Esperanto Information Content

Justin Goodman (jugoodma@umd.edu)
Department of Computer Science, University of Maryland
College Park, MD 20740 USA

Abstract

Natural languages are used almost exclusively in linguistics research, with the exception of "toy" languages. We almost never see corpus analysis performed on non-natural languages. Why? Lack of a sizable corpus, and lack of human applicability. In this work, we challenge this by examining Esperanto – a "lingua franca" with a growing popularity. Specifically, we examine a primary-works Esperanto corpus totaling about 10 million words. We compute correlation statistics modeling a prior work (Piantadosi, Tily, & Gibson, 2011), and hypothesize that Esperanto likely does not hold the claims stated in the aforementioned work.

Keywords: Esperanto; information content; Zipf's Law

Introduction

Linguistics research is intrinsically biased towards popular and naturally evolved languages. This makes the results immediately applicable, at the cost of being hard to test for robustness. Fortunately, we can use certain *auxiliary languages* – those created by an entity for the primary purpose of universal communication – to test our models for robustness. Auxiliary languages, however, need to be *used by humans* to be applicable. Most auxiliary languages are not widely used by humans, and thus are not applicable for robustness testing. Esperanto, however, is one of the most widely used auxiliary languages. While it does not have wide-spread adoption, it can still be useful for linguistic analysis.

In this paper, we analyze a handful of language metrics using an Esperanto corpus. Specifically, we analyze Esperanto from a Zipfian lens. We provide answers and insight to the following questions:

- Does Esperanto hold Zipf's law?
- Is Esperanto word length more correlated with average surprisal or word frequency?
- Does there exist a universal lower bound after which the *n*-gram context approximation does not hold?

The rest of the paper is organized as follows. The next section presents background information on Zipf and word information content. Then we show our methods and results to the aforementioned questions. Finally, we commentate our results and conclude the work in the remaining sections.

Background

In this section, we discuss the difference between natural and constructed languages, and explain why Esperanto should be considered for linguistic study. Then, we discuss corpus analysis techniques, Zipf's law, and information theory.

Natural vs. Constructed Languages

Natural languages are those that have evolved over time through normal human use. Examples include English, Spanish, German, etc. It has been shown (Griffiths & Kalish, 2007) that language learning is a form of iterated learning – the current generation teaches the next generation, and so on. We can compute the starting state of any language using this lens and Markov chains. However, modeling natural human languages in this framework is unsurprisingly difficult. As such, the starting states of natural human languages are unknown. In general, it is still unclear what exactly was the origin of human language (Hewes, 1977). Baron-Cohen argued, however, that theory of mind came about before language (Baron-Cohen, 1999).

Since natural languages are innately human, they are ripe with human complexity and inconsistency. These intricacies make natural languages interesting to study, at the cost of experimental control. On the other hand, constructed languages – those created by an entity instead of naturally evolved – can sometimes return this control to researchers. There are three types of constructed languages – engineered, auxiliary, and artistic languages. *Engineered languages* are very small "languages" created (typically) by researchers to test language acquisition on infants in a controlled manner (Gómez & Gerken, 2000). *Auxiliary languages* are those created for the primary purpose of universal communication. Examples include Esperanto, Ido, and Interlingua. *Artistic languages* are those made for the purposes of art, like those from Tolken's *The Lord of the Rings* or Lucas' *Star Wars*.

We see engineered languages in linguistics research because they offer experimental control, at the expense of being unrealistic. On the other hand, auxiliary languages are often easy to learn since they are intended for universal communication. As a result, their vocabularies are often derived from widely-known languages and their grammars are often highly regular. This gives auxiliary languages more experimental control, while also offering language realism. Unfortunately, auxiliary languages are uncommon, making empirical data hard to both create and obtain.

