Predictive Rate–Distortion of natural language, improving on bounds provided by clustering sequences. Based on the results, we argue that the Predictive Rate–Distortion curve is more useful than the usual notion of statistical complexity for characterizing highly complex processes such as natural language.

Keywords: Predictive Rate–Distortion; natural language; information bottleneck; neural variational inference

1. Introduction

Predicting the future from past observations is a key problem across science. Constructing models that predict future observations is a fundamental method of making and testing scientific discoveries. Understanding and predicting dynamics has been a fundamental goal of physics for centuries. In engineering, devices often have to predict the future of their environment to perform efficiently. Living organisms need to make inferences about their environment and its future for survival.

Real-world systems often generate complex dynamics that can be hard to compute. A biological organism typically will not have the computational capacity to perfectly represent the environment. In science, measurements have limited precision, limiting the precision with which one can hope to make predictions. In these settings, the main goal will typically be to get good prediction at low computational cost.

This motivates the study of models that try to extract those key features of past observations that are most relevant to predicting the future. A general information-theoretic framework for this problem is provided by Predictive Rate–Distortion [1,2], also known as the past-future information bottleneck [3]. The Predictive Rate–Distortion trade-off seeks to find an encoding of past observations that is maximally informative about future observations, while satisfying a bound on the amount of information that has to be stored. More formally, this framework trades off prediction loss in the future, formalized as cross-entropy, with the cost of representing the past, formalized as the mutual information between the past observations and the compressed representations of the past. Due to its information-theoretic nature, this framework is extremely general and applies to processes across vastly different domains. It has been applied to linear dynamical systems [3,4], but is equally meaningful for discrete dynamical systems [2]. For biological systems that make predictions about their environment, this corresponds to placing an information-theoretic constraint on the computational cost used for conditioning actions on past observations [5].

The problem of determining encodings that optimize the Predictive Rate–Distortion trade-off has been solved for certain specific kinds of dynamics, namely for linear dynamic systems [3] and for processes whose predictive dynamic is represented exactly by a known, finite Hidden Markov Model [2]. However, real-world processes are often more complicated. When the dynamics are known, a representing Hidden Markov Model may still be extremely large or even infinite, making general-purpose automated computation difficult. Even more importantly, the underlying dynamics are often not known exactly. An organism typically does not have access to the exact dynamics of its surroundings. Similarly, the exact distribution of sentences in a natural language is not known exactly, precluding the application of methods that require an exact model. Such processes are typically only available implicitly, through a finite sample of trajectories.

Optimal Causal Filtering (OCF, Still et al. [6]) addresses the problem of estimating Predictive Rate–Distortion from a finite sample of observation trajectories. It does so by constructing a matrix of observed frequencies of different pairs of past and future observations. However, such a method faces a series of challenges [2]. One is the curse of dimensionality: Modeling long dependencies requires storing an exponential number of observations, which quickly becomes intractable for current computation methods. This exponential growth is particularly problematic when dealing with processes with a large state space. For instance, the number of distinct words in a human language as found in large-scale text data easily exceeds 1×10^5, making storing and manipulating counts of longer word sequences very challenging. A second challenge is that of overfitting: When deploying a predictive model constructed via OCF on new data to predict upcoming observations, such a model can only succeed when the past sequences occurred in the sample to which OCF was applied. This is because OCF relies on counts of full past and future observation sequences; it does not generalize to unseen past sequences.

Extrapolating to unseen past sequences is possible in traditional time series models representing processes that take continuous values; however, such methods are less easily applied to discrete sequences such as natural language. Recent research has seen a flurry of interest in using flexible nonlinear function approximators, and in particular recurrent neural networks, which can handle sequences with discrete observations. Such machine learning methods provide generic models of sequence data. They are the basis of the state of the art by a clear and significant margin for prediction in natural language [7–10]. They also have been successfully applied to modeling many other kinds of time series found across disciplines [11–18].

We propose Neural Predictive Rate–Distortion (NPRD) to estimate Predictive Rate–Distortion when only a finite set of sample trajectories is given. We use neural networks both to encode past observations into a summary code, and to predict future observations from it. The universal function approximation capabilities of neural networks enable such networks to capture complex dynamics, with computational cost scaling only linearly with the length of observed trajectories, compared to the exponential cost of OCF. When deploying on new data, such a neural model can generalize seamlessly to unseen sequences, and generate appropriate novel summary encodings on the fly. Recent advances in neural variational inference [19,20] allow us to construct predictive models that provide almost optimal predictive performance at a given rate, and to estimate the Predictive Rate–Distortion trade-off from such networks. Our method can be applied to sample trajectories in an off-the-shelf manner, without prior knowledge of the underlying process.

In Section 2, we formally introduce Predictive Rate–Distortion, and discuss related notions of predictive complexity. In Section 3, we describe the prior method of Optimal Causal Filtering (OCF). In Section 4, we describe our method NPRD. In Section 5, we validate NPRD on processes whose

Predictive Rate–Distortion is analytically known, showing that it finds essentially optimal predictive models. In Section 6, we apply NPRD to data from five natural languages, providing the first estimates of Predictive Rate–Distortion of natural language.

2. Predictive Rate–Distortion

We consider stationary discrete-time stochastic processes $(X_t)_{t \in \mathbb{Z}}$, taking values in a state space S. Given a reference point in time, say $T = 0$, we are interested in the problem of predicting the future of $X_{t \geq 0} = (X_0, X_1, X_2, ...)$ from the past $X_{t<0} = (..., X_{-2}, X_{-1})$. In general—unless the observations X_t are independent—predicting the future of the process accurately will require taking into account the past observations. There is a trade-off between the accuracy of prediction, and how much information about the past is being taken into account. On one extreme, not taking the past into account at all, one will not be able to take advantage of the dependencies between past and future observations. On the other extreme, considering the entirety of the past observations $X_{t \leq 0}$ can require storing large and potentially unbounded amounts of information. This trade-off between information storage and prediction accuracy is referred to as Predictive Rate–Distortion (PRD) [2]. The term rate refers to the amount of past information being taken into account, while distortion refers to the degradation in prediction compared to optimal prediction from the full past.

The problem of Predictive Rate–Distortion has been formalized by a range of studies. A principled and general formalization is provided by applying the Information Bottleneck idea [2,6,21]: We will write $\overleftarrow{X}$ for the past $X_{<0}$, and $\overrightarrow{X}$ for the future $X_{\geq 0}$, following [2]. We consider random variables Z, called codes, that summarize the past and are used by the observer to predict the future. Formally, Z needs to be independent from the future $\overrightarrow{X}$ conditional on the past $\overleftarrow{X}$: in other words, Z does not provide any information about the future except what is contained in the past. Symbolically:

$$Z \perp \overrightarrow{X} | \overleftarrow{X}. \tag{1}$$

This is equivalent to the requirement that $Z \leftrightarrow \overleftarrow{X} \leftrightarrow \overrightarrow{X}$ be a Markov chain. This formalizes the idea that the code is computed by an observer from the past, without having access to the future. Predictions are then made based only on Z, without additional access to the past $\overleftarrow{X}$.

The rate of the code Z is the mutual information between Z and the past: $\mathrm{I}[Z, \overleftarrow{X}]$. By the Channel Coding Theorem, this describes the channel capacity that the observer requires in order to transform past observations into the code Z.

The distortion is the loss in predictive accuracy when predicting from Z, relative to optimal prediction from the full past $\overleftarrow{X}$. In the Information Bottleneck formalization, this is equivalent to the amount of mutual information between past and future that is *not* captured by Z [22]:

$$\mathrm{I}[\overleftarrow{X}, \overrightarrow{X} | Z]. \tag{2}$$

Due to the Markov condition, the distortion measure satisfies the relation

$$\mathrm{I}[\overleftarrow{X}, \overrightarrow{X} | Z] = \mathrm{I}[\overleftarrow{X}, \overrightarrow{X}] - \mathrm{I}[Z, \overrightarrow{X}], \tag{3}$$

i.e., it captures how much less information Z carries about the future $\overrightarrow{X}$ compared to the full past $\overleftarrow{X}$. For a fixed process $(X_t)_t$, choosing Z to minimize the distortion is equivalent to maximizing the mutual information between the code and the future:

$$\mathrm{I}[Z, \overrightarrow{X}]. \tag{4}$$

We will refer to (4) as the predictiveness of the code Z.

The rate–distortion trade-off then chooses Z to minimize distortion at bounded rate:

$$\min_{Z:\mathrm{I}[\overleftarrow{X},Z]\leq d} \mathrm{I}[\overleftarrow{X},\overrightarrow{X}|Z] \tag{5}$$

or—equivalently—maximize predictiveness at bounded rate:

$$\max_{Z:\mathrm{I}[\overleftarrow{X},Z]\leq d} \mathrm{I}[Z,\overrightarrow{X}]. \tag{6}$$

Equivalently, for each $\lambda \geq 0$, we study the problem

$$\max_{Z} \left(\mathrm{I}[Z,\overrightarrow{X}] - \lambda \cdot \mathrm{I}[\overleftarrow{X},Z]\right), \tag{7}$$

where the scope of the maximization is the class of all random variables Z such that $Z \rightarrow \overleftarrow{X} \rightarrow \overrightarrow{X}$ is a Markov chain.

The objective (7) is equivalent to the Information Bottleneck [21], applied to the past and future of a stochastic process. The coefficient λ indicates how strongly a high rate $\mathrm{I}[\overleftarrow{X},Z]$ is penalized; higher values of λ result in lower rates and thus lower values of predictiveness.

The largest achievable predictiveness $\mathrm{I}[Z,\overrightarrow{X}]$ is equal to $\mathrm{I}[\overleftarrow{X},\overrightarrow{X}]$, which is known as the excess entropy of the process [23]. Due to the Markov condition (1) and the Data Processing Inequality, predictiveness of a code Z is always upper-bounded by the rate:

$$\mathrm{I}[Z,\overrightarrow{X}] \leq \mathrm{I}[\overleftarrow{X},Z]. \tag{8}$$

As a consequence, when $\lambda \geq 1$, then (7) is always optimized by a trivial Z with zero rate and zero predictiveness. When $\lambda = 0$, any lossless code optimizes the problem. Therefore, we will be concerned with the situation where $\lambda \in (0,1)$.

2.1. Relation to Statistical Complexity

Predictive Rate–Distortion is closely related to Statistical Complexity and the ϵ-machine [24,25]. Given a stationary process X_t, its causal states are the equivalence classes of semi-infinite pasts $\overleftarrow{X}$ that induce the same conditional probability over semi-infinite futures $\overrightarrow{X}$: Two pasts $\overleftarrow{X}$, $\overleftarrow{X}'$ belong to the same causal state if and only if $P(x_{1\ldots k}|\overleftarrow{X}) = P(x_{1\ldots k}|\overleftarrow{X}')$ holds for all finite sequences $x_{0\ldots k}$ ($k \in \mathbb{N}$). Note that this definition is not measure-theoretically rigorous; such a treatment is provided by Löhr [26].

The causal states constitute the state set of a a Hidden Markov Model (HMM) for the process, referred to as the ϵ-machine [24]. The statistical complexity of a process is the state entropy of the ϵ-machine. Statistical complexity can be computed easily if the ϵ-machine is analytically known, but estimating statistical complexity empirically from time series data are very challenging and seems to at least require additional assumptions about the process [27].

Marzen and Crutchfield [2] show that Predictive Rate–Distortion can be computed when the ϵ-machine is analytically known, by proving that it is equivalent to the problem of compressing causal states, i.e., equivalence classes of pasts, to predict causal states of the backwards process, i.e., equivalence classes of futures. Furthermore, [6] show that, in the limit of $\lambda \rightarrow 0$, the code Z that optimizes Predictive Rate–Distortion (7) turns into the causal states.

2.2. Related Work

There are related models that represent past observations by extracting those features that are relevant for prediction. Predictive State Representations [28,29] and Observable Operator Models [30] encode past observations as sets of predictions about future observations. Rubin et al. [31] study agents that trade the cost of representing information about the environment against the reward they can obtain by choosing actions based on the representation. Relatedly, Still [1] introduces a Recursive Past Future Information Bottleneck where past information is compressed repeatedly, not just at one reference point in time.

As discussed in Section 2.1, estimating Predictive Rate–Distortion is related to the problem of estimating statistical complexity. Clarke et al. [27] and Still et al. [6] consider the problem of estimating statistical complexity from finite data. While statistical complexity is hard to identify from finite data in general, Clarke et al. [27] introduces certain restrictions on the underlying process that make this more feasible.

3. Prior Work: Optimal Causal Filtering

The main prior method for estimating Predictive Rate–Distortion from data are Optimal Causal Filtering (OCF, Still et al. [6]). This method approximates Predictive Rate–Distortion using two approximations: first, it replaces semi-infinite pasts and futures with bounded-length contexts, i.e., pairs of finite past contexts ($\overset{M\leftarrow}{X} := X_{-M} \dots X_{-1}$) and future contexts ($\overset{\rightarrow M}{X} := X_0 \dots X_{M-1}$) of some finite length M.(It is not crucial that past and future contexts have the same lengths, and indeed Still et al. [6] do not assume this). (We do assume equal length throughout this paper for simplicity of our experiments, though nothing depends on this). The PRD objective (7) then becomes (9), aiming to predict length-M finite futures from summary codes Z of length-M finite pasts:

$$\max_{Z:Z\perp \overset{\rightarrow M}{X} | \overset{M\leftarrow}{X}} \left(\mathrm{I}[Z, \overset{\rightarrow M}{X}] - \lambda \cdot \mathrm{I}[\overset{M\leftarrow}{X}, Z] \right). \tag{9}$$

Second, OCF estimates information measures directly from the observed counts of ($\overset{M\leftarrow}{X}$), ($\overset{\rightarrow M}{X}$) using the plug-in estimator of mutual information. With such an estimator, the problem in (9) can be solved using a variant of the Blahut–Arimoto algorithm [21], obtaining an encoder $P(Z|\overset{M\leftarrow}{X})$ that maps each observed past sequence $\overset{M\leftarrow}{X}$ to a distribution over a (finite) set of code words Z.

Two main challenges have been noted in prior work: first, solving the problem for a finite empirical sample leads to overfitting, overestimating the amount of structure in the process. Still et al. [6] address this by subtracting an asymptotic correction term that becomes valid in the limit of large M and $\lambda \to 0$, when the codebook $P(Z|\overleftarrow{X})$ becomes deterministic, and which allows them to select a deterministic codebook of an appropriate complexity. This leaves open how to obtain estimates outside of this regime, when the codebook can be far from deterministic.

The second challenge is that OCF requires the construction of a matrix whose rows and columns are indexed by the observed past and future sequences [2]. Depending on the topological entropy of the process, the number of such sequences can grow as $|A|^M$, where A is the set of observed symbols, and processes of interest often do show this exponential growth [2]. Drastically, in the case of natural language, A contains thousands of words.

A further challenge is that OCF is infeasible if the number of required codewords is too large, again because it requires constructing a matrix whose rows and columns are indexed by the codewords and observed sequences. Given that storing and manipulating matrices greater than $10^5 \times 10^5$ is currently not feasible, a setting where $\mathrm{I}[\overleftarrow{X}, Z] > \log 10^5 \approx 11.5$ cannot be captured with OCF.

4. Neural Estimation via Variational Upper Bound

We now introduce our method, Neural Predictive Rate–Distortion (NPRD), to address the limitations of OCF, by using parametric function approximation: whereas OCF constructs a codebook mapping between observed sequences and codes, we use general-purpose function approximation estimation methods to compute the representation Z from the past and to estimate a distribution over future sequences from Z. In particular, we will use recurrent neural networks, which are known to provide good models of sequences from a wide range of domains; our method will also be applicable to other types of function approximators.

This will have two main advantages, addressing the limitations of OCF: first, unlike OCF, function approximators can discover generalizations across similar sequences, enabling the method to calculate good codes Z even for past sequences that were not seen previously. This is of paramount importance in settings where the state space is large, such as the set of words of a natural language. Second, the cost of storing and evaluating the function approximators will scale *linearly* with the length of observed sequences both in space and in time, as opposed to the exponential memory demand of OCF. This is crucial for modeling long dependencies.

4.1. Main Result: Variational Bound on Predictive Rate–Distortion

We will first describe the general method, without committing to a specific framework for function approximation yet. We will construct a bound on Predictive Rate–Distortion and optimize this bound in a parametric family of function approximators to obtain an encoding Z that is close to optimal for the nonparametric objective (7).

As in OCF (Section 3), we assume that a set of finite sample trajectories $x_{-M} \dots x_{M-1}$ is available, and we aim to compress pasts of length M to predict futures of length M. To carry this out, we restrict the PRD objective (7) to such finite-length pasts and futures:

$$\max_{Z:Z\perp \overset{\rightarrow M}{X} | \overset{M\leftarrow}{X}} \left(\mathrm{I}[Z, \overset{\rightarrow M}{X}] - \lambda \cdot \mathrm{I}[\overset{M\leftarrow}{X}, Z] \right). \tag{10}$$

It will be convenient to equivalently rewrite (10) as

$$\min_{Z:Z\perp \overset{\rightarrow M}{X} | \overset{M\leftarrow}{X}} \left[\mathrm{H}[\overset{\rightarrow M}{X} | Z] + \lambda \cdot \mathrm{I}[\overset{M\leftarrow}{X}, Z] \right], \tag{11}$$

where $\mathrm{H}[\overset{\rightarrow M}{X} |Z]$ is the prediction loss. Note that minimizing prediction loss is equivalent to maximizing predictiveness $\mathrm{I}[\overset{\rightarrow M}{X}, Z]$.

When deploying such a predictive code Z, two components have to be computed: a distribution $P(Z| \overset{M\leftarrow}{X})$ that encodes past observations into a code Z, and a distribution $P(\overset{\rightarrow M}{X} |Z)$ that decodes the code Z into predictions about the future.

Let us assume that we already have some encoding distribution

$$Z \sim P_\phi(Z| \overset{M\leftarrow}{X}), \tag{12}$$

where ϕ is the encoder, expressed in some family of function approximators. The encoder transduces an observation sequence $\overset{M\leftarrow}{X}$ into the parameters of the distribution $P_\phi(\cdot| \overset{M\leftarrow}{X})$. From this encoding distribution, one can obtain the optimal decoding distribution over future observations via Bayes' rule:

$$P(\overset{\rightarrow M}{X} |Z) = \frac{P(\overset{\rightarrow M}{X}, Z)}{P(Z)} = \frac{\mathbb{E}_{\overset{M\leftarrow}{X}} P(\overset{\rightarrow M}{X}, Z| \overset{M\leftarrow}{X})}{\mathbb{E}_{\overset{M\leftarrow}{X}} P_\phi(Z| \overset{M\leftarrow}{X})} =^{(*)} \frac{\mathbb{E}_{\overset{M\leftarrow}{X}} P(\overset{\rightarrow M}{X} | \overset{M\leftarrow}{X}) P_\phi(Z| \overset{M\leftarrow}{X})}{\mathbb{E}_{\overset{M\leftarrow}{X}} P_\phi(Z| \overset{M\leftarrow}{X})}, \tag{13}$$

where $(*)$ uses the Markov condition $Z\perp \overset{\rightarrow M}{X} | \overset{M\leftarrow}{X}$. However, neither of the two expectations in the last expression of (13) is tractable, as they require summation over exponentially many sequences, and algorithms (e.g., dynamic programming) to compute this sum efficiently are not available in general. For a similar reason, the rate $\mathrm{I}[\overset{M\leftarrow}{X}, Z]$ of the code $Z \sim P_\phi(Z| \overset{M\leftarrow}{X})$ is also generally intractable.

