Compression and the origins of Zipf's law of abbreviation

R. Ferrer-i-Cancho[1,*], C. Bentz[2] & C. Seguin[1]

[1]Complexity & Quantitative Linguistics Lab, LARCA Research Group,
Departament de Ciències de la Computació.
Universitat Politècnica de Catalunya,
Campus Nord, Edifici Omega Jordi Girona Salgado 1-3.
08034 Barcelona, Catalonia (Spain).

[2] Department of Theoretical and Applied Linguistics,
Faculty of Modern & Medieval Languages (MML).
University of Cambridge,
Sidgwick Avenue, Cambridge, CB3 9DA, United Kingdom.

(*) Author for correspondence, rferrericancho@cs.upc.edu

Abstract

Languages across the world exhibit Zipf's law of abbreviation, namely more frequent words tend to be shorter. The generalized version of the law - an inverse relationship between the frequency of a unit and its magnitude - holds also for the behaviours of other species and the genetic code. The apparent universality of this pattern in human language and its ubiquity in other domains calls for a theoretical understanding of its origins. To this end, we generalize the information theoretic concept of mean code length as a mean energetic cost function over the probability and the magnitude of the types of the repertoire. We show that the minimization of that cost function and a negative correlation between probability and the magnitude of types are intimately related.

Keywords:: Zipf's law of abbreviation, compression, information theory, language, animal behaviour.

1 Introduction

Zipf's law of abbreviation, the tendency of more frequent words to be shorter [1], holds in every language for which it was tested [1, 2, 3, 4, 5, 6, 7, 8, 9] (Fig. 1 (a)), suggesting that language universals are not necessarily a myth [10]. A generalized version of the law, i.e. a negative correlation between the frequency of a type from a repertoire and its magnitude (e.g., its length, size or duration), has been found in the behaviour of other species [11, 12, 13, 9, 14] (Fig. 1 (b)) and in the genetic code [15]. This is strong evidence for a general tendency of more frequent units to be smaller, i.e. less cost-intensive. The robustness and recurrence of this pattern calls for a theoretical understanding of the mechanisms that give rise to it.

The common interpretation of the law as an indication of the efficiency of language and animal behavior [1, 16, 13] suffers from Kirby's problem of linkage, i.e. the lack of a strong connection between potential processing constraints and the proposed universal [17]. Here we address the problem of linkage for the law of abbreviation with the help of information theory.

Information theory sheds light on the origins of many regularities of natural language, e.g., duality of patterning [18], Zipf's law for word frequencies [19, 20], Clark's principle of contrast [21], a vocabulary learning bias in children [21] and the exponential decay of the distribution of dependency lengths [22, 23]. Those examples suggest that the solutions of information theory to communication problems can be informative for natural languages too, though they might look different in some cases [24].

Here we investigate the law of abbreviation in the light of the problem of compression from standard information theory [25, 26]. In this context, the mean code length is defined as

$$L = \sum_{i=1}^{V} p_i l_i, \qquad (1)$$

where p_i and l_i are, respectively, the probability and the length in symbols of the i-th type of a repertoire of size V. In the case of human language, the types could be words, the symbols could be letters and the repertoire would be a vocabulary. Solving the problem of compression provides word lengths that minimize L when the p_i's are given. An optimal coding of types by using strings of symbols (under the wide scheme of uniquely decipherable codes) satisfies [25]

$$l_i \propto \lceil -\log p_i \rceil, \qquad (2)$$

which is indeed a particular case of Zipf's law of abbreviation.

Eq. 1 can also be interpreted as an energetic cost function where the cost of every unit is exactly its length. Based on this assumption, we will generalize the problem of compression in two ways. First, we put forward a cost function Λ, i.e.

$$\Lambda = \sum_{i=1}^{V} p_i \lambda_i, \tag{3}$$

where λ_i is the energetic cost of the i-th type. In his pioneering research, G. K. Zipf already proposed a particular version of Eq. 3 to explain the origins of the law of abbreviation using qualitative arguments [1, p. 59]. Here we will address Kirby's problem of linkage [17] showing how the minimization of Λ can produce the law of abbreviation.

We assume that the energetic cost of a unit is a monotonically increasing function of its length $\lambda_i = g(l_i)$, for instance the energy that is needed to articulate the sounds of a string of length l_i. Second, we generalize l_i as a magnitude (a positive real number). This way, l_i can indicate not only the length in syllables of a word [1] or the number of strokes of a Japanese kanji [6] but also the duration in time of a vocalisation [26] or the amount of information of a codon that is actually relevant for coding an amino acid [15]. Durations are important in the case of human language because words that have the same length in letters, and even the same number of phonemes, can have different durations[27]. A review of costs associated with the length or duration of a unit in human language and animal behaviour can be found in [26].

Under these two assumptions Λ becomes

$$\Lambda = \sum_{i=1}^{V} p_i g(l_i). \tag{4}$$

Λ is equivalent to L when g is the identity function. We assume that $g(l)$ is a strictly monotonically increasing function of l. The same assumption has been made for the cost of a syntactic dependency as a function of its length in word order models [28].

Assuming that g is the identity function, an equivalence between the minimization of Λ and Zipf's law of abbreviation is suggested by statistical analyses showing that any time that Λ is significantly small, the correlation between frequency and magnitude is significant (and negative) and *vice versa* [26]. Furthermore, theoretical arguments indicate that the law of abbreviation follows from minimum Λ when the empirical distribution of type lengths and type frequencies is constant and g is the identity function [26].

This is indirect evidence of a connection between compression and the law of abbreviation. Here we present direct connections between the minimization of Λ and a generalized law of abbreviation by means of various correlation metrics applied to the probability of a type and its cost. Namely, we will show that Λ is inherent to the Pearson correlation. We will also illustrate the link between minimization of Λ and the maximisation of the concordance with the law of abbreviation using Pearson and Kendal correlation. In consequence, our research is in the spirit of recent studies on the origins of Zipf's law for word frequencies through optimisation principles [19, 20, 29].

2 Predicting the law of abbreviation

A generalized law of abbreviation is defined often simply as a negative correlation between the frequency of a unit and its magnitude [12, 13, 9]. Here we consider three measures of correlation: Pearson correlation (r), Spearman rank correlation (ρ) and Kendall rank correlation (τ) [30]. While Pearson is a measure of linear association, ρ and τ are measures of both linear and non-linear association [31, 32]. Alternatives to r are necessary because the functional dependency between frequency and length is modelled by means of non-linear functions [4] and the actual g may not be linear.

r and ρ have been used in previous research on the generalized law of abbreviation [12, 13]. Here we introduce τ [30]. The non-parametric approach offered by ρ and τ allows one to remain agnostic about the actual functional dependency between frequency and magnitude [4] and avoids common problems of assuming concrete functions for linguistic laws [33, 34].

2.1 The relationship between Λ and Pearson's r

The Pearson correlation between the probability of a unit (p) and its energetic cost (λ) is

$$r = \frac{E[p\lambda] - E[p]E[\lambda]}{\sigma[p]\sigma[\lambda]}, \tag{5}$$

where $E[x]$ and $\sigma[x]$ are, respectively, the expectation and the standard deviation of a random variable x. $E[x]$ can be regarded as the average value of x obtained when drawing types uniformly at random from the repertoire. Knowing

$$E[p\lambda] = \frac{1}{V}\sum_{i=1}^{V} p_i\lambda_i = \frac{\Lambda}{V} \tag{6}$$

$$E[p] = \frac{1}{V}\sum_{i=1}^{V} p_i = \frac{1}{V}, \tag{7}$$

Eq. 5, can be expressed as

$$r = \frac{\Lambda - E[\lambda]}{V\sigma[p]\sigma[\lambda]}. \tag{8}$$

Therefore, r is a function of Λ.

Note that in quantitative linguistics, the term *type* is used to refer to a string of symbols and the term *token* is used to refer to an occurrence of a type [35]. The term type is used with the same meaning in quantitative studies of animal behavior [36]. $E[\lambda]$ and $\sigma[\lambda]$ are, respectively, the mean cost and the standard deviation of the cost of the types.