Esperanto was created by L. L. Zamenhof in 1887 for the purpose of universal communication (The Editors of Encyclopaedia Britannica, 2021). It is a constructed language, and is touted as "easy to learn" due to its highly regular grammar and relatively small root-word vocabulary (Kiselman, 2008). It is (likely) the most popular auxiliary language, with (although estimates are varied) an active speaker population of around 60,000 (Libera Folio, 2017). One study (Manaris, Pellicoro, Pothering, & Hodges, 2006) shows that Esperanto has statistical properties similar to natural non-constructed languages. Given the relatively large speaker population, and the realism associated with more natural languages, Esperanto is a ripe candidate for linguistic study. While there does exist a body of research on Esperanto (Pereltsvaig, 2017), less research has been done on Esperanto *corpora*.

Corpus Analysis

Linguistics research has a heavy bias towards "popular" languages, like English and German (McEnery & Hardie, 2012; Eberhard, Simons, & Fennig, 2021). This makes sense, since linguistics researchers mostly do research in a language they understand, or is understood by most of the world. As evidence to this, we see numerous English corpora with ranging quantities up to trillions of words (Paul & Baker, 1992; Brants & Franz, 2006; Michel et al., 2011)[1]. On the other hand, less popular languages are naturally less likely to have available or large corpora. For example, as far as we know, there are only two Esperanto corpora available: Tekstaro de Esperanto (Esperantic Studies Foundation, 2020), and Wiki-Trans (Bick, 2011). The first contains books written in Esperanto totaling almost 10 million words, and is maintained by the Esperantic Studies Foundation. The second is a translation of English Wikipedia to Esperanto, and is proportional to the size of Wikipedia. Though the second corpus is large, it might contain errors and bias due to machine translation. Furthermore, the actual Esperanto Wikipedia site [2] might be too small of a corpus. We leave these data sources open for exploration in future work, and limit this paper's scope to the Tekstaro since it contains first-hand Esperanto sources.

Given a large enough corpus, we can answer interesting questions about a language's statistics. For example, we may ask if the language satisfies a Zipfian distribution – that is, we ask if the language displays a high inverse correlation between word frequency and word rank. Zipf's law tells us that many types of empirical data show an inverse rank-frequency distribution (Zipf, 1999). It is well known that human language satisfies Zipf's law. At this point, we may ask (1) why does this Zipfian distribution hold for language, and (2) does Zipf's law hold for non-natural (say, auxiliary) languages. For the first, Zipf offered the principle of least effort (Zipf, 2016), saying that the Zipfian language distribution is due to humans compromising effort minimization and communication maximization. This is supported in other areas – for example, in physics, Fermat's principle tells us that light travels the path that takes the least amount of time. Yet, it is still unclear *why* language follows a Zipfian distribution. Piantadosi discusses numerous theories attempting to explain the relationship between word frequency and length, but explains that the cognitive sciences have yet to prove or disprove any theory (Piantadosi, 2014). Regardless, we *can* answer our second question so long as our corpus is large enough.

Though we see a clear relationship between word rank and word frequency across languages, we often see a slight correlation between word *length* and word frequency. It is still unclear why this occurs. In one study (Sigurd, Eeg-Olofsson, & Van Weijer, 2004), researchers find that the frequency distributions of word lengths follow a gamma distribution – low frequency at length 2, high frequency around length 4, and exponentially decreasing frequency at greater lengths. They agree with Zipf's ideas, that this distribution is explained by the trade-off between not enough conveyed information (short words/sentences) and too many possible communication choices (long words/sentences).

Interestingly, prior work has shown that there are better statistics to correlate with word length than word frequency. It was shown in (Piantadosi et al., 2011) that *information content* had a higher Spearman correlation coefficient with word length. Their study defined information content as the surprisal of a given word averaged across its contexts, and used an n-gram model (with $n = 2,3,4$) to approximate the context of words. Related, (Meylan & Griffiths, 2017) use the previous work to show that the length-frequency correlation is a special case of the distinctiveness-frequency correlation. The authors define distinctiveness as the inverse of how similar a word is to other words, and has an obvious link to word length. Finally, (Mahowald, Fedorenko, Piantadosi, & Gibson, 2013) show that shorter words (those that convey less information) are used more often in predictive contexts. They define a predictive context as one that offers information about the shorter word.