Our method will be to introduce additional functions, also expressed using function approximators that approximate some of these intractable quantities: first, we will use a parameterized probability distribution q as an approximation to the intractable marginal $P(Z) = \mathbb{E}_{\overset{M\leftarrow}{X}} P_\phi(Z| \overset{M\leftarrow}{X})$:

$$q(Z) \text{ approximates } P(Z) = \mathbb{E}_{\overset{M\leftarrow}{X}} P_\phi(Z| \overset{M\leftarrow}{X}). \tag{14}$$

Second, to approximate the decoding distribution $P(\overset{\rightarrow M}{X} |Z)$, we introduce a parameterized decoder ψ that maps points $Z \in \mathbb{R}^N$ into probability distributions $P_\psi(\overset{\rightarrow M}{X} |Z)$ over future observations $\overset{\rightarrow M}{X}$:

$$P_\psi(\overset{\rightarrow M}{X} |Z) \text{ approximates } P(\overset{\rightarrow M}{X} |Z) \tag{15}$$

for each code $Z \in \mathbb{R}^N$. Crucially, $P_\psi(\overset{\rightarrow M}{X} |Z)$ will be easy to compute efficiently, unlike the exact decoding distribution $P(\overset{\rightarrow M}{X} |Z)$.

If we fix a stochastic process $(X_t)_{t\in\mathbb{Z}}$ and an encoder ϕ, then the following two bounds hold for *any* choice of the decoder ψ and the distribution q:

Proposition 1. *The loss incurred when predicting the future from Z via ψ upper-bounds the true conditional entropy of the future given Z, when predicting using the exact decoding distribution (13):*

$$-\mathbb{E}_X \mathbb{E}_{Z\sim\phi(X)} \left[\log P_\psi(\overset{\rightarrow M}{X} |Z)\right] \geq \mathrm{H}[\overset{\rightarrow M}{X} |Z]. \tag{16}$$

Furthermore, equality is attained if and only if $P_\psi(\overset{\rightarrow M}{X} |Z) = P(\overset{\rightarrow M}{X} |Z)$.

Proof. By Gibbs' inequality:

$$\begin{aligned} -\mathbb{E}_X \mathbb{E}_{Z\sim\phi(X)} \left[\log P_\psi(\overset{\rightarrow M}{X} |Z)\right] &\geq -\mathbb{E}_X \mathbb{E}_{Z\sim\phi(X)} \left[\log P(\overset{\rightarrow M}{X} |Z)\right] \\ &= \mathrm{H}[\overset{\rightarrow M}{X} |Z]. \end{aligned}$$

□

Proposition 2. *The KL Divergence between $P_\phi(Z| \overset{M\leftarrow}{X})$ and $q(Z)$, averaged over pasts $\overset{M\leftarrow}{X}$, upper-bounds the rate of Z:*

$$\begin{aligned} \mathbb{E}_{\overset{M\leftarrow}{X}} \left[\mathrm{D_{KL}}(P_\phi(Z| \overset{M\leftarrow}{X}) \parallel q(Z))\right] &= \mathbb{E}_{\overset{M\leftarrow}{X}} \mathbb{E}_{Z|\overset{M\leftarrow}{X}} \log \frac{P_\phi(Z| \overset{M\leftarrow}{X})}{q(Z)} \\ &\geq \mathbb{E}_{\overset{M\leftarrow}{X}} \mathbb{E}_{Z|\overset{M\leftarrow}{X}} \log \frac{P_\phi(Z| \overset{M\leftarrow}{X})}{P(Z)} \\ &= \mathrm{I}[\overset{M\leftarrow}{X}, Z]. \end{aligned} \tag{17}$$

Equality is attained if and only if $q(Z)$ is equal to the true marginal $P(Z) = \mathbb{E}_{\overset{M\leftarrow}{X}} P_\phi(Z| \overset{M\leftarrow}{X})$.

Proof. The two equalities follow from the definition of KL Divergence and Mutual Information. To show the inequality, we again use Gibbs' inequality:

$$-\mathbb{E}_{\overset{M\leftarrow}{X}} \mathbb{E}_{Z|\overset{M\leftarrow}{X}} \log q(Z) = -\mathbb{E}_Z \log q(Z) \geq -\mathbb{E}_Z \log P(Z) = -\mathbb{E}_{\overset{M\leftarrow}{X}} \mathbb{E}_{Z|\overset{M\leftarrow}{X}} \log P(Z).$$

Here, equality holds if and only if $q(Z) = P(Z)$, proving the second assertion. □

We now use the two propositions to rewrite the Predictive Rate–Distortion objective (18) in a way amenable to using function approximators, which is our main theoretical result, and the foundation of our proposed estimation method:

Corollary 1 (Main Result)**.** *The Predictive Rate–Distortion objective (18)*

$$\min_{Z:Z\perp \overset{\rightarrow M}{X} | \overset{M\leftarrow}{X}} \left[\mathrm{H}[\overset{\rightarrow M}{X} |Z] + \lambda\, \mathrm{I}[\overset{M\leftarrow}{X}, Z] \right] \tag{18}$$

is equivalent to

$$\min_{\phi,\psi,q} \left[\mathbb{E}_{\overset{M\leftarrow}{X}, \overset{\rightarrow M}{X}} \left[-\mathbb{E}_{Z\sim\phi(\overset{M\leftarrow}{X})} \left[\log P_\psi(\overset{\rightarrow M}{X} |Z) \right] + \lambda \cdot \mathrm{D}_{\mathrm{KL}} \left[(P_\phi(Z|\overset{M\leftarrow}{X}) \parallel q(Z)) \right] \right] \right], \tag{19}$$

where ϕ, ψ, q range over all triples of the appropriate types described above.

From a solution to (19), one obtains a solution to (18) by setting $Z \sim P_\phi(\cdot| \overset{M\leftarrow}{X})$. The rate of this code is given as follows:

$$\mathrm{I}[Z, \overset{M\leftarrow}{X}] = \mathbb{E}_{\overset{M\leftarrow}{X}} \left[\mathrm{D}_{\mathrm{KL}}(P_\phi(Z|\overset{M\leftarrow}{X}) \parallel q(Z)) \right] \tag{20}$$

and the prediction loss is given by

$$\mathrm{H}[\overset{\rightarrow M}{X} |Z] = -\mathbb{E}_{X_{\overset{M\leftarrow}{X}, \overset{\rightarrow M}{X}}} \mathbb{E}_{Z\sim P_\phi(\overset{M\leftarrow}{X})} \left[\log P_\psi(\overset{\rightarrow M}{X} |Z) \right]. \tag{21}$$

Proof. By the two propositions, the term inside the minimization in (19) is an upper bound on (18), and takes on that value if and only if $P_\phi(\cdot| \overset{M\leftarrow}{X})$ equals the distribution of the Z optimizing (18), and ψ, q are as described in those propositions. □

Note that the right-hand sides of (20) and (21) can both be estimated efficiently using Monte Carlo samples from $Z \sim P_\phi(\overset{M\leftarrow}{X})$.

If ϕ, ψ, q are not exact solutions to (19), the two propositions guarantee that we still have bounds on rate and prediction loss for the code Z generated by ϕ:

$$\mathrm{I}[Z, \overset{M\leftarrow}{X}] \leq \mathbb{E}_{\overset{M\leftarrow}{X}} \left[\mathrm{D}_{\mathrm{KL}}(P_\phi(Z|\overset{M\leftarrow}{X}) \parallel q(Z)) \right], \tag{22}$$

$$\mathrm{H}[\overset{\rightarrow M}{X} |Z] \leq -\mathbb{E}_{X_{\overset{M\leftarrow}{X}, \overset{\rightarrow M}{X}}} \mathbb{E}_{Z\sim P_\phi(\overset{M\leftarrow}{X})} \left[\log P_\psi(\overset{\rightarrow M}{X} |Z) \right]. \tag{23}$$

To carry out the optimization (19), we will restrict ϕ, ψ, q to a powerful family of parametric families of function approximators, within which we will optimize the objective with gradient descent. While the solutions may not be exact solutions to the nonparametric objective (19), they will still satisfy the bounds (22) and (23), and—if the family of approximators is sufficiently rich—can come close to turning these into the equalities (20) and (21).

4.2. Choosing Approximating Families

For our method of Neural Predictive Rate–Distortion (NPRD), we choose the approximating families for the encoder ϕ, the decoder ψ, and the distribution q to be certain types of neural networks that are known to provide strong and general models of sequences and distributions.

For ϕ and ψ, we use recurrent neural networks with Long Short Term Memory (LSTM) cells [32], widely used for modeling sequential data across different domains. We parameterize the distribution

$P_\phi(Z|\overset{M\leftarrow}{X})$ as a Gaussian whose mean and variance are computed from the past $\overset{M\leftarrow}{X}$: We use an LSTM network to compute a vector $h \in \mathbb{R}^k$ from the past observations $\overset{M\leftarrow}{X}$, and then compute

$$Z \sim \mathcal{N}(W_\mu h, (W_\sigma h)^2 I_{k\times k}), \tag{24}$$

where $W_\mu, W_\sigma \in \mathbb{R}^{k\times k}$ are parameters. While we found Gaussians sufficiently flexible for ϕ, more powerful encoders could be constructed using more flexible parametric families, such as normalizing flows [19,33].

For the decoder ψ, we use a second LSTM network to compute a sequence of vector representations $g_t = \psi(Z, X_{0\ldots t-1})$ ($g_t \in \mathbb{R}^k$) for $t = 0, \ldots M-1$. We compute predictions using the softmax rule

$$P_\psi(X_t = s_i | X_{1\ldots t-1}, Z) \propto \exp((W_o g_t)_i) \tag{25}$$

for each element s_i of the state space $S = \{s_1, ..., s_{|S|}\}$, and $W_o \in \mathbb{R}^{|S|\times k}$ is a parameter matrix to be optimized together with the parameters of the LSTM networks.

For q, we choose the family of Neural Autoregressive Flows [20]. This is a parametric family of distributions that allows efficient estimation of the probability density and its gradients. This method widely generalizes a family of prior methods [19,33,34], offering efficient estimation while surpassing prior methods in expressivity.

4.3. Parameter Estimation and Evaluation

We optimize the parameters of the neural networks expressing ϕ, ψ, q for the objective (19) using Backpropagation and Adam [35], a standard and widely used gradient descent-based method for optimizing neural networks. During optimization, we approximate the gradient by taking a single sample from Z (24) per sample trajectory $X_{-M}, \ldots, X_{M-1}$ and use the reparametrized gradient estimator introduced by Kingma and Welling [36]. This results in an unbiased estimator of the gradient of (19) w.r.t. the parameters of ϕ, ψ, q.

Following standard practice in machine learning, we split the data set of sample time series into three partitions (training set, validation set, and test set). We use the training set for optimizing the parameters as described above. After every pass through the training set, the objective (19) is evaluated on the validation set using a Monte Carlo estimate with one sample Z per trajectory; optimization terminates once the value on the validation set does not decrease any more.

After optimizing the parameters on a set of observed trajectories, we estimate rate and prediction loss on the test set. Given parameters for ϕ, ψ, q, we evaluate the PRD objective (19), rate (22), and the prediction loss (23) on the test set by taking, for each time series $X_{-M}...X_{M-1} = \overset{M\leftarrow}{X}\overset{\rightarrow M}{X}$ from the test set, a single sample $z \sim \mathcal{N}(\mu, \sigma^2)$ and computing Monte Carlo estimates for rate

$$\mathbb{E}_X \left[\mathrm{D}_{\mathrm{KL}}(P_\phi(Z|\overset{M\leftarrow}{X}) \parallel q(Z)) \right] \approx \frac{1}{N} \sum_{X_{-M...M}\in TestData} \log \frac{p_{\mathcal{N}(\mu,\sigma^2)}(z)}{q(z)}, \tag{26}$$

where $p_{\mathcal{N}(\mu,\sigma^2)}(z)$ is the Gaussian density with μ, σ^2 computed from $\overset{M\leftarrow}{X}$ as in (24), and prediction loss

$$-\mathbb{E}_{Z\sim\phi(\overset{M\leftarrow}{X})} \left[\log P_\psi(\overset{\rightarrow M}{X}|Z) \right] \approx -\frac{1}{N} \sum_{X_{-M...M}\in TestData} \log P_\psi(\overset{\rightarrow M}{X}|Z). \tag{27}$$

Thanks to (22) and (23), these estimates are guaranteed to be *upper bounds* on the true rate and prediction loss achieved by the code $Z \sim \mathcal{N}(\mu, \sigma^2)$, up to sampling error introduced into the Monte Carlo estimation by sampling z and the finite size of the test set.

It is important to note that this sampling error is different from the overfitting issue affecting OCF: Our Equations (26) and (27) provide *unbiased* estimators of upper bounds, whereas overfitting *biases* the values obtained by OCF. Given that Neural Predictive Rate–Distortion provably provide upper bounds on rate and prediction loss (up to sampling error), one can objectively compare the quality of different estimation methods: among methods that provide upper bounds, the one that provides the lowest such bound for a given problem is the one giving results closest to the true curve.

Estimating Predictiveness

Given the estimate for prediction loss, we estimate predictiveness $\mathrm{I}[Z, \overset{\rightarrow M}{X}]$ with the following method. We use the encoder network that computes the code vector h (24) to also estimate the marginal probability of the past observation sequence, $P_\eta(\overset{M\leftarrow}{X})$. P_η has support over sequences of length M. Similar to the decoder ψ, we use the vector representations $f_t \in \mathbb{R}^k$ computed by the LSTM encoder after processing $X_{-M\dots t}$ for $t = -M, \dots, -1$, and then compute predictions using the softmax rule

$$P_\eta(X_t = s_i | X_{1\dots t-1}, Z) \propto \exp((W_{o'} f_t)_i), \tag{28}$$

where $W_{o'} \in \mathbb{R}^{|S| \times k}$ is another parameter matrix.

Because we consider stationary processes, we have that the cross-entropy under P_η of $\overset{\rightarrow M}{X}$ is equal to the cross-entropy of $\overset{M\leftarrow}{X}$ under the same encoding distribution: $\mathbb{E}_{X_{\overset{M\leftarrow}{X}\overset{\rightarrow M}{X}}}\left[\log P_\eta(\overset{\rightarrow M}{X})\right] = -\mathbb{E}_{X_{\overset{M\leftarrow}{X}\overset{\rightarrow M}{X}}}\left[\log P_\eta(\overset{M\leftarrow}{X})\right]$. Using this observation, we estimate the predictiveness $\mathrm{I}[Z, \overset{\rightarrow M}{X}] = \mathrm{H}[\overset{\rightarrow M}{X}] - \mathrm{H}[\overset{\rightarrow M}{X} | Z]$ by the difference between the corresponding cross-entropies on the test set [37]:

$$-\mathbb{E}_{X_{\overset{M\leftarrow}{X}\overset{\rightarrow M}{X}}}\left[\log P_\eta(\overset{\rightarrow M}{X}) - \log P_\psi(\overset{\rightarrow M}{X} | Z)\right], \tag{29}$$

which we approximate using Monte Carlo sampling on the test set as in (26) and (27).

In order to optimize parameters for estimation of P_η, we add the cross-entropy term $-\mathbb{E}_{X_{\overset{M\leftarrow}{X}\overset{\rightarrow M}{X}}}\left[\log P_\eta(\overset{M\leftarrow}{X})\right]$ to the PRD objective (19) during optimization, so that the full training objective comes out to:

$$\min_{\phi,\psi,q,\eta}\left[\mathbb{E}_{\overset{M\leftarrow}{X}, \overset{\rightarrow M}{X}}\left[-\mathbb{E}_{Z\sim\phi(\overset{M\leftarrow}{X})}\left[\log P_\psi(\overset{\rightarrow M}{X} | Z)\right] + \lambda \cdot \mathrm{D}_{\mathrm{KL}}\left[(P_\phi(Z | \overset{M\leftarrow}{X}) \parallel q(Z))\right] - \log P_\eta(\overset{M\leftarrow}{X})\right]\right]. \tag{30}$$

Again, by Gibbs' inequality and Propositions 1 and 2, this is minimized when P_η represents the true distribution over length-M blocks $P(\overset{M\leftarrow}{X})$, $P_\phi(Z | \overset{M\leftarrow}{X})$ describes an optimal code for the given λ, q is its marginal distribution, and $P_\psi(\overset{\rightarrow M}{X} | Z)$ is the Bayes-optimal decoder. For approximate solutions to this augmented objective, the inequalities (22) and (23) will also remain true due to Propositions 1 and 2.

4.4. Related Work

In (19), we derived a variational formulation of Predictive Rate–Distortion. This is formally related to a variational formulation of the Information Bottleneck that was introduced by [38], who applied it to neural-network based image recognition. Unlike our approach, they used a fixed diagonal Gaussian instead of a flexible parametrized distribution for q. Some recent work has applied similar approaches to the modeling of sequences, employing models corresponding to the objective (19) with $\lambda = 1$ [39–42].

In the neural networks literature, the most commonly used method using variational bounds similar to Equation (19) is the Variational Autoencoder [36,43], which corresponds to the setting where $\lambda = 1$ and the predicted output is equal to the observed input. The β-VAE [44], a variant of the Variational Autoencoder, uses $\lambda > 1$ (whereas the Predictive Rate–Distortion objective (7) uses $\lambda \in (0, 1)$), and has been linked to the Information Bottleneck by [45].

5. Experiments

We now test the ability of our new method NPRD to estimate rate–distortion curves. Before we apply NPRD to obtain the first estimates of Predictive Rate–Distortion for natural language in Section 6, we validate the method on processes whose trade-off curves are analytically known, and compare with OCF.

5.1. Implementation Details

OCF

As discussed in Section 3, OCF is affected by overfitting, and will systematically overestimate the predictiveness achieved at a given rate [6]. To address this problem, we follow the evaluation method used for NPRD, evaluating rate and predictiveness on held-out test data. We partition the available time series data into a training and test set. We use the training set to create the encoder $P(Z|\overset{M\leftarrow}{X})$ using the Blahut–Arimoto algorithm as described by Still et al. [6]. We then use the held-out test set to estimate rate and prediction loss. This method not only enables fair comparison between OCF and NPRD, but also provides a more realistic evaluation, by focusing on the performance of the code Z when deployed on new data samples. For rate, we use the same variational bound that we use for NPRD, stated in Proposition 2:

$$\frac{1}{N} \sum_{\overset{M\leftarrow}{X} \in \text{Test Data}} \mathrm{D}_{\mathrm{KL}}(P(Z|\overset{M\leftarrow}{X}) \parallel s(Z)), \tag{31}$$

where $P(Z|\overset{M\leftarrow}{X})$ is the encoder created by the Blahut–Arimoto algorithm, and $s(Z)$ is the marginal distribution of Z on the training set. N is the number of sample time series in the test data. In the limit of enough training data, when $s(Z)$ matches the actual population marginal of Z, (31) is an unbiased estimate of the rate. We estimate the prediction loss on the future observations as the empirical cross-entropy, i.e., the variational bound stated in Proposition 1:

$$\frac{1}{N} \sum_{\overset{M\leftarrow}{X}, \overset{\rightarrow M}{X} \in \text{Test Data}} \mathbb{E}_{Z\sim P(\cdot|\overset{M\leftarrow}{X})} \log P(\overset{\rightarrow M}{X}|Z), \tag{32}$$

where $P(\overset{\rightarrow M}{X}|Z)$ is the decoder obtained from the Blahut–Arimoto algorithm on the training set. Thanks to Propositions 1 and 2, these quantities provide upper bounds, up to sampling error introduced by finiteness of the held-out data. Again, sampling error does not bias the results in either direction, unlike overfitting, which introduces a systematic bias.

