Open Peer Review on Qeios

Modeling the structure and evolution of cultural information as Quasispecies

David Stevenson

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

We present a novel model of culture that directly relates biological evolution with modern aspects of cultural evolution. The model considers the high rate of error in communication and builds on structural and evolutionary similarities between biological molecules and written language. Firstly, both written language and biological molecules are modular. Within RNA and polypeptide molecules there are structural domains that may be recombined while maintaining their function. Likewise, sentences are structured as combinations of clauses, in which each clause contains a domain of information. The clausal structure permits the recombination of information to adopt different meanings, while allowing each unit to retain its identity. Secondly, we show that some, but not all, aspects of communicated culture have a high error rate, ensuring that information exists as rapidly evolving clouds within the population. Through their intrinsically high rate of mutation, clouds of cultural information are analogous to viral quasispecies and may be modelled as such. We then integrate these ideas with the application of Shannon Diversity Index to produce a more holistic view of culture that is centered on the evolution of information. Re-imagining culture, as evolving clouds of information, unifies the mode in which information is stored culturally and biologically, and opens up new avenues of comparative analysis.

David S Stevenson

Carlton le Willows Academy

Wood Lane

Gedling

NG4 4AA

UK

Keywords: Information; language; communication; quasispecies, diversity.

1. Introduction

Cultural evolution has a rich history, extending back to Charles Darwin (reviewed in[1]). While the mechanisms of cultural evolution have been discussed and contended extensively [1][2][3][4][5][6][7][8][9], there has been a tendency of the community to split into Darwinist and non-Darwinist camps, with the latter citing alleged differences in the manner in which cultural information spreads as a negation of the underlying mechanism [3][10]. For example, Claidière *et al.* [2] contend that a student copying a lecture and correcting the lecturer's mistake partly invalidates straightforward analogies between biological and cultural evolution. However, this is untrue. Firstly, in reading and transcribing the lecture notes, the student is producing an internal copy: the information has reproduced.

However, critically, the information has also undergone what we can call "error-prone repair" using another, accurate copy of the information held in the student's memory. In biological, error-prone repair of DNA a stock template is used in one or more ways to provide, erroneously, a copy which replaces a damaged section of information [11]. In the case of the student, it is more likely that they will erroneously copy a lecture note, thereby, generating a mutant copy of the original information. Moreover, as we are aware of multiple "possible", if erroneous, spellings of some words, our memories have or can generate a population of possible spellings. Finally, the internal cognitive systems of the student must select, then express, the correct copy of the information, thereby, producing another verbal or written copy. Therefore, in every sense, the flow of cultural information is analogous to that observed in biological systems.

Smith (2011) was one of the first to quantify the error rate in human communication. His, and later work (Methods) shows that the rate of error commonly exceeds that in RNA viral replication. As such, the term "quasispecies" is applicable when describing the evolution of communicated information. Quasispecies was initially coined to describe populations of molecules in solution. In 1971 Eigen applied the term to describe populations of closely-related RNA molecules [12][13][14][15][16][17]. A quasispecies has some fundamental features that make it applicable to the study of areas as diverse as bacterial populations [18], language evolution [9][19] and, here, as a component of a cultural evolution model.

The critical feature of the quasispecies model is that sequences experience a high error rate in replication, so that any particular sequence of information (allele) may be recreated at non-trivial rates by mutation from other, related sequences. That means that all sequences exist as a fast-evolving cloud of related sequences. Sequences evolve to form a landscape of iterations of varying "fitness", with the fittest replicating at the highest rates. In the viral quasispecies, mutation rates are sufficiently high that the rate of forward and backward creation of sequences by mutation are comparable. Rather than discrete peaks in the fitness landscape, clouds of the fittest sequences form broader plateaus [12][16][19][20]. The observed pattern of fitness is often colloquially referred to as "survival of the flattest" [12]. In viral quasispecies, the fittest variants can, therefore, emerge rapidly and be selected by the environment, so that, for example, drug-resistant variants rapidly come to dominate populations. Likewise, in aspects of culture, where there is rapid replication of information, coupled to a high underlying rate of mutation, variants can arise readily which will come to dominate that subcultural population's language, given appropriate selection. In the absence of selection, the large pool of variants may reasonably be described as a cultural pangenome [21].

Here, we show that the continuity of mutation rates from zero to the high rates seen in social media (below), alters the spread of that information. We quantify many of the key variables necessary for the comparison and discuss the relationship between our model and pragmatic models of information [22][23][24][25]. Finally, we discuss how the manner in the way information is structured facilitates its evolution coupled to a growth in Shannon entropy [26].

2. Methodology: identification of key variables in social evolution of text

2.1. Mutation rates

Underpinning the quasispecies model is a high rate of mutation. Here, we obtain error rates in communication from a variety of different sources: task completion (Smith, 2011); Tweets [27][28]; Facebook [29]; Texting [30][31]; and finally, second language learning [32].

Table 1. A synopsis of the error frequency and type reproduced from Smith [33].

Error Type	Error Frequency Per Task
Read Errors	
Read analogue display wrongly	0.005 (0.5%)
Read digital display wrongly	0.006 (0.6%)
Read checklist incorrectly	0.001 (0.1%)
Read 10-digit number incorrectly	0.006 (0.6%)
Read alphanumeric (single character) incorrectly	0.0002 (0.02%)
Read (clear) 5-letter word incorrectly	0.0003 (0.03%)
Read (unclear) 5-letter word incorrectly	0.03 (3%)
Write-Errors	
Record information wrongly	0.01 (1%)
Type a character wrongly	0.01 (1%)
Enter 10 digits in calculator incorrectly	0.05 (5%)
Dial 10 digits incorrectly	0.06 (6%)
Response Errors	
Fail to respond to annunciation	0.0001 (0.01%)
Wrongly carry out visual inspection	0.003 (0.3%)
Fail to act after 1 minute in an emergency	0.9 (90%)

If the mutation (error) rate is significant, then the frequency of errors should be proportional to the lexicon size (our language's "genome"), as follows:

$$L < 1/(1 - q) \qquad (1)$$

Here, L is the size of the lexicon and q is the probability of an error arising in its replication. The size of an organism's genome may be no greater than the reciprocal of the mutation rate of information contained in it. [19][34].

2.2. Replication rates of information

Reproduction of information occurs through internal duplication of information during thought processes and external duplication in media as diverse as speech, printed written works, and on various social media platforms, to name a few. In the former, different pieces of information held in memory may be duplicated and recombined – as well as becoming scrambled or deleted over time [35]. Quantifying the number of internal replications is difficult, but will depend on how frequently information is recalled.

External replication is more readily quantifiable and will vary from 10^1 copies per year for some books and printed works of art, to several 10^6 for viral Tweets [27][36]. Rates of reproduction (per year) will, therefore, fall in the range 10^1-10^7/annum.

2.3. Sentences modularity, relationship with biomolecules and effect on mutation viability

Biological units of inheritance are alleles. Here, we employ the term*cultural alleles* to discuss the idea of units of cultural information. As with biological information, cultural alleles can be envisaged as descriptors of cultural items or processes, analogous to a functional gene that encodes a polypeptide or functional RNA [37][38][39][40]. As with genes, the sequence of information in sentence may be sub-divided into units, clauses, each of which encodes a domain of information. Clauses can be exchanged between sentences and retain their meaning, as long as they do not disrupt the grammatical structure of that sentence [26]. However, such recombination can alter the meaning of the sentence as a whole. Each clause will be subject to its own selection pressures, determining how likely that variant will be retained. Moreover, these distinct structural features can interact with one another in a manner analogous to social networks [40][41].

2.4. Sentence length and Hamming Distance

The Hamming Distance is the number of differences between related sequences of information and, in effect, conveys how easy it is to change one sequence to another [16]. For the work presented, here, we use one of the following determinations of Hamming Distance: either each word difference or each clause difference (per sentence) has a Hamming Distance of one. An alternative but related measure is the Levenshtein Distance (Holman et al., 2011; Wichmann and Holman, 2022). The Hamming distance is a more general measure of variation; however, the model could be adapted to include the alternative.

The following pair of related words illustrates the difference in the output when using Hamming Distance and Levenshtein Distance: *flaw* and *lawn*. Here, the Levenshtein distance (LD) equals*two*, because there are two-character differences (*f* versus *n*), between each of the two words. However, the Hamming distance is*four* because the letters at each *position* in the four-letter word are different to the corresponding letter in the other member of the pair.

Sentences may have any length, as long as they follow certain grammatical structures: typically, one noun and one verb and some connectives and descriptives. Sentences that are less than 10 words in length sound very clunky. Conversely, sentences longer than 35 words are demanding of memory and interpretation [42][43]. Therefore, while variation in sentence length leads to more interesting and readable text, typically the average sentence, in any body of text, consists of the order of 20 words [44]. For the purposes of this work, we assume a mean sentence length of 20 words (L = 20), comprising two clauses. Table 2 gives some examples.

Table 2. Two, simple examples of diversity affecting sentence meaning by changing one clause within it.

Initial sentence	Plausible variants with the same meaning	Plausible variants with an alternative meaning
The cat sat on the matt, purring contentedly.	The cat *lay* on the matt, purring contentedly.	The cat sat on the matt, *hissing angrily.*
The dog ate the ice cream and was sick.	The dog *consumed* the ice cream and was sick.	The dog ate the ice cream and *was happy.*

In terms of Hamming Distance, each of the variants in columns two and three (Table 1), has a difference of one clause, with the variants in column 2 having also a single word variant; while column three has two words differences. Depending on what unit we are judging the Hamming Distance is one or two, relative to the original sentences.

2.5. Selection

Selection coefficients describe the effect of the environment on the relative fitness of an organism, its phenotype or specific genic sequences (alleles) within it [45][46]. Here, we consider selection coefficients as describing the propensity of a sequence of cultural information to replicate [46].

There is a substantial body of literature describing the identification of selection pressures and their effects in culture (e.g., [47][48]). Therefore, in order to avoid further expansion of the current work, we only note that selection coefficients may be included in quasispecies models [20] and that there are a number of different methodologies available to determine the value of these coefficients in behavioral settings [35][49][50][51][52] or in literature reviews [53].

In Latané’s work (1981; refs [49][50][51]), the scaling constant, *t*, would have a value less than 1[1]. Rewriting Latané’s equation, we assume that the base response to emulate an act is its fitness. These changes give us equation 5a, and its linear equivalent, 5b, as follows:

$$I = cN^{s} \qquad (2a)$$

$$lnI = slnN + lnC \qquad (2b)$$

In the converse situation, where we wish to measure the impact of one person's actions or words on a group, the equation can be rearranged as follows:

$$I = \frac{c}{N^s} \qquad (3a)$$

$$lnI = lnc - slnN \qquad (3b)$$

While we are not pursuing this approach further, here, we wish to note its functionality for in subsequent analysis, where observational data is available.