For this paper, we aim to replicate (Piantadosi et al., 2011) using the Tekstaro. In the prior work, the authors used a corpus with about a trillion words (Brants & Franz, 2006). The Tekstaro has about 10 million words, which is significantly smaller. We need to have a large enough corpus for the n-gram word context approximation to hold. Unfortunately, current linguistics research provides no empirical or analytical bounds on corpus size for this assumption. The closest works we could find simply discuss model improvements for smaller corpus sizes (Hacioglu & Ward, 2001; Pickhardt et al., 2014). We thus cannot give a minimum corpus size that guarantees model accuracy under the n-gram assumption. Likely though, it is larger than the Tekstaro.

Results

To start, every corpus we interacted with was cleaned. Our cleaning was not robust – we simply deleted miscellaneous

[1] https://varieng.helsinki.fi/CoRD/corpora/index.html lists numerous English corpora.

[2] https://eo.wikipedia.org

characters including punctuation and numbers. This can sometimes leave strange artifacts in the corpus (e.g.: one-character "words"). Furthermore, cleaning the Esperanto corpus is less straightforward since we cannot comprehend the language. Regardless, our simple cleaning is likely good enough to provide a close approximation of properly-cleaned data. Our code is available on GitHub[3].

After cleaning our data, we computed a log-log rank-frequency plot of the full Tekstaro. We present this plot in figure 1. This plot appears to hold Zipf's law, and thus partially confirms the results discussed prior in (Manaris et al., 2006).

Next, we aimed to answer whether Esperanto satisfies the higher length-surprisal correlation as given in (Piantadosi et al., 2011). To do this, we followed the method as in the paper. However, we did not cross-reference the most frequent words with another corpus since no other Esperanto corpus really exists. We also did not compute the partial correlations. For the following three sets, we used the 25,000 most frequent words in the Tekstaro and computed the Spearman rank correlation coefficient ρ. First, we computed the correlation between frequency (in bits) and word length (number of characters, in bits) yielding $\rho = 0.1975$ ($p < 0.001$). Next, we computed the correlation between average information content (with a bi-gram approximation) and word length yielding $\rho = 0.1476$ ($p < 0.001$). Finally, we computed the correlation between average information content (with a tri-gram approximation) and word length yielding $\rho = 0.0386$ ($p < 0.001$). We also plotted all three sets of data points, presented in figure 2.

To be sure our code was correct, we used a set of English corpora to test our correlations against those in (Piantadosi et al., 2011). We used free samples from an online[4] website, totaling about 160 million words. We first computed a log-log rank-frequency plot, and indeed verified that Zipf's law holds (figure 3). Then, we computed the same correlations as before. For frequency-length, we found $\rho = 0.1279$ ($p < 0.001$). For information-length (bi-gram), we found $\rho = 0.1739$ ($p < 0.001$). For information-length (tri-gram), we found $\rho = 0.1456$ ($p < 0.001$). See figure 4. These results seem to hold the same ("qualitatively") as the BNC correlations from (Piantadosi et al., 2011). In the paper, the authors note,

> The numerical pattern of correlations differs somewhat from the Google data, likely because the BNC contains only 100 million words, only one 10,000th the size of the Google dataset for English. (Piantadosi et al., 2011)

This left us to search for analytical or numerical results regarding the size of which a corpus must be for the n-gram context assumption to hold. Unfortunately, to the best of our knowledge, no such work exists. As such, we cannot truly verify the validity of the Esperanto correlations until such a lower-bound is shown.

