

The Challenges of Large-Scale, Web-based Language Datasets: Word Length and  
Predictability Revisited

Stephan C. Meylan<sup>1,2</sup> and Thomas L. Griffiths<sup>3</sup>

<sup>1</sup>Department of Brain and Cognitive Science, Massachusetts Institute of Technology,  
Cambridge MA

<sup>2</sup>Department of Psychology and Neuroscience, Duke University, Durham NC

<sup>3</sup>Department of Psychology, Princeton University, Princeton, NJ

Author Note

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under grant no. DGE-1106400 to S.M and NSF grant no. SMA-1228541 to T.L.G. Special thanks to Dr. Steven Piantadosi for sharing materials and lexical information content estimates, to Terry Regier for helpful commentary on early drafts, and members of the Computational Cognitive Science Lab at UC Berkeley for valuable discussion. We also wish to thank Dr. Marc Brysbaert, Dr. Ramon Ferrer-i-Cancho, and two anonymous reviewers for their helpful criticisms and recommendations. Address all correspondence to [smeylan@mit.edu](mailto:smeylan@mit.edu).

## Abstract

Language research has come to rely heavily on large-scale, web-based datasets. These datasets can present significant methodological challenges, requiring researchers to make a number of decisions about how they are collected, represented, and analyzed. These decisions often concern long-standing challenges in corpus-based language research, including determining what counts as a word, deciding which words should be analyzed, and matching sets of words across languages. We illustrate these challenges by revisiting “Word lengths are optimized for efficient communication” (Piantadosi, Tily, & Gibson, 2011), which found that word lengths in 11 languages are more strongly correlated with their average predictability (or average information content) than their frequency. Using what we argue to be best practices for large-scale corpus analyses, we find significantly attenuated support for this result, and demonstrate that a stronger relationship obtains between word frequency and length for a majority of the languages in the sample. We consider the implications of the results for language research more broadly and provide several recommendations to researchers regarding best practices.

*Keywords:* corpus linguistics; noisy channel communication; linguistic universals; information theory; compression; uniform information density; n-gram models

# The Challenges of Large-Scale, Web-based Language Datasets: Word Length and Predictability Revisited

## 1 Introduction

The increasing importance of web-mediated communication in everyday life has produced a wealth of naturalistic data capable of shedding light on longstanding questions in language research. These new datasets include wide-coverage web scrapes of written material (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009; Davies & Fuchs, 2015), as well as more targeted datasets focusing on specific formats and document types, *e.g.*, movie subtitles (Brysbaert, New, & Keuleers, 2012; Tiedemann, 2012), encyclopedia articles (Denoyer & Gallinari, 2006), news articles (Chelba et al., 2013), and web forums (Zhang, Culbertson, & Paritosh, 2017). The scale of these datasets is particularly useful for language research, in that many linguistic phenomena of interest are vanishingly rare, and making robust inferences about such phenomena requires large samples (Tomasello & Stahl, 2004). These web-based datasets also provide researchers with access to a larger and more typologically diverse set of languages, in principle ideal for evaluating claims about linguistic universals (Bochkarev, Solovyev, & Wichmann, 2014; Mahowald, Dautriche, Gibson, & Piantadosi, 2018; Petersen, Tenenbaum, Havlin, Stanley, & Perc, 2012; Piantadosi et al., 2011). Further, these datasets have played a critical role in training recent language model architectures which require larger-scale inputs than traditional linguistic corpora can provide (Devlin, Chang, Lee, & Toutanova, 2019; Lewis et al., 2020).

However, the use of these large-scale web-based datasets for corpus analysis and language model training may introduce previously unknown or under-appreciated challenges. First, the construction of these corpora often depends heavily on automated processing, *e.g.*, automatic language identification to separate scraped web pages into different linguistic sub-corpora. These pre-processing steps may include many non-trivial (and potentially under-documented) design decisions that have implications for downstream research. Different decisions — including reasonable ones — in data analysis

may lead to substantively different results (Silberzahn et al., 2018).

Second, the true generative process for these datasets is generally unknown, and the resulting datasets may include a large proportion of content that was not produced by people, but rather by bots or other automated methods. While responsible research with these datasets entails conducting many tests<sup>1</sup> to validate assumptions regarding the nature of datasets, the scale of these datasets means that such checks cannot be exhaustive. As such, it is imperative that researchers carefully consider the integrity and appropriateness of a linguistic dataset for the research questions at hand.

In the current work, we revisit a well-known result regarding the relationship between communicative efficiency and the observed pattern of word lengths in natural languages presented in Piantadosi et al. (2011). We find that the results obtained in that paper depend on a number of analytical choices, including the choice of text encoding, the inclusion of strings that are not conventional wordforms in the relevant language communities, and the use of rank-ordered word frequency to select the set of word types in the over which correlations are computed. We propose and motivate three specific best practices relevant to the target work: maintaining the appropriate text encoding throughout the analysis, conducting the analysis on the subset of words in each language that are found in large dictionary, and choosing words from a matched set of semantic concepts to minimize the potentially spurious effect of typological variation among languages. Rather than low-level points regarding implementation, we demonstrate that these three points speak to longstanding, basic challenges in language research: deciding how to represent words for computational analyses, defining what constitutes a valid word, and selecting an appropriate sample of words. The scale of new datasets does not resolve these challenges; rather, they remain steadfastly relevant. We additionally emphasize the importance of *multiverse* analysis (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016), where researchers test their hypotheses across the combinatorial space of possible

---

<sup>1</sup>Historically known as “sanity checks”, we encourage the adoption of the term “canary.”

Table 1

*Overview of challenges and best practices in working with large-scale, web-based, and multilingual linguistic corpora*

Challenge	Best Practice	Caveats	Sections
Identifying conventional words in a language community	Filter by a natural language dictionary, e.g., Aspell	May exclude some proper nouns; depends on quality of dictionary (esp. for inflected forms)	2.1 & 5.2.1
Representing words	Maintain UTF-8 (or UTF-16) throughout preprocessing & analysis	None	2.2 & 5.2.2
Selecting a matched set of words across languages	Use a matched set of words like the Intercontinental Dictionary Series (Key & Comrie, 2015)	Selects for higher frequency, shorter words	2.3 & 5.2.3
All of the above [and many more]	Evaluate the robustness of the result over a range of reasonable preprocessing and analysis choices	None	5.2.4

analytical decisions.

In conducting a reproduction and extension of Piantadosi et al. (2011) using these best practices for working with web-based language data, we find substantially attenuated support for their finding that word lengths are more strongly correlated with in-context information content than with word frequency. This result clarifies the burden of proof required for claims regarding cross-linguistic correlational designs. By highlighting best practices in using web-based corpora and their consequences for the analysis (Table 1), we bring attention to several general problems in the emerging use of web-based corpora in language research, as well as corpus-based methods more generally.

## 2 Case Study: Word lengths and predictability

The complex relationship between the structural regularities of languages and the cognitive processes supporting their use remains an active area of investigation in cognitive science, linguistics, psychology, and allied fields. Along with learnability, efficiency – the minimization of effort on the part of the speaker and listener to successfully transmit a message – has been posited to be one of the major pressures that determines the regularities observed across natural languages (Ferrer-i-Cancho, 2018; Ferrer-i-Cancho et al., 2013; Gibson, Bergen, & Piantadosi, 2013; Gibson et al., 2019; Menzerath, 1954; Zipf, 1949). Under this view, commonalities across the world’s languages may not be attributed to a shared, language-specific genetic endowment (*e.g.*, a language acquisition device like that proposed by Chomsky, 1965), but rather reflect “stable design solutions” that are likely to emerge through the process of language evolution to better serve the communicative function of language (Evans & Levinson, 2009).

One such commonality that has been attributed to communicative efficiency is the robust relationship between word length and word frequency: across the world’s languages, frequent words tend to be short. Though already well established by the end of the 19th century (von der Gabelentz, 1891), the relationship is now most widely known from the work of George Kingsley Zipf (Zipf, 1935), and has been attributed to him as “Zipf’s Law of Abbreviation” or “Zipf’s Law of Brevity”.<sup>2</sup> Zipf posited that this relationship of length and frequency emerges from speakers’ proclivity to minimize total articulatory effort by using the shortest form with those words that are used most often, following what he called the *Principle of Least Effort* (Zipf, 1949). Subsequent work has rigorously evaluated the robustness of the Law of Abbreviation in humans (Bentz & Ferrer-i-Cancho, 2016), documented its emergence in iterative experimental designs where participants alter miniature artificial languages (Kanwal, Smith, Culbertson, & Kirby, 2017), and established

---

<sup>2</sup>The better-known Zipf’s Law presented in the same work, concerns the power-law distribution of word frequencies.

its mathematical foundation based on information theory and recent extensions (Ferrer-i-Cancho, Bentz, & Seguin, 2020; Ferrer-i-Cancho et al., 2013). Another line of research demonstrates that Zipf’s law of abbreviation extends, to varying degrees, to the communication systems of other non-human primates (Gustison, Semple, Ferrer-i-Cancho, & Bergman, 2016; Heesen, Hobaiter, Ferrer-i-Cancho, & Semple, 2019; Huang, Ma, Ma, Garber, & Fan, 2020).