Now, let us assume a constancy condition: V, $E[\lambda]$, $\sigma[p]$ and $\sigma[\lambda]$ are constant (see the supplementary online information for a justification). If that symplifying condition holds, Eq. 8 indicates that the minimization of Λ is equivalent to the minimization of r, which in turn maximizes the concordance with the law of abbreviation because $g(l)$ is a monotonically increasing function of l. The same conclusion can be reached with a simpler but less general constancy condition, namely, that the multiset of p_i's and the multiset of λ_i's are constant. This simpler condition, which we refer shortly as multiset constancy, has been used to test the significance of Λ [26] and is the central assumption of correlation tests such as the ones used to test for the law of abbreviation [34].

$r < 0$ can be regarded as concordance with the law of abbreviation. To know if $r < 0$, it is not necessary to actually calculate r with Eq. 8. On the right hand side of this equation, the denominator is positive (since V and the standard deviations are positive). Hence, the sign of r is determined by the sign of the numerator. Therefore, $r < 0$ if and only if

$$\Lambda = \sum_{i=1}^{V} p_i\lambda_i < E[\lambda] = \frac{1}{V}\sum_{i=1}^{V}\lambda_i, \tag{9}$$

i.e. the expected energetic cost of types when selecting them according to their probability (p) is smaller than the expected energetic cost of types picking them uniformly at random from the repertoire. Eq. 9 tell us that a negative sign of r is equivalent to a mean energetic cost of *tokens* that does not exceed the mean energetic cost of *types*.

A limitation of the connection between r and Λ above is not only the validity of the constancy condition but also that r is a measure of linear association. A priori, we do not know if the relationship between p and λ is linear. For this reason it is vital to explore a connection between measures of correlation that can capture non-linear dependencies.

2.2 The relationship between Λ and τ

Here we will unravel a strong dependency between the minimization of Λ and the minimization of Kendall's τ, a rank measure of correlation between p_i and l_i under multiset constancy. τ is based on the concept of concordant and discordant pairs of values. In our case, a pair (p_i, l_i) and a pair (p_j, l_j) are

- Concordant if either $p_i > p_j$ and $l_i > l_j$ or $p_i < p_j$ and $l_i < l_j$
- Discordant if either $p_i > p_j$ and $l_i < l_j$ or $p_i < p_j$ and $l_i > l_j$
- Non concordant if $p_i = p_j$ or $l_i = l_j$.

n_c and n_d are defined, respectively, as the number of concordant and discordant pairs. n_0 is the total number of pairs, i.e.

$$n_0 = \frac{V(V-1)}{2}. \tag{10}$$

Then Kendall's τ is defined as a normalized difference between n_c and n_d, i.e.

$$\tau = \frac{n_c - n_d}{n_0}. \tag{11}$$

If the agreement with Zipf's law of abbreviation was perfect, i.e. if we had only discordant pairs, then we would have $\tau = -1$.

We will investigate the consequences of choosing a pair of indices at random, i and j ($i \neq j$) and swapping either p_i and p_j or l_i and l_j. A prime will be used to indicate the value of a quantity or measure after the swap. For instance, τ' and l'_i indicate, respectively, the value of τ and that of l_i after the swap. $\Delta_\tau = \tau' - \tau$ and $\Delta_\Lambda = \Lambda' - \Lambda$ indicate the discrete derivative of τ and Λ, respectively. For instance, $\Delta_\tau < 0$ indicates that the concordance with the law of abbreviation increases after one swap. If p_i's are real probabilities (not frequencies from a sample) and the l_i's are durations [27, 13] (not discrete lengths), ties are unlikely. For this reason we assume that there are no ties (and therefore non concordant pairs are missing) to investigate the relationship between Λ and τ. This has the further advantage of simplifying the mathematical arguments. A careful analysis shows that (see supplementary online information for further details)

- if the pairs (p_i, l_i) and (p_j, l_j) are concordant, then $\Delta_\Lambda, \Delta_\tau < 0$.
- if the pairs (p_i, l_i) and (p_j, l_j) are discordant, then $\Delta_\Lambda, \Delta_\tau > 0$.

The results above can be summarized as

$$\Delta_\Lambda \Delta_\tau > 0, \tag{12}$$

This results means that if one of the two changes (e.g., Λ), the other (e.g., τ) also changes in the same direction (and *vice versa*).

Bear in mind that removing all concordant pairs can lead to a drastic reduction of Λ but does not warrant that the coding is optimal in an information theoretic sense. Suppose that magnitudes are string lengths (in symbols), and that there are neither probability ties nor length ties, i.e. $n_c - n_d = n_0$. After removing all concordant pairs by swapping we get $n_c = 0$, and thus Eq. 11 gives $\tau = -1$, the strongest negative correlation possible. However, optimal coding with discrete units implies length ties for $V > 2$ (supplementary online information).

2.3 The relationship between Λ and ρ

Indirect relationships between ρ and Λ follow from those between ρ and τ. On the one hand, ρ satisfies [37]

$$\frac{1}{2}(3\tau - 1) \leq \rho \leq \frac{1}{2}(1 + 3\tau). \tag{13}$$

On the other hand, ρ satisfies [38]

$$\frac{(1+\tau)^2}{2} - 1 \leq \rho \leq 1 - \frac{(1-\tau)^2}{2}, \tag{14}$$

as illustrated in Fig. 2. See [39] for further relationships between ρ and τ.

3 Discussion

Assuming some constancy conditions on probabilities and magnitudes, we have shown an intimate relationship between the minimization of Λ and the minimization of various measures of correlation between the probability of a type and its magnitude. Notice that minimization of the correlation is equivalent to the maximisation of the concordance with the law of abbreviation. This potentially explains the ubiquity of a generalized version of Zipf's law of abbreviation in human language and also in the behaviour of other species.

More specifically, we have shown that Pearson's r contains Λ in its definition (Eq. 5). Our mathematical results shed light on previous results on the law of abbreviation involving r.

First, r has been used to investigate a generalized version of Zipf's law of abbreviation in dolphins surface behavioural patterns [12] and the vocalisations of Formosan macaques [13]. The magnitude of dolphin surface behavioural patterns was measured in elementary behavioural units while the magnitude of Formosan macaque vocalisations was measured by their duration. The mathematical connections presented in Section 2.1 predict a significantly low mean cost Λ for Formosan macaques [13] and dolphins from the significant negative r found in both species [13, 12]. Only two assumptions are required: multiset constancy for both the correlation test and the test of significance of Λ, and that the energetic cost of a signal is proportional to its magnitude. This prediction is confirmed by the analysis of the significance of Λ in dolphin surface behavioural patterns and the vocalisation of Formosan macaques [26].

Second, the same kind arguments predict a significant negative Pearson correlation between frequency and magnitude from the significantly low Λ that has been found in various languages [26], although that correlation was not investigated for them.

In spite of the predictive power of the minimization of Λ, we do not mean that the law of abbreviation is inevitable. Exceptions are known in other species [9, 14]. This is not surprising from the perspective of information theory. Solving the problem of compression is in conflict with the problem of transmission: redundancy must be added in a controlled fashion to combat noise in the channel [25, p. 184]. Consistently, the law of abbreviation prevails in short range communication [9, 14].

3.1 Compression versus random typing

Simple mechanisms such as random typing [40, 41, 42, 43], can reproduce the law of abbreviation [44]. Random typing models produce "words" by concatenating units, one of them behaving as "word" delimiter. Some researchers regard random typing as not involving any optimisation at all [42, 43]. However, its conceivable that the dynamical rules of random typing arise or are reinforced or stabilised by compression, given

- The equivalence between the law of abbreviation and compression outlined above.

- The optimality of the nonsingular coding scheme in the definition of random typing models (see supplementary online information for further details).

In that case, random typing could be seen as a special manifestation of compression. Another connection between optimisation and random typing is that stringing subunits to form "words" as in random typing is a convenient strategy to combat noise in communication [18].