Held-out test data may contain sequences that did not occur in the training data. Therefore, we add a pseudo-sequence ω and add pseudo-observations $(\omega, \overset{\rightarrow M}{X})$, $(\overset{M\leftarrow}{X}, \omega)$, (ω, ω) for all observed sequences $\overset{M\leftarrow}{X}\overset{\rightarrow M}{X}$ to the matrix of observed counts that serves as the input to the Blahut–Arimoto algorithm. These pseudo-observations were assigned pseudo-counts γ in this matrix of observed counts; we found that a wide range of values ranging from 0.0001 to 1.0 yielded essentially the same results. When evaluating the codebook on held-out data, previously unseen sequences were mapped to ω.

Neural Predictive Rate–Distortion

For all experiments, we used $M = 15$. Neural networks have hyperparameters, such as the number of units and the step size used in optimization, which affect the quality of approximation depending on properties of the dataset and the function being approximated. Given that NPRD provably provides upper bounds on the PRD objective (18), one can in principle identify the best hyperparameters for a given process by choosing the combination that leads to the lowest estimated upper bounds. As a computationally more efficient method, we defined a range of plausible hyperparameters based both on experience reported in the literature, and considerations of computational efficiency. These parameters are discussed in Appendix A. We then randomly sampled, for each of the processes that we experimented on, combinations of λ and these hyperparameters to run NPRD on. We implemented the model using PyTorch [46].

5.2. Analytically Tractable Problems

We first test NPRD on two processes where the Predictive Rate–Distortion trade-off is analytically tractable. The Even Process [2] is the process of 0/1 IID coin flips, conditioned on all blocks of consecutive ones having even length. Its complexity and excess entropy are both ≈ 0.63 nats. It has infinite Markov order, and Marzen and Crutchfield [2] find that OCF (at $M = 5$) performs poorly. The true Predictive Rate–Distortion curve was computed in [2] using the analytically known ϵ-machine. The Random Insertion Process [2] consists of sequences of uniform coin flips $X_t \in \{0, 1\}$, subject to the constraint that, if X_{t-2} was a 0, then X_t has to be a 1.

We applied NPRD to these processes by training on 3M random trajectories of length 30, and using 3000 additional trajectories for validation and test data. For each process, we ran NPRD 1000 times for random choices of $\lambda \in [0, 1]$. Due to computational constraints, when running OCF, we limited sample size to 3000 trajectories for estimation and as held-out data. Following Marzen and Crutchfield [2], we ran OCF for $M = 1, ..., 5$.

The resulting estimates are shown in Figure 1, together with the analytical rate–distortion curves computed by Marzen and Crutchfield [2]. Individual runs of NPRD show variation (red dots), but most runs lead to results close to the analytical curve (gray line), and strongly surpass the curves computed by OCF at $M = 5$. Bounding the trade-off curve using the sets of runs of NPRD results in a close fit (red line) to the analytical trade-off curves.

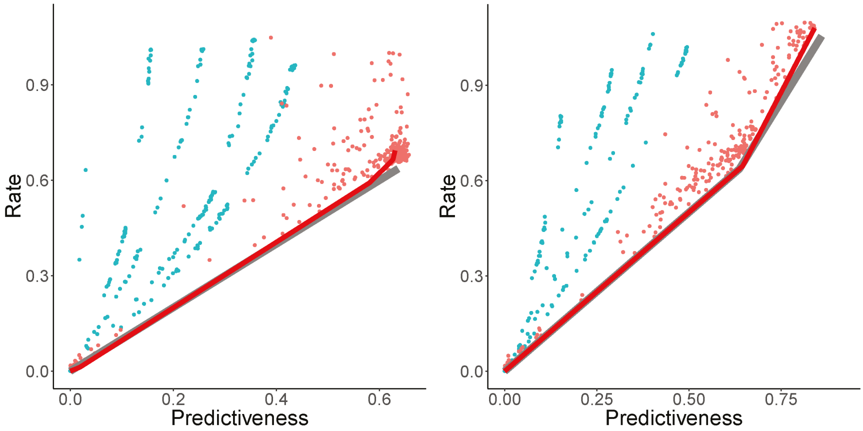


Figure 1. Rate–Distortion for the Even Process (**left**) and the Random Insertion Process (**right**). Gray lines: analytical curves; Red dots: multiple runs of NPRD; red line: trade-off curve computed from NPRD runs; blue: OCF for $M \leq 5$.

Recovering Causal States

Does NPRD lead to interpretable codes Z? To answer this, we further investigated the NPRD approximation to the Random Insertion Process (RIP), obtained in the previous paragraph.

The ϵ-machine was computed by Marzen and Crutchfield [2] and is given in Figure 2 (left). The process has three causal states: State A represents those pasts where the future starts with $1^{2k}0$ ($k = 0, 1, 2, \ldots$)—these are the pasts ending in either 001 or 10111^m ($m = 0, 1, 2, \ldots$). State B represents those pasts ending in 10—the future has to start with 01 or 11. State C represents those pasts ending in either 00 or 01—the future has to start with $1^{2k+1}0$ ($k = 0, 1, 2, \ldots$).

The analytical solution to the Predictive Rate–Distortion problem was computed by Marzen and Crutchfield [2]. At $\lambda > 0.5$, the optimal solution collapses A and B into a single codeword, while all three states are mapped to separate codewords for $\lambda \leq 0.5$.

Does NPRD recover this picture? We applied PCA to samples from Z computed at two different values of λ, $\lambda = 0.25$ and $\lambda = 0.6$. The first two principal components of Z are shown in Figure 3. Samples are colored by the causal states corresponding to the pasts of the trajectories that were encoded into the respective points by NPRD. On the left, obtained at $\lambda = 0.6$, the states A and B are collapsed, as expected. On the right, obtained at $\lambda = 0.25$, the three causal states are reflected as distinct modes of Z. Note that, at finite M, a fraction of pasts is ambiguous between the green and blue causal states; these are colored in black and NPRD maps them into a region between the modes corresponding to these states.

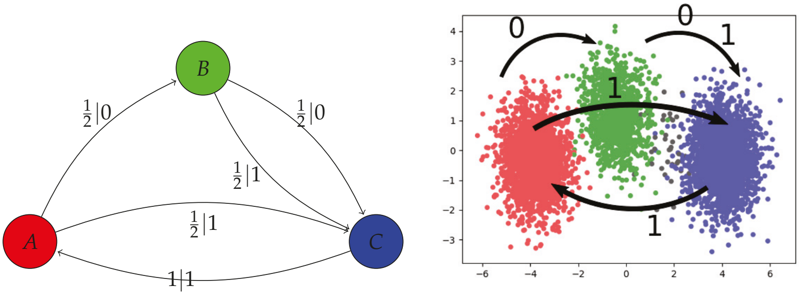


Figure 2. Recovering the ϵ-machine from NPRD. **Left**: The ϵ-machine of the Random Insertion Process, as described by [2]. **Right**: After computing a code Z from a past $x_{-15\ldots-1}$, we recorded which of the three clusters the code moves to when appending the symbol 0 or 1 to the past sequence. The resulting transitions mirror those in the ϵ-machine.

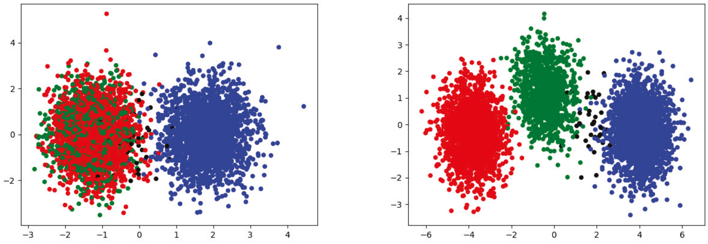


Figure 3. Applying Principal Component Analysis to 5000 sampled codes Z for the Random Insertion Process, at $\lambda = 0.6$ (**left**) and $\lambda = 0.25$ (**right**). We show the first two principal components. Samples are colored according to the states in the ϵ-machine. There is a small number of samples from sequences that, at $M = 15$, cannot be uniquely attributed to any of the states (ambiguous between A and C); these are indicated in black.

In Figure 2 (right), we record, for each of the three modes, to which cluster the distribution of the code Z shifts when a symbol is appended. We restrict to those strings that have nonzero probability for

RIP (no code will ever be needed for other strings). For comparison, we show the ϵ-machine computed by Marzen and Crutchfield [2]. Comparing the two diagrams shows that NPRD effectively recovers the ϵ-machine: the three causal states are represented by the three different modes of Z, and the effect of appending a symbol also mirrors the state transitions of the ϵ-machine.

A Process with Many Causal States

We have seen that NPRD recovers the correct trade-off, and the structure of the causal states, in processes with a small number of causal states. How does it behave when the number of causal states is very large? In particular, is it capable of extrapolating to causal states that were never seen during training?

We consider the following process, which we will call COPY3: $X_{-15}, ..., X_{-1}$, are independent uniform draws from $\{1, 2, 3\}$, and $X_1 = X_{-1}, ..., X_{15} = X_{-15}$. This process deviates a bit from our usual setup since we defined it only for $t \in \{-15, ..., 15\}$, but it is well-suited to investigating this question: the number of causal states is $3^{15} \approx 14$ million. With exactly the same setup as for the EVEN and RIP processes, NPRD achieved essentially zero distortion on unseen data, even though the number of training samples (3 Million) was far lower than the number of distinct causal states. However, we found that, in this setup, NPRD overestimated the rate. Increasing the number of training samples from 3M to 6M, NPRD recovered codebooks that achieved both almost zero distortion and almost optimal rate, on fresh samples (Figure 4). Even then, the number of distinct causal states is more than twice the number of training samples. These results demonstrate that, by using function approximation, NPRD is capable of extrapolating to unseen causal states, encoding and decoding appropriate codes on the fly.

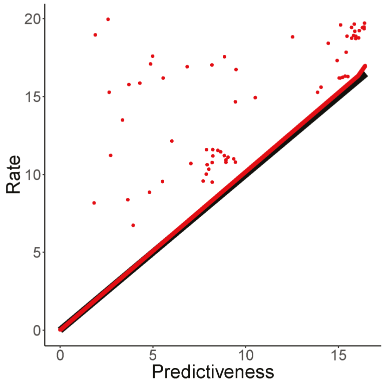


Figure 4. Rate–Distortion for the COPY3 process. We show NPRD samples, and the resulting upper bound in red. The gray line represents the anaytical curve.

Note that one could easily design an optimal decoder and encoder for COPY3 by hand—the point of this experiment is to demonstrate that NPRD is capable of inducing such a codebook purely from data, in a general-purpose, off-the-shelf manner. This contrasts with OCF: without optimizations specific to the task at hand, a direct application of OCF would require brute-force storing of all 14 million distinct pasts and futures.

6. Estimating Predictive Rate–Distortion for Natural Language

We consider the problem of estimating rate–distortion for natural language. Natural language has been a testing ground for information-theoretic ideas since Shannon's work. Much interest has been devoted to estimating the entropy rate of natural language [10,47–49]. Indeed, the information density of language has been linked to human processing effort and to language structure. The word-by-word information content has been shown to impact human processing effort as measured both by per-word reading times [50–52] and by brain signals measured through EEG [53,54]. Consequently, prediction is a fundamental component across theories of human language processing [54]. Relatedly, the Uniform Information Density and Constant Entropy Rate hypotheses [55–57] state that languages order information in sentences and discourse so that the entropy rate stays approximately constant.

The relevance of prediction to human language processing makes the *difficulty* of prediction another interesting aspect of language complexity: Predictive Rate–Distortion describes how much memory of the past humans need to maintain to predict future words accurately. Beyond the entropy rate, it thus forms another important aspect of linguistic complexity.

Understanding the complexity of prediction in language holds promise for enabling a deeper understanding of the nature of language as a stochastic process, and to human language processing. Long-range correlations in text have been a subject of study for a while [58–63]. Recently, Dębowski [64] has studied the excess entropy of language across long-range discourses, aiming to better understand the nature of the stochastic processes underlying language. Koplenig et al. [65] shows a link between traditional linguistic notions of grammatical structure and the information contained in word forms and word order. The idea that predicting future words creates a need to represent the past well also forms a cornerstone of theories of how humans process sentences [66,67].

We study prediction in the range of the words in individual sentences. As in the previous experiments, we limit our computations to sequences of length 30, already improving over OCF by an order of magnitude. One motivation is that, when directly estimating PRD, computational cost has to increase with the length of sequences considered, making the consideration of sequences of hundreds or thousands of words computationally infeasible. Another motivation for this is that we are ultimately interested in Predictive Rate–Distortion as a model of memory in human processing of grammatical structure, formalizing psycholinguistic models of how humans process individual sentences [66,67], and linking to studies of the relation between information theory and grammar [65].

6.1. Part-of-Speech-Level Language Modeling

We first consider the problem of predicting English on the level of Part-of-Speech (POS) tags, using the Universal POS tagset [68]. This is a simplified setting where the vocabulary is small (20 word types), and one can hope that OCF will produce reasonable results. We use the English portions of the Universal Dependencies Project [69] tagged with Universal POS Tags [68], consisting of approximately 586 K words. We used the training portions to estimate NPRD and OCF, and the validation portions to estimate the rate–distortion curve. We used NPRD to generate 350 codebooks for values of λ sampled from $[0, 0.4]$. We were only able to run OCF for $M \leq 3$, as the number of sequences exceeds 10^4 already at $M = 4$.

The PRD curve is shown in Figure 5 (left). In the setting of low rate and high distortion, NPRD and OCF (blue, $M = 1, 2, 3$) show essentially identical results. This holds true until $\mathrm{I}[Z, \overrightarrow{X}] \approx 0.7$, at which point the bounds provided by OCF deteriorate, showing the effects of overfitting. NPRD continues to provide estimates at greater rates.

Figure 5 (center) shows rate as a function of $\log \frac{1}{\lambda}$. Recall that λ is the trade-off-parameter from the objective function (7). In Figure 5 (right), we show rate and the mutual information with the future, as a function of $\log \frac{1}{\lambda}$. As $\lambda \to 0$, NPRD (red, $M = 15$) continues to discover structure, while OCF (blue, plotted for $M = 1, 2, 3$) exhausts its capacity.

Note that NPRD reports rates of 15 nats and more when modeling with very low distortion. A discrete codebook would need over 3 million distinct codewords for a code of such a rate, exceeding the size of the training corpus (about 500 K words), replicating what we found for the Copy3 process: Neural encoders and decoders can use the geometric structure of the code space to encode generalizations across different dimensions, supporting a very large effective number of distinct possible codes. Unlike discrete codebooks, the geometric structure makes it possible for neural encoders to construct appropriate codes 'on the fly' on new input.

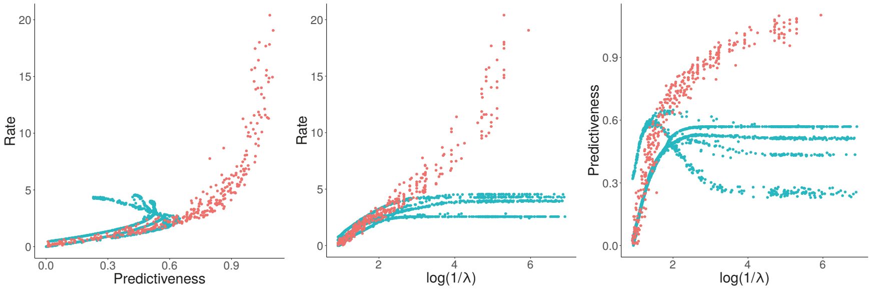


Figure 5. Left: Rate-Predictiveness for English POS modeling. Center and right: Rate (**Center**) and Predictiveness (**Right**) on English POS Modeling, as a function of $-\log\lambda$. As $\lambda \rightarrow 0$, NPRD (red, $M = 15$) continues to discover structure, while OCF (blue, plotted for $M = 1, 2, 3$) exhausts its capacity.

6.2. Discussion

Let us now consider the curves in Figure 5 in more detail. Fitting parametric curves to the empirical PRD curves in Figure 5, we find a surprising result that the statistical complexity of English sentences at the POS level appears to be unbounded.

The rate-predictiveness curve (left) shows that, at low rates, predictiveness is approximately proportional to the rate. At greater degrees of predictiveness, the rate grows faster and faster, whereas predictiveness seems to asymptote to ≈ 1.1 nats. The asymptote of predictiveness can be identified with the mutual information between past and future observations, $E_0 := \mathrm{I}[\overset{M\leftarrow}{X}, \overset{\rightarrow M}{X}]$, which is a lower bound on the excess entropy. The rate should asymptote to the statistical complexity. Judging by the curve, natural language—at the time scale we are measuring in this experiment—has a statistical complexity much higher than its excess entropy: at the highest rate measured by NPRD in our experiment, rate is about 20 nats, whereas predictiveness is about 1.1 nats. If these values are correct, then—due to the convexity of the rate-predictivity curve—statistical complexity exceeds the excess entropy by a factor of at least $\frac{20}{1.1}$. Note that this picture agrees qualitatively with the OCF results, which suggest a lower-bound on the ratio of at least $\frac{2.5}{0.6} > 5$.

Now, turning to the other plots in Figure 5, we observe that rate increases at least linearly with $\log\frac{1}{\lambda}$, whereas predictiveness again asymptotes. This is in qualitative agreement with the picture gained from the rate-predictiveness curve.

Let us consider this more quantitatively. Based on Figure 5 (center), we make the ansatz that the map from $\log\frac{1}{\lambda}$ to the rate $R := \mathrm{I}[\overset{M\leftarrow}{X}, Z]$ is superlinear:

$$R = \alpha \left(\log \frac{1}{\lambda}\right)^{\beta}, \tag{33}$$

with $\alpha > 0, \beta > 1$. We fitted $R \approx \left(\log\frac{1}{\lambda}\right)^{1.7}$ ($\alpha = 1, \beta = 1.7$). Equivalently,

$$\frac{1}{\lambda} = \exp\left(\frac{1}{\alpha^{1/\beta}} R^{1/\beta}\right). \tag{34}$$

From this, we can derive expressions for rate $R := \mathrm{I}[\overset{M\leftarrow}{X}, Z]$ and predictiveness $P := \mathrm{I}[Z, \overset{\rightarrow M}{X}]$ as follows. For the solution of Predictive Rate–Distortion (10), we have

$$\frac{\partial P}{\partial \theta} - \lambda \frac{\partial R}{\partial \theta} = 0, \tag{35}$$

where θ is the codebook defining the encoding distribution $P(Z|\overset{M\leftarrow}{X})$, and thus

$$\lambda = \frac{\partial P}{\partial R}. \tag{36}$$

Our ansatz therefore leads to the equation

$$\frac{\partial P}{\partial R} = \exp\left(-\frac{1}{\alpha^{1/\beta}} R^{1/\beta}\right). \tag{37}$$

Qualitatively, this says that predictiveness P asymptotes to a finite value, whereas rate R—which should asymptote to the statistical complexity—is unbounded.

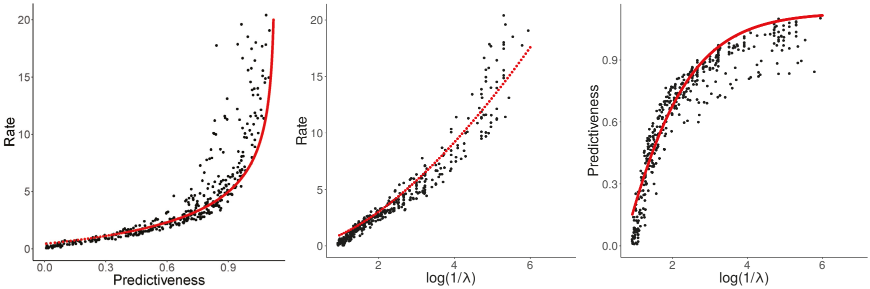


Figure 6. Interpolated values for POS-level prediction of English (compare Figure 5).