Selection of information may be expected to depend on a number of factors, including, but not limited to: the presence of pre-existing schemas; the health and well-being of the person; the cultural availability of information and the availability of free-energy [54][55][56][57]. Selection of information (based on its usefulness) then determines in part, whether received information will be retained or forgotten [58].

The subsequent decision to express information will strongly depend on external selection pressures, most likely whether expression will result in positive feedback from those in the surrounding peer-group (s) or wider culture (conformity and other forms of positive social pressure); but also, where there is an absence of negative feedback, or negative selection [52].

In a cultural setting, rather than through examination of the retention, replication of elimination of cultural information, we may derive fitness and/or selection from the relationship between the number of stimuli required to elicit a response and the strength of that response: effectively, conformity. Biological selection of information reflects the release of various neurotransmitters in recipients in response to stimuli [59][60][61][62]. We suggest that, while not examined here, analysis of neurological activity would provide a useful window on biological selection of information in future work.

2.6. Application of the quasispecies model with variables

The following equations describe the quasispecies model:

$$w_{ij} = A_j q_{ij} \qquad (4)$$

Here, w_{ij} is the expected fraction of i and j variants that arise in replication as the product of the replication rate,A_j of sequence j and the mutation probability q_{ij} that sequence j will mutate to sequence i. Here, w_{ij} describes the propensity to replicate and, therefore, (potentially) survive. In the following equation, the mutation rate per bit is μ; (standardized, in terms of character, μ from p in Wilke, 2005). The difference between sequences (the Hamming distance,H_{ij}) is described above; while L is the overall length of the sequence in question: L could describe a sentence, Tweet or other length of discourse. Therefore, q_{ij} gives the probability of change, per length of information sequence and is dependent on the number of differences between the starting and finishing sequence: i.e.:

$$q_{ij} = \mu^{H_{ij}}(1 - \mu)^{L - H_{ij}} \qquad (5)$$

If we wish to determine population size, *n*, of an information variant at time, t, the following quasispecies equation, may be used:

$$\frac{dx_i}{dt} = \sum_{j=1}^{n} a_j q_{ij} x_j - \phi x_i \qquad (6)$$

The “death function” ϕx_i is inserted to maintain a fixed population size and represents the removal of sequences that are in excess of a prescribed limit, so that dx_i/dt ($\dot{x}_i$) is the population of allele x_i present at a point in time. In our model, this function may be viewed as the cap in numbers imposed by the local carrying capacity [63], so that $\dot{x}_i$ is less than or equal to the carrying capacity, k and $\dot{x}_i$ is the population size, *N (t),* of x_i at a specified time, t, [64].

Substituting equation 4 into equation 6 gives us:

$$\dot{x}_i = \sum_{j=1}^{n} a_j \mu^{H_{ij}} (1-\mu)^{L-H_{ij}} x_j - \phi x_i \qquad (7)$$

Where “x-dot” is the differential dx/dt, in equation 6. The quasispecies equation determines the population size of alleles x_i, and its derivative through mutation, x_j. From the methods, the mutation rate is 10^{-2} to 10^{-4}; the length, *L*, of a typical sentence as 20 words, with two clauses; the Hamming Distance, *H*, set at one and the replication rate, a, set in the range 10^2-10^6 per cycle.

Wilke (2005) illustrates how the quasispecies model is analogous to aspects of population genetics as follows. If we consider a single locus with two alleles, *a* and *A*, then in the absence of mutation, μ, but presence of selection, the following equation can be derived from the quasispecies model (derived from equations 1 and 3 in Wilke [20]):

$$x_A = s x_A(t) \left[1 - x_A(t) \right] \qquad (8a)$$

This version of the logistic function is analogous to [65]:

$$x_{n+1} = r x_n (1 - x_n) \qquad (8b)$$

Where x is a fraction of the maximum value. Then x_n is the population in one year and x_{n+1} is the population in the next time interval.

The family of equations (8a, b) is the standard logistic function, describing the growth of a beneficial allele to fixation in a population, in the absence of mutation but in the presence of selection.

Solving the quasispecies equation in the presence of mutations, μ, and selection, *s*, allows us to determine the instantaneous value of x_i with the following function (equation 6; Wilke, 2005):

$$x_i = 0.5 \left[1 - \mu - \frac{2\mu}{s} + \sqrt{\left(\left(1 - \mu - \frac{2\mu}{s} \right)^2 + \frac{4\mu}{s} \right)} \right] \qquad (9a)$$

$$x_i = 0.5\left[1 - \mu - 2\mu + \sqrt{\left((1 - \mu - 2\mu)^2 + 4\mu\right)}\right] \qquad (9b)$$

Equation 9a reduces to 9b in the absence of selection.

The quasispecies model then predicts that when the mutation rate is low, the fastest-replicating variant takes over the population. However, when the mutation rate is high, a cloud of variants, a quasispecies, dominates. As the mutation rate, *μ,* increases, the concentration of x_i decreases, while x_j increases. Positive values of *μ* mean that allele *j* will constantly be generated even as selection acts to remove it. The action of mutation and selection leads to an equilibrium concentration of the information variants in the population.

For our purposes, equations 8(a, b) and 9(a, b) will be used to model the initial emergence and rise of a cloud of variant forms of information; then the eventual dominance of one variant that is propagated by re-posting, re-Tweeting or otherwise digitally-duplicating a variant. Here, the very low bit error rate can be ignored, as the rate is far lower than the reciprocal of the population size in question (> 10^{-9} versus a population of ca. 10^4 - 10^6).

2.7. Lexical variation and entropy

We discussed the Shannon entropy of language in an earlier work[26], but as the concept relates closely to the outcomes of the current work, we provide an overview here. The entropy of each sentence is related to the number of viable replacements (word or clause) that leave the sentence grammatically correct. The number of viable replacements will vary from culture to culture (or sub-culture to sub-culture) as words are created, modified or lost from each cultural unit. The entropy of a sentence is then related to the frequency of word variants in a cultural group that can function in the sentence to maintain its identity and/or its meaning [26][65][66][67][68][69][70]. Ideally, the number of viable replacements is that which recreates a grammatically-correct sentence – and those replacements should be in context (i.e., constrained by the context of the sentence, paragraph, or remainder) of the work in which it resides, formally linking the entropy of the sentence to the thermodynamic entropy [71]. However, various alternative descriptors, which remain valid, are described in Stevenson [26].

In relation to the presented work, the entropy of the sentence is related to the Hamming Distance as follows. If we set the Hamming Distance to one, then the number of viable one-word (or one clause) replacements (alleles) determines the entropy of that sentence, so that the Shannon Entropy related to the observed frequency of the variant allele in cultural use as [72][73]:

$$H = -\sum p_i ln p_i \qquad (10)$$

Where *p* is the probability of a variant in a population of sentences and In is the natural log.

We now consider how the diversity of language used in a group of people will alter through the effects of population growth and mutation. Here, mutation refers to the formation of novel language variants (alleles), such as new words, or changes in meaning and context of information that is transferred between groups. We consider two diversity indices, derived from the measure of Shannon entropy and heterozygosity, with the following derivations adapted from Chao et al. (2015). These are applied using the Infinite Allele Model (IAM).

We describe the variable θ as equivalent to 4Nμ, where N is the population size and μ is the mutation rate. Where θ is greater than 2 (e.g., where the mutation rate is 10^{-6} for overall activity and 10^{-2} - 10^{-4} for read-write errors), the expected Shannon entropy is approximately a linear function of the logarithm of 4Nμ (θ). Here, the "population" is the number of variants of a particular piece of information. Shannon entropy (^{1}H) is described in equation 10, therefore,*p* is the probability of encountering an allele in the population of those terms.

The heterozygosity (^{2}H) of the information in a population is given by:

$$^{2}H = \frac{\theta}{\theta + 1} \qquad (11)$$

Taking equations 10 and 11 as our initial functions, we can derive the following pair of diversity indices.

Shannon's Diversity Index (SDI; ^{1}H) is given by [74]:

$$^{1}D = -exp \sum p_i ln p_i \qquad (12)$$

Which may be reduced to the following forms, linking the mutation rate and population size to the diversity and heterozygosity – broadly the number of information alleles in the population:

$$^{1}D \approx e^{0.5772}(\theta + 0.5) = e^{0.5772}(4N\mu + 0.5) = 1.781\left(^{2}D - 0.5\right) \qquad (13)$$

Here, ^{2}D is the heterozygosity index, where:

$$^{2}D = \frac{1}{1 - {}^{2}H} = \theta + 1 = 4N\mu + 1 \qquad (14)$$

Alternative models are presented in Chao et al (2015), which include the effects of migration and multiple sub-populations. However, these models are also incomplete with regard to the movement of information, which is a lot faster than the migration of people (or other organisms) that their models consider. Moreover, the IAM model that is adapted, above, produces outputs that are close to the observations that the authors show in their work (Chao et al., 2015). Therefore, we consider the IAM model as applicable, if imperfect.

We encourage interested parties to read Chao's work to consider how our model may be adapted to more realistic scenarios, where migration of information is coupled to migration of people, as well as the "mutational" loss or gain of information alleles that we consider, here. That is currently, beyond the scope of the work presented.

2.8. Model Outline

To model the effect of the mode and rate of transmission of information we use a news headline, posted digitally, from the BBC: "A Chinese couple plotted to set up a mini-state on the Marshall Islands in the Pacific, bribing MPs and officials along the way, US prosecutors say." (Frances Mao, BBC News; 08/09/22). The headline is 123 characters long; 148 characters, including spaces. The headline consists of one sentence with three clauses, ignoring the location qualifier, "in the Pacific".

Table 3.

Unit	Headline Count
Length, *L* / Characters	123
Length, *L* / Words	26*
Length, *L* / Clauses	3**
Mutation rate (read rate) / μ	0.03% {m/10^{-4}/}
Mutation rate (write rate) / μ	1% {m/10^{-2}/}
Mutation rate (re-posting) / μ	10^{-9}
Hamming Distance, *H* (arbitrary)	1
Replication rate, *A* (non-viral)	10^{3}
Replication Rate, *A* (viral)	10^{6}

** "mini-state" is counted as one word; ** The length in clauses, ignores the qualifier, "in the Pacific". The Hamming Distance of 1 is an appropriate value for a sentence of this length, given the mutation rate. A maximum value would be 2 for a mutation rate of 0.31% (~0.003).*

For quasispecies models, we use a continuous range of mutation rates covering three orders of magnitude and with different population sizes (equations 5-9). We then model changes in information content using equations 11-14, with a viral population of 10^{6}. Here, the information from the headline is shared verbally amongst a social group, with the associated read error rate {m/10^{-4}/} and a population size of 100. Different error rates are then compared with regard to the change in information diversity.