Having said this, it is unlikely that random typing is the mediator between compression and the law of abbreviation in human languages. A serious limitation of random typing models is that the probability of a word is totally determined by its composition. In simple versions of the model [41, 42, 43], the probability of a word is determined by its length (the characters constituting the words are irrelevant), i.e. the length of a word (l) is a decreasing linear function of its probability (p)

$$l = a \log p + b, \tag{15}$$

where a and b are constants ($a < 0$). To see it, notice that the probability of a "word" w is [45, p. 838]

$$p(w) = \left(\frac{1-p_s}{N}\right)^l \frac{p_s}{(1-p_s)^{l_0}}, \tag{16}$$

where l is the length of w, p_s is the probability of producing the word delimiter, N is the size of the alphabet that the words consist of ($N > 1$) and l_0 is a binary parameter indicating the minimum word length ($l_0 \in \{0, 1\}$). Eq. 16 allows one to express l as a function of $p(w)$. Rearranging the terms of Eq. 16, taking logarithms, and replacing $p(w)$ by p, one recovers Eq. 15 with

$$a = \left(\log \frac{1-p_s}{N}\right)^{-1} \tag{17}$$

and

$$b = a \log \frac{(1-p_s)^{l_0}}{p_s}. \tag{18}$$

Another limitation of random typing is that it is not a plausible model for human language from a psychological perspective [16] and also from a social perspective: the "words" produced by random typing are not constrained by a predetermined vocabulary of words whose meanings have been agreed upon by social interaction among individuals as in human language [46].

Also, comparing the statistical properties of random typing against real languages reveals striking differences:

- The distribution of "word" frequencies deviates from that of actual word frequencies significantly [16].
- While random typing yields a geometric distribution of "word" lengths, the actual distribution of word lengths is not a monotonically decreasing function of l [47, 48].
- In the simple versions of the model, words of the same length are equally likely, a property that real languages do not satisfy [49, 50].

Another challenge for random typing are homophones: words that have the same composition (i.e. the same sequence of phonemes, and thus, the same length in phonemes) but can have different frequencies. Interestingly, given a pair of homophones the more frequent one tends to have a shorter duration in time [27]. This is impossible in random typing but explainable by the minimization of Λ. However, random typing might be relevant for the finding of the law of abbreviation in bird song [44], where the need for social consensus about the meaning of a song is missing or secondary while pressure for song diversity is crucial to maximise the chances of mating [51].

3.2 Optimisation and correlation in community detection

We have shown above an intimate relationship between Λ and r. A straightforward relationship between a function to optimise and correlation is also found in methods for community detection in networks [52]. A central concept in those methods is Q, a measure of the quality of a partitioning into communities of a network that must be maximised [52]. Interestingly, Q is intimately related with Fisher's intraclass correlation [53], another correlation coefficient that should not be confused with the popular Pearson interclass correlation r that we have considered above. To see it in detail, suppose that m is the number of edges of a network, k_i is the degree of the i-th vertex, A is the adjacency matrix and c_i is the community to which the i-th vertex belongs, Q can be defined as [54, p. 224]

$$Q = \frac{1}{2m} \sum_i \sum_j \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j), \tag{19}$$

where $\delta(x, y)$ is the Kronecker delta ($\delta(x, y) = 1$ if $x = y$; $\delta(x, y) = 0$ otherwise). The intraclass correlation that is connected with Q is defined between the communities at both ends of an edge. If x_i is a scalar quantity associated to the i-th vertex, the intraclass correlation between x_i and x_j over edges is $r_{intra} = cov(x_i, x_j)/\sigma_x^2$, where σ_x is the standard deviation of the x_i's and $cov_{intra}(x_i, x_j)$ is the intraclass covariance between x_i and x_j over edges, i.e. [54, p. 228]

$$cov_{intra}(x_i, x_j) = \frac{1}{2m} \sum_i \sum_j \left(A_{ij} - \frac{k_i k_j}{2m} \right) x_i x_j. \quad (20)$$

The similarity between Q and $cov_{intra}(x_i, x_j)$ is strong: the only difference is that $\delta(c_i, c_j)$ in Q is replaced by $x_i x_j$ in $cov_{intra}(x_i, x_j)$ [54, p. 228]. For the case of only two communities, one may define a variant of Q, namely

$$Q' = \frac{1}{2m} \sum_i \sum_j \left(A_{ij} - \frac{k_i k_j}{2m} \right) s_i s_j, \quad (21)$$

where $s_i \in \{-1, +1\}$, indicates the community of the i-th vertex. Then, $Q' = cov(s_i, s_j)$. Thus, the maximisation of Q' is fully equivalent to the maximisation of r_{intra}.

3.3 Applications beyond the law of abbreviation

The results presented in this article go beyond Zipf's law of abbreviation. For instance, the online memory cost of a sentence of n words can be defined as [28, 55]

$$D = (n-1) \sum_{d=1}^{n-1} p(d) g(d), \quad (22)$$

where $n-1$ is the number of edges of a syntactic dependency tree of n vertices, $p(d)$ is the proportion of dependencies of length d and $g(d)$ is the cognitive cost of a dependency of length d. Recently, it has been argued that $g(d)$ may not be a monotonically decreasing function of d as commonly believed [23]. The minimization of D can be regarded as particular case of Λ where $V = n-1$, p_i is $p(d)$ and $g(l_i)$ is $g(d)$ and thus a negative correlation between $p(d)$ and $g(d)$ is predicted applying the arguments employed in this article. Finally, knowing that $p(d)$ is a decreasing function of d in real syntactic dependencies [22, 56] and under the null hypothesis that the words of a sentence are arranged linearly at random [22], a positive correlation between d and $g(d)$ follows.

3.4 Causality

A challenge for our theoretical arguments is the extent to which the minimization of Λ is a causal force for the emergence of Zipf's law of abbreviation.

The apparent universality of the law of abbreviation in languages and the multiple theoretical connections between compression and the law suggest that the minimization of Λ is indeed a causal force. Additional support for compression as cause may come from the investigation of its predictions in grammaticalisation, a process of language change by which words become progressively more frequent and shorter [57, 58]. In the worst case (i.e. compression is not the driving force), compression would still illuminate the optimality of the law of abbreviation and its stability once it was reached through a mechanism unrelated to compression.

Authors' contributions

RFC conceived the mathematical work. RFC and CB drafted the manuscript. RFC and CS performed the mathematical work. CS and CB revised critically the article. All authors gave final approval for publication.

Acknowledgements

We thank M. E. J. Newman for indicating the connection between modularity and correlation, and S. Semple, G. Agoramoorthy and M.J. Hsu for the opportunity to use the macaque data for Fig. 1 (b). Special thanks to M. Arias and L. Debowski for helping us to strengthen some of the mathematical proofs. We are also grateful to M. Arias, N. Ay, L. Debowski, M. Gustison, A. Hernández-Fernández and S. Semple for valuable discussions. RFC is funded by the grants 2014SGR 890 (MACDA) from AGAUR (Generalitat de Catalunya) and also the APCOM project (TIN2014-57226-P) from MINECO (Ministerio de Economia y Competitividad). CB is funded by an Arts and Humanities Research Council (UK) doctoral grant and Cambridge Assessment (reference number: RG 69405), as well as a grant from the Cambridge Home and European Scholarship Scheme. CS is funded by an Erasmus Mundus master scholarship granted by the Education, Audiovisual and Culture Executive Agency of the European Commission.

References

[1] Zipf GK. Human behaviour and the principle of least effort. Cambridge (MA), USA: Addison-Wesley; 1949.

[2] Bates E, D'Amico S, Jacobsen T, Székely A, Andonova E, Devescovi A, et al. Timed picture naming in seven languages. Psychonomic Bulletin & Review. 2003;10:344–380.

[3] Sigurd B, Eeg-Olofsson M, van Weijer J. Word length, sentence length and frequency - Zipf revisited. Studia Linguistica. 2004;58(1):37–52.

[4] Strauss U, Grzybek P, Altmann G. Word length and word frequency. In: Grzybek P, editor. Contributions to the science of text and language. Dordrecht: Springer; 2007. p. 277–294.

[5] Jedličková K, Nemcová E. Word length and word frequency in Slovak. Glottotheory. 2008;1(1):25–30.

[6] Sanada H. Investigations in Japanese historical lexicology. Göttingen: Peust & Gutschmidt Verlag; 2008.

[7] Jayaram BD, Vidya MN. The relationship between word length and frequency in Indian languages. Glottotheory. 2009;2(2):62–69.

[8] Piantadosi ST, Tily H, Gibson E. Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences. 2011;108(9):3526–3529.

