Originally published in: Cognitive Science vol. 46 (2022) no. 6, article e13090.
DOI: *https://doi.org/10.1111/cogs.13090*

Alexander Koplenig,[a]
Marc Kupietz,[b]
Sascha Wolfer[a]

[a]Department of Lexical Studies, Leibniz-Institute for the German Language (IDS)
[b]Department of Digital Linguistics, Leibniz-Institute for the German Language (IDS)

Testing the Relationship between Word Length, Frequency, and Predictability Based on the German Reference Corpus

Abstract:

In a recent article, Meylan and Griffiths (Meylan & Griffiths, 2021, henceforth, M&G) focus their attention on the significant methodological challenges that can arise when using large-scale linguistic corpora. To this end, M&G revisit a well-known result of Piantadosi, Tily, and Gibson (2011, henceforth, PT&G) who argue that average information content is a better predictor of word length than word frequency. We applaud M&G who conducted a very important study that should be read by any researcher interested in working with large-scale corpora. The fact that M&G mostly failed to find clear evidence in favor of PT&G's main finding motivated us to test PT&G's idea on a subset of the largest archive of German language texts designed for linguistic research, the German Reference Corpus consisting of ∼43 billion words. We only find very little support for the primary data point reported by PT&G.

Keywords: compression - corpus linguistics - information theory - large-scale corpora - *N*-gram modeling - uniform information density

1. Introduction

We thank M&G for an important study that we recommend to any researcher interested in working with large-scale corpora, as it highlights important methodological challenges that can arise when using large-scale linguistic corpora. Also, their study provides several recommendations for researchers conducting such analyses. When applying those best practices, M&G demonstrate that there is "substantially attenuated support for the claim [of

Correspondence should be sent to Alexander Koplenig, Department of Lexical Studies, Leibniz-Institute for the German Language (IDS), 68161 Mannheim, Germany. E-mail: koplenig@ids-mannheim.de

PT&G] that word lengths are more strongly correlated with average information content than with frequency" (M&G, p. 5). We believe that the finding of PT&G is important both for cognitive science and quantitative linguistics, as it provides an information-theoretic explanation for Zipf's law of abbreviation—the tendency of more frequent words to be shorter, a potentially universal property of human languages (Bentz & Ferrer-i-Cancho, 2016). Therefore, we decided to test PT&G's main finding[1] on a subset of the (presumably) largest archive of contemporary German language texts specifically designed for linguistic research, the German Reference Corpus (henceforth, DeReKo, for details, see Kupietz et al., 2010; Kupietz et al., 2018).

2. Data and preprocessing

DeReKo currently contains more than 50 billion [henceforth, b] tokens and comprises a multitude of genres, such as (a large number of) newspaper texts, fiction, or specialized texts, with a current growth rate of ∼3b words per year (Kupietz et al., 2018). Tokenization was carried out using the KorAP tokenizer (Kupietz et al., 2021), the deterministic finite automaton scanning rules of which are based on those of the Apache Lucene tokenizer. Part-of-speech tagging and lemmatization is based on TreeTagger (Schmid, 1994). We do not impose any frequency threshold regarding the inclusion of *n*-grams. For example, only *n*-grams with a token frequency of at least 40 are included in the Google 1T datasets (Brants & Franz, 2006) that is used both by M&G and in PT&G. The same threshold was used in the second main data source of M&G, the Google Books 2012 datasets (Michel et al., 2011). We believe that this threshold severely limits the usability of the Google datasets for most quantitative linguistic research. For example, in our sample of DeReKo consisting of 43,139,394,275 tokens,[2] only 4.03% of all 107,834,517 types occur with a token frequency of more than 40. In a similar vein, only 1.16% of all 6,843,888,373 3-gram types occur more than 40 times in DeReKo.[3]

First, we converted all characters to lower case. The average information content value of each word type based on an unsmoothed 3-gram model was estimated as was done in both PT&G and M&G. In addition, we generated a smoothed 3-gram model (see Appendix for details on statistical modeling). Following M&G's suggested best practices, we then used a basic lemma list of the New High German standard language generated by Stadler (2014) to identify a set of conventionalized word forms. The list is based on DeReKo and consists of more than 325 thousand [henceforth, k] entries. While the lemma list was generated with the help of automatic methods (Stadler, 2014), it meets highest quality standards, since it uses a headword list from a lexicological-lexicographic project based at the Institute for the German Language (IDS) as input where all contained 300k headwords were manually checked for consistency and, if necessary, errors were deleted from the list (Schnörch, 2015; Storjohann, 2016). The list does not contain proper names, abbreviations, or other nonwords (Stadler, 2014) and we also excluded strings that contain cardinal numbers since orthographic word length for cardinal numbers naturally does not match actual production times. To find the set of word forms realized for each lemma (i.e., inflected forms), we first identified all word types where a corresponding lemma is recognized by the TreeTagger. We then merged this list with Stadler's basic lemma list and kept all word forms where the corresponding lemma is also available in the lemma list. The final list consists of $N = 841{,}435$ different word types.