While Zipf proposed a speaker-driven explanation for the relationship between word frequency and length, subsequent work has posited a greater role for the listener in shaping language properties, especially in light of an increasing appreciation of the role of prediction in language comprehension (*e.g.*, Altmann & Kamide, 1999; Levy, 2008). Information theory provides a way to characterize speech in terms of an idealized *noisy channel*, where a message is encoded by a speaker, sent over a channel which may introduce errors, and decoded by a listener (Shannon, 1948); further, this framework provides a formal basis for quantifying a message’s *information content* as a function of its probability under the listener’s expectations. This allows for a principled characterization of maximal efficiency as the optimal use of *channel capacity*, or transmitting the maximum amount of information per unit time while maintaining a given probability of transmission error (though see Ferrer-i-Cancho, 2017 for a critique of this formalization). To maximize communicative efficiency, speakers should systematically vary their productions (*e.g.* word choice or speech rate) to achieve an information profile that maximizes channel capacity while taking into account the expectations of the listener. First articulated by Fenk and Fenk (1980), this basic logic holds for two proposals, known as *entropy rate constancy* (Aylett & Turk, 2004; Genzel & Charniak, 2002) and *uniform information density* (UID, Levy and Jaeger 2007). An alternative proposal offers that this systematic variation may emerge from more general principles of compression (*e.g.*, Ferrer-i-Cancho et al., 2013), and that more specific optimization accounts are not necessary (Ferrer-i-Cancho, 2017).

Systematic variation maximizing communicative efficiency could arise from intentional

lexical selection by speakers, who could choose a shorter wordform whenever a word is more predictable from context, *e.g.*, *rhino* vs. *rhinoceros*. Indeed, Mahowald, Fedorenko, Piantadosi, and Gibson (2013) found exactly such a preference in a web-based behavioral study where participants were prompted to provide a continuation for given high- or low-predictability contexts. In time, a shorter wordform may come to be the dominant one in the language (*e.g.*, *autobus* is now almost entirely eclipsed by *bus* in most English dialects). Alternatively, words may undergo more subtle phonetic processes, in terms of duration (Seyfarth, 2014) or fidelity to articulatory targets (degree of phonetic reduction). In either case, the acoustic form of a word may vary as a function of listener’s expectations in the contexts in which it appears, above and beyond language-wide word frequency statistics.

Piantadosi et al. (2011) tested the hypothesis that listener-oriented efficiency influences lexical forms by examining whether word lengths are better predicted by average in-context predictability than by word frequency. As a measure of predictability, they computed average Shannon information content (or context-conditional surprisal) across the contexts in which a word appears, and comparing that with word frequency. For consistent scaling, they operationalized frequency as the negative log probability of each word without reference to conditioning context, or *unigram surprisal* (see Methods of the current work for further details). They conducted this analysis using web-scale corpora from eleven European languages, and found that average average information content exhibits a stronger rank correlation with word lengths than unigram surprisal with word lengths across the languages in the sample.

This result has attracted criticism on its theoretical merits. Ferrer-i-Cancho and del Prado Martín (2011) demonstrated that the same result could emerge under a random typing model, and thus did not constitute strong evidence of communicative optimization in which listeners use the preceding context (though see Piantadosi, Tily, & Gibson, 2013 for several counterarguments). Further, the cost function for optimization is not explicitly



stated (compare with, e.g., i Cancho, Bentz, & Seguin, 2019). In the present work, we open a separate line of critique, that several of the methodological choices in Piantadosi et al. (2011) limit the generality and robustness of its conclusions. In conducting a reproduction and extension of the main study from that work, we find substantially attenuated support for the claim that word lengths are more strongly correlated with average information content than with frequency.

We begin by investigating three specific methodological choices made by Piantadosi et al. (2011) and their implications: the lexical status of “words” in the analysis (2.1), the treatment of character encoding (2.2), and the role of morphological variation on the composition of the wordlist under analysis (2.3). In the first two sections, we test these methodological choices independently, keeping all other aspects of the analysis constant (the critique raised by Section 2.3 rests on theory rather than an empirical demonstration). We then present Methods, Results, and Discussion for a newly-conducted analysis in which we cumulatively apply the proposed methodological improvements.

## 2.1 Items in Analysis: Words or Strings?

A classic problem in corpus-based language research is distinguishing linguistic from non-linguistic material. Piantadosi et al. (2011) took an inclusive approach, analyzing the relationship between word predictability and word length among the 25,000 most frequent strings in each language in the Google 1T datasets that appear one or more times in the OPUS corpus (Tiedemann, 2012). The authors motivated this inclusion criterion as appropriate for evaluating the correspondence between string length and predictability in context across a broad range of different kinds of linguistic materials, including speech, book contents, and web page contents. However, as a consequence of forgoing additional filtering steps on an extremely noisy web scrape, a relatively high proportion of strings in their analysis are of questionable linguistic status with respect to the nominal language of the dataset. This includes the possible inclusion of extensive linguistic content from

languages other than the target language, especially English in the other datasets. While language contact and exchange, including the gradual process of loanword adaptation, are standard processes of language evolution (Kenstowicz, 2007), we argue that the Google 1T dataset—web scrapes where the source language was identified by a relatively unsophisticated machine learning method—and OPUS—crowd-generated movie subtitles—allow for high levels of data pollution. Moreover, these words may have distinctive profiles in terms of their length, frequency, and average information content, such that their inclusion may drive the obtained results.

Here we use a dictionary-based filter to identify the set of words under analysis<sup>3</sup> While this approach may erroneously exclude some words from informal registers (e.g., *rebonjour* and *waouh* in French in Table 2), it should identify a set of core conventionalized word forms among which the relationship between average information content, word frequency, and word length can be evaluated. According to the theoretical proposal outlined in Piantadosi et al. (2011), a stronger relationship between average information content and word length *should also* hold among this set of word types.

To investigate the composition of the words analyzed in Piantadosi et al. (2011), we use Aspell dictionaries to sort strings into *in dictionary*, *out of dictionary* and *English* categories (for the English dataset, word types were sorted only into the first two categories). Aspell is a UNIX command line utility for spell checking that supports language-specific dictionary- and rule-based- spell checking; in that it serves as the shared spell checking backend for many computer applications, the vocabularies are up to date and extremely large by comparison to traditional dictionaries. Because it implements rule-based grammars, Aspell can also evaluate words with complex affixal morphology. We use the following coding procedure: Strings found in the relevant dictionary are marked as “In Dictionary.” Strings in a language’s dictionary that are also present in English are marked as “In Dictionary”, e.g., Spanish *pan* (bread). Strings that do not appear in a

---

<sup>3</sup>We do *not* use the dictionary-based filter to restrict the input to the  $n$ -gram models used to compute in-context predictability

specific language’s dictionary, but do appear in English, are marked as “From English.” Strings appearing in neither the specific language dictionary nor the English dictionary are marked as “Not in Dictionary.”

Assigning the strings in the original wordlists used by Piantadosi et al. (2011) to these three categories reveals that a substantial proportion of strings are *not* among the commonly accepted word types in the relevant languages (Figure 1). For the Czech, French, Portuguese, Spanish, and Swedish items used by Piantadosi et al. (2011), more than 10% of words are found in an English dictionary, but *not* in the dictionary of the nominal language. Examples of such items, sorted by frequency in the corpus, are presented in Table 2. While loanwords are expected for samples of most languages, these rates are significantly higher than those found for Google Books 2012, where the same analysis yielded rates of English-only types from 0 to 3% of word types in non-English datasets. A second issue is that a high proportion of strings in the analysis are found in *neither* the relevant language-specific dictionary nor English. These types include a high proportion of misspellings, low-frequency proper nouns, acronyms, emoji, and HTML tags which were not removed by the preprocessing procedure in Piantadosi et al. (2011).

Admitting these two categories of strings into the analysis could have minimal effects on the overall results if their members behave in a relatively similar fashion to words widely accepted as belonging to the language. However, an analysis of the proposed relationship (information content vs. length) *within* each of these categories shows that this is not the case. We find that the correlations computed over the in-dictionary subsets across languages (Figure 2, column 2) do not reproduce the global pattern for each language found in (Piantadosi et al., 2011) (column 1). While a stronger correlation between length and in-context information content obtains for within-dictionary types for Dutch, English, French, and Portuguese, we find the opposite preference among the remainder of the languages. Out-of-dictionary types show no strong preference (column 3), and types from English (for non-English datasets) show a stronger relationship between

Table 2

*A random sample of “word” types in the analysis from Piantadosi et al. (2011) that are not found in the Aspell dictionary corresponding to the nominal language of the dataset.*

Czech	muhu iuventus terezo vitali zdn farmaceutik cvm latine mauser ujeme beggin johne nikolase lidmila mooney hakuna kreditka carnevale hansì homeo
Dutch	uurs klotejaar neuman heee jame roodlicht delige kzal ern euhm plaatsers hej jaah twas tif gdc ehhl ehk tnx haw
English	itat ordinator semitism cest montel eet svensk roget ppb xn stang tep nge ance rrr ahh hahaha hmmm eam hoh
French	djamena ouaip rebonjour waouh ahhl ndt deposez antioche pauley riverdance lls pffff andro ahk naire iwa adios argh pof alembert
German	hallihallo jaaaa hler nunja siert tigen lasset tive duane nagut achja iich rapie denkste bengt anm uhh pssst genaugenommen nder
Italian	goditela konishi iume piuttosto poiizia ehmm signorinella dipartenza tinkertoy hiruta niene larkspur problema matrimony amram moiti iotta kittle paico oscorp
Polish	ehm kapler ciany jaaaa nienajlepiej rbit fredzia taaaak taaa danke yep sadzac upi sasz pomyslcie tce archie niw tyros mumin
Spanish	racias rankeado iah demotores cest mins erlich lbs although bles khz holaaa tante ghz mph uff prentice btu however nidad
Portuguese	metai grupo caezinhos khz lidade tenny lsm tante eev wez clack aquanaut oberst mins btus reconciliacao braddock lft descon although
Romanian	fives keeling bodin vampira georgette raisin rupeee wigs dba minimizeaza factiunea grasule rezultate sottotitoli encounters however tortoise soren notitele inches
Swedish	ningar nepp nyliga tjaa podnapisi tionen derna terna snyggve prenum pyjamaser mln gaeliska oops etruskiska ligen raken neeej jupp ivoir

frequency and word length (column 4).