[9] Ferrer-i-Cancho R, Hernández-Fernández A. The failure of the law of brevity in two New World primates. Statistical caveats. Glottotheory. 2013;4(1).

[10] Evans N, Levinson SC. The myth of language universals: language diversity and its importance for cognitive science. Behavioral and Brain Sciences. 2009;32:429–492.

[11] Hailman JP, Ficken MS, Ficken RW. The 'chick-a-dee' calls of *Parus atricapillus*: a recombinant system of animal communication compared with written English. Semiotica. 1985;56:121–224.

[12] Ferrer-i-Cancho R, Lusseau D. Efficient coding in dolphin surface behavioral patterns. Complexity. 2009;14(5):23–25.

[13] Semple S, Hsu MJ, Agoramoorthy G. Efficiency of coding in macaque vocal communication. Biology Letters. 2010;6:469–471.

[14] Luo B, Jiang T, Liu Y, Wang J, Lin A, Wei X, et al. Brevity is prevalent in bat short-range communication. Journal of Comparative Physiology A. 2013;199:325–333.

[15] Naranan S, Balasubrahmanyan VK. Information theory and algorithmic complexity: applications to linguistic discourses and DNA sequences as complex systems. Part I: Efficiency of the genetic code of DNA. J Quantitative Linguistics. 2000;7(2):129–151.

[16] Ferrer-i-Cancho R, Elvevåg B. Random texts do not exhibit the real Zipf's-law-like rank distribution. PLoS ONE. 2009;5(4):e9411.

[17] Kirby S. Function, selection and innateness. The emergence of language universals. Oxford: Oxford University Press; 1999.

[18] Plotkin J, Nowak M. Language Evolution and Information Theory. J theor Biol. 2000;205:147–159.

[19] Prokopenko M, Ay N, Obst O, Polani D. Phase transitions in least-effort communications. J Stat Mech. 2010;p. P11025.

[20] Ferrer i Cancho R. Zipf's law from a communicative phase transition. European Physical Journal B. 2005;47:449–457.

[21] Ferrer-i-Cancho R. The optimality of attaching unlinked labels to unlinked meanings. http://arxivorg/abs/13105884. 2013;.

[22] Ferrer-i-Cancho R. Euclidean distance between syntactically linked words. Physical Review E. 2004;70:056135.

[23] Alday P. Be careful when assuming the obvious. Commentary on "The placement of the head that minimizes online memory: a complex systems approach". Language Dynamics and Change. 2015;5(1):147–155.

[24] Christiansen MH, Chater N. The now-or-never bottleneck: a fundamental constraint on language. Behavioral and Brain Sciences. 2015;p. in press.

[25] Cover TM, Thomas JA. Elements of information theory. New York: Wiley; 2006. 2nd edition.

[26] Ferrer-i-Cancho R, Hernández-Fernández A, Lusseau D, Agoramoorthy G, Hsu MJ, Semple S. Compression as a universal principle of animal behavior. Cognitive Science. 2013;37(8):1565–1578.

[27] Gahl S. "Thyme" and "Time" are not homophones. Word durations in spontaneous speech. Language. 2008;84:474–496.

[28] Ferrer-i-Cancho R. The placement of the head that minimizes online memory: a complex systems approach. Language Dynamics and Change. 2015;5:114–137.

[29] Dickman R, Moloney NR, Altmann EG. Analysis of an information-theoretic model for communication. Journal of Statistical Mechanics: Theory and Experiment. 2012;2012(12):P12022.

[30] Conover WJ. Practical nonparametric statistics. New York: Wiley; 1999. 3rd edition.

[31] Zou K, Tuncali K, Silverman SG. Correlation and simpler linear regression. Radiology. 2003;227:617–628.

[32] Kruskal WH. Ordinal Measures of Association. Journal of the American Statistical Association. 1958;53:814–861.

[33] Altmann EG, Gerlach M. Statistical laws in linguistics. http://arxivorg/abs/150203296. 2015;.

[34] Ferrer-i-Cancho R, Hernández-Fernández A, Baixeries J, Dębowski Ł, Mačutek J. When is Menzerath-Altmann law mathematically trivial? A new approach. Statistical Applications in Genetics and Molecular Biology. 2014;13:633–644.

[35] Covington MA, McFall JD. Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR). Journal of Quantitative Linguistics. 2010;17:94–100.

[36] McCowan B, Hanser SF, Doyle LR. Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. Anim Behav. 1999;57:409–419.

[37] Daniels HE. Rank correlation and population models. Journal of the Royal Statistical Society, Series B. 1950;12:171–81.

[38] Durbin J, Stuart A. Inversions and rank correlations. Journal of the Royal Statistical Society, Series B. 1951;13:303–309.

[39] Fredricks GA, Nelsen RB. On the relationship between Spearman's rho and Kendall's tau for pairs of continuous random variables. Journal of Statistical Planning and Inference. 2007;137(7):2143 – 2150.

[40] Conrad B, Mitzenmacher M. Power laws for monkeys typing randomly: the case of unequal probabilities. IEEE Transactions on Information Theory. 2004;50(7):1403–1414.

[41] Miller GA. Some effects of intermittent silence. Am J Psychol. 1957;70:311–314.

[42] Miller GA, Chomsky N. Finitary models of language users. In: Luce RD, Bush R, Galanter E, editors. Handbook of Mathematical Psychology. vol. 2. New York: Wiley; 1963. p. 419–491.

[43] Li W. Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution. IEEE T Inform Theory. 1992;38(6):1842–1845.

[44] Ficken MS, Hailman JP, Ficken RW. A model of repetitive behaviour illustrated by chickadee calling. Animal Behaviour. 1978;26(2):630–631.

[45] Ferrer-i-Cancho R, Gavaldà R. The frequency spectrum of finite samples from the intermittent silence process. Journal of the American Association for Information Science and Technology. 2009;60(4):837–843.

[46] Baronchelli A, Felici M, Caglioti E, Loreto V, Steels L. Sharp Transition towards Shared Vocabularies in Multi-Agent Systems. Journal of Statistical Mechanics. 2006;p. P06014.

[47] Newman MEJ. Power laws, Pareto distributions and Zipf's law. Contemporary Physics. 2005;46:323–351.

[48] Manin DY. Mandelbrot's model for Zipf's law: can Mandelbrot's model explain Zipf's law for language. Journal of Quantitative Linguistics. 2009;16(3):274–285.

[49] Leopold E. Frequency Spectra within Word-Length Classes. J Quantitative Linguistics. 1998;5(3):224–231.

[50] Ferrer i Cancho R, Solé RV. Zipf's law and random texts. Advances in Complex Systems. 2002;5:1–6.

[51] Catchpole CK, Slater PJB. Bird Song: Biological Themes and Variations. Cambridge, UK: Cambridge University Press; 1995.

[52] Fortunato S. Community detection in graphs. Physics Reports. 2010;486(3âĂŞ5):75 – 174.

[53] Koch GG. Intraclass correlation coefficient. In: Kotz S, Johnson NL, editors. Encyclopedia of Statistical Sciences. vol. 4. New York: John Wiley & Sons; 1982. p. 213–217.

[54] Newman MEJ. Networks. An introduction. Oxford: Oxford University Press; 2010.

[55] Ferrer-i-Cancho R. A stronger null hypothesis for crossing dependencies. Europhysics Letters. 2014;108:58003.

[56] Liu H. Probability distribution of dependency distance. Glottometrics. 2007;15:1–12.

[57] Heine B, Kuteva T. Genesis of grammar: a reconstruction. Oxford: Oxford University Press; 2007.

[58] Bybee JL, Perkins R, Pagliuca W. The evolution of grammar: tense, aspect and modality in the language of the world. Chicago: University of Chicago Press; 1994.

[59] Mandelbrot B. Information theory and psycholinguistics: a theory of word frequencies. In: Lazarsfield PF, Henry NW, editors. Readings in mathematical social sciences. Cambridge: MIT Press; 1966. p. 151–168.

[60] Cocho G, Flores J, Gershenson C, Pineda C, Sánchez S. Rank Diversity of Languages: Generic Behavior in Computational Linguistics. PLoS ONE. 2015 04;10(4):e0121898.