Importantly, note that, like in M&G (note 3), those preprocessing steps were conducted *after* the estimation of average information content values, that is, both the unsmoothed and smoothed language models were estimated for the whole corpus, but words that are not part of the generated list of word types were not used to calculate the association between information content and word length.

Word lengths are calculated as the number of Unicode characters of a word type. In addition, we use production times (i.e., average word pronunciation durations) and syllable counts from the Bavarian Archive for Speech Signals (Schiel, 2010).[4]

3-gram data are freely available in an IDS Repository at hdl.handle.net/10932/00-057D-0921-30F0-F201-D under a non-commerical academic license. To test the reliability of the resulting statistical associations, we generated three versions of the corpus: (i) a full version consisting of all articles (∼43.1b word tokens); (ii) a half version where we randomly chose half of all available articles (∼21.6b tokens); and (iii) a quarter version where we randomly chose one quarter of all articles (∼10.8b tokens).[5] Smoothed and unsmoothed average information contents for all ∼107.8 million [henceforth, m] word types, commented Stata 14.2 (StataCorp, 2015) code, and a replication in R (R Core Team, 2021) are available in a repository on the Open Science Foundation https://osf.io/NREBJ/.

3. Results

Fig. 1 presents the results for the association between orthographic word length and predictability. PT&G include the 25k most frequent words in their analyses.[6] For all three corpus sizes and both for the unsmoothed and smoothed language models, our results do not support PT&G's main finding in this frequency range: in all six scenarios, the 1-gram model (i.e., negative log 1-gram probability as in PT&G, see Appendix) is the better predictor of word length, both in isolation and when the other predictor is held constant.

The same is true if we compute associations for the 12.5k most frequent words. Like M&G (p. 13), we emphasize that the theoretical argumentation in PT&G implies that there actually should be the expected *stronger* relationship between word length and predictability than between word length and frequency for these subsets of words: for example, Fig. 1 shows that the 12.5k most frequent in-dictionary types already account for ∼65% of all 41.1b word tokens in the corpus. We, therefore, believe that this questions the theoretical idea of a communicative efficiency maximization principle as discussed by PT&G.

For the 50k most frequent types, only the unsmoothed quarter model supports a *stronger* relationship between word length and predictability than between word length and frequency. There is support for PT&G's main finding if we include 100k, 200k, 400k, or 800k most frequent in-dictionary word types for the three unsmoothed models. However, the smoothed models, which presumably provide less noisy estimates of information content (see PT&G, Supporting Information Text), only replicate this finding for the 100,000 or 200,000 most