At what corpus size, then, does the n-gram context assumption accurately approximate the true word context? Furthermore, is there a way to analytically quantify how close the approximation will be? We leave these as open questions, with some commentary. Almost certainly, the bound will be corpus-dependent. A corpus that only contains one word, sampled from a widely varied language, will not produce usable results. We should assume, then, that a given corpus holds a Zipfian distribution of words sampled from the true language usage patterns. Under this assumption, is there a universal lower bound?

One potential method to test this is to examine the *growth rate* of the aforementioned correlations. We show how this might work in our English corpora. We ran 10 trials for varying corpus sizes ranging from 5 million to 80 million words. For each trial, and for each corpus size, we took random chunks of contiguous documents to form a sub-corpus. To each sub-corpus, we computed the three correlation metrics from before. Then, we plotted the mean (with standard deviation bars) for each sub-corpus. This yielded figure 5. There does seem to be some form of logarithmic or logistic growth for the info-length correlations, whereas the frequency-length correlations slightly decrease. In (Piantadosi et al., 2011), their English correlations at the 100 million corpus size are $\rho = 0.121$ (freqency-length), $\rho = 0.161$ (info-length bi-gram), and $\rho = 0.168$ (info-length tri-gram). The correlations are better likely due to the authors' use of smoothing and training techniques. Their English correlations from the Google data (Brants & Franz, 2006) (1 trillion corpus size) are $\rho \approx 0.10$ (freqency-length), $\rho = 0.21$ (info-length bi-gram), and $\rho = 0.30$ (info-length tri-gram). All together, these seem to hold along the growth pattern illustrated in figure 5.

We repeated the same process for the Tekstaro. We cannot extend the sub-corpus size as far out for Esperanto as we did for English since the Tekstaro is one 10$^{\text{th}}$ the size of our English database. The result is shown in figure 6.

Discussion

In sum, we conclude that our code correctly computes the approximated average information content, and our Esperanto corpus size is not large enough to make any substantive conclusions. However, we provide some commentary on our hypotheses.

The Tekstaro shows a Zipfian distribution, though it is unclear whether naturally-spoken Esperanto would as well. Likely, it does. We have no estimate for how much data are needed for Zipf's law to show a strong correlation. We leave estimating this empirically for future work. Potentially, a similar growth rate chart as in figures 5 and 6 could be helpful.

Using the English correlations, we see that the information content correlation with bi-gram approximation at the 20 million word corpus size undoubtedly surpasses the frequency

[3] https://github.com/jugoodma/828f-project

[4] https://www.corpusdata.org/formats.asp

correlation. Even at the 10 million word level, the two correlations are very close. We do not see the same effect from the Tekstaro. The difference in correlations is much higher with Esperanto at the same corpus size (even the 5 million size). We suspect that Esperanto truly does not hold a higher correlation between word information content and word length. Esperanto is a constructed language – it is a melting pot of language vocabularies, mixed with a highly regular grammar. These together likely flatten out the amount of information conveyed from each word. Our hypothesis would indeed imply our Tekstaro correlation results, but unfortunately our corpus size is not large enough to conclude the converse. For future work, we suggest implementing smoothing techniques and proper cleaning on the Tekstaro. Furthermore, we recommend investigating the various Esperanto Wikipedias mentioned before for expanding the corpus size. We also recommend investigating neural networks for estimating word information content or word contexts. Finally, if Esperanto is truly easy to learn, then we should expect frequently used words to convey *more information*. Thus, we should likely expect a higher correlation between word frequency (as opposed to length) and word information content. This is tangentially related to this work, but may lead to interesting results across langauges.

We challenge future work to find an empirical relationship between the size of a corpus and its n-gram context approximations for information content accuracy. This is likely incredibly difficult. We mentioned before that indeed it depends on the corpus itself. There may be no relationship, making it entirely language-dependent. If that is the case, then can we use it as another statistic for language classification? If that is not the case, then is there a way to derive an analytical relationship?