[61] Gerlach M, Altmann EG. Stochastic model for the vocabulary growth in natural languages. Physical Review X. 2013;3:021006.

[62] Petersen AM, Tenenbaum J, Havlin S, Stanley HE, Perc M. Languages cool as they expand: Allometric scaling and the decreasing need for new words. Scientific Reports. 2012;2(943).

[63] Ferrer i Cancho R, Solé RV. Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited. Journal of Quantitative Linguistics. 2001;8(3):165–173.

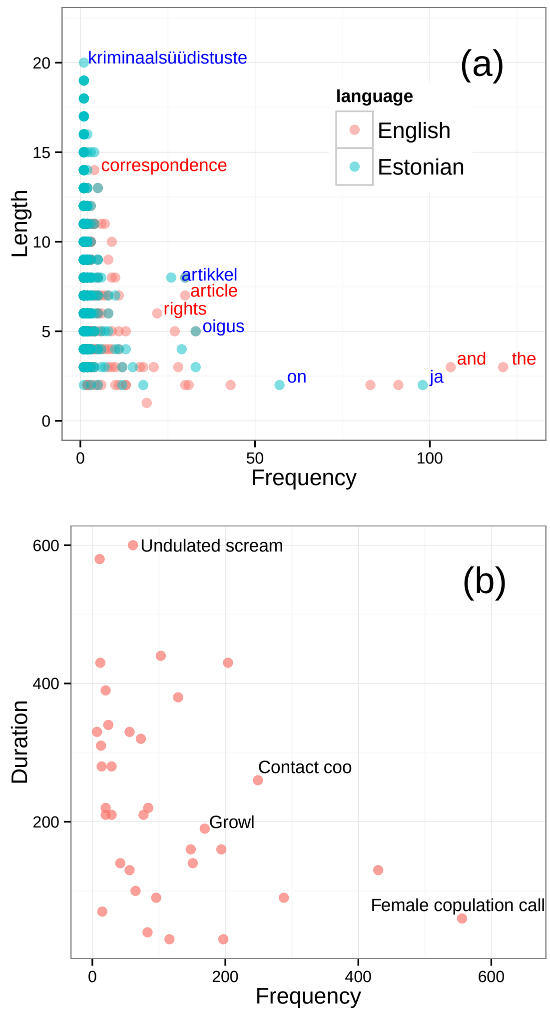

Figure 1: The law of abbreviation in human languages and Formosan macaques. (a) The relationship between word frequency and word length (in characters) for the English and Estonian translations of Universal Declaration of Human Rights (`http://www.unicode.org/udhr/`). The Estonian words chosen are *ja* (and), *on* (the), *oigus* (right), *artikkel* (article) and *kriminaalsüüdistuste* (criminal prosecutions). (b) The relationship between call type frequency call type mean duration (in ms) in Formosan macaques (data borrowed from [13]).

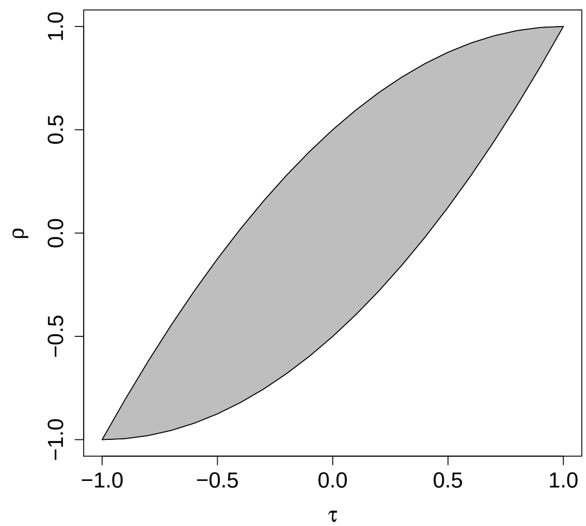

Figure 2: The gray region covers the values of Spearman ρ satisfying Eq. 14 as a function of Kendall τ.

SUPPLEMENTARY ONLINE INFORMATION

A Optimal coding

We have put forward a cost function Λ, i.e.

$$\Lambda = \sum_{i=1}^{V} p_i \lambda_i, \tag{23}$$

where p_i and λ_i are, respectively, the probability and the energetic cost of the i-th type. We have assumed that the energetic cost of a type is a monotonically increasing function of its magnitude l_i, i.e. $\lambda_i = g(l_i)$. When $g(l_i) = l_i$ and l_i is the length in symbols of the alphabet, Λ becomes L, the mean code length of standard information theory [25].

Here investigate the minimization of Λ when the $p_i{}'s$ are given.

A.1 Optimal coding without any constraint

The solution to the minimization of Λ when no further constraint is imposed is that all types have minimum magnitude, i.e.

$$l_i = l_{min} \text{ for } i = 1, 2, ..., V. \tag{24}$$

If l_i is the length of the i-th symbol with $l_i \geq 1$, then $l_{min} = 1$. If confusion between types has to be avoided, the unconstrained minimisation of Λ implies that N, the size of the alphabet used to build strings of symbols, cannot be smaller than V.

A.2 Optimal coding with nonsingular codes

Standard information theory bears on the elementary assumption that different types cannot be represented by the same string of symbols [25]. Under the wide scheme of uniquely decipherable codes, standard information tell us that the minimization of L leads to [25]

$$l_i \propto \lceil -\log p_i \rceil, \tag{25}$$

which is indeed a particular case of Zipf's law of abbreviation.

Here we investigate the optimal coding using nonsingular codes, a superset of uniquely decipherable codes [25, p. 106]. The function to minimize is Λ, a generalization of L the mean code length of standard coding theory [25].

We consider the set of all the strings of symbols that can be built with an alphabet of size N. Suppose that we sort the strings by increasing length (the relative ordering of strings of the same length is arbitrary), thus the strings in positions 1 to N have length 1, the strings in positions $N+1$ to $N+N^2$ have length 2, and so on. Suppose that types to be coded are sorted by decreasing probability, i.e.

$$p_1 \geq p_2 \geq ... \geq p_V, \tag{26}$$

being p_i the probability of the i-th type. Suppose that we assign the i-th string to the i-th type for $i = 1, 2, ...V$. An example of this coding are random typing models [42, 59, 43, 40]. We will show that this coding method, which we refer to as method A is optimal. We will proceed in two steps. First, recall that a the pairs (p_i, l_i) and (p_j, l_j) are concordant if and only if $p_i < p_j$ and $l_i < l_j$ or $p_i > p_j$ and $l_i > l_j$. If the pairs are concordant pair then swapping l_i and l_j will decrease Λ strictly (as explained in the main article). Thus, a requirement for optimal coding is the absence of concordant pairs. Notice that, by definition, method A produces no concordant pairs. Second, suppose that

$$\Lambda^x = \sum_{i=1}^{V} p_i \lambda_i^x, \tag{27}$$

where λ_i^x is the cost of the i-th type according to some coding method x. We will show by induction on V that for any alternative method B that is based on nonsingular codes and that does not produce concordant pairs, $\Lambda^B \geq \Lambda^A$.

- *Setup.* Since method A only produces pairs that are either concordant or discordant and probabilities obey Eq. 26, we have

  $$\lambda_1^A \leq ... \leq \lambda_i^A ... \leq \lambda_V^A. \tag{28}$$

  Suppose that method B has not produced any probability tie. Then

  $$\lambda_1^B \leq ... \leq \lambda_i^B ... \leq \lambda_V^B \tag{29}$$

  for the same reason as method A. If method B has produced probability ties, Eq. 29 may not hold. Suppose $V = 3$, $p_1 = p_2 = p_3$ and $l_1^B = 2$, $l_2^B = l_3^B = 1$. This coding lacks concordant pairs but does not satisfy Eq. 29. However, any coding produced by method B can be converted into one that satisfies Eq. 29 with the same Λ by sorting all magnitudes increasingly in every probability tie. We assume that the codings produced by method B have been rearranged in this fashion. This is crucial for the inductive step.