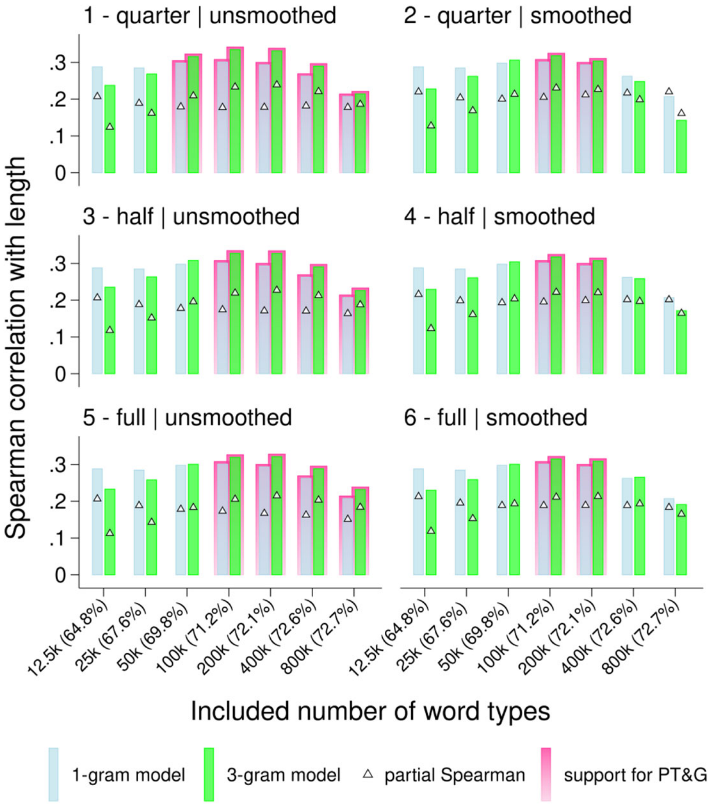

Fig 1. Spearman correlations between orthographic word length and average information content for three different corpus sizes (quarter: ∼10.8b tokens; half: ∼21.6b tokens; and full: ∼41.1b tokens) based on unsmoothed and smoothed 1-gram (blue) and 3-gram models (lime); see Appendix for details on statistical modeling. Triangles show partial Spearman correlations (with the other predictor residualized out). Values on the *x*-axis indicate the number of word types included in the analysis (sorted in descending order by frequency). Values in parentheses represent the cumulated relative frequency of word tokens in the sample. Pink borders indicate results that support PT&G, that is, the Spearman correlation based on the 3-gram model is significantly higher than the Spearman correlation based on the 1-gram model; likewise for the partial Spearman correlations (to determine statistical significance, we bootstrapped the difference between the two Spearman correlation coefficients with 10,000 replications each; likewise for the partial Spearman correlations. A "significant" result indicates that the corresponding bias-corrected 99% confidence interval does not contain zero).

frequent word types. Fig. 1 also demonstrates that the statistical associations are qualitatively highly comparable across varying corpus sizes, which suggests that they are reliable and not the result of estimation errors.

PT&G further theorize that "the amount of information conveyed by a word should be linearly related to the amount of time it takes to produce" (p. 3526). Using actual average production time statistics, we show in Fig. 2 that when partialling out the effect of word frequency, there is effectively no support for any statistical association between production time and predictability. In a similar vein, if we compute associations between information

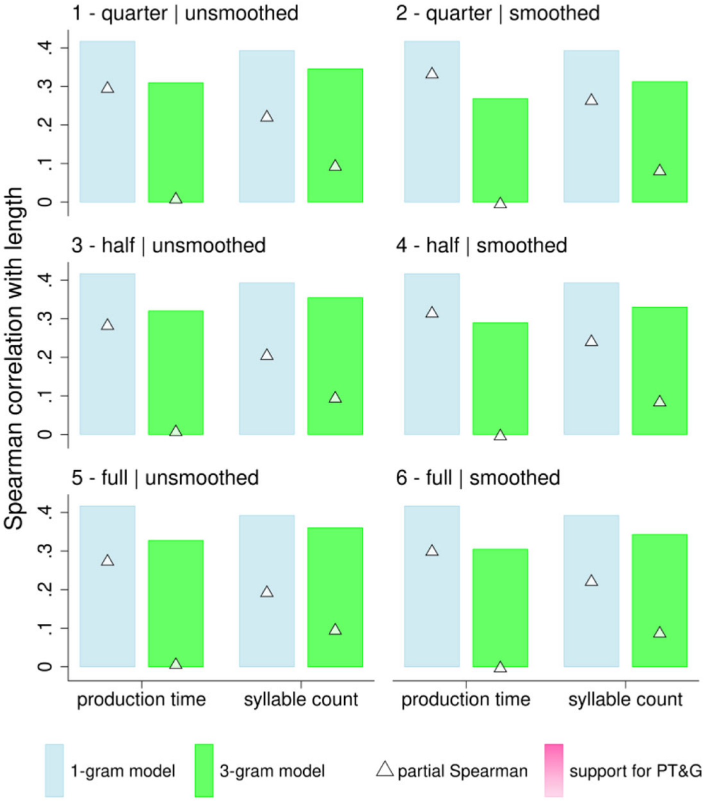

Fig 2. Spearman correlation between average information content and average production time/syllable count for the three different corpus sizes (included word types N = 13,356 for the full corpus and N = 13,555 for the quarter/half corpus, cumulated relative frequency: 57.0% for all three sizes). Other plotting conventions are the same as in Fig. 1.

content/frequency and the number of syllables as a measure of word length, there is no support for PT&G's main finding in neither of the six scenarios.