The change in Shannon entropy is determined for each of these population models, where the Shannon diversity index is given by $1.781(4N\mu + 0.5)$ (equation 13); and the heterozygosity index by $4N\mu + 1$ (equation 14).

3. Results

In the following sections we quantify the data from the above methodological components.

3.1. Mutation rate and lexicon size

In biological systems, RNA viruses have the highest mutation rates on the order of 10^{4} - 10^{-5} with genomes that are typically 10^{3}-10^{4} nucleotides long [75][76][77][78][79]. Observed error rates in verbal and written communication (reading and copying) are higher, on the order of 10^{-2}-10^{-4} per task (Table 1, 28). Smith (2011) estimated that, at least in engineering -

the context of his work - that an error rates of 10^{-6} per task would be considered reasonable. Likewise, Tesdell [32] reported similar rates of error in ESL (English as a Second Language) users, leading us to conclude that error rates for the use of written and spoken English are within one order of magnitude of 10^{-2}-10^{-4}.

The error rate in Tweets is reported at 0.56% (> 10^{-3}), based on incorrect spelling or deviant word use[27][28]; while Facebook users have a near-equivalent error rate of 0.31% (> 10^{-3}; 31). Texting errors are similarly high (0.4%; 0.004), although the overall rate is lower in smartphone messages compared to pre-existing alphanumeric keypad messages [31]. These error rates agree with those of Smith (2011). A lower error rate in smartphones, as compared to older, alphanumeric keypads, illustrates an effect of error-repair. The older keypads required up to four presses per letter to select that required. Smartphones have individual letter keys, removing insufficient depressing or over-pressing of the key as an error-source. Smartphones also employ error repair in the form of predictive text.

Tweets have pre-set character limits and the observed error rate is proportional to Tweet length (10^2 characters for a ca. 0.5% (0.005) error rate). Moreover, when the character limit was increased, the observed error rate was lower [28].

The observed relationship between error rate and Tweet length accords with the broader relationship between error rate and genome length ([19][34]; equation 1). If we extrapolate the overall error rate of 10^6 (Smith, 2011), we would expect the English lexicon to be approximately 10^5-10^6 words, as observed (58,85,86. 87, 88). Moreover, a person's lexicon is on the order of 10^4-10^5 lemmas [80][81][82][83]. Given differences in the manner and accuracy in which language is stored in memory, compared with printed or digitally-stored text, the size of the lexicons agrees with the predictions of equation 1.

Of note, but unfortunately, quantified in a different manner, the aviation industry has documented errors of varying linguistic classes [84][85] that match expectations of meme-models for social evolution [86][87][88][89]. Unfortunately, the frequency of these errors was not recorded (Drury, pers comm). Likewise, Rabøl *et al.* [90] and Topcu *et al.*, [91] also illustrate similar types of error in hospitals as seen in the aviation industry; but again, the data does not include rate.

Table 4. Examples of classes of read and hearing error adapted from Drury and Ma [84][85].

Language Category	ASRS	IATA
Language/Accent	47	5
Partial or Improper Readback	24	8
Dual Language Switching	23	2
Unfamiliar Terminology	17	4
Speech Acts	9	0
False Assumptions or Inference	7	23
Homophony	5	1
Unclear Hand-off	4	3
Repetition across Languages	3	2
Uncertain Addressee	1	13
Lexical Inference	0	
Lexical Confusion (speed/heading/runway/altitude)	4	
Mistakes (unexplained)	3	
Total	**152**	**68**

Lexical inferencing involves making informed guesses based on neighbouring lexical cues Haastrup[92]*; while lexical confusion refers to the meaning of a word in one dialect or language being confused with a different meaning in another, such as "pants" meaning an item of underwear in the UK and trousers in the US. These errors are analogous to those seen in the medical profession: e.g., Rabøl et al* [90]*. ASRS Aviation Safety Reporting System; IATA International Air Transport Association.*

These data clearly illustrate an error frequency that is considerably greater than seen in biological systems. The high error rate necessitates that information in the above modes of communication must exist as a cloud of variants, where reproduction rates are high.

At the opposite end of the mutational spectrum, the digital (bit) error rate per bit is on the order of 10^{-9}-10^{-13} [93][94]. This error rate is vastly lower than the human error rate and is, therefore, not worthy of consideration in terms of cultural evolution, at present. However, one would expect that the lexicon that digital systems could have would be larger than any human language – should (in the future) digital systems be able to reproduce. Such evolution of digital language could be expected with a large number of digital devices working in parallel or very large information systems with data storage and use exceeding ca.10^{13} bytes. Here, $\mu N > 1$ [95].

3.2. Abbreviations and Tweets are analogous to Defective Interfering Variants

In biological systems, deletion mutants often emerge that are fitter than the original, longer sequence[13][75][96][97][98]. By "fitter" we mean replicate at a higher rate. In viral systems, these, shortened RNA molecules are called defective-interfering or DI variants. These abbreviated variants are transmitted along with the intact virus, but faster replication means that they can come to dominate the population [13][75][96][97][98][99]. These shortened molecules compete for replication in subsequent rounds of infection and contribute to inefficient transmission of viruses [13][75][76][96].

Likewise, Texts [30][100][101], Tweets [102][103] and, recently emerged Emoji-based languages [104] represent natural evolutionary processes that permit more rapid replication, by the loss of information that is non-critical for the integrity of the selected sequence [27][28][104]. By “integrity” we mean that the message remains functional in terms of conveying information, however, the context or subordinate meanings may not be present.

As with co-replication of viruses and DI particle genomes, transmission of these DI languages would also be expected to reduce transmission of the parent lexicon, through competition for available free energy [54][55]. Tapia et al [105] show that upwards of 10^4-10^5 times the number of DI copies per full-length viral genome per ml of fluid in tissue culture, with a concomitant decrease in transmission of the full viral genome compared to DI variants. If our analogy is correct, then we would expect an enhanced transmission of Tweets compared with the equivalent, full-length news items. Indeed, enhanced transmission of information in Tweets is observed in a number of different scenarios, where rapid replication of the information is favored over transmission of the full and more nuanced message [36][98][99][102].

Aside from shortened messages in the form of Tweets, text abbreviations, in general, constitute lexical analogies of DI variants. Take the phrase, “U R L8”: practically, every English-speaking person will recognize the meaning of that phrase despite missing six characters, not including spaces [30][94]. This deletion-message is clearly functional and is a lot easier to type than the full sentence. The truncated version permits faster replication and transmission, with a lower use of free energy than the full-length message. Therefore, in terms of the functional content of the media, there may be no reduction, as long as the simplified text conveys the same meaning [23][25].

3.3. Quasispecies modeling of sentence variation

Table 3 listed the initial parameters that were used when modeling quasispecies, using equations 11 and 12. We employed a range of mutation rates, ranging from 10^{-1} to 10^{-4}. These fully encompass the range expected for human discourse. Otherwise, the following summarize the parameters used:

- Sentence length, *L*: 123 characters; 26 words; 3 clauses
- Mutation rate, *μ*: variable, as above
- Hamming Distance, *h*: 1 or 2 (one or two words or clauses)
- Replication rate, A_{ij}: 10^2 and 10^6

No selection is applied in these models, so that s = 1 in equations 11 and 12. Table 5 illustrates the expected range of variants for a particular set of mutation rates.

Table 5. summary of the parameters for the three models we consider.

Model	Mutation Rate / *μ*	Replication Rate / h^{-1}	Variants / h^{-1}
1 - re-posting	10^{-9}	10^{6}	10^{-3}
2 - verbal share	10^{-4}	10^{2}	10^{-2}
3 - re-write share	10^{-2}	10^{6}	10^{4}

Figure 1 illustrates the effect of changing the mutation rate on the fitness,*w*, of variant *i* with two different replication rates, r = 10^2 versus r = 10^6 and a Hamming distance of 1. The effect of altering mutation rates is non-linear, with fitness increasing at low mutation rates, but then declining steeply at rates above 10^{-2}, which are typical of social media and common discourse (above). As might be expected, high rates of mutation (here, changes in the content of the transmitted message) will be associated with the loss of the meaning of the original message.

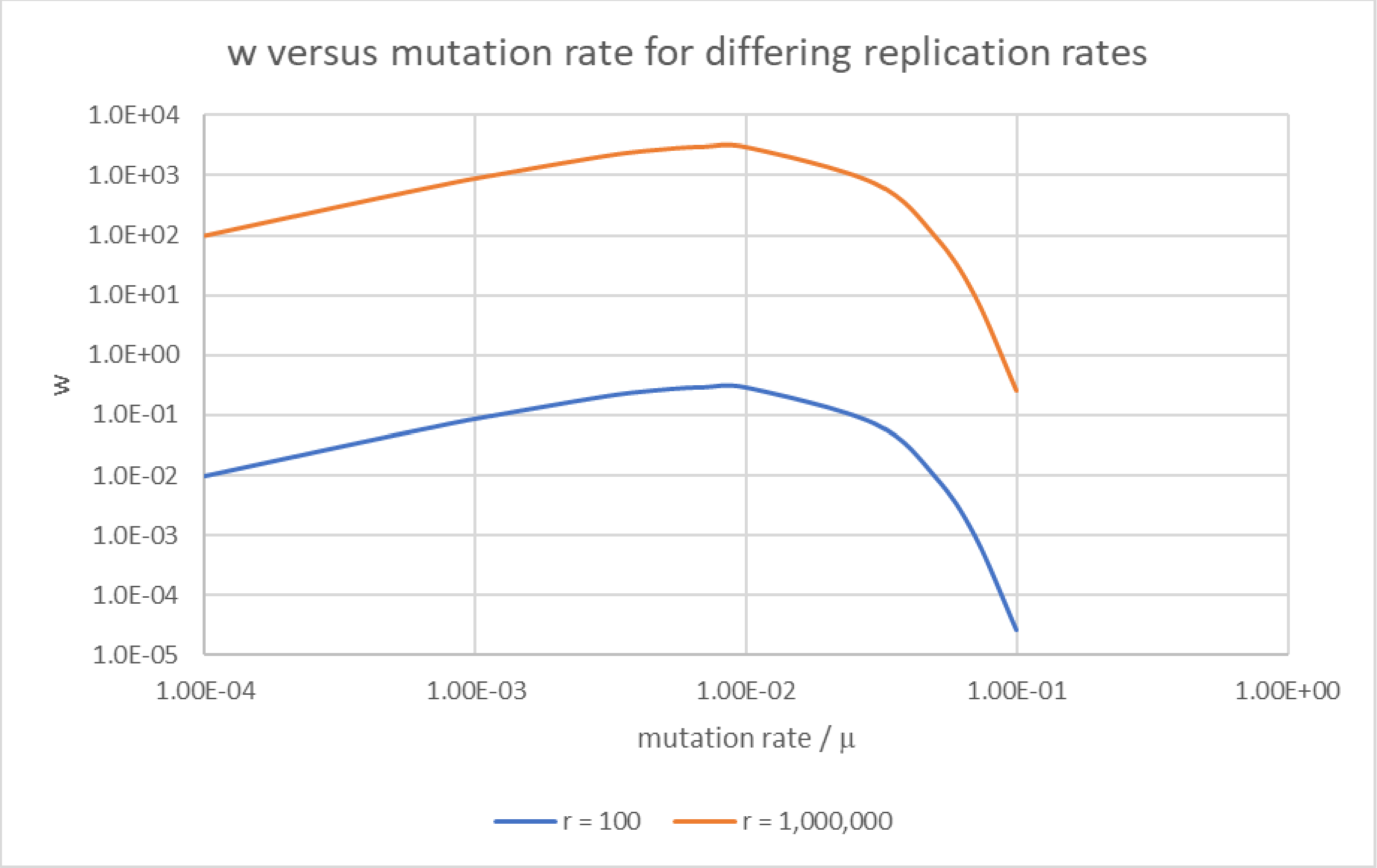

Figure 1.