Equation (37) has the solution

$$P = C - \alpha\beta \cdot \Gamma\left(\beta, (R/\alpha)^{1/\beta}\right), \tag{38}$$

where Γ is the incomplete Gamma function. Since $\lim_{R\to\infty} P = C$, the constant C has to equal the maximally possible predictiveness $E_0 := \mathrm{I}[\overset{M\leftarrow}{X}, \overset{\rightarrow M}{X}]$.

Given the values fitted above ($\alpha = 1$, $\beta = 1.7$), we found that $E_0 = 1.13$ yielded a good fit. Using (33), this can be extended without further parameters to the third curve in Figure 5. Resulting fits are shown in Figure 6.

Note that there are other possible ways of fitting these curves; we have described a simple one that requires only two parameters $\alpha > 0$, $\beta > 1$, in addition to a guess for the maximal predictiveness E_0. In any case, the results show that natural language shows an approximately linear growth of predictiveness with a rate at small rates, and exploding rates at diminishing returns in predictiveness later.

6.3. Word-Level Language Modeling

We applied NPRD to the problem of predicting English on the level of part-of-speech tags in Section 6.1. We found that the resulting curves were described well by Equation (37). We now consider the more realistic problem of prediction at the level of words, using data from multiple languages. This problem is much closer to the task faced by a human in the process of comprehending text, having to encode prior observations so as to minimize prediction loss on the upcoming words. We will examine whether Equation (37) describes the resulting trade-off in this more realistic setting, and whether it holds across languages.

For the setup, we followed a standard setup for recurrent neural language modeling. The hyperparameters are shown in Table A1. Following standard practice in neural language modeling, we restrict the observation space to the most frequent 10^4 words; other words are replaced by their part-of-speech tag. We do this for simplicity and to stay close to standard practice in natural language processing; NPRD could deal with unbounded state spaces through a range of more sophisticated techniques such as subword modeling and character-level prediction [70,71].

We used data from five diverse languages. For English, we turn to the Wall Street Journal portion of the Penn Treebank [72], a standard benchmark for language modeling, containing about 1.2 million tokens. For Arabic, we pooled all relevant portions of the Universal Dependencies treebanks [73–75]. We obtained 1 million tokens. We applied the same method to construct a Russian corpus [76], obtaining 1.2 million tokens. For Chinese, we use the Chinese Dependency Treebank [77], containing 0.9 million tokens. For Japanese, we use the first 2 million words from a large processed corpus of Japanese business text [78]. For all these languages, we used the predefined splits into training, validation, and test sets.

For each language, we sampled about 120 values of $\log \frac{1}{\lambda}$ uniformly from $[-6, 0]$ and applied NPRD to these. The resulting curves are shown in Figures 7 and 8, together with fitted curves resulting from Equation (37). As can be seen, the curves are qualitatively very similar across languages to what we observed in Figure 6: In all languages, rate initially scales linearly with predictiveness, but diverges as the predictiveness approaches its supremum E_0. As a function of $\log \frac{1}{\lambda}$, rate grows at a slightly superlinear speed, confirming our ansatz (33).

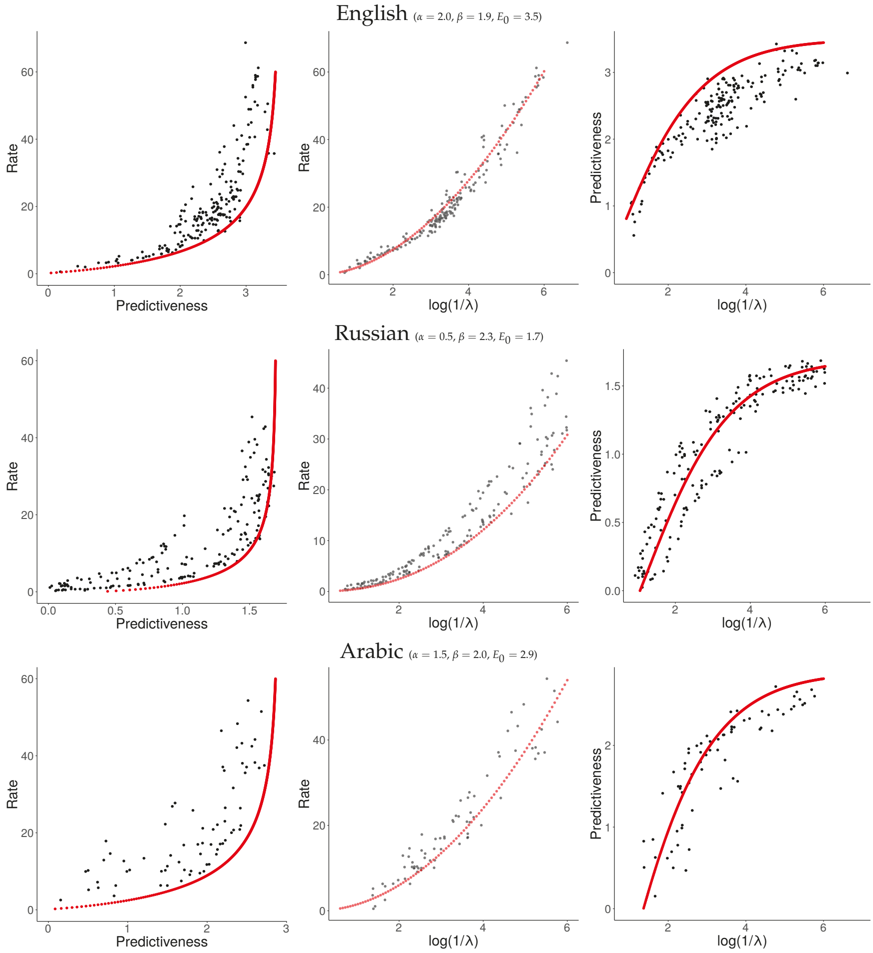


Figure 7. Word-level results.

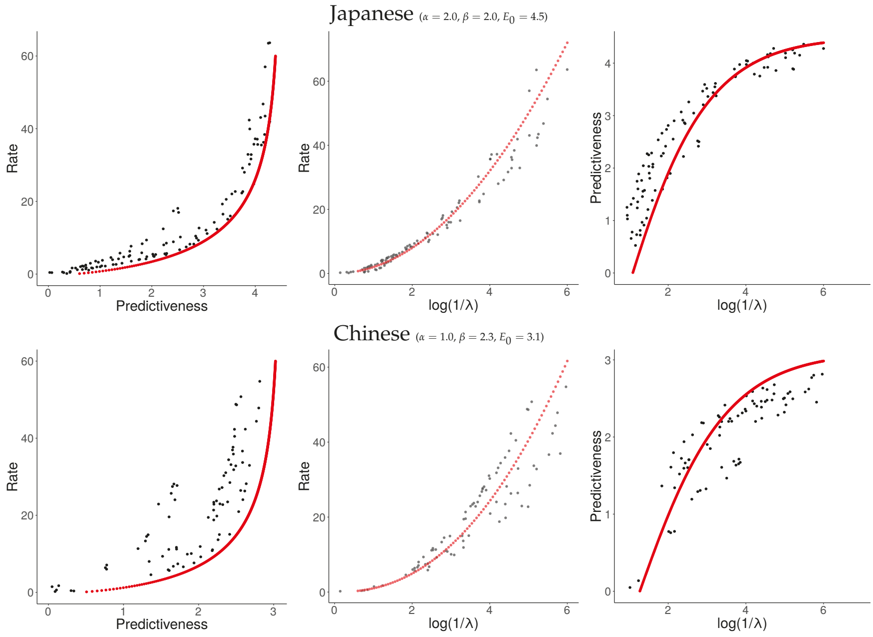


Figure 8. Word-level results (cont.).

These results confirm our results from Section 6.1. At the time scale of individual sentences, Predictive Rate–Distortion of natural language appears to quantitatively follow Equation (37). NPRD reports rates up to $\approx$ 60 nats, more than ten times the largest values of predictiveness. On the other hand, the growth of rate with predictiveness is relatively gentle in the low-rate regime. We conclude that predicting words in natural language can be approximated with small memory capacity, but more accurate prediction requires very fine-grained memory representations.

6.4. General Discussion

Our analysis of PRD curves for natural language suggests that human language is characterized by very high and perhaps infinite statistical complexity, beyond its excess entropy. In a similar vein, Dębowski [64] has argued that the excess entropy of connected texts in natural language is infinite (in contrast, our result is for isolated sentences). If the statistical complexity of natural language is indeed infinite, then statistical complexity is not sufficiently fine-grained as a complexity metric for characterizing natural language.

We suggest that the PRD curve may form a more natural complexity metric for highly complex processes such as language. Among those processes with infinite statistical complexity, some will have a gentle PRD curve—meaning that they can be well-approximated at low rates—while others will have a steep curve, meaning they cannot be well-approximated at low rates. We conjecture that, although natural language may have infinite statistical complexity, it has a gentler PRD curve than other processes with this property, meaning that achieving a reasonable approximation of the predictive distribution does not require inordinate memory resources.

7. Conclusions

We introduce Neural Predictive Rate–Distortion (NPRD), a method for estimating Predictive Rate–Distortion when only sample trajectories are given. Unlike OCF, the most general prior method, NPRD scales to long sequences and large state spaces. On analytically tractable processes, we show that it closely fits the analytical rate–distortion curve and recovers the causal states of the process. On part-of-speech-level modeling of natural language, it agrees with OCF in the setting of low rate and short sequences; outside these settings, OCF fails due to combinatorial explosion and overfitting, while NPRD continues to provide estimates. Finally, we use NPRD to provide the first estimates of Predictive Rate–Distortion for modeling natural language in five different languages, finding qualitatively very similar curves in all languages.

All code for reproducing the results in this work is available at https://github.com/m-hahn/predictive-rate--distortion.

Author Contributions: Methodology, M.H.; Writing—original draft, M.H. and R.F.

Funding: This research received no external funding.

Acknowledgments: We thank Dan Jurafsky for helpful discussion, and the anonymous reviewers for helpful comments.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NPRD	Neural Predictive Rate–Distortion
OCF	Optimal Causal Filtering
POS	parts of speech
PRD	Predictive Rate–Distortion
LSTM	Long Short Term Memory
VAE	Variational Autoencoder

Appendix A. Hyperparameters

All hyperparameter choices are shown in Table A1. We defined separate hyperparameter ranges for the three Sections 5.2, 6.1, and 6.3. Guided by the fact that the analytically known processes in Section 5.2 are arguably less complex than natural language, we allowed larger and more powerful models for modeling of natural language (Sections 6.1 and 6.3), in particular word-level modeling (Section 6.3).

In Table A1, hyperparameters are organized into four groups. The first group of parameters is the dimensions of the input embedding and the recurrent LSTM states. The second group of parameters is regularization parameters. We apply dropout [79] to the input and output layers. The third group of parameters is related to the optimization procedure [35]. Neural Autoregressive Flows, used to approximate the marginal q, also have a set of hyperparameters [20]: the length of the flow (1, 2, ...), the type (DSF/Deep Sigmoid Flow or DDSF/Deep Dense Sigmoid Flow), the dimension of the flow (an integer) and the number of layers in the flow (1,2, ...). Larger dimensions, more layers, and longer flows lead to more expressive models; however, they are computationally more expensive.

Training used Early Stopping using the development set, as described in Section 4.3. Models were trained on a TITAN Xp graphics card. On Even Process and Random Insertion Process, NPRD took a median of 10 min to process 3M training samples. OCF took less than one minute at $M \leq 5$; however, it does not scale to larger values of M. On English word-level modeling, training took a median number of nine epochs (max 467 epochs) and five minutes (max 126 min).

Table A1. NPRD Hyperparameters. See Appendix A for description of the parameters.

	Section 5.2	**Section 6.1**	**Section 6.3**
Embedding Dimension	50, 100	50, 100, 200, 300	150
LSTM Dimension	32	64, 128, 256, 512	256, 512
Dropout rate	0.0, 0.1, 0.2	0.0, 0.1, 0.2	0.1, 0.4
Input Dropout	0	0.0, 0.1, 0.2	0.2
Adam Learning Rate	$\{1,5\} \cdot 10^{-4}$, $\{1,2,4\} \cdot 10^{-3}$	$\{1,5\} \cdot 10^{-4}$, $\{1,2,4\} \cdot 10^{-3}$	0.00005, 0.0001, 0.0005, 0.001
Batch Size	16, 32, 64	16, 32 64	16, 32, 64
Flow Length	1, 2	1, 2, 3, 4, 5	1, 2, 3, 4, 5
Flow Type	DSF, DDSF	DSF, DDSF	DSF, DDSF
Flow Dimension	32, 64, 128, 512	512	512
Flow Layers	2	2	2

Appendix B. Alternative Modeling Choices

In this section, we investigate the trade-offs involved in alternative modeling choices. It may be possible to use simpler function-approximators for ϕ, ψ, and q, or smaller context windows sizes M, without harming accuracy too much.

First, we investigated the performance of a simple fixed approximation to the marginal q. We considered a diagonal unit-variance Gaussian, as is common in the literature on Variational Autoencoders [36]. We show results in Figure A1. Even with this fixed approximation to q, NPRD continues to provide estimates not far away from the analytical curve. However, comparison with the results obtained from full NPRD (Figure 1) shows that a flexible parameterized approximation still provides considerably better fit.

Second, we investigated whether the use of recurrent neural networks is necessary. As recurrent models such as LSTMs process input sequentially, they cannot be fully parallelized, posing the question of whether they can be replaced by models that can be parallelized. Specifically, we considered Quasi-Recurrent Neural Networks (QRNNs), which combine convolutions with a weak form of recurrence, and which have shown strong results on language modeling benchmarks [80]. We replaced the LSTM encoder and decoder with QRNNs and fit NPRD to the Even Process and the Random Insertion Process. We found that, when using QRNNs, NPRD consistently fitted codes with zero rate for the Even Process and the Random Insertion Process, indicating that the QRNN was failing to extract useful information from the past of these processes. We also found that the cross-entropies of the estimated marginal distribution $P_\eta(\overset{M\leftarrow}{X})$ were considerably worse than when using LSTMs or simple RNNs. We conjecture that, due to the convolutional nature of QRNNs, they cannot model such processes effectively in principle: QRNNs extract representations by pooling embeddings of words or n-grams occurring in the past. When modeling, e.g., the Even Process, the occurrence of specific n-grams is not informative about whether the length of the last block is even or odd; this requires some information about the positions in which these n-grams occur, which is not available to the QRNN, but which is generally available in a more general autoregressive model such as an LSTM.

Third, we varied M, comparing $M = 5, 10$ to the results obtained with $M = 15$. Results are shown in Figure A2. At $M = 5$, NPRD provides estimates similar to OCF (Figure 1). At $M = 10$, the estimates are close to the analytical curve; nonetheless, $M = 15$ yields clearly more accurate results.

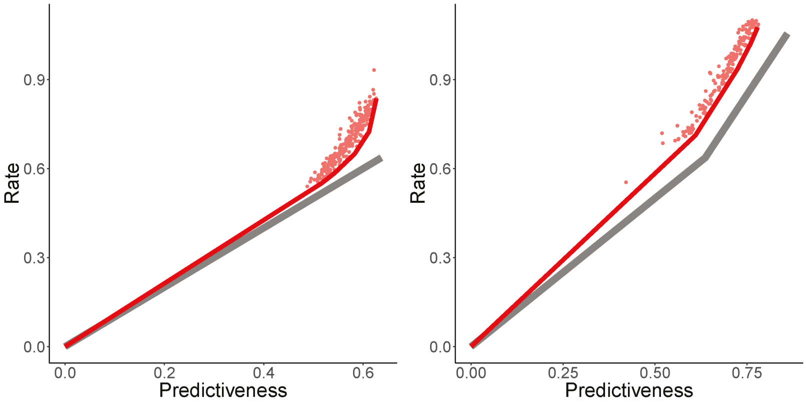


Figure A1. Rate–Distortion for the Even Process **(left)** and the Random Insertion Process (**right**), estimated using a simple diagonal unit Gaussian approximation for q. Gray lines: analytical curves; red dots: multiple runs of NPRD (>200 samples); red line: trade-off curve computed from NPRD runs. Compare Figure 1 for results from full NPRD.

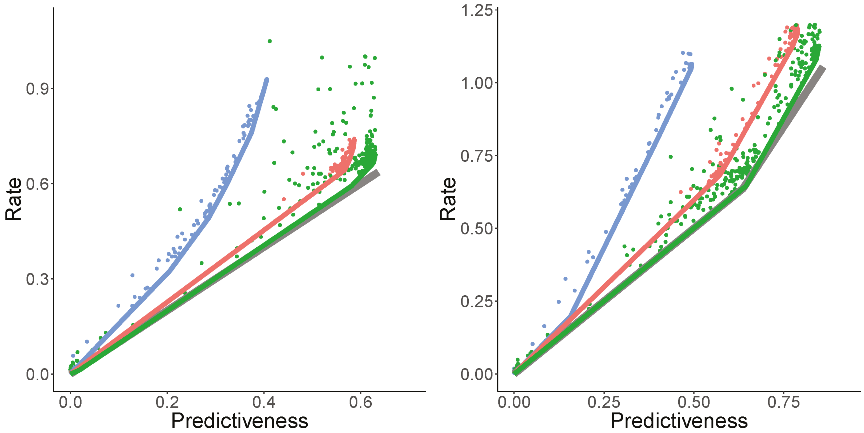
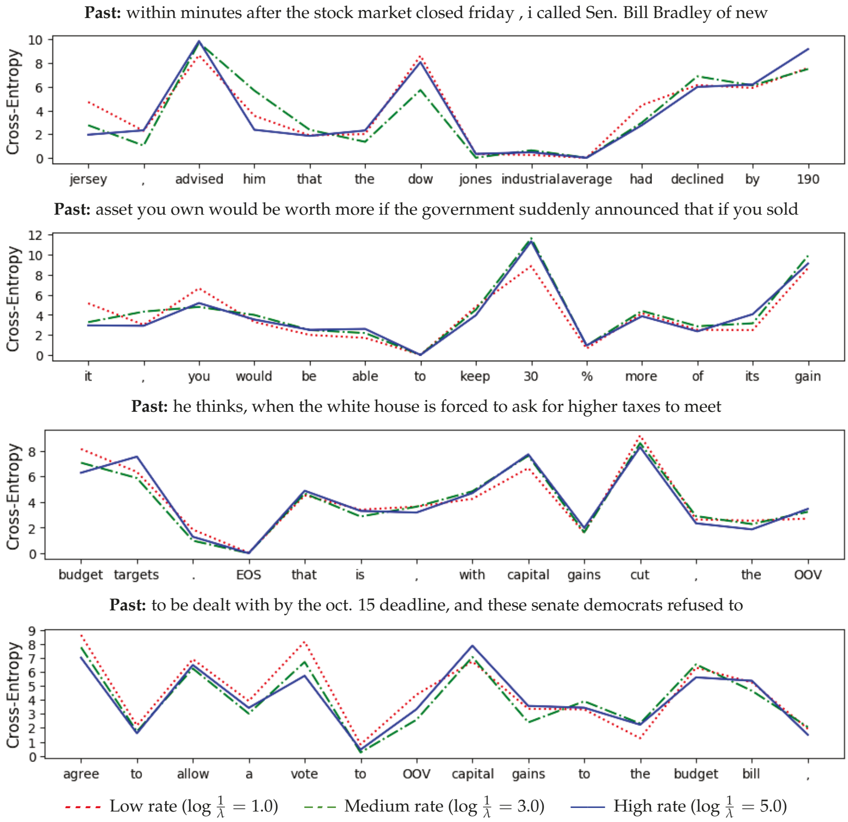


Figure A2. Rate–Distortion for the Even Process (**left**) and the Random Insertion Process (**right**), varying $M = 5$ (blue), 10 (red), 15 (green); gray lines: analytical curves; red dots: multiple runs of NPRD; red line: trade-off curve computed from NPRD runs. Compare Figure 1 for results from full NPRD.