4. Concluding remarks

While several of the methodological challenges mentioned by M&G regarding large-scale corpora also arise for the German Reference Corpus, DeReKo has the advantage that it does not consist of web-gathered material, but was specifically designed as an empirical basis for linguistic research (Kupietz et al., 2010), resulting in a higher quality standard and thus a potentially better testbed for evaluating hypotheses about structural properties of language. The same is true for the basic lemma list that we used to identify conventionalized word forms (Stadler, 2014). Since—like M&G—we are not able to consistently reproduce PT&G's systematically higher correlation between average information content and word length than between word frequency and word length, we tentatively conclude that there is, as of now, not enough support for PT&G's main finding, at least for German.[8]

Open Research Badges

This article has earned Open Data and Open Materials badges. Data are available at hdl.handle.net/10932/00-057D-0921-30F0-F201-D and materials are available at https://osf.io/NREBJ/.

Notes

[1] As clarified by PT&G (Piantadosi, Tily, & Gibson, 2013: 2), their main finding is "that a word's average in-context surprisal predicted word length better than frequency predicts word length." In what follows, we directly test this primary data point.

[2] The basis for this sample was DeReKo-2020-I (Leibniz-Institute for the German Language, 2020) from which we excluded corpora with high proportions of foreign-language passages, corpora with high proportions of nonredacted texts, and corpora not sufficiently licensed for external use.

[3] To illustrate the effect of truncation in the present context, see Fig. S1, where we replicated the analyses based on truncated corpora, where all 3-grams with a token frequency of less than 40 were excluded from the raw data.

[4] Available at https://www.bas.uni-muenchen.de/forschung/Bas/BasPHONSTATeng.html

[5] Available files are compressed with *xz*. The repository consists of 16 3-gram frequency lists. Each list was generated based on 1/16 of all selected DeReKo articles (randomly assigned). 3-grams are stored using integer IDs for each word type, that is, each word type is mapped to a unique integer (Brants, Popat, Xu, Och, & Dean, 2007). Each list consists of four TAB-separated columns, where the first three columns represent IDs for each of the three words. The fourth column gives the frequency count for the corresponding 3-gram. The available list DATA mapping key ("case_ignore_padded_key.tsv") contains the mapping of words to the corresponding IDs. For the quarter version/half version, we used the first four/eight lists.

[6] For each analyzed language, PT&G included the 25k most frequent words that also occurred in the OpenSubtitles corpus of the OPUS Corpus (Tiedemann, 2012).

[7] We include 3-grams that run across sentence boundaries, but not 3-grams that run across article boundaries. To make the 3-gram model a true probability distribution (Jurafsky & Martin, 2021), we include begin- and end-of article markers («START» and «END») and compute counts accordingly, that is, C(«START» «START» die) represents the frequency of articles that begin with definite article "die."

[8] We thank Carolin Müller-Spitzer, Peter Fankhauser and three anonymous reviewers for input and feedback. We also thank Peter Meyer for helpful conversations regarding our statistical language model, Denis Arnold for creating an IDS Repository that hosts the DeReKo 3-gram data, Oliver Schonefeld for IT support and Sarah Signer for proofreading.

References

Bentz, C., & Ferrer-i-Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In C. Bentz, G. Jäger, & I. Yanovich (Eds.), *Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics*. Tübingen: University of Tübingen.

Brants, T., & Franz, A. (2006). Web 1T 5-gram, 10 European Languages Version 1. Linguistic Data Consortium. https://doi.org/10.35111/MESN-FV79. https://catalog.ldc.upenn.edu/LDC2009T25.

Brants, T., Popat, A. C., Xu, P., Och, F. J. & Dean, J. (2007). Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 858–867). Prague: Association for Computational Linguistics.

Chen, S. F., & Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In 34th Annual Meeting of the Association for Computational Linguistics (pp. 310–318). Santa Cruz, CA: Association for Computational Linguistics.

Cleary, J., & Witten, I. (1984). Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, *32*(4), 396–402. https://doi.org/10.1109/TCOM.1984.1096090

Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing*. 3rd ed.

Knoll, B., & Freitas, N. D. (2012). A machine learning perspective on predictive coding with PAQ8. In *2012 Data Compression Conference* (pp. 377–386). Snowbird, UT: IEEE.

Kupietz, Marc, Cyril Belica, Holger Keibel & Andreas Witt. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In Nicoletta Calzolari, Daniel Tapias, Mike Rosner, Stelios Piperidis, Jan Odjik, Joseph Mariani, Bente Maegaar & Khalid Choukri (eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation. International Conference on Language Resources and Evaluation (LREC-10)*, (pp. 1848–1854). Valetta, Malta: European Language Resources Association (ELRA).