The effect of altering the unit of length (character versus word count or clause count) has only a marginal effect on fitness (figure 2).

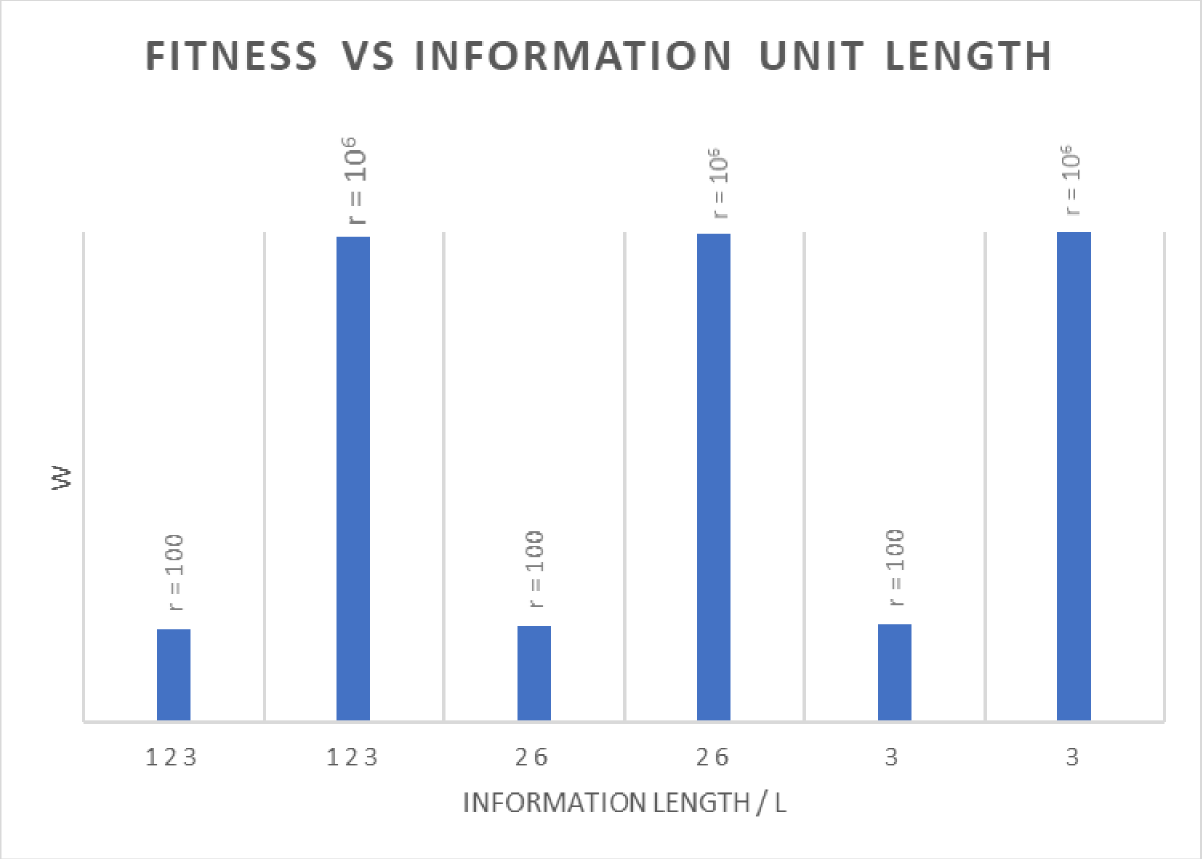

Figure 2. Changes to information length (character, word or clause) has little influence on the fitness of the sequence. The replication number (as would be expected) is the dominant influencer on fitness.

Doubling the Hamming distance decreases the probability of the change being observed in the population of alleles (sentence variants) by 1,000-fold (not shown); a result that is irrespective of the chosen mutation rate, length of information or the replication rate.

We also measure the abundance of variants (x_i and x_j) in our hypothetical population for differing values of length, Hamming distance, replication rate and mutation rate. In order to simplify the output, we determine the Shannon Diversity Index, ^{2}D (equation 4; [74]) for the modelled populations. Our simplified model has only two *alleles* – two variants of the sentence so that $x_i + x_j = 1$. In reality, we would assume a more complex population, but the presented model is for illustrative purposes only.

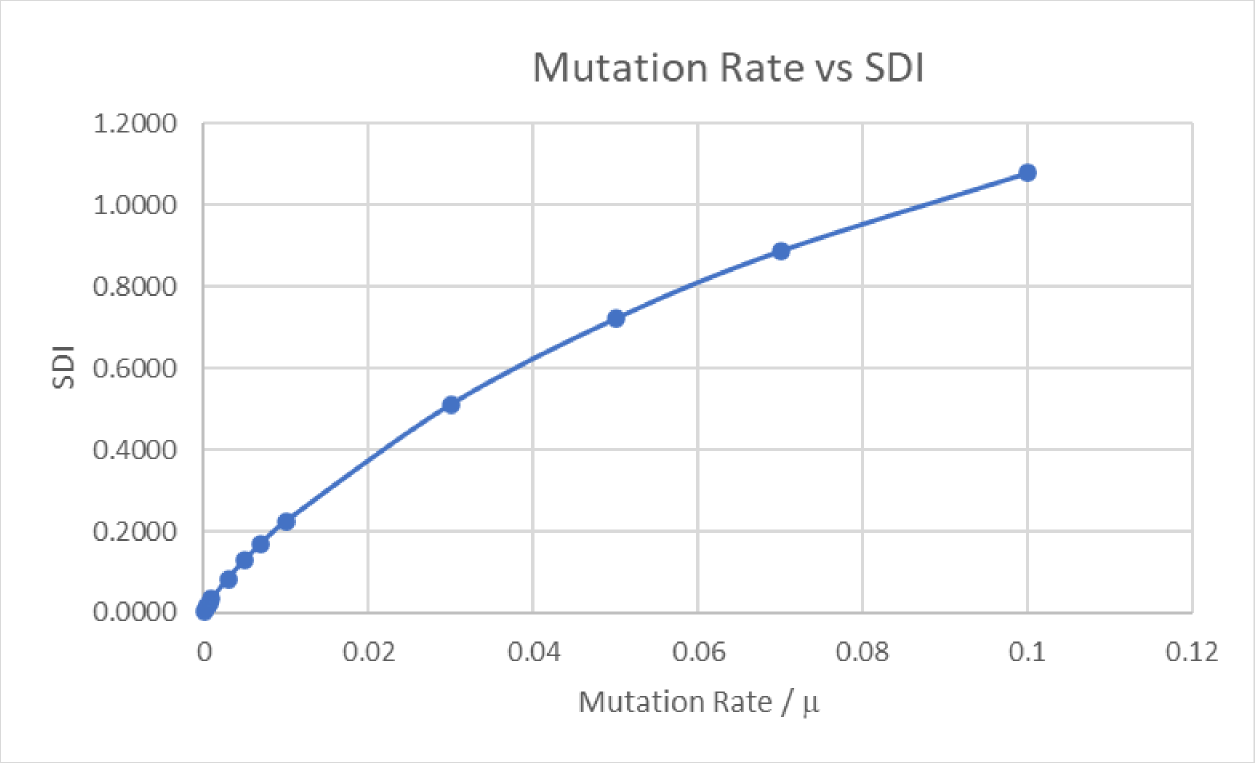

Figure 3. illustrates the effect of increasing mutation rate on the diversity of sequences in the quasispecies.

Doubling the Hamming Distance; altering the replication rate or length of unit does not alter the Diversity index,^{2}D (not shown). Finally, in figure 4 we use the Diversity Index, ^{2}D, to illustrate the change in abundance of the original variant,x_i, with decreasing mutation rates.

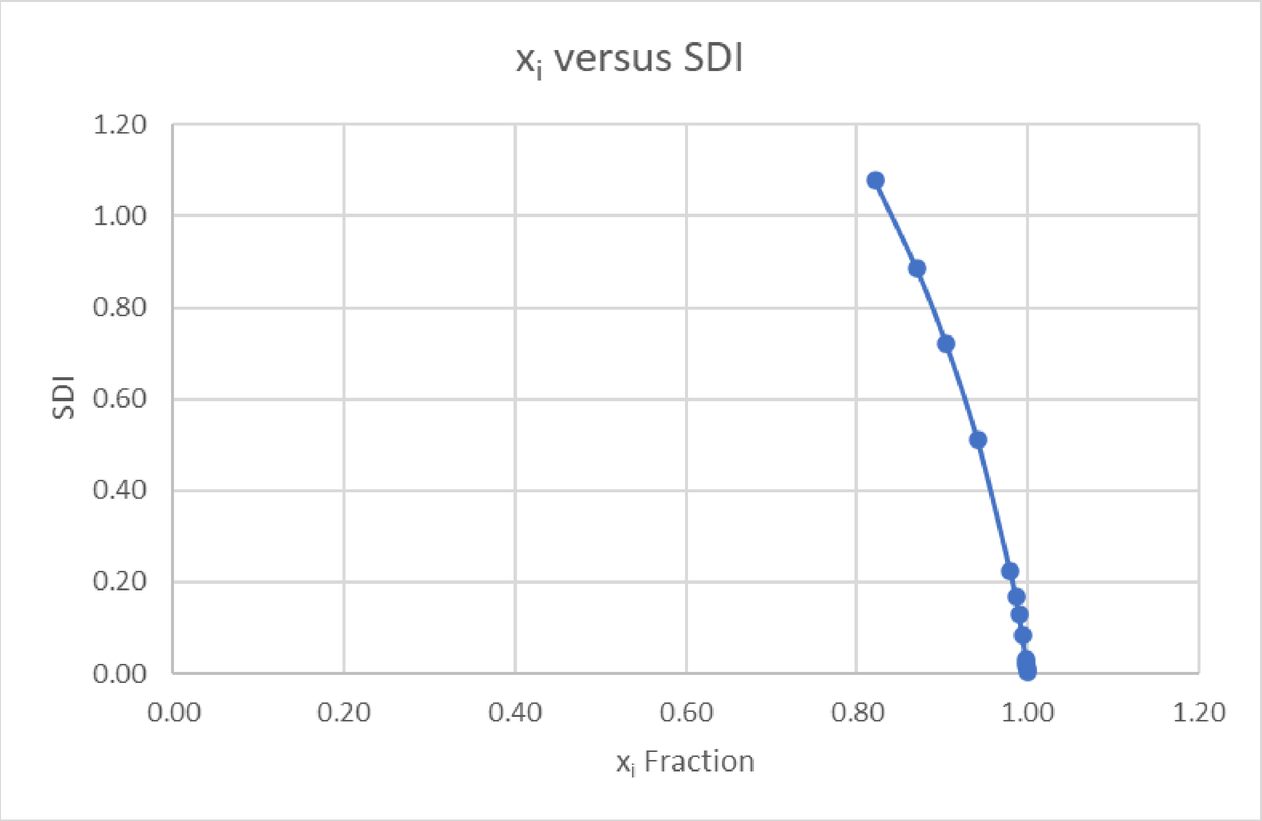

Figure 4. Variation in sentence diversity and x_i fraction, assuming only two alleles (sentence variants) exist in the population. There is no effect of changing Hamming distance, length or replication rate on the diversity index.