Using this tripartite categorization, we can investigate how the somewhat unintuitive correlations in Figure 2 emerge from the three subsets across the languages in the sample. This analysis suggests that the key result in Piantadosi et al. (2011) reflects instances of Simpson’s paradox, where correlations in the aggregate dataset obscure correlations among important subgroups in the sample (Blyth, 1972). Global correlations (i.e., correlations computed over word types in all three categories) are high between mean in-context information content and word length, and low between word frequency and word length. Global correlations are high in the first case because out-of-dictionary strings and words from English are on average *shorter* than in-dictionary words, and have lower average information content (red and green density plots in Figure 3, right), while in-dictionary

words are *longer* and have higher average information content (blue density plots in Figure 3, right). The positions of these three different categories make for a strong global correlation, but much weaker per-category correlations as observed in Figure 2.

In the second case of Simpson’s paradox, the strength of the relationship between unigram surprisal and word length is depressed by the inclusion of all three categories. Word types from English have a similar distribution of frequency (and thus negative log probability) to in-dictionary words, and those not found in either dictionary tend to have a *higher* negative log probability than those from the language (Figure 3, left). This disguises a stronger correlation between unigram surprisal and word length among in-dictionary types.

Said otherwise, for many of the language in the analysis, words from English and those found in neither dictionary — short and highly predictable, yet relatively infrequent — deflate the correlation between frequency and character length, and inflate the correlation between information content and character length. Taken together, these analyses suggest that the obtained correlations in Piantadosi et al. (2011) are highly dependent on the choice of strings in the analysis, and that in-context information content is no more predictive than frequency of word length when the correlation is evaluated on words found in the dictionary for the relevant language.

## 2.2 Text Encoding: UTF-8 vs. ASCII

A second standard problem in computational analyses of language is choosing *how to represent words*, or how to reflect the orthographic conventions of a community in the distinctions between words. The representation of words as strings on modern computing systems requires the specification of a “character encoding” standard, which maps binary sequences (interpretable as numerical *code units*) to human-readable onscreen character representations, well-known examples of which include ASCII and UTF-8. Piantadosi et al. (2011) converted word tokens from the Google 1T datasets (Brants & Franz, 2006, 2009) to

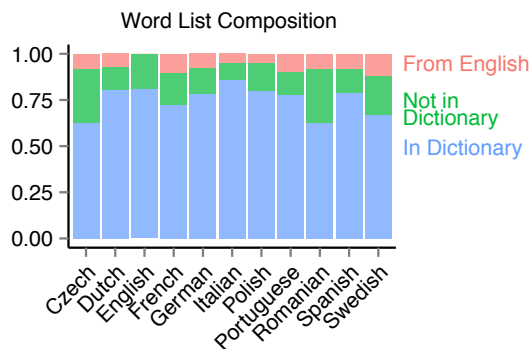


Figure 1. A significant proportion of the word types in Piantadosi et al. (2011) are not found in the relevant language’s dictionary. English word types are especially common in the datasets that are nominally representative of other languages.

closest ASCII equivalents. For English, this is an inconsequential data processing choice: almost all English words can be encoded losslessly using ASCII character representations (*i.e.*, representing them with the letters a-z), and operations with ASCII representations are significantly faster—often an order of magnitude so—than those with UTF-8 encoded text (*e.g.*, ö, ô, ò, ò, ø). But while this mundane data processing choice has minimal implications for English, it is highly consequential in languages that make use of non-ASCII orthography, especially diacritical marks. Psycholinguistic research suggests that accented letters have separate representation in languages where the phonology of accented letters diverges from that of unaccented letters (Chetail & Boursain, 2019; Perea, Fernández-López, & Marcet, 2020). Converting UTF-8 to ASCII may thus merge multiple distinct forms, *e.g.*, Spanish *si* (approximately English “if”) and *sí* (“yes”), in effect collapsing the statistical profiles of multiple word forms. This may have complex effects on both frequency and average information content estimates. An analysis of average information content estimated from UTF-8 encoded datasets vs. ASCII-encoded datasets reveals that this data processing step is indeed consequential for some languages (see Methods for additional details). While the two measures are almost perfectly correlated in

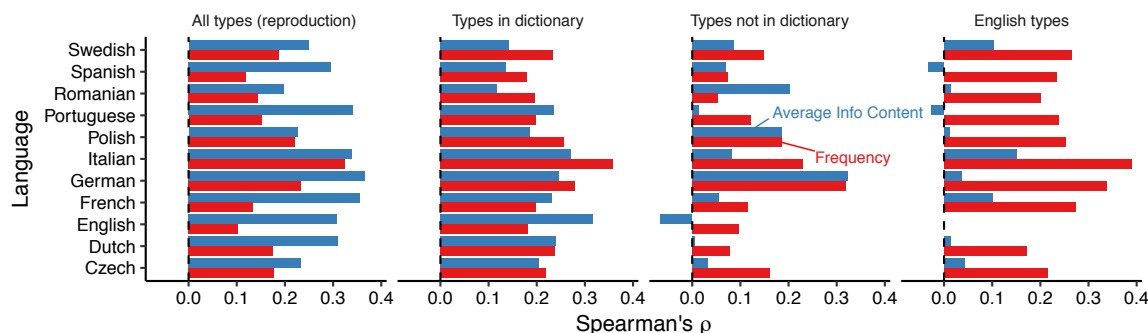
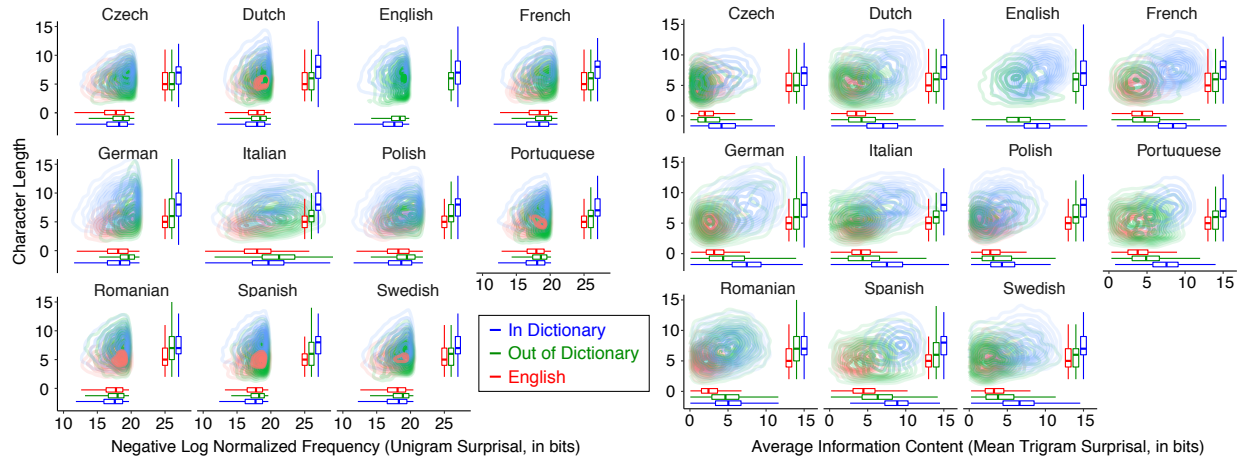


Figure 2. Piantadosi et al. (2011) found higher global correlations between word length and in-context predictability (average information content) as measured by mean trigram surprisal (blue bars in panel 1) than between word length and frequency (treated as unigram surprisal), red bars in panel 1). The pattern is substantially weakened or reversed within each of three sub-groups of word types: those that are found in the relevant dictionary (panel 2), those not found in a dictionary (panel 3) and those found in English (panel 4).

English ( $R^2 = .994$  in the 25,000 most frequent words; Figure 4, left), languages that make use of a broader inventory of characters in their orthography have substantially more unexplained variance across words (*e.g.*, Czech,  $R^2 = .464$ ; see Figure 4, center). Further, an examination of the distribution of the differences in average information content estimates in Czech (Figure 4, right) shows a significant rightwards skew, indicating that many words are found to have inflated average information content estimates when computed from the ASCII-encoded dataset.