Conclusion

In this paper, we have discussed auxiliary languages for testing model robustness. Specifically, we have analyzed Esperanto from a Zipfian lens. We found that likely Esperanto does not hold previously found correlations based on information theory, but remark that our analyzed corpus is too small to provide a substantive conclusion. Our analysis leaves more questions than answers, and we look forward to future analysis.

References

Baron-Cohen, S. (1999). *The evolution of a theory of mind.*

Bick, E. (2011). Wikitrans: the english wikipedia in esperanto. In *Constraint grammar applications, workshop proceedings at nodalida* (Vol. 14, pp. 8–16).

Brants, T., & Franz, A. (2006). *Web 1T 5-gram Version 1 LDC2006T13.* Philadelphia: Linguistic Data Consortium.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2021). *Ethnologue: Languages of the world* (Twenty-fourth ed.). Online version: http://www.ethnologue.com. Dallas, Texax: SIL International.

Esperantic Studies Foundation. (2020). *Tekstaro de esperanto.* http://tekstaro.com.

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in cognitive sciences*, *4*(5), 178–186.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive science*, *31*(3), 441–480.

Hacioglu, K., & Ward, W. (2001). Dialog-context dependent language modeling combining n-grams and stochastic context-free grammars. In *2001 ieee international conference on acoustics, speech, and signal processing. proceedings (cat. no.01ch37221)* (Vol. 1, p. 537-540 vol.1). doi: 10.1109/ICASSP.2001.940886

Hewes, G. W. (1977). chapter 1 - language origin theories. In D. M. Rumbaugh (Ed.), *Language learning by a chimpanzee* (p. 3-53). Academic Press. Retrieved from https://www.sciencedirect.com/science/article/pii/B9780126018509500087 doi: https://doi.org/10.1016/B978-0-12-601850-9.50008-7

Kiselman, C. (2008). Esperanto: its origins and early history. *Prace Komisji Spraw Europejskich PAU*, *2*, 39–56.

Libera Folio. (2017). *Nova takso: 60.000 parolas esperanton.* https://www.liberafolio.org/2017/02/13/nova-takso-60-000-parolas-esperanton/.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313-318. Retrieved from https://www.sciencedirect.com/science/article/pii/S0010027712002107 doi: https://doi.org/10.1016/j.cognition.2012.09.010

Manaris, B. Z., Pellicoro, L., Pothering, G., & Hodges, H. (2006). Investigating esperanto's statistical proportions relative to other languages using neural networks and zipf's law. In *Artificial intelligence and applications* (pp. 102–108).

McEnery, T., & Hardie, A. (2012). *Corpus linguistics : Method, theory and practice.* Cambridge University Press.

Meylan, S. C., & Griffiths, T. L. (2017). *Word forms - not just their lengths- are optimized for efficient communication.*

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, … Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. Retrieved from https://science.sciencemag.org/content/331/6014/176 doi: 10.1126/science.1199644

Paul, D. B., & Baker, J. M. (1992). The design for the wall street journal-based csr corpus. In *Proceedings of the workshop on speech and natural language* (p. 357–362). USA: Association for Computational Linguistics. Retrieved from https://doi.org/10.3115/1075527.1075614 doi: 10.3115/1075527.1075614

Pereltsvaig, A. (2017). Esperanto linguistics: State of the art [Journal Article]. *Language Problems* *and Language Planning*, *41*(2), 168-191. Retrieved from https://www.jbe-platform.com/content/journals/10.1075/lplp.41.2.06per doi: https://doi.org/10.1075/lplp.41.2.06per

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, *21*(5), 1112–1130.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Pickhardt, R., Gottron, T., Körner, M., Wagner, P. G., Speicher, T., & Staab, S. (2014). A generalized language model as the combination of skipped n-grams and modified kneser-ney smoothing. *arXiv preprint arXiv:1404.3377*.