3.4. Conditional growth in Shannon entropy

Using equations 5 and 6, we compute the Shannon Diversity and Heterozygozity of our model populations of sentence variants. The quasispecies models (above) assume only two variants, while the models presented in figure 5 shows a more realistic picture of the number of variants that will arise in a population of differing sizes. Here, the term “population” refers to the variety of information present, rather than people who transmit it. We do not include the effects of selection in this analysis. Tables 6a and 6b list values for Shannon Diversity and Heterozygosity reflecting different modes of communication, while figure 5 illustrates the variation in Shannon Diversity with population size.

Table 6a. Variation in the Shannon Diversity Index for a variety of errors rates taken from the literature.

Population Size	Read Mutation Rate/ Checklist	Write mutation rate	Tweet Write Error	Text Error Rate	^{1}D Read Errors	^{1}D Write Errors	^{1}D Tweet Errors	^{1}D Text Errors
10,000,000	0.0001	0.01	0.0056	0.004	4000	400000	224000	160000
1,000,000	0.0001	0.01	0.0056	0.004	400	40000	22400	16000
100,000	0.0001	0.01	0.0056	0.004	40.5	4000	2240	1600
10,000	0.0001	0.01	0.0056	0.004	4.5	400.5	224.5	160.5
1,000	0.0001	0.01	0.0056	0.004	0.9	40.5	22.9	16.5
100	0.0001	0.01	0.0056	0.004	0.54	4.5	2.74	2.1
10	0.0001	0.01	0.0056	0.004	0.504	0.9	0.724	0.66

Given the high error rates in communication, the diversity within the population of information is expected to be high, except where simple re-Tweeting or sharing/copying of unmodified information is dominant. Note, values in excess of 10^3 are rounded, removing the "+0.5" from the determined values (equation 5).

Table 6b. Variation in heterozygosity, ^{1}H, for the same variety of references error rates used in table 6a.

Population Size	Read Mutation Rate/ Checklist	Write mutation rate	Tweet Write Error	Text Error Rate	^{1}H / Read	^{1}H / Write	^{1}H / Tweet	^{1}H / Text
10,000,000	0.0001	0.01	0.0056	0.004	400	40000	22400	16000
1,000,000	0.0001	0.01	0.0056	0.004	40	4000	2240	1600
100,000	0.0001	0.01	0.0056	0.004	41	4001	2241	1601
10,000	0.0001	0.01	0.0056	0.004	5	401	225	161
1,000	0.0001	0.01	0.0056	0.004	1.4	41	23.4	17
100	0.0001	0.01	0.0056	0.004	1.04	5	3.24	2.6
10	0.0001	0.01	0.0056	0.004	1.004	1.4	1.224	1.16

Given the high error rates in communication, heterozygosity is expected to be high, except where simple re-Tweeting or sharing/copying of unmodified information is dominant. Note: values in excess of 10^3 are rounded, removing the "+1.0" from the determined values (equation 6).

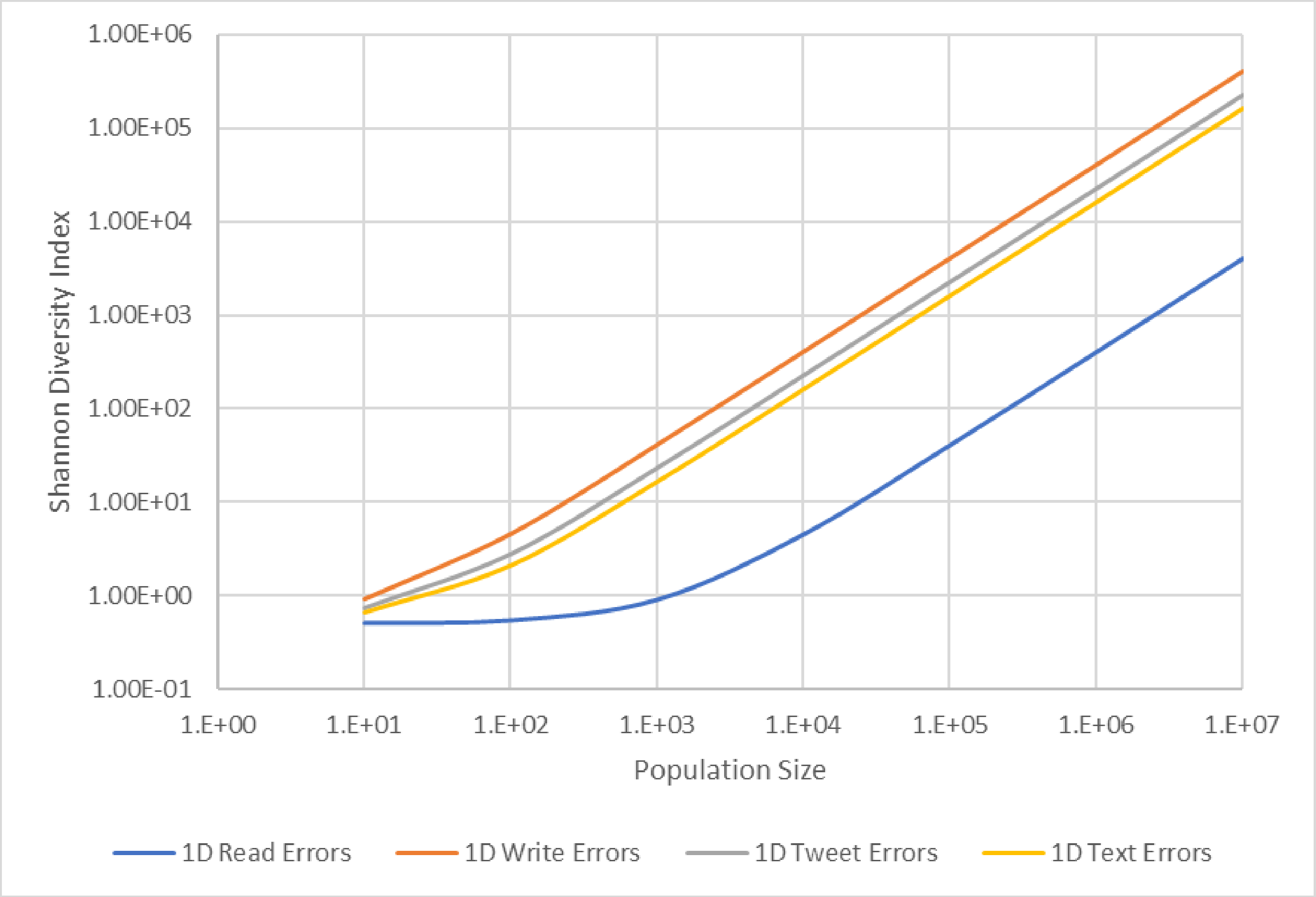

Figure 5. Graphical representation of the data from table 6a.

4. Discussion

In this paper we illustrate how cultural evolution should be considered within the framework of the evolution of information, which includes biological evolution. We present a series of models that illustrate the kinds of processes that lead to diversification of information in certain social contexts. We show that variation in populations of cultural information can be modelled in the same manner as biological information, in the form of genetic material.

In this work, we show that there are rates of variation in cultural information that match and, in some cases, grossly exceed that seen in biological systems [19][33][31]. The implication of these high rates of variation will be an abundance of variation in populations of information that circulates in the human population. As such, the abundance of these variants is analogous to quasispecies in RNA viral populations [96][97][98][99]. Moreover, the clausal structure of sentences, facilitates their transmission as modular units [11][37][38][39]; with modularity permitting recombination and further evolution. Human behavior is particularly prone to driving error-prone replication, either because we mis-hear and mis-read information, before passing it on [33][84][85][90][91]; or we merely do not remember it, accurately. We also fit information to pre-existing schemas, thereby facilitating its change and recombination with other information in our memory [58]. Fitting of data to expectation may also promote lexical inferencing [92].

Moreover, the formation of "textisms" [31] or Emoji languages [104] can be thought of as analogous to defective interfering viral particles, where the compressed form of social communication replicates more efficiently (and, therefore, has a higher fitness) than the original language [27].

We compute differing measures of Shannon Diversity and Heterozygosity, either based on simplified quasispecies models with two alleles (information variants), or more broadly with a range of observed error rates in different modes of communication. The two allele (quasispecies) model shows declining diversity with declining mutation rate (as might be expected); but no effect of altering the Hamming Distance, Length of message (as the type of unit assessed) or the replication. Diversity is solely dependent on the mutation rate in these models.

In the second series of models, diversity is judged by the Shannon methodology and by another biological method: heterozygosity. Both of these measures allow free-reign on diversification in the model populations and each illustrates how the mutation rate is the primary determinant of diversification. Each also illustrate how different modes of communication affect the rise in diversity.

We do not show the effects of selection on information spread, nor the long-term persistence of information, once it has been disseminated. A more accurate measure of fitness (w) is to include survivability with reproduction rate, then determine relative fitness of one variant to another [46][105][106]. For the purposes of brevity, in what is already a far-reaching article, we propose that this measure of fitness (w) is used in future analysis, with suitable time-frames used to judge the effect. However, the key idea is that information is simply information, irrespective of how it is stored. Cultural information in text, music, computerized data, or artifacts such as Art or architecture, is still information. The mode of transmission of information then determines how it evolves [107][108]. Art and architectural artifacts primarily change through the effects of light and weather, while written text evolves in our perceptions of it and the largely imperceptible effects of bit error, where that information is stored digitally.