## 2.3 Confounds from Typological Variation

A third common challenge in corpus-based language research is selecting an appropriate sample of words for analysis. This can be particularly challenging when words, constructions, or syntactic structures must be compared across languages. Of particular interest here, languages vary in their richness of their morphological processes, or the ways



*Figure 3.* Density plots depict the relationship between word length and unigram surprisal (negative log probability, a transformation of normalized frequency) and average information content (mean negative log probability under a trigram model). Words are stratified into three categories: those in each language’s Aspell dictionary (blue), those in a spelling dictionary for English (red), and those not in either (green). Marginal boxplots show the median, inter-quartile range (IQR), and  $1.5 \times \text{IQR}$  for each of the groups. Densities are normalized per category. This analysis suggests that English and Out-of-dictionary words may have different profiles than In-Dictionary words, and that the higher correlation between average information content and word length is in large part driven by their inclusion in a case of Simpson’s paradox.

in which wordforms can be constructed from units smaller than the word. Among the languages present in the sample in Piantadosi et al. (2011), there is substantial variation in various dimensions of morphological complexity. This includes the degree of inflectional synthesis of verb forms (Bickel & Nichols, 2005), propensity for pronominal forms to attach to verb forms (e.g., pronominal clitics in Spanish), and the richness of case-marking systems for nouns, which range from two cases in English (Quirk, Greenbaum, Leech, & Svartvik, 1985) to 6+ cases in Polish (Bielec, 1998). This variation can have substantial consequences for the composition of the set of word types under analysis. For example,



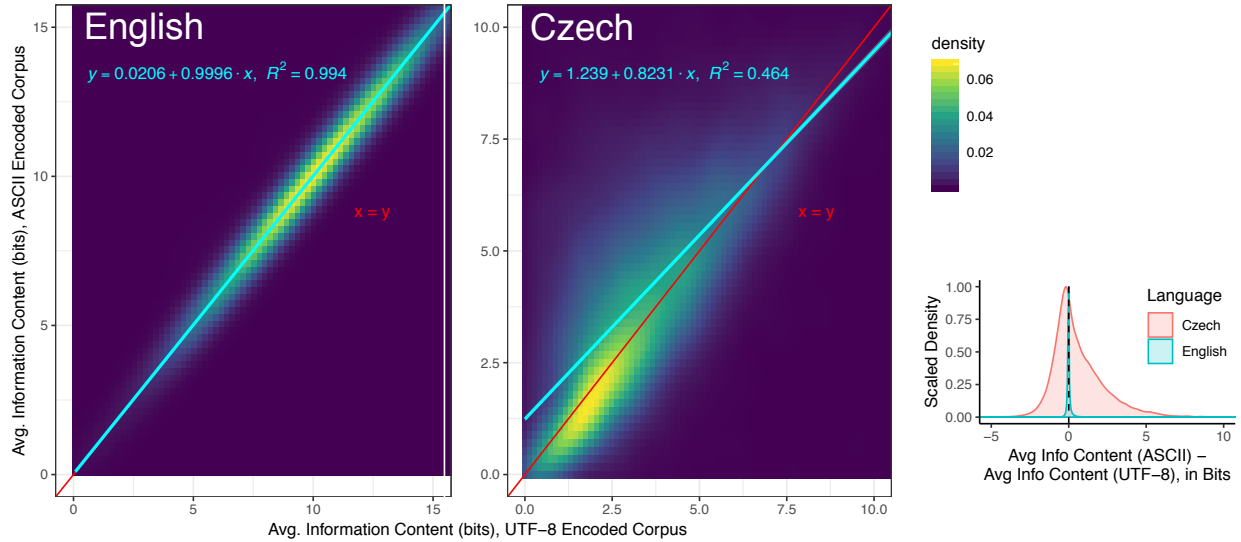


Figure 4. Correlation between average information content estimates computed from ASCII and UTF-8 encoded datasets for the English language (left) and Czech (right). Red indicates the line of identity  $y = x$ , expected if ASCII encoding has no effect on computed information content estimates. Cyan indicates the best linear fit for the language. The right panel shows the distribution of the difference in the two datasets. The subplot shows the distribution of the differences for the two languages. This analysis suggests that converting UTF-8 strings to closest ASCII equivalents introduces significant variance and may obfuscate underlying relationships in some languages, like Czech.

whereas the English 1T dataset has six inflections of the lemma *sell* (e.g., *sell*, *sells*, *sold*, *selling*, *seller*, *sellers*), Spanish, a language with much richer tense system for verbs and pronominal clitics, has 22 in the top 25,000 most frequent words, e.g., *véndemelo*, or *you (informal) sell (imperative) me it (masculine)*. Depending on what parts of speech have high morphological complexity, large differences may emerge in the composition of the word list under analysis across languages: Spanish may have a preponderance of verb forms from a few verbs with high lemma frequency, while a language with extensive nominal case marking would have more of the wordlist coming from nouns with high lemma frequencies (e.g., Russian, per Wade 1992). Again, it is unclear what bias may be introduced by this

sort of variation in the list of word types under analysis: the implications may vary by language and interact in complex ways with the two issues above (text encoding and the choice of the set of word types under analysis). We introduce a method below to address this problem by selecting items from a list of semantically-matched concepts across languages.

### 3 Methods

The current work follows the basic methodology of Piantadosi et al. (2011) in 1) selecting a set of word forms to analyze in each language 2) computing word lengths from orthographic representations, and unigram surprisal and average information content for each of these words and 3) computing the correlation between these two measures for each language in the sample. However, we make three substantively different methodological choices than the preceding work, as motivated above: limiting the set of strings under analysis to those conventionally accepted as belonging to the language’s lexicon, enforcing UTF-8 text encoding throughout, and computing the correlation over a matched sample of lexical concepts rather than the highest frequency tokens in the sample. We also extend these analyses to an additional cross-linguistic dataset that has since been made available: Google Books 2012 (Michel et al., 2011). As the product of scanned books, Google Books may be subject to lower levels of data pollution than the internet-derived Google 1T corpus.

#### 3.1 Datasets for Frequency and Information Content Estimates

We obtained Google Web 1T datasets from the Linguistic Data Consortium (Brants & Franz, 2006, 2009), the Google Books 2012 datasets from [storage.googleapis.com/books/ngrams/books/datasetv2.html](http://storage.googleapis.com/books/ngrams/books/datasetv2.html) (Michel et al., 2011), and OPUS (2013) from [opensubtitles.org](http://opensubtitles.org) (Tiedemann, 2012). We discard all *n*-grams consisting only of punctuation, and remove punctuation appearing in other (non-punctuation) words, with the exception of apostrophes. Following Piantadosi et al. (2011) we make the simplifying assumption that the tokenized orthographic forms

correspond to psychologically “real” words in the lexicon of speakers; while the lexical status of orthographic words is an active area of investigation (*e.g.*, German compound nouns, Lorenz & Zwitterlood, 2016), it holds for the vast majority of word forms in the analysis. We convert all characters to lowercase using the relevant POSIX locale. In the case of Google Books 2012, records from earlier than 1800 are discarded, in light of a higher error rate in optical character recognition among these older documents. Counts are stored using ZS, a specialized file format for efficient retrieval of  $n$ -gram counts (Smith, Submitted).

### 3.2 Estimating In-Context Information Content

Following Piantadosi et al. (2011), we first analyze a word list constructed from the 25,000 most frequent words in each dataset, filtered by the tokens present in the OPUS subtitle corpus.<sup>4</sup> For each word type  $w$ , we compute a measure of frequency, negative log probability (unigram surprisal). We also compute a measure of average in-context predictability, or average information content, as the negative log trigram probability weighted by context frequency,  $-\frac{1}{N} \sum_{i=1}^N \log P(W = w | C = c_i)$ , where  $c_i$  is the context for the  $i$ th occurrence of  $w$  and  $N$  is the frequency of word  $w$  in the dataset. Consistent with Piantadosi et al. (2011), we do not employ a smoothing scheme for the  $n$ -gram models.

We then analyze a second set of 25,000 word types for each language, maintaining the OPUS filter while additionally limiting to in-dictionary word types. We limit to the word types recognized by the UNIX utility Aspell for locale corresponding to each dataset.<sup>5</sup> U.S. English and European Portuguese are used for English and Portuguese, respectively. This step removes proper nouns including person and place names, acronyms, and loanwords from other languages.

---

<sup>4</sup>We use the 2013 release of the dataset.

<sup>5</sup>We allow uppercase forms for German, which capitalizes all nouns by convention.

### 3.3 Matching Wordlist Composition Across Languages

Here, we propose a control that limits the number of wordforms per each lemma in the analysis. To do this, we conduct the analysis over a subset of wordforms from each language intended to match semantic content to the degree possible. Specifically we use elicited labels for a matched set of concepts from the Intercontinental Dictionary Series (Key & Comrie, 2015). Similar to Swadesh lists (Swadesh, 1971), IDS datasets contain sets of synonyms matched on conceptual content across a broad sample of languages. Unlike Swadesh lists, IDS datasets include a larger set of approximately 1300 concepts, of which a subset (800-1200) are present in a given language. The IDS datasets contain the unmarked form of the concept, e.g., the singular in the case of an English count noun, or the infinitive in the case of a Spanish verb. Conducting the analysis on the IDS subset thus results in a more similar set of lemmas in the analysis, while reducing the effect of morphological complexity on the composition of the wordlist. Additionally, it permits the treatment of concept as a factor in modeling.