Sigurd, B., Eeg-Olofsson, M., & Van Weijer, J. (2004). Word length, sentence length and frequency–zipf revisited. *Studia linguistica*, *58*(1), 37–52.

The Editors of Encyclopaedia Britannica. (2021). *L.l. zamenhof.* https://www.britannica.com/biography/L-L-Zamenhof. Encyclopedia Britannica.

Zipf, G. K. (1999). *The psycho-biology of language: An introduction to dynamic philology* (Vol. 21). Psychology Press.

Zipf, G. K. (2016). *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books.

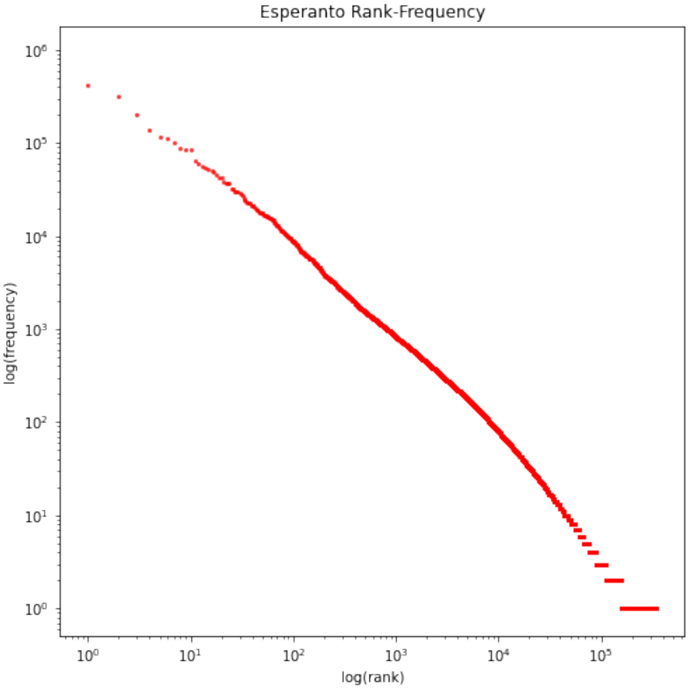
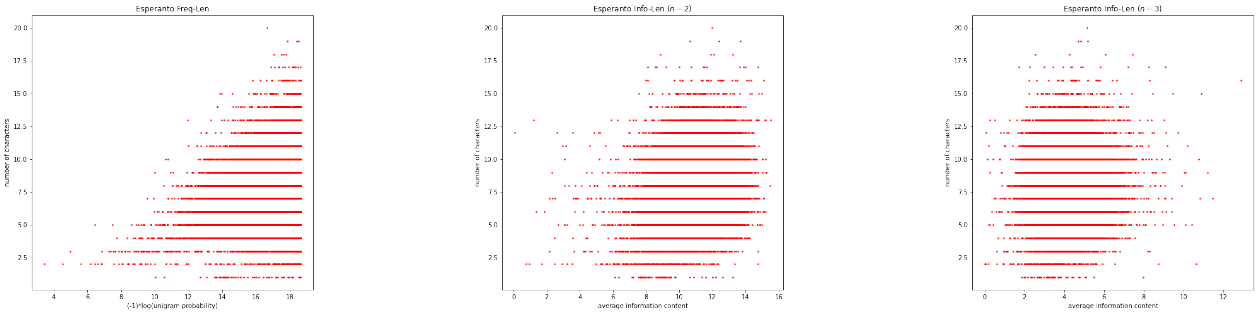


Figure 1: Rank-Frequency distribution of the Tekstaro.



(a) Against word frequency, transformed to units of bits.

(b) Against average information content using a bi-gram context approximation.

(c) Against average information content using a tri-gram context approximation.

Figure 2: Esperanto word length plots against word frequency and average information content.