Generally, it is in the faster-paced world of social media and conversation, that one should expect rapid evolution of information. The imperfect acquisition, storage and transmission of information by people is the bedrock through which cultural information will evolve. While the principle of equivalence in genetic (and epigenetic) and cultural information is not new, it remains contentious [2]. However, in this work we illustrate directly how processes central to biological evolution are measurable in cultural evolution. Moreover, we show that conversation and social media may be modelled as quasispecies – rapidly evolving populations of information that rapidly sample the cultural landscape in a manner equivalent to the biological sampling of the underlying fitness landscape [45][109].

While we have not looked more broadly at cultural context, within the BBC news quote that we analyzed, there was a potential source of cultural error: "…bribes ranging from $7,000 to $22,000 (£6,100 to £19,000)". Here, a simple conversion of "$7,000 to $22,000" to "£7,000 to £22,000", would increase the value of the bribe, significantly. Indeed, on the week beginning March 20th 2023 there were reports of a new gold find in China. One report had the estimated value as $3 billion[2], while another article had the same find valued at $3 trillion[3]. Upon checking the estimated yield of the deposit and the current market price, the lower value turned out to be correct. Therefore, we note that the social context of information is a source of further error and is seen in incidents involving airplanes (Table 4: 88) and hospital [89][90].

5. Conclusions and perspectives

We discuss a few implications of the model that allow it to be tested, but also illustrate the impact of the quasispecies model on communication and cultural evolution.

5.1. Relationship between the Pragmatic Theory of Information and this work

The Pragmatic Theory of Communication is an attempt to reformulate the information content of a communication in terms is the actions it has on the recipients [23][24][25]. As such, the theory converges with that presented, as it posits that the useful information content of a message is determined by the extent of selection by its recipients, as follows:

- *Two messages are equivalent when they lead to the same actions;*
- *Equivalent messages, of different sizes, can have the same information content;*
- *The same message has different information content when used in different decision contexts*

The first bullet is, in effect, a statement of convergent evolution, where selection drives the same outcome, irrespective of the source. The second statement is analogous to neutral mutations, where addition, deletion or alteration of genic information may have no effect on phenotype – but where we assume that larger messages carry a greater energy burden to transmit/replicate. The third statement is the more interesting of the three and may be considered as a function of the “clausal structure” of genes (exons; domains) and sentences (clauses).

Error-prone replication can alter verbs and nouns, but clauses provide a context, which depending on recipients, may have different impacts. When information is interpreted, the nature of the clauses that are used to add structure will affect how it is interpreted. Adding additional clauses with emotional or less direct meaning, will increases the number of interpretations and hence the rate of divergence in its interpretations. Reducing the number of clauses reduces the number of targets for differential selection. Therefore, where there is a need to keep a message’s intent fixed, limiting the number of clauses in a sentence is important.

To summarize, if two different messages lead to the same action, the selection acting on each is identical (convergent evolution). Secondly, where additional content is added to speech or text, such additional text need not provide a useful target for selection. In essence, “waffle” may be ignored. Thirdly, where there is a sufficiently-rich piece of information, different components of that information form targets for selection. Therefore, the context in which that information is perceived can lead to distinct units within it being selected by different individuals or groups. In summary, if you wish there to be one outcome from a communication, keep your message simple.

5.2. Practical implications of the growth in Shannon entropy of cultural information

Shannon entropy may be used as a measure of information complexity. It follows from The Pragmatic Theory, given error-prone human communication – whether by interpretation, memory deficiencies or intent – the Shannon entropy of societal information will grow over time, in a manner dependent on the availability of free energy (manuscript in preparation). Each time information is heard, memorized and disseminated, the opportunity for mis-replication and the generation of variants, increases. As cultural information is modular - and sentence structure is modular – the units on which error-prone replication and subsequent selection act are those modules. Therefore, in keeping with Weisenberg [25] and the principles, above, shorter and less modular sentences are less prone to drift in interpretation or memorization.

5.3. A shrinking population means less cultural information, but no change in evolutionary pace

As we begin to depart from a world where population growth changes from a seemingly exponential path to a logistic one [110], it is worth asking what will and must happen to information and culture in the future.

The growth of the internet has matched population growth with a 10:1 ratio of information growth to people[111][112]. Moreover, while technology continues to grab more and more information about our universe and allow us to generate our own interpretations of it, there has already come a time when the flow of information exceeds our capacity to use it [110][111]. How does a culture and its Shannon entropy change when the influx of information grossly exceeds the population that can use it?

In such a world, we might expect the following:

- The fraction of information that is held and actively expressed must decline
- A decreasing fraction of total cultural information will be held in the human population
- Divergence in cultural information and decreasing human population should encourage cultural fragmentation, in the absence of selection
- Machine and AI come to dominate the utilization of information, with humans increasingly serving as spectators

With regard to Shannon entropy, a couple of observations can be made. Firstly, the information held in any one person will increase over their lifetime to whatever capacity they are able to hold. Such growth in information means that in any one person, the Shannon entropy of the information will increase, in step with the volume of the information they hold and its decay with an aging memory.

Culturally, the amount of information that can be expressed by humans, assuming that a growing AI does not express it, must decline as the number of people who are present declines (figure 5). Within the declining population, each person continues to grow their pocket of information entropy, but overall, the loss of human population means that a smaller proportion will be expressed and evolve over time. Therefore, rather obviously, the richness of information expressed in humanity will be proportional to its size and will decline once the human population declines.

An obvious outcome, will be the wholesale loss of expressed languages with a shrinking global population. As an aside, an interesting gedankenexperiment considered the evolution of language on long interstellar flights [113]. In essence McKenzie and Punske [113] applied the concept of island biogeography to an isolated population of humans, replete with language “founder effects”.

We should also consider music as a language[114][115][116], one which has a very high rate of evolution, despite the constraints imposed by its “grammatical system”. One only has to look at the evolution of music in the punk and post-punk eras [117]. Here, you see how one idea begat another and led to a rapid proliferation of styles; an expansion which was further facilitated by the availability of novel technologies [118]. Music, is thus the ideal arena in which to further probe language evolution through the lens of information evolution.

If we are intent on preserving and growing the richness of our culture, evolving and expanding AI systems will need to take over the role. Moreover, if that interest extends to developing cultures within these AI systems, we will have to build in error-prone systems to generate the kinds of evolutionary diversification biological systems experience. While the inclusion of such processes could lead to the extinction of some human culture, it will bring evolutionary change to AI culture, thereby linking it to the evolution of information in the biosphere.

Acknowledgements

The work is solely that of the author. However, I would like to thank Colin G. Drury for some discussion on the nature and frequency of communication errors in the aviation industry.

Competing Financial Interest

There are no competing interests, financial or otherwise.

Footnotes

[1] Latané’s original equation is I = sN^t, therefore, we have changed “s” to “c” to avoid confusion with the selection coefficient, s, in genetics; while t has been changed to w, to reflect the fitness of the response.

[2] https://news.cgtn.com/news/2023-03-19/China-discovers-huge-gold-deposit-worth-3-trillion-1ij0YJRKiXK/index.html noting that the URL has the incorrect $3 trillion value, while the article is correct.

[3] https://www.msn.com/en-xl/news/other/china-discovers-huge-gold-deposit-worth-3-trillion/ar-AA18QoyB

Other References

- Stevens, SS. On the psychophysical law. Psychological Review; 1957. 64, 153-181.
- Linzen T and Jaeger TF. Investigating the role of entropy in sentence processing. Proceedings of the 2014 ACL Workshop on Cognitive Modeling and Computational Linguistics; 2014. 10–18

- Chao A, Jost L, Hsieh TC, Ma KH, Sherwin, WB, Rollins LA. Expected Shannon Entropy and Shannon Differentiation between Subpopulations for Neutral Genes under the Finite Island Model. PLoS ONE; 2015. 10(6): e0125471. doi:10.1371/journal.pone.0125471
- Lyddy F, Farina F, Hanney J, Farrell L, Kelly O'Neill N. An analysis of language in university students' text messages. J Comput-Mediat Commun.; 2014. 19:546–561

References

1. a, b*Richerson PJ and Boyd R. The Darwinian theory of human cultural evolution and gene–culture coevolution. In: Evolution Since Darwin: The First 150 Years. Bell MA, Futuyma DJ, Eanes WF, and Levinton JS (eds).; 2010. Sinauer, 561-588.*
2. a, b, c*Claidière N, Scott-Phillips TC, Sperber D. How Darwinian is cultural evolution? Phil. Trans. R. Soc. B; 2014. 369: 20130368*
3. a, b*Creanza N, Kolodny O, Feldman MW Cultural evolutionary theory: How culture evolves and why it matters. Proceedings of the National Academy of Sciences; 2017a. 114 (30) 7782-7789*
4. ^*Harton HC and Bullock M Dynamic Social Impact: A Theory of the Origins and Evolution of Culture. Social and Personality Psychology Compass; 2007. 1/1: 521–540*
5. ^*Mesoudi A Pursuing Darwin's curious parallel: Prospects for a science of cultural evolution. Proc Natl Acad Sci USA; 2017. 114:7853–7860*
6. ^*Stanley S Cultural Evolutionary Theory and the Significance of the Biology-Culture Analogy. Philosophy of the Social Sciences; 2021. 51(2) 193–214*
7. ^*Vegvari, C and Foley RA. High selection pressure promotes increase in cumulative adaptive culture. PloS one; 2014. 9(1), e86406. https://doi.org/10.1371/journal.pone.0086406*
8. ^*Waring TM, Wood ZT Long-term gene–culture coevolution and the human evolutionary transition. Proc. R. Soc. B; 2021. 288: 20210538*
9. a, b*Witzany G Can mathematics explain the evolution of human language? Communicative & Integrative Biology; 2011. 4(5); 516-520*
10. ^*Creanza N, Kolodny O and Feldman MF Greater than the sum of its parts? Modeling population contact and interaction of cultural repertoires. J. R. Soc. Interface; 2017b. 14; 20170171*
11. a, b*Muñoz E, Park J-M, and Deem MW. Quasispecies theory for Horizontal Gene Transfer and Recombination. Phys Rev E Stat Nonlin Soft Matter Phys; 2008. 78(6 Pt 1): 061921. doi: 10.1103/PhysRevE.78.061921*
12. a, b, c*Bull JJ, Meyers LA, Lachmann M. Quasispecies Made Simple. PLoS Comput Biol.; 2005. 1(6): e61.*
13. a, b, c, d*Donohue RC, Pfaller CK and Cattaneo R. Cyclical adaptation of measles virus quasispecies to epithelial and lymphocytic cells: To V, or not to V. PLoS Pathog., 2019. 15(2); e1007605*
14. ^*Escarmís C, Lázaro E and Manrubia SC. Population Bottlenecks in Quasispecies Dynamics. In: Domingo E. (eds) Quasispecies: Concept and Implications for Virology. Current Topics in Microbiology and Immunology; 2006. 299. Springer, Berlin, Heidelberg*