### 3.4 Code Availability

Information content estimates and analyses presented here are available in a repository on the Open Science Foundation [osf.io/np2bs](https://osf.io/np2bs). Lexical information content estimates were calculated with the `ngrawk2` library, [github.com/smeylan/ngrawk2](https://github.com/smeylan/ngrawk2)

## 4 Results

### 4.1 Maintaining Character Encoding

We first reproduce the analysis from Piantadosi et al. (2011), but recompute the frequencies and average information content estimates over UTF-8 representations. We use the same filtering rule of including only wordforms that appear at least once in the OPUS subtitle corpus (*i.e.*, not using a dictionary-based filter). In addition to the Google 1T dataset used by Piantadosi et al. (2011), we also evaluate the correlation among datasets in

Google Books 2012. Even with this minimal data processing manipulation, we find attenuated support for the pattern of results found by Piantadosi et al. (2011) (Figure 5). Among the Google 1T languages, the difference in correlations for Romanian fails to reach significance, and Czech and Polish demonstrate a statistically significant higher correlation between frequency and word length than between average information content and word length. In Google Books 2012, Hebrew shows a stronger correlation between frequency and word length, while Spanish shows no significant difference. While the above languages diverge from the previously-obtained results, English, French, German, and several other languages demonstrate robustly higher correlations for average information content and length, consistent with Piantadosi et al. (2011).

## 4.2 Analyzing Word Types in Dictionaries

Using UTF-8 encodings adopted in the previous section, we next investigate the consequences of enforcing a stronger constraint on words entering the analysis. The enforcement of this stronger constraint on the word types under analysis results in a substantive change in the pattern of results (Figure 6). We again emphasize that the theoretical arguments in Piantadosi et al. (2011) (p. 3526) imply that there should be a *stronger* relationship between word length and predictability than word length and frequency among this subset of words. If the correlation emerges from qualitatively different behaviors among different types of strings (in-dictionary vs. out-of-dictionary), then the theory must be amended to account for this fact. Among the eleven languages from the Google 1T corpora, only two (English and French) show significantly higher correlations for in-context predictability and word length than frequency and word length. Only two of the eleven 1T languages, English and French, show the pattern found by Piantadosi et al. (2011). 5 of 11 languages (Czech, Dutch, Polish, Spanish, and Swedish) show the opposite pattern, with a significantly higher correlation for frequency and word length. Neither predictor is significantly stronger among three remaining languages

(German, Italian, and Portuguese).

### 4.3 Addressing Confounds From Typological Variation

In the third analysis, we limit the set of types under consideration to those in the Intercontinental Dictionary Series (IDS). This limits the number of word types in the analysis associated with any one lemma, in that morphologically unmarked forms are elicited to construct the dictionary entry for each language. Matching semantic content to the items in the IDS means that the identity of each concept can be used as a control variable. We include average information content and frequency estimates from OPUS for this analysis because the words included on the IDS word lists are sufficiently high frequency that data sparsity is less problematic, such that smaller datasets such as those in OPUS may provide relatively accurate average information content estimates. For each dataset (OPUS, Google 1T, and Google Books 2012) we compare two mixed-effects regression models that predict word length, one using frequency (again operationalized as unigram surprisal) and the other in-context predictability (again operationalized as average information content under a trigram model). Unigram surprisal, average information content, and word length were all Z-scored within each language. Language and concept were both treated as random intercepts. This corresponds to the intuition that different concepts have different average word lengths across languages, and that some languages may tend towards longer orthographic representations than others.

Does frequency or in-context predictability better predict the pattern of word lengths seen across languages for this matched set? We compare two models. In Model 1, average information content of IDS words in each language is treated as the sole fixed effect. In Model 2, unigram surprisal of IDS words in each language is treated as the sole fixed effect. Both models are fit with maximum likelihood estimation (vs. restricted maximum likelihood estimation) because they differ in fixed effects. Across the three datasets, Model 2 demonstrates a better fit according to model log likelihood, AIC and BIC (Table 3). In

Table 3

*Comparison of mixed effects models predicting word length for word types in the Intercontinental Dictionary Series (IDS) in three datasets. Average Information Content (average trigram surprisal under a trigram model), unigram surprisal (negative log probability), and word length are each Z-scored within each language. Concepts (entries in the IDS) and languages are treated as random effects.*

	OPUS		Google 1T		Google Books 2012	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Fixed Effects						
Average Information Content	0.251		0.202		0.248	
	(0.008)		(0.007)		(.012)	
Negative log probability		0.395		0.392		0.348
		(0.009)		(0.009)		(.014)
Intercept	0.021	−0.001	0.033	- 0.002	.004	−.014
	(0.018)	(0.018)	(.019)	(.018)	(.021)	(.021)
Random Effects						
# of Concepts	1,303	1,303	1,302	1,302	1,292	1,288
Concepts Standard Deviation	0.605	0.590	.625	.605	.643	.657
# of Languages	12	12	11	11	5	5
Languages Standard Deviation	$< 10^{-10}$	$10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$
Model Fits						
Log Likelihood	−23,878	<b>−23,411</b>	−20,914	<b>−20,379</b>	−9,976	<b>−9,879</b>
AIC	47,766	<b>46,833</b>	41,876	<b>40,806</b>	19,962	<b>19,768</b>
BIC	47,805	<b>46,872</b>	41,837	<b>40,768</b>	19,996	<b>19,803.0</b>

each case, an ANOVA reveals this difference to be significant, in favor of the model using negative log frequency as the better predictor of word length in characters. This suggests that log word frequency is the better predictor of word length among the word forms used for this matched set of concepts and referents.

## 5 Discussion

Large-scale web-based linguistic corpora introduce new methodological challenges. We investigate the implications of methodological choices made in a prominent recent work regarding the relationship between communicative pressures and the properties of natural language lexica, Piantadosi et al. (2011). After implementing a set of methodological improvements, we find that the results are robust in some languages (*e.g.*, English in

Google Books and Google 1T), but we fail to find clear cross-linguistic evidence of a stronger correlation between in-context predictability and word length than word frequency and word length. In that these methodological choices — inclusion criteria for linguistic material, treatment of text encoding, and composition of the wordlist under analysis — are common to many large-scale, cross-linguistic corpus analyses, we encourage researchers to take heed regardless of the implications for research on communicative efficiency, a point that we take on after considering the specific implications of our results.

### 5.1 Implications for research on communicative efficiency

In the current work, we fail to reproduce the systematically higher correlations between in-context predictability and word length than frequency and word length, as obtained by Piantadosi et al. (2011). Subsequent experimental and corpus work has provided some convergent evidence for the theoretical proposal in Piantadosi et al. (2011) concerning communicative efficiency and language structure, namely a preference for shorter referring expressions when supported by context (Mahowald et al., 2013) as well as shorter acoustic durations (Seyfarth, 2014) when a word is more contextually predictable. We leave it to future work to adjudicate between these theoretical viewpoints and to refine theories of listener-oriented language optimization to account for this discrepancy.

One possibility for the difference between the results reported here and those from Piantadosi et al. (2011) is that the pattern remains robust for the language with the largest corpora (English), but that average information content estimates are biased in the smaller corpora. While the Google 1T dataset contains a nominal trillion tokens of English, the Western European languages datasets *sum* to one trillion (i.e., 100B tokens per each of ten languages). This quantity is further reduced in pre-processing, once web page content has been filtered out. These datasets also have minimum token counts necessary for inclusion in the dataset, e.g. that a trigram is included if it is seen more than 40 times in the raw data.

While several billion appears to be a relatively large number of tokens, it may



nonetheless fail to overcome the fact that language is composed, in large part, of exceedingly rare events, and that characterizing information content requires accurately characterizing behavior that is only very rarely observed. This problem is exacerbated by the difficulty of applying smoothing methods to such large count-based corpora. Words with frequency rank 25,000 — the lower limit of frequency among those analyzed in Piantadosi et al. (2011) — are still seen in a relatively small number of contexts in many of the datasets.

We thus conducted a series of analyses on the Google 1T datasets to test the hypothesis that average information content estimates may be less reliable — more biased — in the smaller datasets. We use a binomial sampling method (See Appendix: Binomial Sampling) to estimate average information content for 100 randomly sampled words, using a Swedish Google 1T-sized subsample of the English Google 1T counts. We chose the size of the Swedish corpus for this analysis because it has the smallest token count among the Google 1T languages after filtering. Re-computing average information content on 100 words from English from a 6% subsample of the 1T dataset reveals a an extremely high correlation ( $R^2 > .99$ ; Fig. 7). This suggests that the difference in sizes between English 1T and the other European languages, though roughly 10x (some variability depends on preprocessing), may have a minimal effect on average information content estimates.