- *Basis*: $V = 1$. Then

$$\Lambda^A = p_1 \lambda_1^A = p_1 g(1) \tag{30}$$

$$\Lambda^B = p_1 \lambda_1^B = p_1 g(l_1). \tag{31}$$

  Recalling that g is a strictly monotonically increasing function, Eqs. 30 and 31 indicate that the condition $\Lambda^B \geq \Lambda^A$ is equivalent to $l_1 \geq 1$, which is trivially true by the definition of l_i.

- *Inductive hypothesis.* If Eq. 29 holds then

$$\sum_{i=1}^{V} p_i \lambda_i^B \geq \sum_{i=1}^{V} p_i \lambda_i^A \tag{32}$$

- *Inductive step* We want to show that

$$\sum_{i=1}^{V+1} p_i \lambda_i^B \geq \sum_{i=1}^{V+1} p_i \lambda_i^A \tag{33}$$

  when

$$\lambda_1^B \leq ... \leq \lambda_i^B ... \leq \lambda_{V+1}^B. \tag{34}$$

  Eq. 33 is equivalent to

$$S + p_{V+1}(\lambda_{V+1}^B - \lambda_{V+1}^A) \geq 0, \tag{35}$$

  with

$$S = \sum_{i=1}^{V} p_i(\lambda_i^B - \lambda_i^A). \tag{36}$$

  Let us define l_i^x as the magnitude of the i-th type according to some coding method x. To show that Eq. 35 holds, it suffices to show that

$$l_{V+1}^B \geq l_{V+1}^A, \tag{37}$$

  because $S \geq 0$ by the induction hypothesis (notice that if Eq. 34 holds then Eq. 29 also holds), p_{V+1} is positive by definition and $\lambda_i^x = g(l_i^x)$, where g is a strictly monotonically increasing function. Notice that if

$$l_{V+1}^B < l_{V+1}^A \tag{38}$$

  then method B would not employ nonsingular codes. To see it, notice that l_{V+1}^A is the smallest integer such that

$$V \leq \sum_{l=1}^{l_{V+1}^A} N^l, \tag{39}$$

where N^l are all the strings of length l that can be produced. Thus, method B must assign the same string to different types when Eq. 38 holds.

B The constancy conditions

B.1 Preliminaries: a lower bound for $E[\lambda]$

The optimal coding method presented above allows one to derive a lower bound for

$$E[\lambda] = \frac{1}{V}\sum_{i=1}^{V} \lambda_i. \tag{40}$$

The point is that $E[\lambda]$ can be seen as a particular case of Λ with $p_i = 1/V$. Suppose that l_{max} is the maximum string length needed by the optimal coding above. Obviously, l_{max} is the smallest integer such that

$$V \leq \sum_{l=1}^{l_{max}} N^l. \tag{41}$$

Such an optimal coding requires all strings of length smaller than l_{max} and

$$U = V - \sum_{l=1}^{l_{max-1}} N^l \tag{42}$$

strings of length l_{max}. Therefore, Eq. 40 gives

$$E[\lambda] \geq \frac{1}{V}\left(\sum_{l=1}^{l_{max}-1} g(l)\right) + Ug(l_{max}). \tag{43}$$

B.2 The constancy of V, $E[\lambda]$, $\sigma[\lambda]$ and $\sigma[p]$

The constancy of V could derive from the existence of a core lexicon [60, 61, 62, 63] and social constraints on the addition and propagation of new types [46].

The constancy of $E[\lambda]$ and $\sigma[\lambda]$ could be due to cost-cutting pressures. $VE[\lambda]$ can be regarded as the cost of learning the strings of symbols making the repertoire and storing them in memory. Suppose that every type is assigned a different string of symbols, i.e. the coding scheme is nonsingular [25]. One could use strings of length $\lceil \log_N V \rceil$ to code for every type but this would be a waste. A lower bound for $E[\lambda]$ is given by Eq. 43. We expect that natural systems are attracted towards this lower bound to minimize the cost of storing the repertoire, providing support for the simplifying assumption that $E[\lambda]$ is constant.

The constancy of $\sigma[\lambda]$ can be supported by the need of intermediate values of $\sigma[\lambda]$:

- A small value of $\sigma[\lambda]$ might be difficult or impossible to achieve. Let us consider that $\sigma[\lambda]$ is minimum, i.e. $\sigma[\lambda] = 0$, which is equivalent to all types having the same magnitude k. If $V < N^k$ then some types are not distinguishable (the coding scheme is nonsingular), which is something to avoid. Recall that the unconstrained solution to the minimisation of Λ, i.e. Eq. 24 yields $\sigma[\lambda] = 0$ but sacrificing the distinguishability of types (if $V > N$, distinguishability imposes that $\sigma[\lambda]$ is bounded below by a non-zero value). Although it is possible to code any repertoire with strings of length 1 from an alphabet (one only needs that $V = N$), strings of length greater than one have been shown to be evolutionary advantageous to combat noise [18]. If $V \geq N^k$ then the high cost implied by the value of $E[\lambda]$ (as explained above) turns $\sigma[\lambda] = 0$ unlikely. A further reason against $\sigma[\lambda] = 0$ is that, under pressure to minimize $E[\lambda]$ or Λ close to the optimal coding for nonsingular codes, all lengths up to l_{max} are taken.

- Let us consider that $\sigma[\lambda]$ is large (much greater than the value of $\sigma[\lambda]$ for nonsingular codes). Then very long strings are expected but this is unnecessarily costly and then less likely to happen.

Finally, the constancy of $\sigma[p]$ could arise from mechanisms that shape symbol probabilities independently from Λ but that can still involve cost-cutting factors [20, 47].

C The relationship between Λ and τ

Here we investigate the consequences of swapping p_i and p_j or l_i and l_j on Δ_Λ and Δ_τ, the discrete derivative of Λ and τ, respectively. For mathematical simplicity, we assume that there are no ties. Then i and j define a concordant pair or a discordant pair. Section C.1 presents a result on the discrete derivative of n_c which is crucial to conclude in Section C.2.2 that

- If the pairs (p_i, l_i) and (p_j, l_j) are concordant, $\Delta_\tau < 0$.

- If the pairs (p_i, l_i) and (p_j, l_j) are discordant, $\Delta_\tau > 0$.

Moreover, Section C.2.2 also shows that

- If the pairs (p_i, l_i) and (p_j, l_j) are concordant, $\Delta_\Lambda < 0$.
- If the pairs (p_i, l_i) and (p_j, l_j) are discordant, $\Delta_\Lambda > 0$.

C.1 The discrete derivative of the number of concordant pairs

Hereafter we keep i and j for the subindices of the pairs being swapped and use x and y for the subindices of pairs in general. n_c, the number of concordant pairs, can be defined as a summation over all pairs, i.e.

$$n_c = \sum_{x=1}^{V} \sum_{y=1}^{V} c_{xy}, \tag{44}$$

where c_{xy} is an indicator variable. $c_{xy} = 1$ if $p_x < p_y$ and $l_x < l_y$ for the x-th and the y-th type; $c_{xy} = 0$ otherwise. Thus, the pairs (p_x, l_x) and (p_y, l_y) are concordant if and only if c_{xy} or c_{yx}. c_{xy} can be expressed as a product of indicator variables, i.e.

$$c_{xy} = a_{xy} b_{xy}, \tag{45}$$

where a_{xy} indicates if $p_x < p_y$ and b_{xy} indicates if $l_x < l_y$.

The assumption that there are no ties gives a couple of valuable properties:

- If $x = y$, $a_{xy} = b_{xy} = 0$.
- If $x \neq y$

$$a_{xy} = 1 - a_{yx} \tag{46}$$
$$b_{xy} = 1 - b_{yx} \tag{47}$$

  by symmetry.

We define n'_c as the number of concordant pairs after the swap. A' and B' are the state of matrices $A = \{a_{xy}\}$ and $B = \{b_{xy}\}$ after the swap.