Appendix C. Sample Runs on English Text

In Figure A3, we provide four sample outputs from English word-level modeling with three values of $\log \frac{1}{\lambda}$ (1.0, 3.0, 5.0), corresponding to low (1.0), medium (3.0), and high (5.0) rates (compare Figure 7). We obtained sample sequences by selecting the first 32 sequences $\overset{M \leftarrow}{X} \overset{\rightarrow M}{X}$ (at $M = 15$) from the Penn Treebank validation set, and selected four examples where the variation in cross-entropy values at X_0 was largest between the three models.

Across these samples, models generated at $\log \frac{1}{\lambda} = 5$ show lower cross-entropy on the first future observation X_0, as these codes have higher rates. For instance, in the first sample, the cross-entropy on the first word *Jersey* is lowest for the code with the higher rate; indeed, this word is presumably strongly predicted by the preceding sequence *...Sen. Billd Bradley of New*. Codes with higher rates are better at encoding such predictive information from the past.



Past: within minutes after the stock market closed friday , i called Sen. Bill Bradley of new

Past: asset you own would be worth more if the government suddenly announced that if you sold

Past: he thinks, when the white house is forced to ask for higher taxes to meet

Past: to be dealt with by the oct. 15 deadline, and these senate democrats refused to

Figure A3. Four example outputs from English word-level modeling, with low rate (log $\frac{1}{\lambda} = 1$; red, dotted), medium rate (log $\frac{1}{\lambda} = 3$; green, dashed), high rate (log $\frac{1}{\lambda} = 5$; blue, solid). For each sample, we provide the prior context $\overset{M\leftarrow}{X}$ (**top**), and the per-word cross-entropies (in nats) on the future words $\overset{\rightarrow M}{X}$ (**bottom**).

References

1. Still, S. Information Bottleneck Approach to Predictive Inference. *Entropy* **2014**, *16*, 968–989. [CrossRef]
2. Marzen, S.E.; Crutchfield, J.P. Predictive Rate-Distortion for Infinite-Order Markov Processes. *J. Stat. Phys.* **2016**, *163*, 1312–1338. [CrossRef]
3. Creutzig, F.; Globerson, A.; Tishby, N. Past-future information bottleneck in dynamical systems. *Phys. Rev. E* **2009**, *79*. [CrossRef]
4. Amir, N.; Tiomkin, S.; Tishby, N. Past-future Information Bottleneck for linear feedback systems. In Proceedings of the 54th IEEE Conference on Decision and Control (CDC), Osaka, Japan, 15–18 December 2015; pp. 5737–5742.
5. Genewein, T.; Leibfried, F.; Grau-Moya, J.; Braun, D.A. Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Front. Robot. AI* **2015**, *2*, 27. [CrossRef]
6. Still, S.; Crutchfield, J.P.; Ellison, C.J. Optimal causal inference: Estimating stored information and approximating causal architecture. *Chaos Interdiscip. J. Nonlinear Sci.* **2010**, *20*, 037111. [CrossRef]
7. Józefowicz, R.; Vinyals, O.; Schuster, M.; Shazeer, N.; Wu, Y. Exploring the Limits of Language Modeling. *arXiv* **2016**, arXiv:1602.02410.
8. Merity, S.; Keskar, N.S.; Socher, R. An analysis of neural language modeling at multiple scales. *arXiv* **2018**, arXiv:1803.08240.

9. Dai, Z.; Yang, Z.; Yang, Y.; Cohen, W.W.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. *arXiv* **2019**, arXiv:1901.02860.
10. Takahashi, S.; Tanaka-Ishii, K. Cross Entropy of Neural Language Models at Infinity—A New Bound of the Entropy Rate. *Entropy* **2018**, *20*, 839. [CrossRef]
11. Ogunmolu, O.; Gu, X.; Jiang, S.; Gans, N. Nonlinear systems identification using deep dynamic neural networks. *arXiv* **2016**, arXiv:1610.01439.
12. Laptev, N.; Yosinski, J.; Li, L.E.; Smyl, S. Time-series extreme event forecasting with neural networks at uber. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 11 August 2017; pp. 1–5.
13. Meyer, P.; Noblet, V.; Mazzara, C.; Lallement, A. Survey on deep learning for radiotherapy. *Comput. Biol. Med.* **2018**, *98*, 126–146. [CrossRef] [PubMed]
14. Hundman, K.; Constantinou, V.; Laporte, C.; Colwell, I.; Soderstrom, T. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 387–395.
15. White, G.; Palade, A.; Clarke, S. Forecasting qos attributes using lstm networks. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–8.
16. Woo, J.; Park, J.; Yu, C.; Kim, N. Dynamic model identification of unmanned surface vehicles using deep learning network. *Appl. Ocean Res.* **2018**, *78*, 123–133. [CrossRef]
17. Sirignano, J.; Cont, R. Universal features of price formation in financial markets: perspectives from Deep Learning. *arXiv* **2018**, arXiv:1803.06917.
18. Mohajerin, N.; Waslander, S.L. Multistep Prediction of Dynamic Systems With Recurrent Neural Networks. *IEEE Transa. Neural Netw. Learn. Syst.* **2019**. [CrossRef] [PubMed]
19. Rezende, D.J.; Mohamed, S. Variational inference with normalizing flows. In Proceedings of the 32nd International Conference on International Conference on Machine Learning, Lille, France, 6–11 July 2015; Volume 37, pp. 1530–1538.
20. Huang, C.W.; Krueger, D.; Lacoste, A.; Courville, A. Neural Autoregressive Flows. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 2083–2092.
21. Tishby, N.; Pereira, F.C.; Bialek, W. The Information Bottleneck Method. In Proceedings of the Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 22–24 September 1999; pp. 368–377.
22. Harremoës, P.; Tishby, N. The information bottleneck revisited or how to choose a good distortion measure. In Proceedings of the IEEE International Symposium on Information Theory, Nice, France, 24–29 June 2007; pp. 566–570.
23. Feldman, D.P.; Crutchfield, J.P. Synchronizing to Periodicity: The Transient Information and Synchronization Time of Periodic Sequences. *Adv. Complex Syst.* **2004**, *7*, 329–355. [CrossRef]
24. Crutchfield, J.P.; Young, K. Inferring statistical complexity. *Phys. Rev. Lett.* **1989**, *63*, 105–108. [CrossRef] [PubMed]
25. Grassberger, P. Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.* **1986**, *25*, 907–938. [CrossRef]
26. Löhr, W. Properties of the Statistical Complexity Functional and Partially Deterministic HMMs. *Entropy* **2009**, *11*, 385–401. [CrossRef]
27. Clarke, R.W.; Freeman, M.P.; Watkins, N.W. Application of computational mechanics to the analysis of natural data: An example in geomagnetism. *Phys. Rev. E* **2003**, *67*, 016203. [CrossRef]
28. Singh, S.P.; Littman, M.L.; Jong, N.K.; Pardoe, D.; Stone, P. Learning predictive state representations. In Proceedings of the 20th International Conference on Machine Learning (ICML-03), Washington, DC, USA, 21–24 August 2003; pp. 712–719.
29. Singh, S.; James, M.R.; Rudary, M.R. Predictive state representations: A new theory for modeling dynamical systems. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*; AUAI Press: Arlington, VA, USA, 2004; pp. 512–519.
30. Jaeger, H. *Discrete-Time, Discrete-Valued Observable Operator Models: A Tutorial*; GMD-Forschungszentrum Informationstechnik: Darmstadt, Germany, 1998.
31. Rubin, J.; Shamir, O.; Tishby, N. Trading value and information in MDPs. In *Decision Making with Imperfect Decision Makers*; Springer: Berlin, Germany, 2012; pp. 57–74.

32. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
33. Kingma, D.P.; Salimans, T.; Jozefowicz, R.; Chen, X.; Sutskever, I.; Welling, M. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2016; pp. 4743–4751.
34. Papamakarios, G.; Pavlakou, T.; Murray, I. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2017; pp. 2338–2347.
35. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
36. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.
37. McAllester, D.; Statos, K. Formal Limitations on the Measurement of Mutual Information. *arXiv* **2018**, arXiv:1811.04251.
38. Alemi, A.A.; Fischer, I.; Dillon, J.V.; Murphy, K. Deep Variational Information Bottleneck. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
39. Grathwohl, W.; Wilson, A. Disentangling space and time in video with hierarchical variational auto-encoders. *arXiv* **2016**, arXiv:1612.04440.
40. Walker, J.; Doersch, C.; Gupta, A.; Hebert, M. An uncertain future: Forecasting from static images using variational autoencoders. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin, Germany, 2016; pp. 835–851.
41. Fraccaro, M.; Kamronn, S.; Paquet, U.; Winther, O. A disentangled recognition and nonlinear dynamics model for unsupervised learning. In *Advances in Neural Information Processing Systems*; 2017; pp. 3601–3610.
42. Hernández, C.X.; Wayment-Steele, H.K.; Sultan, M.M.; Husic, B.E.; Pande, V.S. Variational encoding of complex dynamics. *Phys. Rev. E* **2018**, *97*, 062412. [CrossRef] [PubMed]
43. Bowman, S.R.; Vilnis, L.; Vinyals, O.; Dai, A.M.; Jozefowicz, R.; Bengio, S. Generating Sentences from a Continuous Space. In Proceedings of the CoNLL, Berlin, Germany, 11–12 August 2016.
44. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. β-VAE: Learning basic visual concepts with a constrained variational framework. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
45. Burgess, C.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; Lerchner, A. Understanding disentangling in β-VAE. *arXiv* **2018**, arXiv:1804.03599.
46. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic Differentiation in PyTorch 2017. Available online: https://openreview.net/forum?id=BJJsrmfCZ (accessed on 28 June 2019).
47. Shannon, C.E. Prediction and entropy of printed English. *Bell Syst. Tech. J.* **1951**, *30*, 50–64. [CrossRef]
48. Takahira, R.; Tanaka-Ishii, K.; Dębowski, Ł. Entropy rate estimates for natural language—A new extrapolation of compressed large-scale corpora. *Entropy* **2016**, *18*, 364. [CrossRef]
49. Bentz, C.; Alikaniotis, D.; Cysouw, M.; Ferrer-i Cancho, R. The entropy of words—Learnability and expressivity across more than 1000 languages. *Entropy* **2017**, *19*, 275. [CrossRef]
50. Hale, J. A Probabilistic Earley Parser as a Psycholinguistic Model. In Proceedings of the NAACL, Pittsburgh, PA, USA, 1–7 June 2001; Volume 2, pp. 159–166.
51. Levy, R. Expectation-based syntactic comprehension. *Cognition* **2008**, *106*, 1126–1177. [CrossRef]
52. Smith, N.J.; Levy, R. The effect of word predictability on reading time is logarithmic. *Cognition* **2013**, *128*, 302–319. [CrossRef]
53. Frank, S.L.; Otten, L.J.; Galli, G.; Vigliocco, G. Word surprisal predicts N400 amplitude during reading. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 4–9 August 2013; Volume 2, pp. 878–883.
54. Kuperberg, G.R.; Jaeger, T.F. What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* **2016**, *31*, 32–59. [CrossRef]
55. Fenk, A.; Fenk, G. Konstanz im Kurzzeitgedächtnis—Konstanz im sprachlichen Informationsfluß. *Z. Exp. Angew. Psychol.* **1980**, *27*, 400–414.
56. Genzel, D.; Charniak, E. Entropy rate constancy in text. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002.

57. Jaeger, T.F.; Levy, R.P. Speakers optimize information density through syntactic reduction. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 3–6 December 2007; pp. 849–856.
58. Schenkel, A.; Zhang, J.; Zhang, Y.C. Long range correlation in human writings. *Fractals* **1993**, *1*, 47–57. [CrossRef]
59. Ebeling, W.; Pöschel, T. Entropy and long-range correlations in literary English. *EPL (Europhys. Lett.)* **1994**, *26*, 241. [CrossRef]
60. Ebeling, W.; Neiman, A. Long-range correlations between letters and sentences in texts. *Phys. A Stat. Mech. Appl.* **1995**, *215*, 233–241. [CrossRef]
61. Altmann, E.G.; Cristadoro, G.; Degli Esposti, M. On the origin of long-range correlations in texts. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 11582–11587. [CrossRef]
62. Yang, T.; Gu, C.; Yang, H. Long-range correlations in sentence series from A Story of the Stone. *PLoS ONE* **2016**, *11*, e0162423. [CrossRef]
63. Chen, H.; Liu, H. Quantifying evolution of short and long-range correlations in Chinese narrative texts across 2000 years. *Complexity* **2018**, *2018*, 9362468. [CrossRef]
64. Dębowski, Ł. Is natural language a perigraphic process? The theorem about facts and words revisited. *Entropy* **2018**, *20*, 85. [CrossRef]
65. Koplenig, A.; Meyer, P.; Wolfer, S.; Mueller-Spitzer, C. The statistical trade-off between word order and word structure–Large-scale evidence for the principle of least effort. *PLoS ONE* **2017**, *12*, e0173614. [CrossRef]
66. Gibson, E. Linguistic complexity: locality of syntactic dependencies. *Cognition* **1998**, *68*, 1–76. [CrossRef]
67. Futrell, R.; Levy, R. Noisy-context surprisal as a human sentence processing cost model. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3–7 April 2017; pp. 688–698.
68. Petrov, S.; Das, D.; McDonald, R.T. A Universal Part-of-Speech Tagset. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, 23–25 May 2012; pp. 2089–2096.
69. Nivre, J.; Agic, Z.; Ahrenberg, L.; Antonsen, L.; Aranzabe, M.J.; Asahara, M.; Ateyah, L.; Attia, M.; Atutxa, A.; Augustinus, L.; et al. Universal Dependencies 2.1 2017. Available online: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2515 (accessed on 28 June 2019).
70. Kim, Y.; Jernite, Y.; Sontag, D.; Rush, A.M. Character-aware neural language models. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
71. Luong, M.T.; Manning, C.D. Achieving open vocabulary neural machine translation with hybrid word-character models. *arXiv* **2016**, arXiv:1604.00788.
72. Marcus, M.P.; Marcinkiewicz, M.A.; Santorini, B. Building a large annotated corpus of English: The Penn Treebank. *Comput. Linguist.* **1993**, *19*, 313–330.
73. Nivre, J.; de Marneffe, M.C.; Ginter, F.; Goldberg, Y.; Hajic, J.; Manning, C.D.; McDonald, R.T.; Petrov, S.; Pyysalo, S.; Silveira, N.; et al. Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, 23–28 May 2016.
74. Maamouri, M.; Bies, A.; Buckwalter, T.; Mekki, W. The penn arabic treebank: Building a large-scale annotated arabic corpus. In Proceedings of the NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt, 27–29 March 2004; Volume 27, pp. 466–467.
75. Hajic, J.; Smrz, O.; Zemánek, P.; Šnaidauf, J.; Beška, E. Prague Arabic dependency treebank: Development in data and tools. In Proceedings of the NEMLAR Internaional Conference on Arabic Language Resources and Tools, Cairo, Egypt, 22–23 September 2004; pp. 110–117.
76. Dyachenko, P.B.; Iomdin, L.L.; Lazurskiy, A.V.; Mityushin, L.G.; Podlesskaya, O.Y.; Sizov, V.G.; Frolova, T.I.; Tsinman, L.L. Sovremennoe sostoyanie gluboko annotirovannogo korpusa tekstov russkogo yazyka (SinTagRus). *Trudy Instituta Russkogo Yazyka im. VV Vinogradova* **2015**, *10*, 272–300.
77. Che, W.; Li, Z.; Liu, T. *Chinese Dependency Treebank 1.0 LDC2012T05*; Web Download; Linguistic Data Consortium: Philadelphia, PA, USA, 2012.
78. Graff, D.; Wu, Z. *Japanese bUsiness News Text*; LDC95T8; Linguistic Data Consortium: Philadelphia, PA, USA, 1995.

79. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
80. Bradbury, J.; Merity, S.; Xiong, C.; Socher, R. Quasi-recurrent neural networks. In Proceedings of the ICLR 2017, Toulon, France, 24–26 April 2017.

entropy

MDPI

Article

Studying Lexical Dynamics and Language Change via Generalized Entropies: The Problem of Sample Size

Alexander Koplenig *, Sascha Wolfer and Carolin Müller-Spitzer

Department of Lexical Studies, Institute for the German language (IDS), 68161 Mannheim, Germany; wolfer@ids-mannheim.de (S.W.); mueller-spitzer@ids-mannheim.de (C.M.-S.)

* Correspondence: koplenig@ids-mannheim.de; Tel.: +49-621-1581-426

Received: 9 April 2019; Accepted: 30 April 2019; Published: 3 May 2019

Abstract: Recently, it was demonstrated that generalized entropies of order α offer novel and important opportunities to quantify the similarity of symbol sequences where α is a free parameter. Varying this parameter makes it possible to magnify differences between different texts at specific scales of the corresponding word frequency spectrum. For the analysis of the statistical properties of natural languages, this is especially interesting, because textual data are characterized by Zipf's law, i.e., there are very few word types that occur very often (e.g., function words expressing grammatical relationships) and many word types with a very low frequency (e.g., content words carrying most of the meaning of a sentence). Here, this approach is systematically and empirically studied by analyzing the lexical dynamics of the German weekly news magazine *Der Spiegel* (consisting of approximately 365,000 articles and 237,000,000 words that were published between 1947 and 2017). We show that, analogous to most other measures in quantitative linguistics, similarity measures based on generalized entropies depend heavily on the sample size (i.e., text length). We argue that this makes it difficult to quantify lexical dynamics and language change and show that standard sampling approaches do not solve this problem. We discuss the consequences of the results for the statistical analysis of languages.

Keywords: generalized entropy; generalized divergence; Jensen–Shannon divergence; sample size; text length; Zipf's law

1. Introduction

At a very basic level, the quantitative study of natural languages is about counting words: if a word occurs very often in one text but not in a second one, then we conclude that this difference might have some kind of significance for classifying both texts [1]. If a word occurs very often after another word, then we conclude that this might have some kind of significance in speech and language processing [2]. In both examples, we can use the gained knowledge to make informed predictions "with accuracy better than chance" [3], thus leading us to information theory quite naturally. If we consider each word type $i = 1, 2, \ldots, K$ as one distinct symbol, then we can count how often each word type appears in a document or text t and call the resulting word token frequency f_i. We can then represent t as a distribution of word frequencies. In order to quantify the amount of information contained in t, we can calculate the Gibbs–Shannon entropy of this distribution as [4]:

$$H(p) = -\sum_{i=1}^{K} p_i * log_2(p_i) \tag{1}$$

where $p_i = \frac{f_i}{N}$ is the maximum likelihood estimator of the probability of i in t for a database of $N = \sum_{i=1}^{K} f_i$ tokens. In [5], word entropies are estimated for more than 1000 languages. The results are then interpreted in light of information-theoretic models of communication, in which it is argued that word entropy constitutes a basic property of natural languages. $H(p)$ can be interpreted as the average number of guesses required to correctly predict the type of word token that is randomly sampled from the entire text base (more precisely, [4], Section 5.7) show that the expected number of guesses EG satisfies $H(p) \leq EG < H(p) + 1$). In the present paper, we analyze the lexical dynamics of the German weekly news magazine *Der Spiegel* (consisting of N = 236,743,042 word tokens, K = 4,009,318 different word types, and 365,514 articles that were published between 1947 and 2017; details on the database and preprocessing are presented Section 2). If the only knowledge we possess about the database were K, the number of different word types, then we would need on average $H_{\max} = \log_2(K) = \log_2(4{,}009{,}318) \approx 21.93$ guesses to correctly predict the word type, calculating H for our database based on Equation (1) using the corresponding probabilities for each i yields 12.28. The difference between $H_{\max}$ and $H(p)$ is defined as information in [3]. Thus, knowledge of the non-uniform word frequency distribution gives us approximately 9.65 bits of information, or put differently, we save on average almost 10 guesses to correctly predict the word type.