The conclusion that average information content estimates in the much smaller Swedish dataset are minimally biased with respect to an English-sized dataset requires testing a key assumption: that the relationship between sample size and bias is the same between languages. This assumption is, at best, suspect given the ways in which language structure, as characterized by  $n$ -grams, may vary across languages. For example, while languages can generally be characterized in terms of a prototypical word order, the degree to which they deviate from this order may vary, with several languages showing a stronger proclivity for free word order, especially when grammatical role is redundantly marked with a case system. A language with less strict adherence to dominant word order will

*ceteris paribus* demonstrate higher entropy for simple sequence based models like  $n$ -grams.

To test how bias in average information content varies as a function of sample size across languages, we again use binomial sampling (Appendix: Binomial Sampling). First, we sample contexts and continuations representative of a 16 billion word corpus for each of 100 words in each language, reflecting uniform spacing between frequency rank 1 and 25,000. We then estimate average information content on 7 subsamples, reflecting 2 billion word increments sampled from the above-derived 16 billion word subsamples. For each size subsample, average information content is taken as the mean from 10 samples, and we compute the difference of this mean from the average information content estimate from the first, 16 billion word sampled corpus.

This analysis reveals that bias as a function of sample size varies across languages (compare Czech and Dutch in Figure 8). This is corroborated by an ANOVA examining difference as a function of language, corpus size, and language  $\times$  corpus size (all significant,  $p < .001$ ). However, the absolute magnitude of the bias is relatively small in comparison to the average information content estimates themselves. Even in a 12.5% subsample of the most variable language (Dutch), error with respect to the larger dataset is much less than .05 nats for 95% of the words tracked. This suggests that the small corpus sizes in non-English Google 1T datasets (all an order of magnitude smaller than the English 1T dataset) are unlikely to bias average information content estimates in a way that drives the principal pattern of results, and that the same pattern of results would obtain if datasets as large as the English one were available for other languages.

Turning to the third analysis, one possible critique of the use of the Intercontinental Dictionary Series is that the restricted set of lemmas may not be representative of full lexica, and may thus obscure patterns that emerge across the entire lecion. IDS datasets contain a small subset (800-1200 items per language) and tend to be relatively short and frequent with respect to the larger set of words under analysis in Piantadosi et al. (2011). However, analogous to the argument presented above regarding the use of in-dictionary

words, we should expect to see the same pattern of results hold for this subset if indeed length is driven by information content rather than frequency. Further, if a different relationship holds among these words, theories of the interaction of communicative pressures and language structure must account for *why* it does not hold among a sample of high frequency words.

A second possible critique of the IDS-based analysis is that information content estimates may vary according to what communicative contexts are covered by the base form of a word. For example, consider a hypothetical language, English Prime, that differentiates a nominative form *dog* and an accusative form of dog, *doggo*. In English, the information content estimate for “dog” reflects the weighted combination of subject and direct object uses of “dog”; under English Prime, it reflects only the subject usage. Under both languages, the direct object usage is more predictable than subject usage: it is a relatively likely continuation for a small set of transitive verbs. This would mean that *dog* would have a *higher* information content in English Prime because it reflects *only* the less predictable nominative usage of “dog.”

Because of variability in usage, the IDS analysis would ideally take into account *all* inflections of a lemma. Unfortunately, we do not have access to consistent, high-quality lemmatizers across the set of languages in the sample, nor access to datasets with the words in their original sentential frames, as may be required to tag with appropriate lemmas. A related potential source of variation in the information content estimate is which meanings are co-lexified under the base form, *i.e.*, the set of polysemous and homographic uses of the base form in each language. We leave it to future work to match usages more precisely on the basis of word meaning and word form.

## 5.2 Best Practices for working with large-scale linguistic data

The case study presented above demonstrates how preprocessing and analytical choices can influence the obtained results for a corpus analysis of large-scale linguistic data,

particularly those built from language data on the Internet. We enumerate here a set of specific recommendations for researchers conducting (or reviewing) work that relies on such methods (Table 1). All of these recommendations presuppose practices that we hope are already considered standard in the field: that researchers will publish their code and share derived datasets in cases where the cost of computation is high.

**5.2.1 Evaluate the linguistic status of web-gathered material.** A challenge we address above is that of verifying that the collection of text corresponds to conventionalized linguistic forms shared by speakers of a language (Section 2.1). Web-scraped datasets may contain a large proportion of material that is outside of these conventional forms, either as a result of errors in identifying the source language (e.g., the processing pipeline that was used to generate the Google 1T dataset), or because many strings appear on the Internet which are not properly linguistic. If extensive linguistic material from other languages is present in an analysis, it may drive the obtained results, as we found here. We recommend conducting analyses after filtering with a large, up-to-date, rule-based dictionary such as Aspell to identify the conventional wordforms in the language.

**5.2.2 Represent words with care.** Orthographic words — a slippery yet nonetheless useful construct — are surprisingly difficult to represent to machines. Modern systems of text representation rely on text encoding, or maps from code units (typically binary representations) to characters (units of text with human-interpretable semantic content). Researchers should choose and consistently employ a text encoding throughout their preprocessing and analysis that is capable of capturing the minimal differences between wordforms respected by a linguistic community. Outside of English-only research, this effectively means that researchers should use the Unicode standard, or UTF-8 encoding. While the choice of encoding scheme makes little difference for English, electing a less expressive encoding scheme can have profound effects on analyses for languages that require a broader character set to distinguish wordforms, such as the Czech case in Section 2.2.

Researchers should also note that above and beyond the problem of maintaining the appropriate text encoding throughout their analysis, there is the further problem of *text normalization*: more expressive text encoding schemes such as Unicode are able to generate what appear to be the same human-readable character with many different code units. For example, in Unicode, the sequences U+0065, U+0301 (“Latin small letter e”, “combining acute accent”) and U+00E9 (“Latin small letter e with acute”) are both rendered as é. Unless the code units are “normalized” to the same standard (i.e., all multi-character diacritical marks converted to single characters, or the reverse), there is no guarantee of consistency within or across datasets. This detail, while seemingly banal, can have consequences for the computation of lexical statistics, as well as for merging information about words across datasets.

Finally, the process of tokenization — the discretization of utterances into orthographic words — is nontrivial. While whitespace between characters generally works in English,<sup>6</sup> this solution is not robust across languages, many of which use spaces differently (e.g., Thai). A variety of heuristic and statistical methods may be used to tokenize words in such cases; both the effectiveness of the method and its appropriateness for the task at hand should be considered carefully.

For research using orthographic representations of words, we strongly recommend the consistent use of “normalized” UTF-8 representation, and encourage researchers to carefully consider how the dataset was tokenized. Researchers may also consider using learned, generalizable representations from natural language processing that are more robust to these issues, such as byte-pair encoding (Kudo, 2018; Sennrich, Haddow, & Birch, 2016) or embeddings derived from the hidden layer of a convolutional neural network (Jozefowicz, Vinyals, Schuster, Shazeer, & Wu, 2016; Kim, Jernite, Sontag, & Rush, 2016).

**5.2.3 Match the set of words under analysis across languages.** A third challenge of cross-linguistic work looking at the properties of lexica is choosing the set of

---

<sup>6</sup>The obvious exception is punctuation; sequences with high mutual information content such as *whoops-e-daisies* or *South Korea* pose harder theoretical challenges (Church & Hanks, 1990).

words in a given analysis. As argued in Section 2.3, the composition of a wordlist of the  $n$  most frequent words in a language depends heavily on the typological characteristics of a language, especially the complexity of morphological paradigms. We recommend the use of matched word lists – a collection of minimally-inflected / minimally-marked wordforms elicited for a list of concepts such as the Intercontinental Dictionary Series (Key & Comrie, 2015).

**5.2.4 Perform a sensitivity analysis or take a “multiverse” approach.** In addition to the above three recommendations that target specific points of failure, we urge researchers to adopt a general strategy of evaluating hypotheses over a range of analytical choices, rather than committing to one set of choices. While the specific recommendations above constitute best practices that may help avoid common pitfalls, they are not uniformly appropriate, nor do they solve the vast range of possible problems that may emerge. Limiting an analysis to words in a dictionary or using a matched set of words for an analysis introduce their own methodological drawbacks (see 2.1 and 3.3). Rather, we encourage researchers working with large, noisy datasets in cognitive sciences and related fields to employ the general framework and approach of *sensitivity analysis*, in which the robustness of a particular result is evaluated by examining how those results change as a function of models, inputs, and assumptions (Thabane et al., 2013). In line with the “multiverse” analytical approach outlined in Steegen et al. (2016), we also encourage researchers to evaluate their hypotheses over a range of “reasonable” preprocessing and analytical choices, and report on this larger space of analyses whenever possible in corpus-based research.

While sensitivity analysis is often prohibitively expensive or time-consuming for experimental work, computation-first methods like corpus analysis allow for researchers to automate the process of hypothesis testing across a range of models, inputs, and assumptions. Provided that researchers have allocated time beforehand to take into account this added dimensionality, the approach can incur relatively minor additional effort. This

effort is well worth it: rigorous evaluation across a range of preprocessing choices can help establish the robustness of results, or flag them as “brittle” results that only hold under certain assumptions. For some researchers, this practice may already be self-evident; here we appeal to the remainder of the cognitive science, quantitative linguistics, psycholinguistics, and NLP communities to pursue these practices in the name of constructing a robust lattice of scientific results, from which future work can best progress.