Kupietz, Marc, Harald Lüngen, Pawel Kamocki & Andreas Witt. (2018). The German Reference Corpus DeReKo: New Developments–New Opportunities. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, et al. (eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA).

Kupietz, Marc. (2021). KorAP/KorAP-Tokenizer v2.1.0. Zenodo. https://doi.org/10.5281/ZENODO.5040450. https://zenodo.org/record/5040450 (20 July, 2021).

Leibniz-Institute for the German Language 2020. German Reference Corpus DeReKo-2020-I. Retrieved from https://hdl.handle.net/10932/00-04B6-B898-AD1A-8101-4

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Meylan, S. C., & Griffiths, T. L. (2021). The challenges of large-scale, web-based language datasets: Word length and predictability revisited. *Cognitive Science*, *45*(6). https://doi.org/10.1111/cogs.12983.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Verses, A., Gray, M. K., Pickett, J. P., … Eiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(14), 176–182. https://doi.org/10.1126/science.1199644

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. https://doi.org/10.1073/pnas.1012551108

Piantadosi, S., Tily, H. & Gibson, E. 2013. Information content versus word length in natural language: A reply to Ferrer-i-Cancho and Moscoso del Prado Martin [arXiv:1209.1751]. https://dblp.uni-trier.de/rec/journals/corr/PiantadosiTG13.html?view=bibtex

R Core Team. 2021. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Schiel, F. 2010. *BAStat: New Statistical Resources at the Bavarian Archive for Speech Signals*.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing* (pp. 44–49). Manchester.

Schnörch, U. (2015). Die elexiko-Stichwortliste. In U. Haß (Ed.), *Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz* (Schriften des Instituts für Deutsche Sprache, *12*, pp. 71–90). Berlin/New York: De Gruyter.

Stadler, H. (2014). Die Erstellung der Basislemmaliste der neuhochdeutschen Standardsprache aus mehrfach linguistisch annotierten Korpora. In H. Blühdorn, M. Elstermann, & A. Klosa (Eds.), *OPAL - Online publizierte Arbeiten zur Linguistik, 5*. Mannheim: Institut für Deutsche Sprache.

StataCorp. 2015. *Stata Statistical Software*. StataCorp.

Storjohann P. (2017). elexiko: A Corpus-Based Monolingual German Dictionary. *HERMES - Journal of Language and Communication in Business*, *18*(34), 55. https://doi.org/10.7146/hjlcb.v18i34.25800

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *LREC'12 P*roceedings (pp. 2214–2218). Istanbul: ELRA.

Witten, I. H., & Bell, T. C. (1991). The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, *37*(4), 1085–1094. https://doi.org/10.1109/18.87000

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1: Spearman correlations between orthographic word length and average information content for three different corpus sizes

Appendix

For each type of word in our corpus, we first extract all M sequences of three words, where the third word is w and where M represents the total count of w in the corpus. In what follows, we will denote the total count of a word or word sequence X as $C(X)$, so here, $M \equiv C(w)$.[7] Let us then consider three random variables, U for the first word, V for the second word, and W for the last word in the three word sequence. Each random variable can take on different values, for example, U can take on the value "the," V can take on the value "pale," and W can take on the value "king," for simplicity, we write $C(\textit{the pale king})$ for the total count of the 3-gram "the pale king". More generally, we write $C(uvw)$ for the total count of the 3-gram uvw, where $U = u$, $V = v$, and $W = w$. Accordingly, $P(uvw)$ represents the joint probability of uvw and $P(w|uv)$ represents the conditional probability of w given the two preceding words u and v, that is, uv (Jurafsky & Martin, 2021, chapter 3; Manning & Schütze, 1999, chapter 6; Brants, Popat, Xu, Och, & Dean, 2007).