To quantify the (dis)similarity between two different texts or databases, word entropies can be used to calculate the so-called Jensen–Shannon divergence [6]:

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

where p and q are the (relative) word frequencies of the two texts and $p + q$ is calculated by concatenating both texts. From a Bayesian point of view, $D(p,q)$ can be interpreted as the expected amount of gained information that comes from sampling one word token from the concatenation of both texts regarding the question which of the two texts the word token belongs to [7]. If the two texts are identical, $D(p,q) = 0$, because sampling a word token does not provide any information regarding to which text the token belongs. If, on the other side, the two texts do not have a single word type in common, then sampling one word token is enough to determine from which text the token comes, and correspondingly, $D(p,q) = 1$. The Jensen–Shannon divergence has already been applied in the context of measuring stylistic influences in the evolution of literature [8], cultural and institutional changes [9,10], the dynamics of lexical evolution [11,12], or to quantify changing corpus compositions [13].

Perhaps the most intriguing aspect of word frequency distributions is the fact that they can be described remarkably well by a simple relationship that is known as Zipf's law [14]: if one assigns rank $r = 1$ to the most frequent word (type), rank $r = 2$ to the second most frequent word, and so on, then the frequency of a word and its rank r is related as follows:

$$p(r) \propto r^{-\gamma} \tag{3}$$

where the exponent γ is a parameter that has to be determined empirically. An estimation of γ by maximum likelihood (as described in [15]) for our database yields 1.10. However, when analyzing word frequency distributions, the main obstacle is that all quantities basically vary systematically with the sample size, i.e., the number of word tokens in the database [16,17]. To visualize this, we randomly arranged the order of all articles of our database. This step was repeated 10 times in order to create 10 different versions of our database. For each version, we estimate H and γ after every $n = 2^k$ consecutive tokens, where $k = 6, 7, \ldots, log_2(N) \;=\; 28$. Figure 1 shows a Simpson's Paradox [18] for the resulting data: an apparent strong positive relationship between H and γ is observed across all datapoints (Spearman $\rho = 0.99$). However, when the sample size is kept constant, this relationship completely changes: if the correlation between H and γ is calculated for each k, the results indicate a strong negative relationship (ρ ranges between −0.98 and −0.64 with a median of −0.92). The reason for this apparent contradiction is the fact that both H and γ monotonically increase with the sample size. When studying word frequency distributions quantitatively, it is essential to take this dependence on the sample size into account [16].

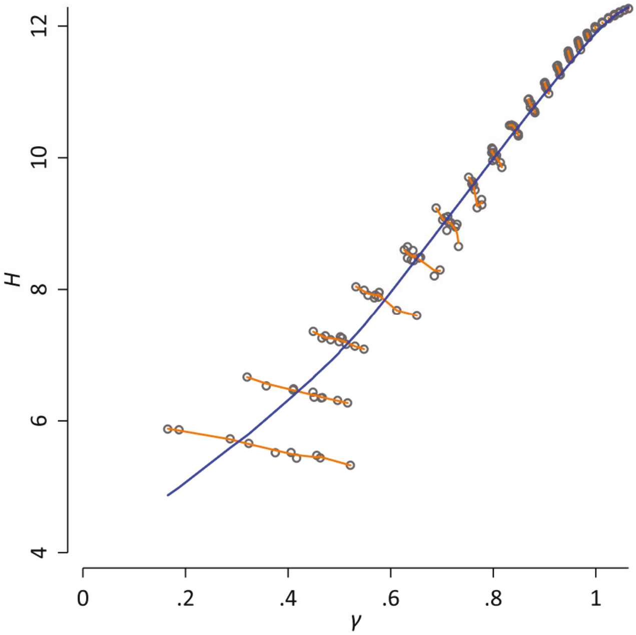


Figure 1. A Simpson's Paradox for word frequency distributions. Here, the word entropy H and the exponent of the Zipf distribution γ are estimated after every $n = 2^k$ consecutive tokens, where $k = 6, 7, \ldots, log_2(N)$ for 10 different random re-arrangements of the database; each dot corresponds to one observed value. The blue line represents a locally weighted regression of H on γ (with a bandwidth of 0.8). It indicates a strong positive relationship between H and γ (Spearman $\rho = 0.99$). However, when the sample size is held constant, this relationship completely changes, as indicated by the orange lines that correspond to separate locally weighted regressions of H on γ for each k. Here, the results indicate a strong negative relationship between H and γ (ρ ranges between −0.98 and −0.64 with a median of −0.92). The reason for this apparent contradiction is the fact that both H and γ monotonically increase with the sample size.

Another important aspect of word distributions is the fact that word frequencies vary by a magnitude of many orders, as visualized in Figure 2. On the one hand, Figure 2a shows that there are very few word types that occur very often. For example, the 100 most frequent word types account for more than 40% of all word occurrences. Typically, many of those word types are function words [16] expressing grammatical relationships, such as adpositions or conjunctions. On the other hand, Figure 2b shows that there are a great deal of word types with a very low frequency of occurrence. For example, more than 60% of all word types only occur once, and less than 3% of all word types have a frequency of occurrence of more than 100 in our database. Many of those low frequency words are content words that carry the meaning of a sentence, e.g., nouns, (lexical) verbs, and adjectives. In addition to the sample size dependence outlined above, it is important to take this broad range of frequencies into account when quantitatively studying word frequency distributions [19].

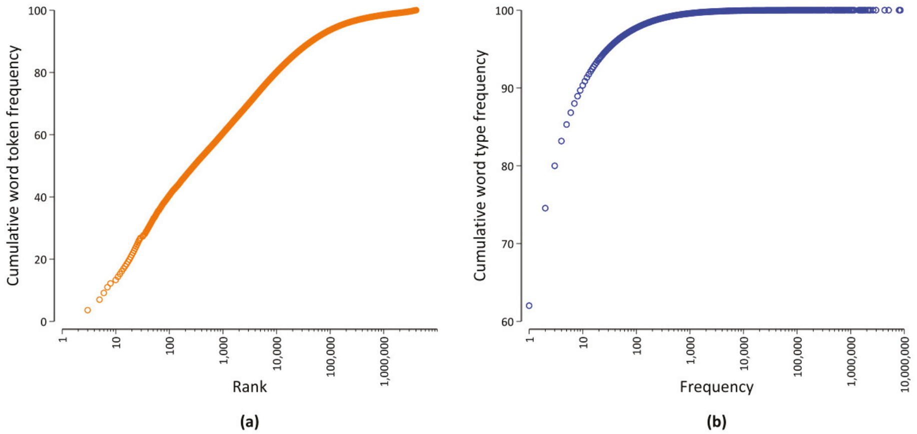


Figure 2. Visualization of the word frequency distribution of our database. Cumulative distribution (in %) as a function of (**a**) the rank and (**b**) the word frequency.

In this context, it was recently demonstrated that generalized entropies of order α, also called Havrda–Charvat–Lindhard–Nielsen–Aczél–Daróczy–Tsallis entropies [20], offer novel and interesting opportunities to quantify the similarity of symbol sequences [21,22]. It can be written as:

$$H_\alpha(p) \;=\; \frac{1}{\alpha - 1}\left(1 - \sum_{i\,=\,1}^{K} p_i^\alpha\right) \tag{4}$$

where $\alpha$is a free parameter. For $\alpha = 1$, the standard Gibbs–Shannon entropy is recovered. Correspondingly, a generalization of the standard Jensen–Shannon divergence (Equation (2)) can be obtained by replacing H (Equation (1)) with H_α (Equation (4)) and thus leading to a spectrum of divergence measures D_α, parametrized by α [22]. For the analysis of the statistical properties of natural languages, this parameter is highly interesting, because, as demonstrated by [21,22], varying the α-parameter allows us to magnify differences between different texts at specific scales of the corresponding word frequency spectrum. If α is increased (decreased), then the weight of the most frequent words is increased (decreased). As pointed out by an anonymous reviewer, a similar idea was already reported in the work of Tanaka-Ishii and Aihara [23], who studied a different formulation of generalized entropy, the so-called Rényi entropy of order α [24]. Because we are especially interested in using generalized entropies to quantify the (dis)similarity between two different texts or databases, following [21,22], we chose to focus on the generalization of Havrda–Charvat–Lindhard–Nielsen–Aczél–Daróczy–Tsallis instead of the formulation of Rényi, because a divergence measure based on the latter can become negative for $\alpha > 1$ [25], while it can be shown that the corresponding divergence measure based on the former formulation is strictly non-negative [20,22]. In addition, $D_\alpha(p,q)$ is the square of a metric for $\alpha \in (0, 2]$, i.e., (i) $D_\alpha(p,q) \geq 0$, (ii) $D_\alpha(p,q) = 0 \iff p = q$, (iii) $D_\alpha(p,q) = D_\alpha(q,p)$, and (iv) $\sqrt{D_\alpha}$ obeys the triangular inequality [7,20,22].

In addition, [21] also estimated the size of the database that is needed to obtain reliable estimates of generalized divergences. For instance, [21] showed that only the 100 most frequent words contribute to H_α and D_α for $\alpha = 2.00$, and all other words are practically irrelevant. This number quickly grows with α. For example, database sizes of $N \approx 10^8$ are needed for a robust estimation of the standard Jensen–Shannon divergence (Equation (2)), i.e., for $\alpha = 1.00$. This connection makes the approach of [21,22] particularly interesting in relation to the systematic influence of the sample size demonstrated above (cf. Figure 1).

In this study, the approach is systematically and empirically studied by analyzing the lexical dynamics of the *Der Spiegel* periodical. The remainder of the paper is structured as follows: In the next section, details on the database and preprocessing are given (Section 2). In Sections 3.1 and 3.2, the dependence of both H_α and D_α on the sample size is tested for different α-parameters. This section is followed by a case study, in which we demonstrate that the influence of sample size makes it difficult to quantify lexical dynamics and language change and also show that standard sampling approaches do not solve this problem (Section 3.3). This paper ends with some concluding remarks regarding the consequences of the results for the statistical analysis of languages (Section 4).

2. Materials and Methods

In the present study, we used all 365,514 articles that were published in the German weekly news magazine *Der Spiegel* between January 1947, when the magazine was first published, and December 2017. To read-in and tokenize the texts, we used the *Treetagger* with a German parameter file [26]. All characters were converted to lowercase. Punctuation and cardinal numbers (both treated as separate words by the Treetagger) were removed. However, from a linguistic point of view, changes in the usage frequencies of punctuation marks and cardinal numbers are also interesting. For instance, a frequency increase of the full stop could be indicative of decreases in syntactic complexity [15]. In Appendix A, we therefore present and discuss additional results in which punctuation and cardinal numbers were not removed from the data.

In total, our database consists of N = 236,743,042 word tokens and K = 4,009,318 different word types.

Motivated by the studies of [21,22], we chose the following six α values to study the empirical behavior of generalized entropies and generalized divergences: 0.25, 0.75, 1.00, 1.50, and 2.00. To highlight that varying α makes it possible to magnify differences between different texts at specific scales of the corresponding word frequency spectrum, we take advantage of the fact that H_α can be written as a sum over different words, where each individual word type i contributes

$$\begin{matrix} \frac{p_i^\alpha - \frac{1}{K}}{\alpha - 1}, \; for \; \alpha \neq 1.00 \\ -p_i * log_2(p_i), \; for \; \alpha \; = \; 1.00 \end{matrix} . \tag{5}$$

In Table 1, we divided the word types into different groups according to their token frequency (column 1). Each group consists of $g = 1, 2, \ldots, G$ word types (cf. column 2). For each group, column 3 presents three randomly chosen examples.

Table 1. Contribution (in %) of word types with different token frequencies as a function of α *.

Token Frequency	Number of Cases	Examples	α = 0.25	α = 0.75	α = 1.00	α = 1.50	α = 2.00
1	2,486,393	koalitionsbündnisse nr.6/1962 bruckner-breitklang	48.65	9.32	2.38	0.00	0.00
2–10	1,135,102	geschlechterschulung unal wiedervereinigungs-prozedur	29.86	10.89	3.65	0.01	0.00
11–100	296,573	hotpants lánský planwirtschaftlichen	13.16	14.03	7.13	0.04	0.00
101–1000	74,791	wanda verbannte mitschnitt	5.83	19.21	14.69	0.28	0.00
1001–10,000	14,388	schüren ablesen vollmachten	1.96	19.81	22.07	1.53	0.06
10,001–100,000	1871	london sitzen beginnen	0.44	13.38	20.68	5.31	0.64
100,001–1,000,000	173	mark frau kaum	0.07	7.38	15.21	17.83	7.12
1,000,001 +	27	es die er	0.02	5.98	14.19	75.02	92.18
	4,009,318		100.00	100.00	100.00	100.00	100.00

* Values are rounded for illustration purposes only throughout this paper.

This implies that the relative contribution $C(g)$ per group can be calculated as (see also ([21], Equation (5))):

$$C(g) = \begin{cases} \frac{\sum_{g=1}^{G} p_g^{\alpha}}{\sum_{i=1}^{K} p_i^{\alpha}}, \ for\ \alpha \neq 1.00 \\ \frac{\sum_{g=1}^{G}(-1)*p_g*log_2(p_{g)}}{\sum_{i=1}^{K}(-1)*p_i*log_2(p_{i)}}, \ for\ \alpha = 1.00 \end{cases}. \tag{6}$$

Columns 4–8 of Table 1 show the relative contribution (in %) for each group to H_α as a function of α. For lower values of α, H_α is dominated by word types with lower token frequencies. For instance, hapax legomena, i.e., word types that only occur once, contribute almost half of $H_{\alpha=0.25}$. For larger values of α, only the most frequent word contributes to H_α. For example, the 27 word types with a token frequency of more than 1,000,000 contribute more than 92% to $H_{\alpha=2.00}$. Because words in different frequency ranges have different grammatical and pragmatic properties, varying α makes it possible to study different aspects of the word frequency spectrum [21].

As written above, we are interested in testing the dependence of both H_α and D_α on the sample size for the different α-values. Let us note that each article in our database can be described by different attributes, e.g., publication date, subject matter, length, category, or author. Of course, this list of attributes is not exhaustive but can be freely extended depending on the research objective. In order to balance the article's characteristics across the corpus, we prepared 10 versions of our database, each with a different random arrangement of the order of all articles. To study the convergence of H_α, we computed H_α after every $n = 2^k$ consecutive tokens for each version, where $k = 6, 7, \ldots, log_2(N) = 27$. For D_α, we compared the first $n = 2^k$ word tokens with the last $n = 2^k$ of each version of our database. Here, $k = 6, 7, \ldots, 26$. For instance for $k = 26$, the first 67,108,864 word tokens are compared with the last 67,108,864 word tokens by calculating the generalized divergence between both "texts" for different α-values. Through the manipulation of the article order, it can be inferred that, random fluctuations aside, any systematic differences are caused by differences in the sample size.

As outlined above, our initial research interest concerned the use of generalized entropies and divergence in order to measure lexical change rates at specific ranges of the word frequency spectrum. To this end, we used the publication date of each article on a monthly basis to create a diachronic version of our database. Figure 3 visualizes the corpus size N_t for each t, where each monthly observation is identified by a variable containing the year y = 1947, 1948, ... , 2017 and the month m = 1, 2, ... ,12.

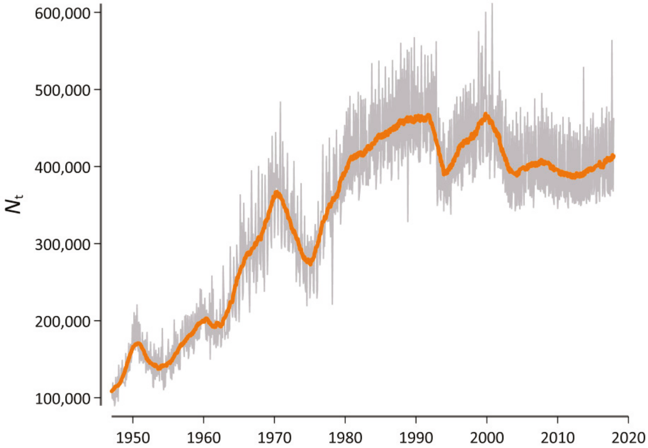


Figure 3. Sample size of the database as a function of time. The gray line depicts the raw data, while the orange line adds a symmetric 25-month window moving-average smoother highlighting the central tendency of the series at each point in time.

Instead of calculating the generalized Jensen–Shannon divergences for two different texts p and q, D_α was calculated for successive moments in time, i.e., $D_\alpha(t,t-1)$, in order to estimate the rate of lexical change at a given time point t [11,12]. For instance, D_α at y = 2000 and m = 1 represents the generalized divergence for a corresponding α-value between all articles that were published in January 2000 and those published in December 1999. The resulting series of month-to-month changes could then be analyzed in a standard time-series analysis framework. For example, we can test whether the series exhibits any large-scale tendency to change over time. A series with a positive trend increases over time, which would be indicative of an increasing rate of lexical change. It would also be interesting to look at first differences in the series, as an upward trend here in addition to an upward trend in the actual series would mean that the rate of lexical change is increasing at an increasing rate.

However, because the sample size clearly varies as a function of time (cf. Figure 3), it was essential to rule out the possibility that this variation systematically influences the results. Therefore, we generated a second version of this diachronic database in which we first randomly arranged the order of each article again. We then used the first $N_{t=1}$ words of this version of the database to generate a new corpus that has the same length (in words) as the original corpus at t = 1 but in which the diachronic signal is destroyed. We then proceeded and used the next $N_{t=2}$ words to generate a corpus that has the same length as the original corpus at t = 2. For example, the length of a concatenation of all articles that where published in *Der Spiegel* in January 1947 is 94,716 word tokens. Correspondingly, our comparison corpus at this point in time also consisted of 94,716 word tokens, but the articles of which it consisted could belong to any point in time between 1947 and 2017. In what follows, we computed all D_α $(t,t-1)$ values for both the original version of our database and for the version with a destroyed diachronic signal. We tentatively called this a "Litmus test", because it determined whether our results can be attributed to real diachronic changes or if there is a systematic bias due to the varying sample sizes.

Statistical analysis: To test if H_α and D_α vary as a function of the sample size without making any assumptions regarding the functional form of the relationship, we used the non-parametric Spearman correlation coefficient denoted as ρ. It assesses whether there is a monotonic relationship between two variables and is computed as Pearson's correlation coefficient on the ranks and average ranks of the two variables. The significance of the observed coefficient was determined by Monte Carlo permutation tests in which the observed values of the sample size are randomly permuted 10,000 times. The null hypothesis is that H_α/D_α does not vary with the sample size. If this is the case, then the sample size becomes arbitrary and can thus be randomly re-arranged, i.e., permuted. Let c denote the number of times the absolute ρ-value of the derived dataset is *greater than or equal to* the absolute ρ-value computed on the original data. A corresponding coefficient was labeled as "statistically significant" if $c < 10$, i.e., $p < 0.001$. In cases where l, i.e., the number of datapoints, was lower than or equal to 7, an exact test for all $l!$ permutations was calculated. Here, let c^* denote the number of times where the absolute ρ-value of the derived dataset is *greater than* the absolute ρ-value computed on the original data. A coefficient was labeled as "statistically significant" if $c^*/l! < 0.001$.