15. ^*Eigen M Self-organization of matter and evolution of biological macromolecules. Naturwissenschaften; 1971. 58:465–523*
16. a, b, c*Eigen M and Schuster P The Hypercyde: A Principle of Natural Self-Organization. Part A Emergence of the Hypercycle. Naturwissenschaften. 1977; 64, 541-565*
17. ^*Eigen M, McCaskill J, and Schuster P. Molecular quasispecies. J. Phys. Chem.; 1988. 92, 24, 6881–6891*
18. ^*Bertels F, Gokhale, CS, and Traulsen A. Discovering Complete Quasispecies in Bacterial Genomes. Genetics; 2017. 206(4), 2149–2157. https://doi.org/10.1534/genetics.117.201160*
19. a, b, c, d, e*Nowak MA. From Quasispecies to Universal Grammar. Z. Phys. Chem.; 2002. 216; 5–20*
20. a, b, c*Wilke CO, Quasispecies theory in the context of population genetics. BMC Evolutionary Biology; 2005. 5: 44, 8pp*
21. ^*Domingo-Sananes MR and McInerney JO. Selection-based model 1 of prokaryote pangenomes. bioRxiv preprint doi: https://doi.org/10.1101/782573; this version posted October 21, 2019. Accessed 23/03/2022.*
22. ^*Andrew F. Duckham M; Goodchild M; Worboys M (eds.). Pragmatic Information Content—How to Measure the Information in a Route Description. Foundations of Geographic Information Science. London: Taylor & Francis; 2003. 47–68. ISBN 0-415-30726-0.*
23. a, b, c*Witzany G. Pragmatic turn in biology: From biological molecules to genetic content operators. World J Biol Chem; 2014. 5(3), 279-285*
24. a, b*Blečić M The notion of 'information' in genetics: a pragmatic model. Journal of Biological Education; 2021. DOI: 10.1080/00219266.2021.2020876*
25. a, b, c, d*Weinberger ED A theory of pragmatic information and its application to the quasi-species model of biological evolution. Biosystems; 2002. 66(3); 105-119. doi: 10.1016/s0303-2647(02)00038-2.*
26. a, b, c, d, e*Stevenson, D. Application of Shannon Entropy Metrics to Cultural Diversity and Language Evolution. Academia Letters. 2021; Article 2503. https://doi.org/10.20935/AL2503*
27. a, b, c, d, e*Altman E and Portilla Y. Geo-linguistic fingerprint and the evolution of languages in Twitter. pp.14. ffhal-00674853v2f; 2012. Accessed on HAL (https://hal.inria.fr/hal-00674853v2/document) on 27/04/21*
28. a, b, c, d*Boot, AB, Sang ETK, Dijkstra, K, Zwaan RA. How character limit affects language usage in tweets. Palgrave Commun.;2019. 5, 76 https://doi.org/10.1057/s41599-019-0280-3*
29. ^*Windels J. https://www.brandwatch.com/blog/research-shows-twitter-is-driving-english-language-evolution/ Accessed 28/04/202153. Willer R, Kuwabara K and Macy MW (2009) The False Enforcement of Unpopular Norms. AJS. 2013; 115(2); 451–490*
30. a, b, c*Thurlow C and Brown. A Generation Txt? The sociolinguistics of young people's text-messaging. 2003. Available at: http://www.shu.ac.uk/daol/articles/v1/n1/a3/thurlow2002003-paper.html Accessed 27/04/2021*
31. a, b, c, d*Kent S and Johnson, G. Differences in the Linguistic Features of Text Messages send with an Alphanumeric Multi-Press Keypad Mobile Phone versus a Full Keypad Touchscreen Smartphone. Scottish Journal of Arts, Social Sciences and Scientific Studies; 2012. ISSN 2047-1278. Available at: https://espace.curtin.edu.au/bitstream/handle/20.500.11937/4431/190699_74459_72356.pdf;jsessionid=9955E30CACA957417C1D5C11CDD9088A?sequence=2 (Accessed 09/17/2022)*
32. a, b*Tesdell, LS. ESL spelling errors: a taxonomy. Retrospective Theses and Dissertations; 1982. 7903.*

https://lib.dr.iastate.edu/rtd/7903: Accessed 11/19/2019.

33. a, b, c *Smith DJ. Reliability, Maintainability and Risk (Eighth Edition). Practical Methods for Engineers including Reliability Centred Maintenance and Safety-Related Systems, Elsevier; 2011. ISBN 978-0-08-096902-2*
34. a, b *Gupta A, LaBar T, Miyagi M and Adami C. Evolution of Genome Size in Asexual Digital Organisms. Sci Rep. 2016; 6, 25786*
35. a, b *Kleider HM, Pezdek K, Goldinger SD, Kirk A. Schema-driven source misattribution errors: Remembering the expected from a witnessed event. Applied Cognitive Psychology; 2008. 22 (1): 1–20*
36. a, b *Milos C, Panagiotidis T, Dergiades T. Does It Matter Where You Search? Twitter versus Traditional News Media. Journal of Money, Credit and Banking; 2021. DOI: 10.1111/jmcb.12805. Available at: https://onlinelibrary.wiley.com/doi/10.1111/jmcb.12805*
37. a, b *Bork P. Shuffled domains in extracellular proteins. FEBS; 1991. 286 (1,2), 44-54 FEBS 09908*
38. a, b *Geary C, Chworos A, Verzemnieks E, Voss NR, and Jaeger L. Composing RNA Nanostructures from a Syntax of RNA Structural Modules. Nano Lett.; 2017. 17 (11), 7095–7101*
39. a, b *Spirov AV Eremeev AV. Modularity in Biological Evolution and Evolutionary Computation. Biology Bulletin Reviews; 2020. 10, 308–323*
40. a, b *Villarreal LP and Witzany G. Rethinking quasispecies theory: From fittest type to cooperative consortia. World J Biol Chem.; 2013. 4(4): 79-90*
41. ^ *Villarreal LP and Witzany G. Social Networking of Quasi-Species Consortia drive Virolution via Persistence. AIMS Microbiology; 2021. 7(2): 138–162.*
42. ^ *Markel M. 2010. Technical Communication, 9th ed. Bedford/St Martin's,*
43. ^ *Garner BA. Legal Writing in Plain English. University of Chicago Press, 2001*
44. ^ *Nordquist R. Sentence Length. ThoughtCo, Feb. 16, 2021. thoughtco.com/sentence-length-grammar-and-composition-1691948. Accessed: 28/11/2021*
45. a, b *Tamuri AU, dos Reis M and Goldstein RA. Estimating the Distribution of Selection Coefficients from Phylogenetic Data Using Sitewise Mutation-Selection Models. Genetics; 2010. 190, 1101–1115*
46. a, b, c *Thurman TJ and Barrett RDH. The genetic consequences of selection in natural populations. Molecular Ecology; 2016. 25, 1429–1448*
47. ^ *Herzog, H. Biology, culture, and the origins of pet-keeping. Animal Behavior and Cognition; 2014. 1(3), 296-308. doi: 10.12966/abc.08.06.2014*
48. ^ *Newberry, M.G., Plotkin, J.B. Measuring frequency-dependent selection in culture. Nat. Hum. Behav. (2022). https://doi.org/10.1038/s41562-022-01342-6*
49. a, b *Latané, B. The Psychology of Social Impact. American Psychologist; 1981. 36 (4): 343–356*
50. a, b *Lambert TA, Kahn AS, Apple KJ. Pluralistic Ignorance and Hooking up. The Journal of Sex Research; 2003. 40(2); 129-133*
51. a, b *Suls J and Green P. Pluralistic Ignorance and College Student Perceptions of Gender-Specific Alcohol Norms. Health Psychology; 2003. 22(5); 479–486*
52. a, b *Brady WJ, Wills JA, Jost JT, Tucker JA, Van Bavel JJ. Emotion shapes the diffusion of moralized content in social*

networks. Proceedings of the National Academy of Sciences; 2017. 114 (28); 7313-7318

53. ^*Smaldino PE, McElreath R. The natural selection of bad science. R. Soc. open sci.; 2016. 3: 160384.*
54. a, b*Campo J and Sarmiento V. The Relationship between Energy Consumption and GDP: Evidence from a Panel of 10 Latin American countries. Latin American Journal of Economics; 2013. 50(2); 233–255*
55. a, b*Liu W-C. The Relationship between Primary Energy Consumption and Real Gross Domestic Product: Evidence from Major Asian Countries. Sustainability; 2020. 12, 2568, 16pp*
56. ^*Reksulak M, Shughart II WF and Tollison RD. Economics and English: Language Growth in Economic Perspective. Southern Economic Journal; 2004. 71(2); 232-259*
57. ^*Topolewski, Ł. Relationship between Energy Consumption and Economic Growth in European Countries: Evidence from Dynamic Panel Data Analysis. Energies; 2021. 14, 3565.*
58. a, b*Tuckey M and Brewer N. The influence of schemas, stimulus ambiguity, and interview schedule on eyewitness memory over time. Journal of Experimental Psychology: Applied; 2003. 9 (2); 101–118*
59. ^*Doll BB, Bath KG, Daw ND and Frank XMJ. Variability in Dopamine Genes Dissociates Model-Based and Model-Free Reinforcement Learning. The Journal of Neuroscience; 2016. 36(4); 1211–1222*
60. ^*Lee Y-A and Goto Y. The Roles of Serotonin in Decision making under Social Group Conditions. Nature Scientific Reports; 2018. 8:10704 | DOI:10.1038/s41598-018-29055-9 3*
61. ^*Rogers RD. The Roles of Dopamine and Serotonin in Decision Making: Evidence from Pharmacological Experiments in Humans. Neuropsychopharmacology Reviews; 2011. 36, 114–132*
62. ^*Sebold M, Garbusow M, Jetzschmann P, Schad DJ, Nebe S, Schlagenhauf F, Heinz A, Rapp M and Romanczuk-Seiferth N. Reward and avoidance learning in the context of aversive environments and possible implications for depressive symptoms. Psychopharmacology; 2019. 236; 2437–2449*
63. ^*Lotka AJ. Elements of Physical Biology; 1925. Reprinted by Dover in 1956 as Elements of Mathematical Biology.*
64. ^*R Hopfenberg, Human Carrying Capacity is Determined by Food Availability: Population and Environment Popul. Environ..; 2003. 25, 2, 109-117.*
65. a, b*May R. Simple mathematical models with very complicated dynamics. Nature; 1976. 261, 459-467*
66. ^*Ippolito D, Grangier D, Eck D, Callison-Burch C. Toward Better Storylines with Sentence-Level Language Models. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020. 7472–7478*
67. ^*Fadyushin SG, Lobodenko AS, Milyaeva CE. Impact of text entropy on the human emotional state. Life Science Journal; 2014. 11; 289-291*
68. ^*Fadyushin S, Lobodenko A, Milyaeva E. Entropy of the Word as a Correction Factor of Addictive Human Behavior. Procedia - Social and Behavioral Sciences; 2015. 214; 797-804*
69. ^*Edmonds P and Hirst G. Near-Synonymy and Lexical Choice. Computational Linguistics; 2002. 28 (2); 105-144.*
70. ^*Cui Y, Che W, Zhang W-N, Liu T, Wang S, Hu G. Discriminative Sentence Modelling for Story Ending Prediction. Proceedings of the AAAI Conference on Artificial Intelligence; 2020. 34 (5), 7602-7609.*
71. ^*Jaynes ET. Gibbs versus Boltzmann entropies. American Journal of Physics; 1964. 33(5), 391-398*
72. ^*Shannon CE. A Mathematical Theory of Communication. The Bell System Technical Journal; 1948. 27; 379–423, and 623–656*