C.2 l_i and l_j are swapped

If only l_i and l_j are swapped, then $A' = A$, but $B' = B$ is not warranted. Interestingly, the changes in B concern only the i-th and the j-th row and the i-th and the j-th column, i.e.

$$b'_{xy} = \begin{cases} b_{ji} & \text{if } x \neq y \text{ and } x = i \text{ and } y = j \\ b_{ij} & \text{if } x \neq y \text{ and } x = j \text{ and } y = i \\ b_{jy} & \text{if } x \neq y \text{ and } x = i \text{ and } y \neq j \\ b_{iy} & \text{if } x \neq y \text{ and } x = j \text{ and } y \neq i \\ b_{xj} & \text{if } x \neq y \text{ and } x \neq j \text{ and } y = i \\ b_{xi} & \text{if } x \neq y \text{ and } x \neq i \text{ and } y = j \\ b_{xy} & \text{otherwise} \end{cases} \tag{48}$$

Then, the 1st derivative of n_c as a function of the number of swaps performed is

$$\Delta_{n_c} = n'_c - n_c. \tag{49}$$

Suppose that γ is sum of the values in the i-th and j-th row as well as in the i-th and j-th column of $C = \{c_{ij}\}$ (if a value is found in both a column and a row it will be summed only once) and γ' as the value of γ after the swap. Then the 1st derivative can also be defined as

$$\Delta_{n_c} = \gamma' - \gamma. \tag{50}$$

By definition,

$$\gamma = S_1 + S_2 + S_3 + S_4 - T \tag{51}$$

with

$$S_1 = \sum_{y=1}^{V} c_{iy} = \sum_{y=1}^{V} a_{iy} b_{iy} \tag{52}$$

$$S_2 = \sum_{y=1}^{V} c_{jy} = \sum_{y=1}^{V} a_{jy} b_{jy} \tag{53}$$

$$S_3 = \sum_{x=1}^{V} c_{xi} = \sum_{x=1}^{V} a_{xi} b_{xi} \tag{54}$$

$$S_4 = \sum_{x=1}^{V} c_{xj} = \sum_{x=1}^{V} a_{xj} b_{xj} \tag{55}$$

and

$$\begin{aligned} T &= c_{ij} + c_{ji} + c_{ii} + c_{jj}. \\ &= a_{ij} b_{ij} + a_{ji} b_{ji} + a_{ii} b_{ii} + a_{jj} b_{jj}. \end{aligned} \tag{56}$$

T is the sum of the values that have been summed twice by S_1, S_2, S_3 and S_4 in Eq. 51. Notice that $S_1 + S_2 + S_3 + S_4 = 2T$ when $V = 2$. Since $a_{xx} = b_{xx} = 0$,

$$T = a_{ij}b_{ij} + a_{ji}b_{ji}. \tag{57}$$

Recalling Eq. 47, S_3 can be expressed as

$$\begin{aligned} S_3 &= \sum_{y=1}^{V}(1 - a_{iy})(1 - b_{iy}) - (1 - a_{ii})(1 - b_{ii}) \\ &= \sum_{y=1}^{V}(1 - a_{iy})(1 - b_{iy}) - 1. \end{aligned} \tag{58}$$

Similarly, S_4 can be expressed as

$$\begin{aligned} S_4 &= \sum_{y=1}^{V}(1 - a_{jy})(1 - b_{jy}) - (1 - a_{jj})(1 - b_{jj}) \\ &= \sum_{y=1}^{V}(1 - a_{jy})(1 - b_{jy}) - 1. \end{aligned} \tag{59}$$

On the one hand,

$$S_1 + S_3 = \sum_{y=1}^{V} a_{iy}(2b_{iy} - 1) + V - \sum_{y=1}^{V} b_{iy} - 1. \tag{60}$$

On the other hand,

$$S_2 + S_4 = \sum_{y=1}^{V} a_{jy}(2b_{jy} - 1) + V - \sum_{y=1}^{V} b_{jy} - 1. \tag{61}$$

Thus, Eq. 51 becomes

$$\begin{aligned} \gamma &= \sum_{y=1}^{V} \left[a_{iy}(2b_{iy} - 1) + a_{jy}(2b_{jy} - 1)\right] - \sum_{y=1}^{V}(b_{iy} + b_{jy}) + \\ &\quad 2(V - 1) - a_{ij}b_{ij} - a_{ji}b_{ji}. \end{aligned} \tag{62}$$

By definition,

$$\gamma' = S'_1 + S'_2 + S'_3 + S'_4 - T' \tag{63}$$

with

$$S'_1 = \sum_{y=1}^{V} c'_{iy} = \sum_{y=1}^{V} a_{iy}b'_{iy} \tag{64}$$

$$S'_2 = \sum_{y=1}^{V} c'_{jy} = \sum_{y=1}^{V} a_{jy}b'_{jy} \tag{65}$$

$$S'_3 = \sum_{x=1}^{V} c'_{xi} = \sum_{x=1}^{V} a_{xi}b'_{xi} \tag{66}$$

$$S'_4 = \sum_{x=1}^{V} c'_{xj} = \sum_{x=1}^{V} a_{xj}b'_{xj} \tag{67}$$

and

$$\begin{aligned} T' &= c'_{ij} + c'_{ji} + c'_{ii} + c'_{jj} \\ &= a_{ij}b'_{ij} + a_{ji}b'_{ji} + a_{ii}b'_{ii} + a_{jj}b'_{jj}. \end{aligned} \tag{68}$$

T' is the sum of the values that have been summed twice by S'_1, S'_2, S'_3 and S'_4 in Eq. 63. Notice that $S'_1 + S'_2 + S'_3 + S'_4 = 2T'$ when $V = 2$. Applying Eq. 48, T' becomes

$$\begin{aligned} T' &= a_{ij}b_{ji} + a_{ji}b_{ij} + a_{ii}b_{ii} + a_{jj}b_{jj} \\ &= a_{ij}b_{ji} + a_{ji}b_{ij} \quad \text{(applying } a_{xx} = b_{xx} = 0) \\ &= a_{ij} + b_{ij} - 2a_{ij}b_{ij}. \quad \text{(applying } a_{ji} = 1 - a_{ij} \text{ and } b_{ji} = 1 - b_{ij}) \end{aligned} \tag{69}$$

Applying Eq. 48, S'_1 can be expressed as

$$\begin{aligned} S'_1 &= \sum_{y=1}^{V} a_{iy}b_{jy} - a_{ii}b_{ji} - a_{ij}b_{jj} + a_{ii}b_{ii} + a_{ij}(1 - b_{ij}) \quad \text{(applying } b_{ji} = 1 - b_{ij}) \\ &= \sum_{y=1}^{V} a_{iy}b_{jy} + a_{ij}(1 - b_{ij}) \quad \text{(applying } a_{xx} = b_{xx} = 0) \end{aligned} \tag{70}$$

while S'_2 can be expressed as

$$\begin{aligned} S'_2 &= \sum_{y=1}^{V} a_{jy}b_{iy} - a_{ji}b_{ii} - a_{jj}b_{ij} + a_{ji}b_{ij} + a_{jj}b_{jj} \\ &= \sum_{y=1}^{V} a_{jy}b_{iy} + (1 - a_{ij})b_{ij}. \end{aligned} \tag{71}$$

Similar arguments give

$$\begin{aligned} S'_3 &= \sum_{x=1}^{V} a_{xi}b_{xj} - a_{ii}b_{ij} - a_{ji}b_{jj} + a_{ii}b_{ii} + a_{ji}b_{ij} \\ &= \sum_{x=1}^{V} a_{xi}b_{xj} + (1 - a_{ij})b_{ij} \qquad (72) \\ S'_4 &= \sum_{x=1}^{V} a_{xj}b_{xi} - a_{ij}b_{ii} - a_{jj}b_{ji} + a_{ij}b_{ji} + a_{jj}b_{jj} \\ &= \sum_{x=1}^{V} a_{xj}b_{xi} + a_{ij}(1 - b_{ij}). \qquad (73) \end{aligned}$$

Recalling Eq. 47, S'_3 can be expressed as

$$\begin{aligned} S'_3 &= \sum_{y=1}^{V}(1 - a_{iy})(1 - b_{jy}) - (1 - a_{ii})(1 - b_{ji}) - (1 - a_{ij})(1 - b_{jj}) + \\ &\quad (1 - a_{ij})b_{ij} \qquad (74) \\ &= \sum_{y=1}^{V}(1 - a_{iy})(1 - b_{jy}) - b_{ij} - (1 - a_{ij}) + b_{ij} - a_{ij}b_{ij} \\ &= \sum_{y=1}^{V}(1 - a_{iy})(1 - b_{jy}) + a_{ij} - a_{ij}b_{ij} - 1. \qquad (75) \end{aligned}$$