Data availability and reproducibility: All datasets used in this study are available in Dataverse (https://doi.org/10.7910/DVN/OP9PRL). For copyright and license reasons, each actual word type is replaced by a unique numerical identifier. Regarding further data access options, please contact the corpus linguistics department at Institute for the German language (IDS) (korpuslinguistik@ids-mannheim.de). In the spirit of reproducible science, one of the authors (A.K.) first analyzed the data using Stata and prepared a draft. Another author (S.W.) then used the draft and the available datasets to reproduce all the results using R. The results of this replication are available and the code (Stata and R) required to reproduce all the results presented in this paper are available in Dataverse (https://doi.org/10.7910/DVN/OP9PRL).

3. Results

3.1. Entropy H_α

To test the sample size dependence of H_α, we computed H_α for the first $n = 2^k$ consecutive tokens, where $k = 6, 7, \ldots, 27$ for the 10 versions of our database (each with a different random article order) and calculated averages. Figure 4A shows the convergence pattern for the five α-values in a superimposed scatter plot with connected dots where the colors of each y-axis correspond to one α-value (cf. the legend in Figure 4, the axes are log-scaled for improved visibility). For values of $\alpha <$ 1.00, there is no indication of convergence, while for $H_{\alpha=1.50}$ and $H_{\alpha=2.00}$, it seems that H_α converges rather quickly. To test the observed relationship between the sample size and H_α for different α-values, we calculated the Spearman correlation between the sample size and H_α for different minimum sample sizes. For example, a minimum sample size of $n = 2^{17}$ indicates that we restrict the calculation to sample sizes ranging between $n = 2^{17}$ and $n = 2^{27}$. For those 11 datapoints, we computed the Spearman correlation between the sample size and H_α and ran the permutation test. Table 2 summarizes the results. For all α-values, except for $\alpha = 2.00$, there is a clear indication for a significant (at $p < 0.001$) strong, positive, monotonic relationship between H_α and the sample size for all the minimum sample sizes. Thus, while Figure 4A seems to indicate that $H_{\alpha=1.50}$ converges rather quickly, the Spearman analysis reveals that the sample size dependence of $H_{\alpha=1.50}$ persists for higher values of k with a minimum ρ of 0.80. Except for the last two minimum sample sizes, all the coefficients pass the permutation test. For $\alpha = 2.00$, H_α starts to converge after $n = 2^{14}$ word tokens. None of the correlation coefficients of higher minimum sample sizes passes the permutation test. In line with the results of [21,22], this suggest $\alpha = 2.00$ as a pragmatic choice when calculating H_α. However, it is important to point out that for $\alpha = 2.00$, the computation of H_α is almost completely determined by the most frequent words (cf. Table 1). For lower values of α, the basic problem of sample size dependence (cf. Figure 1) persists. If it is the aim of a study to compare H_α for databases with varying sizes, this has to be taken into account. Correspondingly, [23] reached similar conclusions for the convergence of Rényi entropy of order $\alpha = 2.00$ for different languages and for different kinds of texts, both on the level of words and on the level of characters. In Appendix B, we have replicated the results of Table 2 based on Rényi's formulation of the entropy generalization. Table A5 shows that the results are almost identical, which is to be expected because the Havrda–Charvat–Lindhard–Nielsen–Aczél–Daróczy–Tsallis entropy is a monotone function of the Rényi entropy [20].

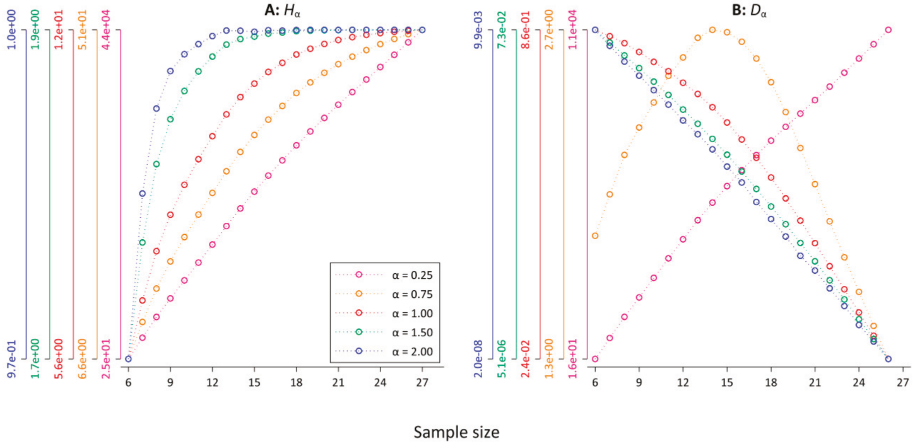


Figure 4. Generalized entropies H_α and divergences D_α as a function of the sample size. (**A**) P_α, (**B**) D_α.

Table 2. Spearman correlation between the sample size and H_α for different α-values *.

Minimum Sample Size	Number of Datapoints	α = 0.25	α = 0.75	α = 1.00	α = 1.50	α = 2.00
2^6	22	1.00 *	1.00 *	1.00 *	1.00 *	0.92 *
2^7	21	1.00 *	1.00 *	1.00 *	1.00 *	0.91 *
2^8	20	1.00 *	1.00 *	1.00 *	1.00 *	0.89 *
2^9	19	1.00 *	1.00 *	1.00 *	1.00 *	0.87 *
2^{10}	18	1.00 *	1.00 *	1.00 *	1.00 *	0.85 *
2^{11}	17	1.00 *	1.00 *	1.00 *	1.00 *	0.82 *
2^{12}	16	1.00 *	1.00 *	1.00 *	1.00 *	0.79 *
2^{13}	15	1.00 *	1.00 *	1.00 *	1.00 *	0.74
2^{14}	14	1.00 *	1.00 *	1.00 *	1.00 *	0.71
2^{15}	13	1.00 *	1.00 *	1.00 *	0.99 *	0.65
2^{16}	12	1.00 *	1.00 *	1.00 *	0.99 *	0.55
2^{17}	11	1.00 *	1.00 *	1.00 *	0.99 *	0.43
2^{18}	10	1.00 *	1.00 *	1.00 *	0.99 *	0.24
2^{19}	9	1.00 *	1.00 *	1.00 *	0.98 *	−0.05
2^{20}	8	1.00 *	1.00 *	1.00 *	0.98 *	−0.17
2^{21}	7	1.00 *	1.00 *	1.00 *	0.96 *	0.25
2^{22}	6	1.00 *	1.00 *	1.00 *	0.94	−0.20
2^{23}	5	1.00 *	1.00 *	1.00 *	0.90	0.10
2^{24}	4	1.00 *	1.00 *	1.00 *	0.80	−0.80

* An asterisk indicates that the corresponding correlation coefficient passed the permutation test at $p < 0.001$. For minimum sample sizes above 2^{20}, an exact permutation test is calculated.

3.2. Divergence D_α

To test the relationship between the sample size and D_α for different α-values, we computed D_α for a "text" that consists of the first $n = 2^k$ word tokens, a "text" that consists of the last $n = 2^k$ word tokens for each version of our database for $k = 6, 7, \ldots, 26$, and took averages. As for H_α above, we then calculated the Spearman correlation between the sample size and D_α for different minimum sample sizes. It is worth pointing out that the idea here is that the "texts" come from the same population, i.e., all *Der Spiegel* articles, so one should expect that with growing sample sizes, D_α should fluctuate around 0 with no systematic relationship between D_α and the sample size. Table 3 summarizes the results, while Figure 4B visualizes the convergence pattern. For all settings, there is a strong monotonic relationship between the sample size and D_α that passes the permutation test in almost every case. For $\alpha = 0.25$, the Spearman correlation coefficients are positive. This seems to be due to the fact that $H_{\alpha=0.25}$ is dominated by word types from the lower end of the frequency spectrum (cf. Table 1). Because, for example, word types that only occur once contribute almost half of $H_{\alpha=0.25}$. Those word types then either appear in the first 2^k or in the last 2^k word tokens.

Table 3. Spearman correlation between the sample size and D_α for different α-values *.

Minimum Sample Size	Number of Datapoints	α = 0.25	α = 0.75	α = 1.00	α = 1.50	α = 2.00
2^6	21	1.00 *	−0.42	−1.00 *	−1.00 *	−1.00 *
2^7	20	1.00 *	−0.54	−1.00 *	−1.00 *	−1.00 *
2^8	19	1.00 *	−0.64	−1.00 *	−1.00 *	−1.00 *
2^9	18	1.00 *	−0.74	−1.00 *	−1.00 *	−1.00 *
2^{10}	17	1.00 *	−0.83 *	−1.00 *	−1.00 *	−1.00 *
2^{11}	16	1.00 *	−0.90 *	−1.00 *	−1.00 *	−1.00 *
2^{12}	15	1.00 *	−0.95 *	−1.00 *	−1.00 *	−1.00 *
2^{13}	14	1.00 *	−0.99 *	−1.00 *	−1.00 *	−1.00 *
2^{14}	13	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{15}	12	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{16}	11	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{17}	10	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{18}	9	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{19}	8	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{20}	7	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{21}	6	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{22}	5	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{23}	4	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{24}	3	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *

* An asterisk indicates that the corresponding correlation coefficient passed the permutation test at $p < 0.001$. For minimum sample sizes above 2^{19}, an exact permutation test is calculated.

The results demonstrate that the larger the sample sizes the larger D_α (cf. the pink line in Figure 4B). For = 0.75, a similar pattern is observed for smaller sample sizes (cf. the orange line in Figure 4 B). However, at around $k = 15$, the pattern changes. For $k \geq 15$, there is a perfect monotonic negative relationship between $D_{\alpha=0.75}$ and the sample size. Surprisingly, there is a perfect monotonic negative relationship for all settings for $\alpha \geq 1.00$, even if we restrict the calculation to relatively large sample sizes. However, the corresponding values are very small. For instance, $D_{\alpha=2.00} = 7.91 \times 10^{-8}$ for $n = 2^{24}$, $D_{\alpha=2.00} = 4.08 \times 10^{-8}$ for $n = 2^{25}$, and $D_{\alpha=2.00} = 1.379 \times 10^{-8}$ for $n = 2^{26}$. One might object that this systematic sample size dependence is practically irrelevant. In the next section, we show that, unfortunately, this is not the case.

3.3. Case Study

As previously outlined, our initial idea was to use generalized divergences to measure the rate of lexical change at specific ranges of the word frequency spectrum. In what follows, we estimate the rate by calculating D_α for successive months, i.e., $D_\alpha(t,t-1)$. To rule out a potential systematical influence of the varying sample size, we also calculated $D_\alpha(t,t-1)$ for our comparison corpus where the diachronic signal was destroyed ("Litmus test").

For α, we chose 2.00 and 1.00. On the one hand, the analyses of [21,22] and our analysis presented above indicate that $\alpha = 2.00$ seems to be the most robust choice. On the other hand, we chose $\alpha = 1.00$, i.e., the original Jensen–Shannon divergence, because, as explained above, it has already been employed in the context of analyzing natural language data without explicitly testing the potential influence of varying sample sizes. Figure 5 shows our results. If we only looked at the plots on the left side (blue lines), the results would look very interesting, as there is a clear indication that the rate of lexical change decreases as a function of time for both $\alpha = 1.00$ and for $\alpha = 2.00$. However, looking at the plots in the middle reveals that a very similar pattern emerges for the comparison data. For our "Litmus test", we destroyed all diachronic information except for the varying sample sizes. Nevertheless, our conclusions would have been more or less identical. Interestingly, the patterns in Figure 5 clearly resemble the pattern of the sample size in Figure 3 (in reverse order) and thus suggest a negative association between $D_\alpha(t,t-1)$ and the sample size. To test this observation, we calculated the Spearman correlation between the sample size and $D_\alpha(t,t-1)$ for both $\alpha = 1.00$ and $\alpha = 2.00$ and ran a permutation test. Table 4, row 1, shows that there is a significant strong negative correlation between the sample size and D_α for both $\alpha = 1.00$ and $\alpha = 2.00$. Rows 2–5 present different approaches to solving the sample size dependence of D_α. In row 2, we extended Equation (2) to allow for unequal sample sizes, i.e., $N_p \neq N_q$ as suggested by ([22], Appendix A); here:

$$\begin{aligned} D^\pi_\alpha(p,q) &= H_\alpha\Big(\pi_p p + \pi_q q\Big) - \pi_p H_\alpha(p) - \pi_q H_\alpha(q) \\ \text{where } \pi_p &= N_p/\Big(N_p + N_q\Big) \text{ and } \pi_q = N_q/\Big(N_p + N_q\Big). \end{aligned} \quad (7)$$

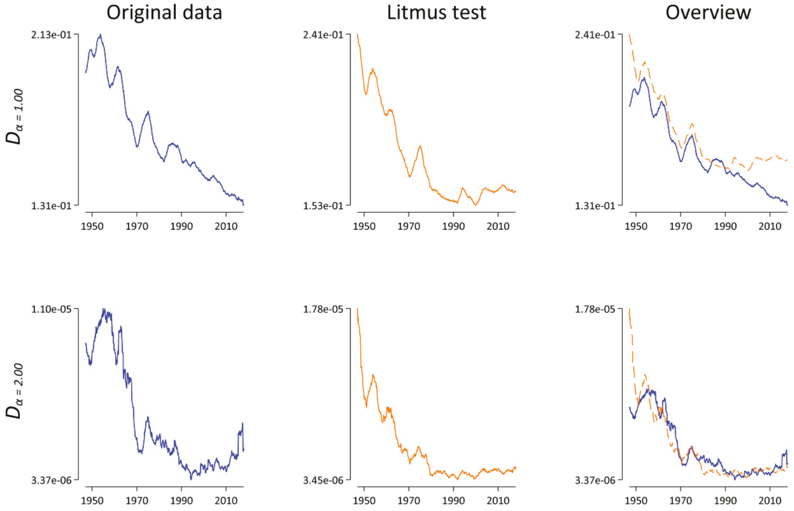


Figure 5. $D_\alpha(t,t-1)$ as a function of time for $\alpha = 1.00$ and $\alpha = 2.00$. Lines represent a symmetric 25-month window moving-average smoother highlighting the central tendency of the series at each point in time. Left: results for the original data in blue. Middle: results for the "Litmus" data in orange. Right: superimposition of both the original and the "Litmus" data.

Row 2 of Table 4 demonstrates that this "natural weights" approach does not qualitatively affect the results; there is still a significant and strong negative correlation between the sample size and D^π_α for both $\alpha = 1.00$ and $\alpha = 2.00$. Another approach is to increase the sample size (if possible). To this end, we aggregated the articles at the annual level instead of the monthly level. On average, the annual corpora are $\overline{N} = 3{,}334{,}409.04$ words long, compared to $\overline{N} = 277{,}867.42$ word tokens for the monthly data. Row 3 of Table 4 shows that increasing the sample size does not help in removing the influence of the sample size either. Another standard approach [15,22] is to randomly draw N_{min} word tokens from the monthly databases, where N_{min} is equal to the smallest of all monthly corpora, here $N_{min} = 75{,}819$ (June 1947). To our own surprise, row 4 of Table 4 reveals that this "random draw" approach also does not break the sample size dependence. While the absolute values of the correlation coefficients for both $\alpha = 1.00$ and $\alpha = 2.00$ are smaller for the original data than for the comparison data, all four coefficients are significantly different from 0 (at $p < 0.001$) and thus indicate that the "random draw" approach fails to pass the "Litmus test". As a last idea, we decided to truncate each monthly corpus after N_{min} word tokens. The difference between this "cut-off" approach and the "random draw" is that the latter approach assumes that words occur randomly in texts, while truncating the data after N_{min} as in the "cut-off" approach respects the syntactical and semantical coherence and the discourse structure at the text level [16,17]. On the one hand, row 5 of Table 4 demonstrates that this approach mostly solves the problem: all four coefficients are small, and only one coefficient is significantly different from zero, but positive. This suggests that the "cut-off" approach passes the "Litmus test". On the other hand, it's worth pointing out that we lose a lot of information with this approach. For example, the largest corpus is N = 507,542 word tokens long (October 2000). With the "cut-off" approach, more than 85% of those word tokens are not used to calculate $D_\alpha(t,t-1)$.

Table 4. Spearman correlation between the sample size and $D_\alpha(t,t-1)$ for the original data and for the "Litmus test" for α = 1.00 and α = 2.00.

Row	Scenario	α	Number of Cases	Original Data	Litmus Test
1	Original	1.00	851	−0.76 *	−0.91 *
		2.00	851	−0.70 *	−0.79 *
2	Natural weights	1.00	851	−0.77 *	−0.90 *
		2.00	851	−0.70 *	−0.79 *
3	Yearly data	1.00	70	−0.74 *	−0.97 *
		2.00	70	−0.46 *	−0.87 *
4	Random draw	1.00	851	−0.16 *	−0.69 *
		2.00	851	−0.50 *	−0.61 *
5	Cut-off	1.00	851	0.12 *	0.08
		2.00	851	0.08	−0.10

* An asterisk indicates that the corresponding correlation coefficient passed the permutation test at $p < 0.001$.

While the resulting pattern in Figure 6 might be indicative of an interesting lexico-dynamical process, especially for α = 1.00, what is more important in the present context is the fact that both blue lines look completely different compared with the corresponding blue lines in Figure 5. Thus, in relation to the analysis above (cf. Section 3.2), we concluded that the systematic sample size dependence of D_α is far from practically irrelevant. On the contrary, the analyses presented in this section demonstrate again why it is essential to account for the sample size dependence of lexical statistics.

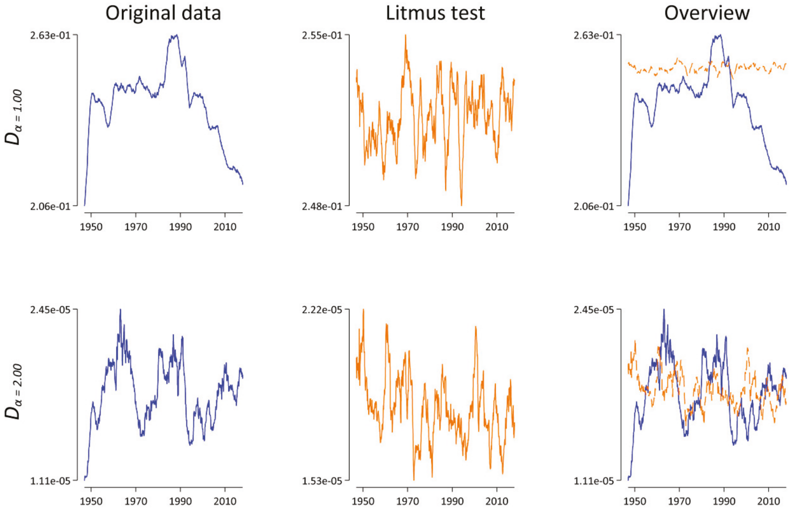


Figure 6. $D_\alpha(t,t-1)$ as a function of time for $\alpha = 1.00$ and $\alpha = 2.00$. Here, each monthly corpus is truncated after N_{min} = 75,819 word tokens. Lines represent a symmetric 25-month window moving-average smoother highlighting the central tendency of the series at each point in time. Left: results for the original data in blue. Middle: results for the "Litmus" data in orange. Right: superimposition of both the original and the "Litmus" data.

4. Discussion

In this paper, we explored the possibilities of using generalized entropies to analyze the lexical dynamics of natural language data. Using the α-parameter in order to automatically magnify differences between different texts at specific scales of the corresponding word frequency spectrum is interesting, as it promises a more objective selection method compared to, e.g., [8], who use a pre-compiled list of content-free words, or [12], who analyzes differences within different part-of-speech classes.