73. ^*Shannon CE. Prediction and Entropy of Printed English. Bell System Technical Journal; 1951. 30(1); 50-64*
74. a, b*Jost L. Partitioning Diversity into Independent Alpha and Beta Components. Ecology; 2007. 88, (10), 2427-2439*
75. a, b, c, d*Moya A, Elena SF, Bracho A, Miralles R, and Barrio E. The evolution of RNA viruses: A population genetics view. PNAS; 2000. 97 (13), 6967–6973*
76. a, b*Miralles R, Moya A and Elena SF. Diminishing Returns of Population Size in the Rate of RNA Virus Adaptation. Journal of Virology; 2000. 74(8); 3566–3571*
77. ^*Michel J-B, Shen YK, Aiden AP, Veres A, Gray MK, The Google Books Team, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J, Pinker S, Nowak MA, and Aiden EL. Quantitative analysis of culture using millions of digitized books. Science; 2011. 331(6014); 176–182*
78. ^*Stauffer Thompson KA and Yin J. Population dynamics of an RNA virus and its defective interfering particles in passage cultures. Virology Journal; 2010. 7; 257-267*
79. ^*Villarreal LP and Witzany G. Viruses are essential agents within the roots and stem of the tree of life. Journal of Theoretical Biology; 2010. 262 (4); 698-736.*
80. ^*Laufer B. The Development of Passive and Active Vocabulary in a Second Language: Same or Different? Applied Linguistic; 1998. 19(2); 255-271*
81. ^*Laufer B and Paribakht TS. The Relationship Between Passive and Active Vocabularies: Effects of Language Learning Context. Language Learning; 1998. 48(3). 365–391*
82. ^*Brysbaert M, Stevens M, Mandera P, Keuleers E. How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. Frontiers in Psychology; 2016. 7 DOI: 10.3389/fpsyg.2016.01116*
83. ^*Brysbaert M, Mandera P, McCormick S & Keuleers E. Word prevalence norms for 62,000 English lemmas Behav Res Methods; 2019. 51(2); 467-479*
84. a, b, c*Drury CG and Ma J. Language Errors in Aviation Maintenance: Year 1 Interim Report. University of Buffalo. Prepared for the Federal Aviation Administration William J. Hughes Technical Center Atlantic City International Airport, NJ 08405; 2003. Dr. William K. Krebs Research Grant #2002-G-025. Available at: https://www.researchgate.net/profile/Colin-Drury/publication/252752897_Language_Errors_in_Aviation_Maintenance_Year_1_Interim_Report/links/545ce2fe0cf295b5615e60c3/Language-Errors-in-Aviation-Maintenance-Year-1-Interim-Report.pdf*
85. a, b, c*Drury CG, Ma J and Marin CV. Language error in aviation maintenance: Data from Asia Proceedings of the Human Factors and Ergonomics Society; 2005. 123-127*
86. ^*Dawkins R. The Selfish Gene. 1976. New York: Oxford University Press*
87. ^*Massad E, Rocha AF, Coutinho FAB and Lopez LF. Modelling the Spread of Memes: How Innovations are Transmitted from Brain to Brain. Applied Mathematical Sciences; 2013. 7 (46), 2295 – 2306*
88. ^*Wang L and Wood BC. An epidemiological approach to model the viral propagation of memes. Applied Mathematical Modelling; 2011. 35; 5442–5447*
89. a, b*Valensise CM, Serra A, Galeazzi A, Etta G, Cinelli M & Quattrociocchi W. Entropy and complexity unveil the landscape of memes evolution. Nature Scientific Reports; 2021. 11:20022*

90. a, b, c, d*Rabøl LI, Andersen ML, Østergaard D, Bjørn B, Lilja B Mogensen T. Republished error management: Descriptions of verbal communication errors between staff. An analysis of 84 root cause analysis-reports from Danish hospitals. BMJ Quality & Safety; 2011. 20, 268-274.*
91. a, b*Topcu I, Turkmen AS, Sahiner NC, Savaser S, Sen H. Physicians' and nurses' medical errors associated with communication failures. J Pak Med Assoc.; 2017. 67(4):600-604.*
92. a, b*Haastrup K. Lexical inferencing procedures or talking about words: Receptive procedures in foreign language learning with special reference to English. Tubingen: Günter Narr, 1991.*
93. ^*Ciciora W, Farmer J, Large, D, Adams M. Modern Cable Television Technology Video, Voice, and Data Communications; 2004. pp1052, Morgan Kaufman (San Francisco), ISBN 978-1-55860-828-3*
94. a, b*Watt H. How Does the Use of Modern Communication Technology Influence Language and Literacy Development? A Review Contemporary Issues in Communication Science and Disorders; 2010. 37; 141–148*
95. ^*van Nimwegen E, Crutchfield JP, Huynen M. Neutral Evolution of Mutational Robustness. Proc Natl Acad Sci USA; 1999. 96; 9716-9720.*
96. a, b, c, d*Alnaji FG, Brooke CB. Influenza virus DI particles: Defective interfering or delightfully interesting? PLoS Pathog.; 2020. 16(5): e1008436*
97. a, b, c*Leontis NB and Westhof E. Self-assembled RNA nanostructures. Science; 2014; 345 (6198) 732-733*
98. a, b, c, d*Pfaller CK, Mastorakos GM, Matchett WE, Ma X, Samuel CE, Cattaneo R. Measles virus defective interfering RNAs are generated frequently and early in the absence of C protein and can be destabilized by adenosine deaminase acting on RNA-1-like hypermutations. J Virology; 2015. 89; 7735–7747*
99. a, b, c*Genoyer E, López CB. Defective viral genomes alter how Sendai virus interacts with cellular trafficking machinery, leading to heterogeneity in the production of viral particles among infected cells. J Virol.; 2019. 93; e01579-18.*
100. ^*Freudenberg K. Thesis: Investigating the impact of SMS speak on the written work of English first language and English second language high school learners. MA Thesis: Philosophy Stellenbosch University; 2009. Accessed 27/04/2021.*
101. ^*Grieve J, Nini A, Guo D. Mapping Lexical Innovation on American Social Media. Journal of English Linguistics; 2018. 46(4); 293-319.*
102. a, b*Neubig G and Duh K. How Much Is Said in a Tweet? A Multilingual, Information-Theoretic Perspective. Analyzing Microtext: Papers from the 2013 AAAI Spring Symposium.*
103. ^*Ghosh R, Surachawala T and Lerman K. Entropy-based Classification of 'Retweeting' Activity on Twitter. 2011 Available at: https://arxiv.org/abs/1106.0346*
104. a, b, c*Kralj NP, Smailović J, Sluban B, Mozetič I. Sentiment of Emojis. PLoS ONE; 2015. 10(12): e0144296. https://doi.org/10.1371/journal.pone.0144296*
105. a, b*Tapia F, Laske T, Wasik MA, Rammhold M, Genzel Y and Reichl U. Production of Defective Interfering Particles of Influenza A Virus in Parallel Continuous Cultures at Two Residence Times—Insights From qPCR Measurements and Viral Dynamics Modeling. Front. Bioeng. Biotechnol.; 2019. 7:275.*
106. ^*Vignuzzi M and López CB. Defective viral genomes are key drivers of the virus–host interaction. Nature Microbiology*

Nature Microbiology; 2019. 4,1075-1087

107. ^*Tseng S-H and Nguyen TS. Agent-Based Modeling of Rumor Propagation Using Expected Integrated Mean Squared Error Optimal Design. Appl. Syst. Innov.; 2020. 3; 48-60*
108. ^*Zhao Z, Zhao J, Sano Y, Levy O, Takayasu H, Takayasu M, Li1 D, Wu J, Havlin S. Fake news propagates differently from real news even at early stages of spreading. EPJ Data Sci.; 2020. 9, 7. https://doi.org/10.1140/epjds/s13688-020-00224-z*
109. ^*van Veelen M, Luo S and Simon B. A simple model of group selection that cannot be analyzed with inclusive fitness. Journal of Theoretical Biology; 2014. 360(7); 279-289*
110. a, b*Grinin L, Markov A; Korotoyev A. On Similarities between Biological and Social Evolutionary Mechanisms: Mathematical Modeling. Cliodynamics; 2013. 4; 185–228.*
111. a, b*Dolgonosov BM. On the reasons of hyperbolic growth in the biological and human world systems. Ecological Modelling; 2010. 221 (13–14), 1702–1709.*
112. ^*Golosovsky M. Hyperbolic Growth of the Human Population of the Earth: Analysis of existing models. History & Mathematics: Processes and Models of Global Dynamics; 2010. 188 –204. Available at: https://www.sociostudies.org/almanac/articles/hyperbolic_growth_of_the_human_population_of_the_earth-_analysis_of_existing_models/*
113. a, b*McKenzie A, Punske J. Language Development During Interstellar Travel. Acta Futura; 2019. 12; 123-132*
114. ^*Ravignani A, Thompson B and Filippi P. The Evolution of Musicality: What Can Be Learned from Language Evolution Research? Front. Neurosci.; 2018. 12:20. doi: 10.3389/fnins.2018.00020*
115. ^*Brown S A. Joint Prosodic Origin of Language and Music. Front. Psychol.; 2017. 8:1894. doi: 10.3389/fpsyg.2017.01894*
116. ^*Jackendoff, R. Parallels and non-parallels between language and music. Music Perception; 2009. 26(3), 195–204, ISSN 0730-7829, Electronic ISSN 1533-8312*
117. ^*Bangs, L. Psychotic Reactions and Carburetor Dung. Anchor Books, a division of Random House, London. 2003.*
118. ^*Vale V and Juno A. RE/Search; 1983. 6/7: Industrial Culture Handbook, RE/Search Publications. ISBN 0-940642-07-7*