## 6 Conclusion

Large-scale web-based linguistic corpora constitute a new set of resources which may help to elucidate longstanding questions in language research. However, they also bring new methodological challenges in data processing and analysis. The best practices we propose offer a way for language researchers to standardize their approach to analyzing large multilingual corpora and highlight the importance of assuring the robustness of analyses.

### Appendix: Binomial Sampling

We use a binomial sampling scheme to evaluate how smaller datasets may bias average information content estimates. This technique allows us to efficiently estimate the average information content of a word in smaller subsamples of an unsmoothed  $n$ -gram dataset. This technique is particularly useful when the original dataset is not available to researchers for sampling (as is true for Google 1T and Google Books 2012). Specifically, to estimate context-weighted trigram surprisal for each word given a subsample of corpus of proportion  $\gamma$ , we begin by drawing new counts for each context  $X_i$  in which a word is a possible continuation according to the original corpus:

$$X_i^s \sim \text{Binom} \left( n = \gamma \cdot C(X), p = \frac{C(X_i)}{C(X)} \right) \quad (1)$$

Where  $C()$  is the count function,  $X$  indicates all bigram contexts, and  $X_i$  indicates this particular bigram context. For example in the process of computing the average information content for the word *car* in a subsample, this would involve retrieving all preceding contexts for which *car* is a continuation, and drawing a new count (possibly 0) for each of those contexts in the new subsample.

We then sample how often the target word is the continuation for each sampled context, using the continuation probabilities from the original dataset:

$$D_i^s \sim \text{Binom} \left( n = X_i^s, p = \frac{C(D_i)}{C(X_i)} \right) \quad (2)$$

Where  $C(D_i)$  indicates the count of the context + target word trigram, and  $C(X_i)$  is the count of the corresponding preceding bigram. For the example of computing the average information content of *car*, this corresponds to drawing a new count for the number of instances where *car* is the continuation for a specific context, e.g. *near the*. Average information content for a word can then be calculated from the sample by evaluating the newly-estimated conditional probability of that word in each of the newly-sampled



contexts, weighting by those context frequencies, where  $X_i^s$  is the sampled count of the  $i$ -th context, and  $X_{\cdot}^s$  is the sum of all sampled contexts.

$$IC^s = - \sum_i \left( \frac{X_i^s}{X_{\cdot}^s} \right) \log \left( \frac{D_i^s}{X_i^s} \right) \quad (3)$$

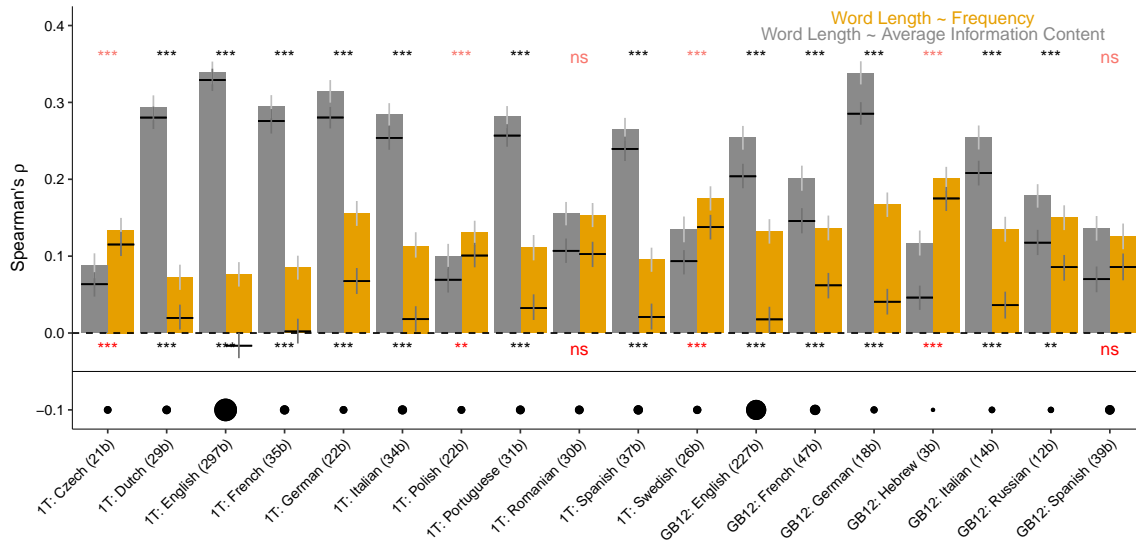


Figure 5. Language-wide Spearman's rank correlation coefficients between average information content and word length (grey) and frequency (normalized and negative log-transformed) and word length (yellow) for the 25,000 most frequent word types in each language. Word types are limited to only those present in the respective OPUS dataset for each language (compare with Figure 6, where the analysis is restricted to word types in the relevant dictionary). Gray error bars indicate 95% bootstrapped confidence intervals. Black horizontal lines indicate the partial correlation (with the other predictor residualized out), again with 95% bootstrapped confidence intervals. Asterisks above the graph indicate whether the difference is statistically significantly different from zero, while those below refer to the partial correlations; red indicates a result inconsistent with Piantadosi et al. (2011). Black circles at the bottom indicate the relative size of the corpus used to estimate frequency and average information content. Parenthetical on the x-axis indicate the number of word tokens in the sample.

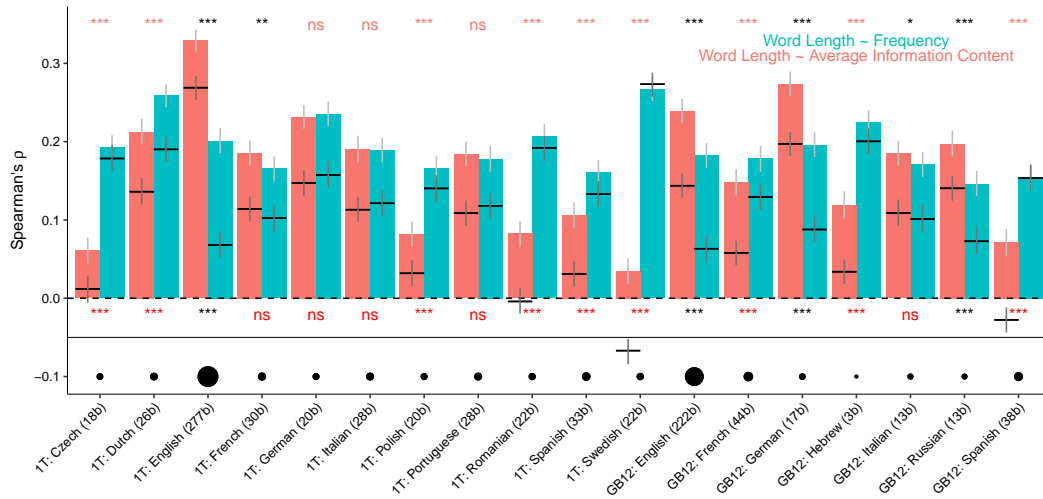


Figure 6. Language-wide Spearman's rank correlation coefficients between average information content and word length (green) and frequency and word length (red) for the 25,000 most frequent word types in each language. Word types are limited to those present in the respective dictionary and OPUS dataset for each language. Plotting conventions are the same as Figure 5.

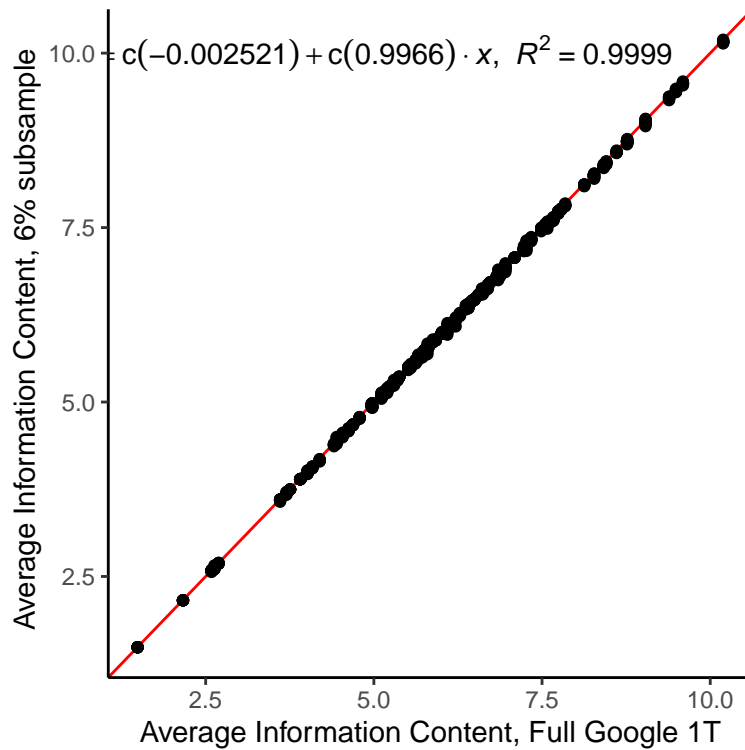


Figure 7. Bias in average information content from a 16B token subsample of the English Google 1T dataset. 16B tokens was chosen as it is the size of the smallest Google 1T dataset for other European languages after preprocessing (Swedish, from a nominal 100B words).