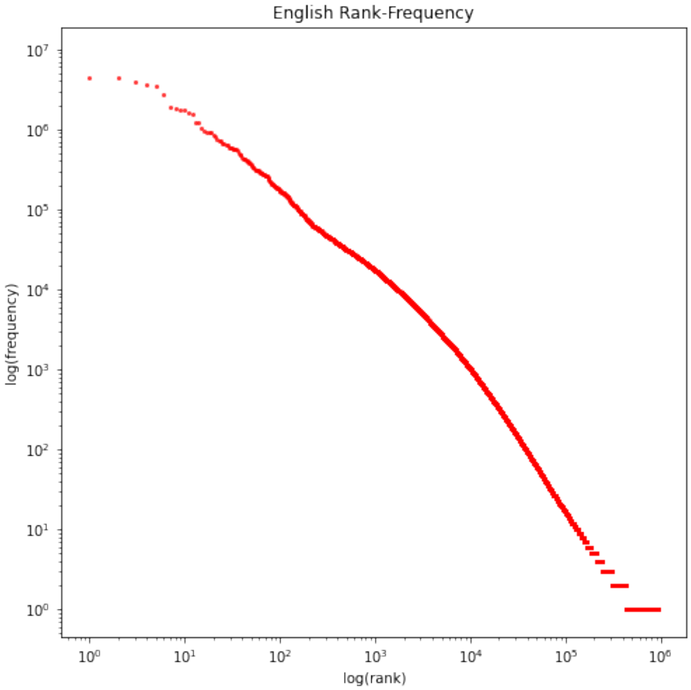
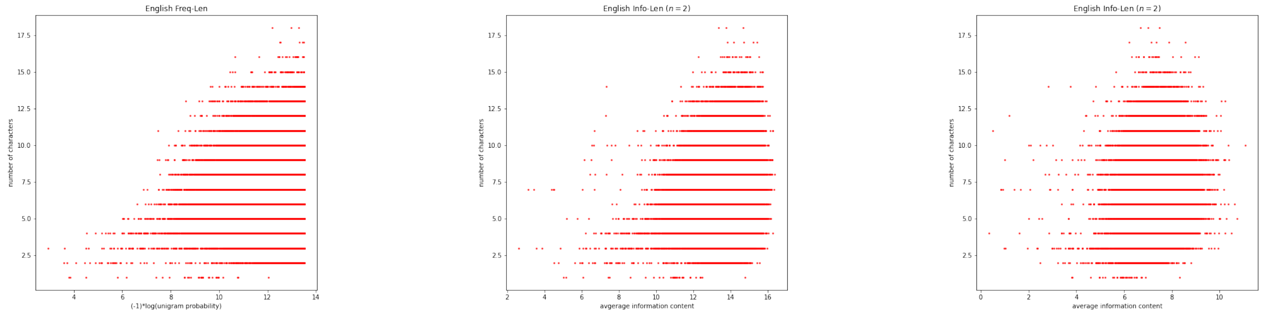


Figure 3: Rank-Frequency distribution of our free sample English corpora.



(a) Against word frequency, transformed to units of bits.

(b) Against average information content using a bi-gram context approximation.

(c) Against average information content using a tri-gram context approximation.

Figure 4: English word length plots against word frequency and average information content.