In line with other studies [17,23,27–29], the results demonstrate that it is essential for the analysis of natural language to always take into account the systematic influence of the sample size. With the exception of $H_{\alpha=2.00}$ for larger sample sizes, all quantities that are based on general entropies seem to strongly covary with the sample size (also see [23] for similar results based on Rényi's formulation of generalized entropies). In his monograph on word frequency distributions, Baayen [16] introduces the two fundamental methodological issues in lexical statistics:

> The sample size crucially determines a great many measures that have been proposed as characteristic text constants. However, the values of these measures change systematically as a function of the sample size. Similarly, the parameters of many models for word frequency distribution [sic!] are highly dependent on the sample size. This property sets lexical statistics apart from most other areas in statistics, where an increase in the sample size leads to enhanced accuracy and not to systematic changes in basic measures and parameters. . . . The second issue concerns the theoretical assumption [. . .] that words occur randomly in texts. This assumption is an obvious simplification that, however, offers the possibility of deriving useful formulae for text characteristics. The crucial question, however, is to what extent this simplifying assumption affects the reliability of the formulae when applied to actual texts and corpora. (p.1)

The main message of this paper is that those two fundamental issues also pose a strong challenge to the application of information theory for the quantitative study of natural language signals. In addition, the results of the case study (cf. Section 3.3) indicate that both fundamental issues in lexical statistics apparently interact with each other. As mentioned above, there are numerous studies that used the Jensen–Shannon divergence or related measures without an explicit "Litmus test". Let us mention two examples from our own research:

(i) In [12], an exploratory data-driven method was presented that extracts word-types from diachronic corpora that have undergone the most pronounced change in frequency of occurrence in a given period of time. To this end, a measure that is approximately equivalent to the Jensen–Shannon divergence is computed and period-to-period changes are calculated as in Section 3.3.

(ii) In [15], the parameters of the Zipf–Mandelbrot law were used to quantify and visualize diachronic lexical, syntactical, and stylistic changes, as well as aspects of linguistic change for different languages.

Both studies are based on data from the Google Books Ngram corpora, made available by [30]. It contains yearly token frequencies for each word type for over 8 million books, i.e., 6% of all books ever published [31]. To avoid a potential systematic bias due to strongly changing corpus sizes, random samples of equal size were drawn from the data in both [12] and [15]. However, as demonstrated in Section 3.3, apparently this simplifying assumption is problematic, because it seems to make a difference if we randomly sample N word tokens or if we keep the first N word tokens for the statistical structure of the corresponding word frequency distribution. It is worth pointing out again that, without the "Litmus test" the interpretation of the results presented in Section 3.3 would have been completely different, because randomly drawing word tokens from the data does not seem to break the sample size dependence. It is an empirical question whether the results presented in [12,15], and comparable other papers would pass a "Litmus test". In light of the results presented in this paper, we are rather skeptical, thus echoing the call of [22] that it is "essential to clarify what is the role of finite-size effects in the reported conclusions, in particular in the (typical) case that database sizes change over time." (p. 8). One could even go so far as to ask whether relative frequencies that are compared between databases of different sizes are systematically affected by varying database sizes. However, the test scheme as we introduced it presupposes access to the full text data. For instance, due to copyright constraints, access to Google Books Ngram data is restricted to token frequencies for all words (and phrases) that occur at least 40 times in the corpus. Thus, an analogous "Litmus test" is not possible. At our institute, we are rather fortunate to have access to the full text data of our database. Notwithstanding, copyright and license reasons are a major issue here, as well [32]. To solve this problem for our study, we replaced each actual word type with a unique numerical identifier as explained in Section 3.3. For our focus of research, using such a pseudonymization strategy is fine. However, there are many scenarios where, depending on the research objective, the actual word strings matter, making it necessary to develop a different access and publication strategy. It goes without saying that, in all cases, full-text access is the best option.

While the peculiarities of word frequency distributions make the analysis of natural language data more difficult compared to other empirical phenomena, we hope that our analyses (especially the "Litmus test") also demonstrate that textual data offer novel possibilities to answer research questions. Or put differently, natural language data contain a lot of information that can be harnessed. For example, two reviewers pointed out that it could make sense to develop a method that recovers an unbiased lexico-dynamical signal by removing the "Litmus test" signal from the original signal. This is an interesting avenue for future research.

Supplementary Materials: The replication results and code (Stata and R) required to reproduce all the results presented in this paper are available in Dataverse (https://doi.org/10.7910/DVN/OP9PRL).

Author Contributions: Conceptualization A.K., S.W., and C.M.-S.; methodology and study design, A.K.; data preparation and management, A.K. and S.W.; data analysis—original analysis, A.K.; data analysis —replication, S.W.; visualization, A.K.; writing—original draft preparation, A.K.; writing—review and editing, A.K., S.W., and C.M.-S. All the authors have read and approved the final manuscript.

Funding: This research received no external funding.

Acknowledgments: We thank Sarah Signer for proofreading. We also thank Gerardo Febres and the two anonymous reviewers for their insightful feedback.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A Inclusion of Punctuation and Cardinal Numbers.

Here, punctuation and numbers are included. This version of the database consists of $N = 286{,}729{,}999$ word tokens and $K = 4{,}056{,}122$ different word types. Table A1 corresponds to Table 1. Because (especially) punctuation symbols have a very high token frequency, the contribution of the highest frequency groups increases when punctuation is not removed from the database. However, the results are still qualitatively very similar. Table A2 corresponds to Table 2. For $\alpha \leq 1.50$, removing punctuation does not qualitatively affect the results. However, for $\alpha = 2.00$, except for $n = 2^{24}$ none of the correlation coefficients pass the permutation test. Again, this indicates that $\alpha = 2.00$ is a pragmatic choice when calculating H_α. However, it also demonstrates that the conceptual decision to remove punctuation/cardinal numbers can affect the results. Table A3 corresponds to Table 3 The results are not qualitatively affected by the exclusion of punctuation/cardinal numbers. The same conclusion can be drawn for Table A4, which corresponds to Table 4.

Table A1. Contribution of word types with different token frequency as a function of α.

Token Frequency	Number of Cases	Examples	$\alpha = 0.25$	$\alpha = 0.75$	$\alpha = 1.00$	$\alpha = 1.50$	$\alpha = 2.00$
1	2,511,837	paragraphenplantage penicillinhaltigen partei-patt	48.51	8.94	2.16	0.00	0.00
2–10	1,148,295	koberten optimis-datenbank gazprom-zentrale	29.82	10.46	3.32	0.00	0.00
11–100	303,049	dunkelgraue stirlings drollig	13.26	13.57	6.54	0.02	0.00
101–1000	76,049	abgemagert irakern aufzugehen	5.86	18.56	13.50	0.15	0.00
1001–10,000	14,710	nord- selbstbestimmung alexandra	1.99	19.35	20.60	0.83	0.02
10,001-100,000	1966	parteien banken entscheidungen	0.46	13.24	19.57	2.86	0.22
100,001-1,000,000	183	wurde würde dieses	0.08	7.47	14.89	10.05	2.66
1,000,001 +	33	auf wie ,	0.03	8.40	19.42	86.09	97.09
	4,056,122		100.00	100.00	100.00	100.00	100.00

Table A2. Spearman correlation between the sample size and H_α for different α-values *.

Minimum Sample Size	Number of Datapoints	$\alpha = 0.25$	$\alpha = 0.75$	$\alpha = 1.00$	$\alpha = 1.50$	$\alpha = 2.00$
2^{6}	23	1.00 *	1.00 *	1.00 *	0.99 *	0.49
2^{7}	22	1.00 *	1.00 *	1.00 *	0.99 *	0.41
2^{8}	21	1.00 *	1.00 *	1.00 *	0.99 *	0.32
2^{9}	20	1.00 *	1.00 *	1.00 *	0.99 *	0.22
2^{10}	19	1.00 *	1.00 *	1.00 *	0.99 *	0.09
2^{11}	18	1.00 *	1.00 *	1.00 *	0.99 *	−0.08
2^{12}	17	1.00 *	1.00 *	1.00 *	0.98 *	−0.28
2^{13}	16	1.00 *	1.00 *	1.00 *	0.98 *	−0.53
2^{14}	15	1.00 *	1.00 *	1.00 *	0.97 *	−0.50
2^{15}	14	1.00 *	1.00 *	1.00 *	0.97 *	−0.45
2^{16}	13	1.00 *	1.00 *	1.00 *	0.96 *	−0.81
2^{17}	12	1.00 *	1.00 *	1.00 *	0.95 *	−0.76
2^{18}	11	1.00 *	1.00 *	1.00 *	0.94 *	−0.71
2^{19}	10	1.00 *	1.00 *	1.00 *	0.95 *	−0.61
2^{20}	9	1.00 *	1.00 *	1.00 *	0.95 *	−0.47
2^{21}	8	1.00 *	1.00 *	1.00 *	0.93	−0.31
2^{22}	7	1.00 *	1.00 *	1.00 *	0.89	0.04
2^{23}	6	1.00 *	1.00 *	1.00 *	0.83	0.66
2^{24}	5	1.00 *	1.00 *	1.00 *	1.00 *	1.00 *

* An asterisk indicates that the corresponding correlation coefficient passed the permutation test at $p < 0.001$. For minimum sample sizes above 2^{20}, an exact permutation test is calculated.

Table A3. Spearman correlation between the sample size and D_α for different α-values *.

Minimum Sample Size	Number of Datapoints	α = 0.25	α = 0.75	α = 1.00	α = 1.50	α = 2.00
2^6	22	1.00 *	−0.51	−1.00 *	−1.00 *	−1.00 *
2^7	21	1.00 *	−0.59	−1.00 *	−1.00 *	−1.00 *
2^8	20	1.00 *	−0.68 *	−1.00 *	−1.00 *	−1.00 *
2^9	19	1.00 *	−0.76 *	−1.00 *	−1.00 *	−1.00 *
2^{10}	18	1.00 *	−0.84 *	−1.00 *	−1.00 *	−1.00 *
2^{11}	17	1.00 *	−0.89 *	−1.00 *	−1.00 *	−1.00 *
2^{12}	16	1.00 *	−0.94 *	−1.00 *	−1.00 *	−1.00 *
2^{13}	15	1.00 *	−0.97 *	−1.00 *	−1.00 *	−1.00 *
2^{14}	14	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{15}	13	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{16}	12	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{17}	11	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{18}	10	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{19}	9	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{20}	8	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{21}	7	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{22}	6	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{23}	5	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *
2^{24}	4	1.00 *	−1.00 *	−1.00 *	−1.00 *	−1.00 *

* An asterisk indicates that the corresponding correlation coefficient passed the permutation test at $p < 0.001$. For minimum sample sizes above 2^{20}, an exact permutation test is calculated.

Table A4. Spearman correlation between the sample size and $D_\alpha(t,t-1)$ for the original data and for the "Litmus test" for $\alpha = 1.00$ and $\alpha = 2.00$.

Row	Scenario	α	Number of Cases	Original Data	Litmus Test
1	Original	1.00	851	−0.77 *	−0.91 *
		2.00	851	−0.63 *	−0.70 *
2	Natural weights	1.00	851	−0.77 *	−0.91 *
		2.00	851	−0.63 *	−0.70 *
3	Yearly data	1.00	70	−0.74 *	−0.98 *
		2.00	70	−0.39	−0.83 *
4	Random draw	1.00	851	−0.29 *	−0.69 *
		2.00	851	−0.45 *	−0.56 *
5	Cut-off	1.00	851	0.07	0.05
		2.00	851	0.11	−0.07

* An asterisk indicates that the corresponding correlation coefficient passed the permutation test at $p < 0.001$.

Appendix B Replication of Table 2 for a Different Formulation of Generalized Entropy.

Here, we replicate Table 2 for a different formulation of generalized entropy, the so-called Rényi entropy of order α [24]; it can be written as:

$$H'_\alpha(p) = \frac{1}{\alpha - 1} log_2\left(\sum_{i=1}^{K} p_i^\alpha\right). \tag{A1}$$

Table A5. Spearman correlation between the sample size and H'_α for different α-values *.

Minimum Sample Size	Number of Datapoints	α = 0.25	α = 0.75	α = 1.00	α = 1.50	α = 2.00
2^{6}	22	1.00 *	1.00 *	1.00 *	1.00 *	0.92 *
2^{7}	21	1.00 *	1.00 *	1.00 *	1.00 *	0.90 *
2^{8}	20	1.00 *	1.00 *	1.00 *	1.00 *	0.89 *
2^{9}	19	1.00 *	1.00 *	1.00 *	1.00 *	0.87 *
2^{10}	18	1.00 *	1.00 *	1.00 *	1.00 *	0.85 *
2^{11}	17	1.00 *	1.00 *	1.00 *	1.00 *	0.82 *
2^{12}	16	1.00 *	1.00 *	1.00 *	1.00 *	0.78
2^{13}	15	1.00 *	1.00 *	1.00 *	1.00 *	0.73
2^{14}	14	1.00 *	1.00 *	1.00 *	1.00 *	0.70
2^{15}	13	1.00 *	1.00 *	1.00 *	0.99 *	0.65
2^{16}	12	1.00 *	1.00 *	1.00 *	0.99 *	0.55
2^{17}	11	1.00 *	1.00 *	1.00 *	0.99 *	0.43
2^{18}	10	1.00 *	1.00 *	1.00 *	0.99 *	0.24
2^{19}	9	1.00 *	1.00 *	1.00 *	0.98 *	−0.05
2^{20}	8	1.00 *	1.00 *	1.00 *	0.98 *	−0.17
2^{21}	7	1.00 *	1.00 *	1.00 *	0.96 *	0.25
2^{22}	6	1.00 *	1.00 *	1.00 *	0.94	−0.20
2^{23}	5	1.00 *	1.00 *	1.00 *	0.90	0.10
2^{24}	4	1.00 *	1.00 *	1.00 *	0.80	−0.80

* An asterisk indicates that the corresponding correlation coefficient passed the permutation test at $p < 0.001$. For minimum sample sizes above 2^{20}, an exact permutation test is calculated.

References

1. Manning, C.D.; Schütze, H. *Foundations of Statistical Natural Language Processing*; MIT Press: Cambridge, MA, USA, 1999; ISBN 978-0-262-13360-9.
2. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*; Pearson Education (US): Upper Saddle River, NJ, USA, 2009; ISBN 978-0-13-504196-3.
3. Adami, C. What is information? *Philos. Trans. R. Soc. A* **2016**, *374*, 20150230. [CrossRef] [PubMed]
4. Cover, T.M.; Thomas, J.A. *Elements of information theory*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2006; ISBN 978-0-471-24195-9.
5. Bentz, C.; Alikaniotis, D.; Cysouw, M.; Ferrer-i-Cancho, R. The Entropy of Words—Learnability and Expressivity across More than 1000 Languages. *Entropy* **2017**, *19*, 275. [CrossRef]
6. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
7. Endres, D.M.; Schindelin, J.E. A new metric for probability distributions. *IEEE Trans. Inf. Theory* **2003**, *49*, 1858–1860. [CrossRef]
8. Hughes, J.M.; Foti, N.J.; Krakauer, D.C.; Rockmore, D.N. Quantitative patterns of stylistic influence in the evolution of literature. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 7682–7686. [CrossRef] [PubMed]
9. Klingenstein, S.; Hitchcock, T.; DeDeo, S. The civilizing process in London's Old Bailey. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 9419–9424. [CrossRef]
10. DeDeo, S.; Hawkins, R.; Klingenstein, S.; Hitchcock, T. Bootstrap Methods for the Empirical Study of Decision-Making and Information Flows in Social Systems. *Entropy* **2013**, *15*, 2246–2276. [CrossRef]
11. Bochkarev, V.; Solovyev, V.; Wichmann, S. Universals versus historical contingencies in lexical evolution. *J. R. Soc. Interface* **2014**, *11*, 20140841. [CrossRef] [PubMed]
12. Koplenig, A. A Data-Driven Method to Identify (Correlated) Changes in Chronological Corpora. *J. Quant. Linguist.* **2017**, *24*, 289–318. [CrossRef]
13. Pechenick, E.A.; Danforth, C.M.; Dodds, P.S. Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLOS ONE* **2015**. [CrossRef]
14. Zipf, G.K. *The Psycho-biology of Language. An Introduction to Dynamic Philology*; Houghton Mifflin Company: Boston, MA, USA, 1935.

15. Koplenig, A. Using the parameters of the Zipf–Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes–a large-scale corpus analysis. *Corpus Linguist. Linguist. Theory* **2018**, *14*, 1–34. [CrossRef]
16. Baayen, R.H. *Word Frequency Distributions*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2001.
17. Tweedie, F.J.; Baayen, R.H. How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Comput. Hum.* **1998**, *32*, 323–352. [CrossRef]
18. Simpson, E.H. The Interpretation of Interaction in Contingency Tables. *J. R. Stat. Soc. Series B* **1951**, *13*, 238–241. [CrossRef]
19. Gerlach, M.; Altmann, E.G. Stochastic Model for the Vocabulary Growth in Natural Languages. *Phys. Rev. X* **2013**, *3*, 021006. [CrossRef]
20. Briët, J.; Harremoës, P. Properties of classical and quantum Jensen-Shannon divergence. *Phys. Rev. A* **2009**, *79*, 052311. [CrossRef]
21. Altmann, E.G.; Dias, L.; Gerlach, M. Generalized entropies and the similarity of texts. *J. Stat. Mech. Theory Exp.* **2017**, *2017*, 014002. [CrossRef]
22. Gerlach, M.; Font-Clos, F.; Altmann, E.G. Similarity of Symbol Frequency Distributions with Heavy Tails. *Phys. Rev. X* **2016**, *6*, 021009. [CrossRef]
23. Tanaka-Ishii, K.; Aihara, S. Computational Constancy Measures of Texts—Yule's K and Rényi's Entropy. *Comput. Linguist.* **2015**, *41*, 481–502. [CrossRef]
24. Rényi, A. On Measures of Entropy and Information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, Berkeley, CA, USA, 20 June–30 July 1960; University of California Press: Berkeley, CA, USA, 1961; pp. 547–561.
25. He, Y.; Hamza, A.B.; Krim, H. A generalized divergence measure for robust image registration. *IEEE Trans. Signal Process.* **2003**, *51*, 1211–1220.
26. Schmid, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK, 1994; pp. 44–49.
27. Köhler, R.; Galle, M. Dynamic aspects of text characteristics. In *Quantitative Text Analysis*; Hřebíček, L., Altmann, G., Eds.; Quantitative linguistics; WVT Wissenschaftlicher Verlag Trier: Trier, Germany, 1993; pp. 46–53. ISBN 978-3-88476-080-2.
28. Popescu, I.-I.; Altmann, G. *Word Frequency Studies*; Quantitative linguistics; Mouton de Gruyter: Berlin, Germany, 2009; ISBN 978-3-11-021852-7.
29. Wimmer, G.; Altmann, G. Review Article: On Vocabulary Richness. *J. Quant. Linguist.* **1999**, *6*, 1–9. [CrossRef]
30. Michel, J.-B.; Shen, Y.K.; Aiden, A.P.; Verses, A.; Gray, M.K.; Google Books Team; Pickett, J.P.; Hoiberg, D.; Clancy, D.; Norvig, P.; et al. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* **2010**, *331*, 176–182. [CrossRef] [PubMed]
31. Lin, Y.; Michel, J.-B.; Aiden, L.E.; Orwant, J.; Brockmann, W.; Petrov, S. Syntactic Annotations for the Google Books Ngram Corpus. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju Island, Korea, 8–14 July 2012; pp. 169–174.
32. Kupietz, M.; Lüngen, H.; Kamocki, P.; Witt, A. The German Reference Corpus DeReKo: New Developments–New Opportunities. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., et al., Eds.; European Language Resources Association (ELRA): Miyazaki, Japan, 2018.

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

Entropy Editorial Office
E-mail: entropy@mdpi.com
www.mdpi.com/journal/entropy

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland

Tel: +41 61 683 77 34
Fax: +41 61 302 89 18

www.mdpi.com

ISBN 978-3-03936-027-7