Similarly, S'_4 can be expressed as

$$\begin{aligned} S'_4 &= \sum_{y=1}^{V}(1 - a_{jy})(1 - b_{iy}) - (1 - a_{ji})(1 - b_{ii}) - (1 - a_{jj})(1 - b_{ij}) + \\ &\quad a_{ij}(1 - b_{ij}) \\ &= \sum_{y=1}^{V}(1 - a_{jy})(1 - b_{iy}) - a_{ij} - (1 - b_{ij}) + a_{ij} - a_{ij}b_{ij} \\ &= \sum_{y=1}^{V}(1 - a_{jy})(1 - b_{iy}) + b_{ij} - a_{ij}b_{ij} - 1. \qquad (76) \end{aligned}$$

On the one hand,

$$S'_1 + S'_3 = \sum_{y=1}^{V} a_{iy}(2b_{jy} - 1) + V - \sum_{y=1}^{V} b_{jy} + s'_{1,3} \qquad (77)$$

with

$$
\begin{aligned}
s'_{1,3} &= a_{ij}(1-b_{ij}) + a_{ij} + b_{ji} - 2 \\
&= a_{ij}(1-b_{ij}) + a_{ij} + 1 - b_{ij} - 2 \\
&= 2a_{ij} - b_{ij} - a_{ij}b_{ij} - 1. \qquad (78)
\end{aligned}
$$

On the other hand,

$$
S'_2 + S'_4 = \sum_{y=1}^{V} a_{jy}(2b_{iy} - 1) + V - \sum_{y=1}^{V} b_{iy} + s'_{2,4} \qquad (79)
$$

with

$$
\begin{aligned}
s'_{2,4} &= a_{ji}(1-b_{ji}) + a_{ji} + b_{ij} - 2 \\
&= (1-a_{ij})b_{ij} + 1 - a_{ij} + b_{ij} - 2 \\
&= -a_{ij} + 2b_{ij} - a_{ij}b_{ij} - 1. \qquad (80)
\end{aligned}
$$

Finally, Eqs. 69, 77 and 79 transform Eq. 63 into

$$
\begin{aligned}
\gamma' &= \sum_{y=1}^{V} \left[a_{iy}(2b_{jy} - 1) + a_{jy}(2b_{iy} - 1)\right] - \sum_{y=1}^{V} (b_{iy} + b_{jy}) + \\
&\quad 2(V - a_{ij}b_{ij} - 1) + a_{ij} + b_{ij} \qquad (81)
\end{aligned}
$$

Combining Eqs. 62 and 81 yields, after some algebra,

$$
\begin{aligned}
\Delta_{n_c} &= \gamma' - \gamma \\
&= 2\sum_{y=1}^{V} \alpha_y \beta_y + 1 \qquad (82)
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha_y &= a_{iy} - a_{jy} \qquad (83) \\
\beta_y &= b_{jy} - b_{iy}. \qquad (84)
\end{aligned}
$$

The final formulae for γ (Eq. 62), γ' (Eq. 81) and Δ_{n_c} (Eq. 82) have been verified with the help of computer simulations. For a given V, the simulation is based on the following algorithm:

- Setup

    1. Generate a vector A of size V containing numbers from 1 to V (in that order).

2. Generate a vector B of size V containing numbers from 1 to V (in that order) if the initial state is $\tau = 1$ and containing those numbers in the reverse order if that state is $\tau = -1$.
3. Calculate n_c and γ with A and B.

- Test: run T times
  1. Choose uniformly at random two integers i and j such that $1 \leq i, j \leq V$ and $i \neq j$.
  2. Swap the i-th and the j-th element of B.
  3. Calculate n'_c and γ' with A and the new B.
  4. Check that
     (a) γ coincides with the value provided by Eq. 62.
     (b) γ' coincides with the value provided by Eq. 81.
     (c) $\Delta_{n_c} = n'_c - n_c$ coincides with the value provided by Eq. 82.
  5. $n_c = n'_c$, $\gamma = \gamma'$.

The algorithm was run successfully for $V = 1$ to 100 with $T = 10^5$ and both initial states.

C.2.1 p_i and p_j are swapped

Notice that this case is equivalent to the case when l_i and l_j are swapped by symmetry. It suffices to exchange the role of the matrices A and B. Now a_{xy} indicates if $l_x < l_y$ and b_{xy} indicates if $p_x < p_y$.

C.2.2 The variation of τ

If ties are missing, $n_d = n_0 - n_c$ and τ becomes

$$\tau = \frac{2n_c}{n_0} - 1. \tag{85}$$

Then the discrete derivative of τ (as a function of the number of swaps) is

$$\Delta_\tau = \frac{n'_c - n_c}{n_0} \tag{86}$$

$$= \frac{\Delta_{n_c}}{n_0}, \tag{87}$$

where Δ_{n_c}, the discrete derivative of n_c, satisfies

$$\Delta_{n_c} = 2\sum_{y=1}^{V}\alpha_y\beta_y + 1 \tag{88}$$

as explained in Section C.1.

Without any loss of generality, suppose that $i < j$. We want to prove that

Statement 1 If the pairs (p_i, l_i) and (p_j, l_j) are concordant then $\Delta_{n_c} < 0$, which is equivalent to

$$\sum_{y=1}^{V}\alpha_y\beta_y < -\frac{1}{2}. \tag{89}$$

Statement 2 If i and j are a discordant pair then $\Delta_{n_c} > 0$, which is equivalent to

$$\sum_{y=1}^{V}\alpha_y\beta_y > -\frac{1}{2}. \tag{90}$$

First, we notice some relevant properties of α and β. Suppose that $p_1 < p_2 < ... < p_k < ... < p_V$. Then it is easy to see that

Property 1 $\alpha_y = 1$ if $i < y < j$ and $\alpha_y = 0$ otherwise.

Now suppose that $l_1 < l_2 < ... < l_k < ... < l_V$. Then it is easy to see that

Property 2 $\beta_y = -1$ if $i < y < j$ and $\beta_y = 0$ otherwise.

If the pairs (p_i, l_i) and (p_j, l_j) i are concordant, Properties 1 and 2 indicate that $\alpha_y\beta_y \in \{-1, 0\}$, giving

$$\sum_{y=1}^{V}\alpha_y\beta_y \leq 0. \tag{91}$$

Adding that $\alpha_i\beta_i = -a_{ji}b_{ji} = -1$ or $\alpha_j\beta_j = -a_{ij}b_{ij} = -1$ when the pairs (p_i, l_i) and (p_j, l_j) are concordant, it follows that

$$\sum_{y=1}^{V}\alpha_y\beta_y \leq -1, \tag{92}$$

which proves Statement 1. If i and j are discordant, Properties 1 and 2 indicate that $\alpha_y\beta_y \in \{0, 1\}$, giving

$$\sum_{y=1}^{V}\alpha_y\beta_y \geq 0, \tag{93}$$

which proves Statement 2.

C.2.3 The variation of Λ

The value of Λ after one of those swaps will be examined considering two cases. First, imagine that l_i and l_j are swapped. The value of Λ after the swap is

$$\Lambda' = \Lambda - (p_i\lambda_i + p_j\lambda_j) + (p_i\lambda_i' + p_j\lambda_j'). \tag{94}$$

Applying $\lambda_i' = \lambda_j$ and $\lambda_j' = \lambda_i$ it is obtained

$$\Delta_\Lambda = (p_i - p_j)(g(l_j) - g(l_i)). \tag{95}$$

The fact that $g(l)$ is a monotonically increasing function of l allows one to conclude that

- If the pairs were concordant then $\Delta_\Lambda < 0$.
- If the pairs were discordant then $\Delta_\Lambda > 0$.

Second, imagine that p_i and p_j are swapped. The value of Λ after the swap is

$$\Lambda' = \Lambda - (p_i\lambda_i + p_j\lambda_j) + (p_i'\lambda_i + p_j'\lambda_j). \tag{96}$$

Applying $p_i' = p_j$ and $p_j' = p_i$ it is obtained again Eq. 95. Thus, the same arguments and conclusions apply to the swap of p_i and p_j.