As suggested by PT&G and replicated by M&G, the average amount of information conveyed by w can be approximated based on all types of different uv sequences that precede w

(i.e., $C(uvw) > 0$) as:

$$\eta^{\textcircled{3}}(w) = -\sum_{u}\sum_{v} P\left(uv|w\right)\log_2 P\left(w|uv\right) \tag{A1}$$

Let us first apply Bayes' theorem to compute $P(uv|w)$, so we can rewrite Eq. 1 as:

$$\eta^{\textcircled{3}}(w) = -\sum_{u}\sum_{v}\frac{P\left(w|uv\right)P\left(uv\right)}{P\left(w\right)}\log_2 P\left(w|uv\right)$$

Since $P(uv)$ can be expanded as $P(v|u)P(u)$, we can write:

$$\eta^{\textcircled{3}}(w) = -\frac{\sum_{u}\sum_{v}P\left(w|uv\right)P\left(v|u\right)P\left(u\right)\log_2 P\left(w|uv\right)}{P\left(w\right)} \tag{A2}$$

To estimate $\eta^{\textcircled{3}}(w)$, we first train an unsmoothed language model by using the maximum likelihood (henceforth, *ml*) estimates of the probabilities in Eq. 2, to approximate $P(w|uv)$, we compute the count of the 3-gram $C(uvw)$ and normalize by the count of the 2-gram $C(uv)$:

$$P_{ml}\left(w|uv\right) = \frac{C\left(uvw\right)}{C\left(uv\right)}. \tag{A3}$$

In a similar vein, we can write:

$$P_{ml}\left(v|u\right) = \frac{C\left(uv\right)}{C\left(u\right)}. \tag{A4}$$

To compute $P(w)$, we write:

$$P_{ml}\left(w\right) = C\left(w\right)/N, \tag{A5}$$

where N denotes the size of the corpus in words; likewise for $P(u)$.

To address the primary finding reported by PT&G, that is, to test if a word's average in-context surprisal better predicts word length than frequency (Piantadosi, Tily & Gibson, 2013), we compare the results of the 3-gram model, that is, $\eta_{ml}^{\textcircled{3}}(w)$, with an unsmoothed 1-gram model, where the average amount of information conveyed by w is approximated as:

$$\eta_{ml}^{\textcircled{1}}(w) = -\log_2 P_{ml}\left(w\right) \tag{A6}$$

In addition, we train a smoothed 3-gram model, where we linearly interpolate all probabilities based on the *prediction by partial matching* algorithm, where escape probabilities are estimated by *method C* (Cleary & Witten, 1984), which is known as *Witten-Bell* smoothing in the language modeling community (Chen & Goodman, 1996; Witten & Bell, 1991). Or put differently, instead of generating a probability distribution that is solely based on 3-gram counts, we blend together predictions of 3-, 2-, and 1-grams and assign higher weights to n-grams of higher order (Knoll & Freitas, 2012), where individual weights are determined by calculating the number of different words observed after a specific n-gram. In a corpus consisting of K different word types, the smoothed (henceforth, *sm*) unigram probability can be written as:

$$P_{sm}\left(w\right) = \lambda_K P_{ml}\left(w\right) + \left(1-\lambda_K\right)\frac{1}{K}$$

where $(1 - \lambda_K) = 1/(N + W)$ and thus:

$$P_{sm}\left(w\right) = \frac{C\left(w\right) + 1}{N + K}. \tag{A7}$$

Likewise for $P(u)$. The smoothed conditional probability of w given v can be calculated as:

$$P_{sm}(w|v) = \lambda_v P_{ml}(w|v) + (1 - \lambda_v) P_{sm}(w),$$

where $\lambda_v = C(v)/(C(v) + \gamma_v)$ is calculated based on γ_v that denotes the number of different words observed after v. Substituting, we can thus write:

$$P_{sm}\left(w|v\right) = \frac{C\left(vw\right) + \gamma_v P_{sm}\left(w\right)}{C\left(v\right) + \gamma_v} \tag{A8}$$

Likewise for $P(v|u)$. In a similar vein, we estimate the probability of w given uv as:

$$P_{sm}\left(w|uv\right) = \frac{C\left(uvw\right) + \gamma_{uv} P_{sm}\left(w|v\right)}{C\left(uv\right) + \gamma_{uv}}, \tag{A9}$$

where γ_{uv}denotes the number of different words observed after uv. Again, we compare the results of the smoothed 3-gram model, that is, $\eta_{sm}^{\textcircled{3}}(w)$, with a smoothed 1-gram model, where the average amount of information conveyed by w is approximated as:

$$\eta_{sm}^{\textcircled{1}}\left(w\right) = -\log_2 P_{sm}\left(w\right) \tag{A10}$$

All language models were generated in Stata/MP on a Linux server (CentOS 7.9.2009) with 756GB of available RAM (more than 90% were needed for generating the language models for the full corpus; computation time: ∼3.6 days).