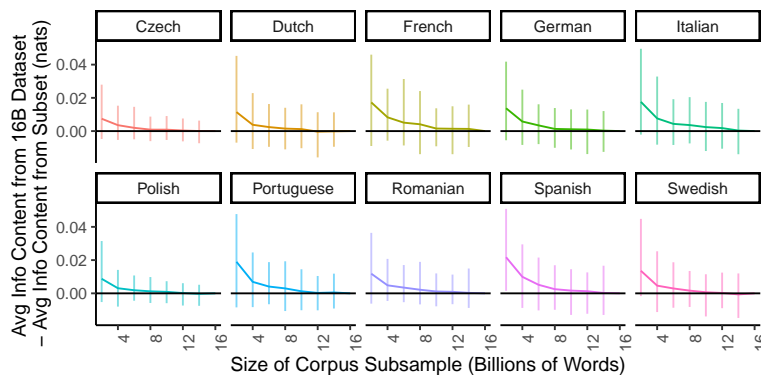


Figure 8. Bias in average information content estimates by language and size of corpus sample (2-16B subsamples of a 16B word corpus for each language).

## References

- Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3), 247–264.
- Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Lang Speech*, 47(1), 31–56.
- Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3), 209–226.
- Bentz, C., & Ferrer-i-Cancho, R. (2016). Zipf’s law of abbreviation as a language universal. In *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. Universitätsbibliothek Tübingen.
- Bickel, B., & Nichols, J. (2005). Inflectional morphology. In T. Shopen (Ed.), *Language typology and syntactic description*. Cambridge, UK: Cambridge University Press.
- Bielec, D. (1998). *Polish: An essential grammar*. London: Routledge.
- Blyth, C. (1972). On simpson’s paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338), 364–366.
- Bochkarev, V., Solovyev, V., & Wichmann, S. (2014). Universals versus historical contingencies in lexical evolution. *J R Soc Interface*, 11(101).
- Brants, T., & Franz, A. (2006). *Web 1T 5-gram Version 1 LDC2006T13*. Philadelphia, PA: Linguistic Data Consortium.
- Brants, T., & Franz, A. (2009). *Web 1T 5-gram, 10 European Languages Version 1 LDC2009T25*. Philadelphia, PA: Linguistic Data Consortium.
- Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, 44(4), 991–997.
- Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2013). One billion word benchmark for measuring progress in statistical language

- modeling. *arXiv preprint arXiv:1312.3005*.
- Chetail, F., & Boursain, E. (2019). Shared or separated representations for letters with diacritics? *Psychonomic Bulletin & Review*, 26(1), 347–352.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: The MIT Press.
- Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22–29.
- Davies, M., & Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide*, 36(1), 1–28.
- Denoyer, L., & Gallinari, P. (2006). The wikipedia xml corpus. In *International Workshop of the Initiative for the Evaluation of XML Retrieval* (pp. 12–19).
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/N19-1423> doi: 10.18653/v1/N19-1423
- Evans, N., & Levinson, S. (2009). The myth of language universals: language diversity and its importance for cognitive science. *Behav Brain Sci*, 32(5), 429–448.
- Fenk, A., & Fenk, G. (1980). Konstanz im kurzzeitgedächtnis - Konstanz im sprachlichen Informationsfluß. In *Zeitschrift für experimentelle und angewandte Psychologie XXVII* (Vol. 3, pp. 400–414).
- Ferrer-i-Cancho, R. (2017). The placement of the head that maximizes predictability. an information theoretic approach. *Glottometrics*, 39, 38–71.
- Ferrer-i-Cancho, R. (2018). Optimization models of natural communication. *Journal of Quantitative Linguistics*, 25(3), 207–237.

- Ferrer-i-Cancho, R., Bentz, C., & Seguin, C. (2020). Optimal coding and the origins of zipfian laws. *Journal of Quantitative Linguistics*, 1–30.
- Ferrer-i-Cancho, R., & del Prado Martín, F. M. (2011). Information content versus word length in random typing. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(12), L12002.
- Ferrer-i-Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M. J., & Semple, S. (2013). Compression as a universal principle of animal behavior. *Cognitive Science*, 37(8), 1565–1578.
- Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 199–206.
- Gibson, E., Bergen, L., & Piantadosi, S. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proc. Natl. Acad. Sci. U.S.A.*, 110(20), 8051–8056.
- Gibson, E., Futrell, R., Piantadosi, S., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*.
- Gustison, M., Semple, S., Ferrer-i-Cancho, R., & Bergman, T. (2016). Gelada vocal sequences follow menzerath’s linguistic law. *Proceedings of the National Academy of Sciences*, 113(19), E2750–E2758.
- Heesen, R., Hobaiter, C., Ferrer-i-Cancho, R., & Semple, S. (2019). Linguistic laws in chimpanzee gestural communication. *Proceedings of the Royal Society B*, 286(1896), 20182900.
- Huang, M., Ma, H., Ma, C., Garber, P., & Fan, P. (2020). Male gibbon loud morning calls conform to zipf’s law of brevity and menzerath’s law: insights into the origin of human language. *Animal Behaviour*, 160, 145–155.
- i Cancho, R. F., Bentz, C., & Seguin, C. (2019). Optimal coding and the origins of zipfian laws. *Journal of Quantitative Linguistics*, in press. Retrieved from <https://arxiv.org/abs/1906.01545> doi: 10.1080/09296174.2020.1778387

- Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.
- Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf’s law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, 165, 45–52.
- Kenstowicz, M. (2007). Salience and similarity in loanword adaptation: a case study from Fijian. *Language Sciences*, 29(2-3), 316–340.
- Key, M., & Comrie, B. (Eds.). (2015). *Intercontinental Dictionary Series (IDS)*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from <http://ids.clld.org/>
- Kim, Y., Jernite, Y., Sontag, D., & Rush, A. (2016). Character-aware neural language models. In *Thirtieth AAAI Conference on Artificial Intelligence*.
- Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177.
- Levy, R., & Jaeger, T. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19* (pp. 849–856). Cambridge, MA: MIT Press.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... Zettlemoyer, L. (2020, July). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7871–7880). Online: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/2020.acl-main.703> doi: 10.18653/v1/2020.acl-main.703
- Lorenz, A., & Zwitserlood, P. (2016). Semantically transparent and opaque compounds in



- german noun-phrase production: Evidence for morphemes in speaking. *Frontiers in Psychology*, 7, 1943.
- Mahowald, K., Dautriche, I., Gibson, E., & Piantadosi, S. (2018). Word forms are structured for efficient use. *Cognitive Science*, 42(8), 3116–3134.
- Mahowald, K., Fedorenko, E., Piantadosi, S., & Gibson, E. (2013). Info/information theory: speakers choose shorter words in predictive contexts. *Cognition*, 126(2), 313–318.
- Menzerath, P. (1954). *Die architektonik des deutschen wortschatzes* (Vol. 3). F. Dümmler.
- Michel, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176–182. doi: 10.1126/science.1199644
- Perea, M., Fernández-López, M., & Marcet, A. (2020). What is the letter é? *Scientific Studies of Reading*, 24(5), 434–443.
- Petersen, A. M., Tenenbaum, J. N., Havlin, S., Stanley, H. E., & Perc, M. (2012). Languages cool as they expand: allometric scaling and the decreasing need for new words. *Sci Rep*, 2, 943.
- Piantadosi, S., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proc. Natl. Acad. Sci. U.S.A.*, 108(9), 3526–9.
- 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 preprint arXiv:1307.6726*.
- Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
- Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1715–1725).
- Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, 133(1), 140–155.

- Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423.
- Silberzahn, R., Uhlmann, E., Martin, D., Anselmi, P., Aust, F., Awtrey, E., ... others (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3), 337–356.
- Smith, N. (Submitted). ZS: A file format for efficiently distributing, using, and archiving record-oriented data sets of any size.
- Steege, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5), 702–712.
- Swadesh, M. (1971). *The origin and diversification of language*. Chicago: Aldine.
- Thabane, L., Mbuagbaw, L., Zhang, S., Samaan, Z., Marcucci, M., Ye, C., ... others (2013). A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC Medical Research Methodology*, 13(1), 92.
- Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.
- Tomasello, M., & Stahl, D. (2004). Sampling children's spontaneous speech: How much is enough? *Journal of Child Language*, 31(1), 101–121.
- von der Gabelentz, G. (1891). *Die sprachwissenschaft, ihre aufgaben, methoden und bisherigen ergebnisse*. Leipzig: Weigel.
- Wade, T. (1992). *A Comprehensive Russian Grammar*. Oxford: Blackwell.
- Zhang, A., Culbertson, B., & Paritosh, P. (2017). Characterizing online discussion using coarse discourse sequences. In *Proceedings of the 11th International AAAI conference on Weblogs and Social Media*.
- Zipf, G. (1935). *The Psychobiology of Language*. Houghton-Mifflin.
- Zipf, G. (1949). *Human Behaviour and the Principle of Least-Effort*. Cambridge, MA: Addison-Wesley.