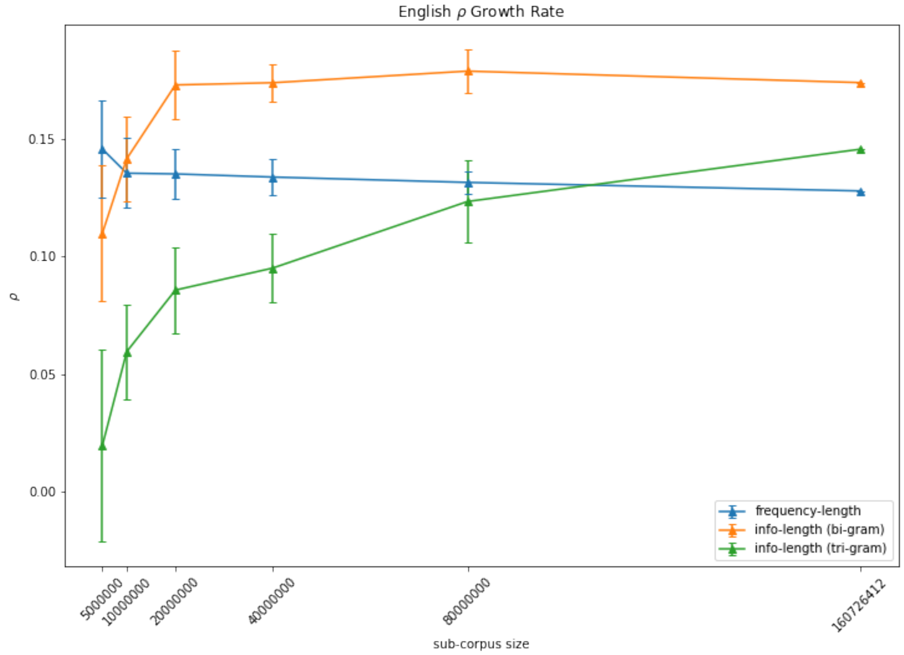


Figure 5: Spearman correlations plotted against sub-corpus size. Each data point represents the mean of 10 trials, with standard deviation error bars. The right-most data point is the entire English corpus, so does not have error bars. Each correlation was statistically significant with $p < 0.001$. However, one trial at the 5 million sub-corpus size had $p \approx 0.0023$.

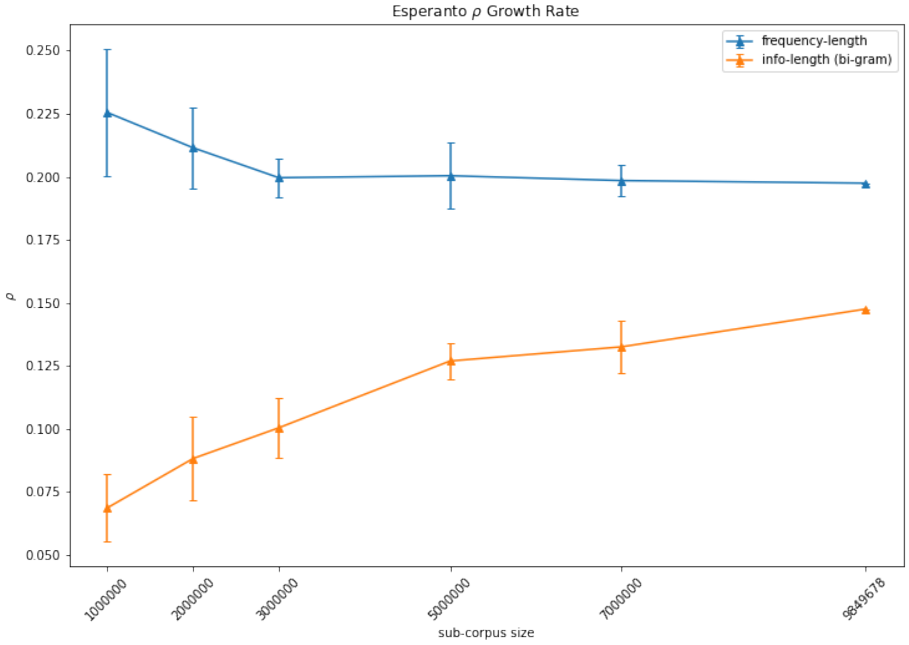


Figure 6: Spearman correlations plotted against sub-corpus size. Each data point represents the mean of 10 trials, with standard deviation error bars. The right-most data point is the entire Esperanto corpus, so does not have error bars. Each correlation was statistically significant with $p < 0.001$. Note the x-axis difference in this plot compared to figure 5.