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

Stephan C. Meylan, a,b Thomas L. Griffiths c

a Department of Brain and Cognitive Science, Massachusetts Institute of Technology
b Department of Psychology and Neuroscience, Duke University

Received 2 November 2019; received in revised form 16 March 2021; accepted 7 April 2021

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

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; for example, 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 that 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 underappreciated challenges. First, the construction of these corpora often depends heavily on automated processing; for example, automatic language identification to separate scraped web pages into different linguistic subcorpora. These preprocessing steps may include many nontrivial (and potentially undocumented) 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\(^1\) 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 word forms in the relevant language communities, and the use of rank-ordered word frequency to select the set of word types in the order in 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 a 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 long-standing, 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
Table 1
Overview of challenges and best practices in working with large-scale, web-based, and multilingual linguistic corpora

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

Additionally emphasize the importance of *multiverse* analysis (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016), where researchers test their hypotheses across the combinatorial space of possible 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 et al., 2013; Gibson, Bergen, & Piantadosi, 2013; Gibson et al., 2019; Ferrer-i-Cancho, 2018; 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 Zipf (1935), and it has been attributed to him as “Zipf’s Law of Abbreviation” or “Zipf’s Law of Brevity.” Zipf posited that this relationship of length and frequency emerges from speakers’ proclivity to minimize total articulatory effort 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 its mathematical foundation based on information theory and recent extensions (Ferrer-i-Cancho et al., 2013; Ferrer-i-Cancho, Bentz, & Seguin, 2020). Another line of research demonstrates that Zipf’s law of abbreviation extends, to varying degrees, to the communication systems of other nonhuman primates (Heesen, Hobaiter, Ferrer-i-Cancho, & Semple, 2019; Huang, Ma, Ma, Garber, & Fan, 2020; Gustison, Semple, Ferrer-i-Cancho, & Bergman, 2016).

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 & 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 word form whenever a word is more predictable from context; for example, rhino versus 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 word form 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 11 European languages and found that 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 it 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., Ferrer-i-Cancho, Bentz, & Seguin, 2020). 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 (Section 2.1), the treatment of character encoding (Section 2.2), and the role of morphological variation on the composition of the word list under analysis (Section 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
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

<table>
<thead>
<tr>
<th>Language</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>muhu iuventus terezoi zdn farmaceutik cvm latine mauser ujeme beggin johns nikolase lidsmila mooney hakuna kreditka carnevale hansy homeo</td>
</tr>
<tr>
<td>Dutch</td>
<td>uurs kloetejaar neuman heee jame rooldicht delige kzl ern euhm plaatsers hej jaah twas tif gdc ehhh ehh txh haw</td>
</tr>
<tr>
<td>English</td>
<td>itat ordenator semitism cest montel eet svensk roget ppb xn stang tep nge ane rrr ahh hahaha hmmm eam hoh</td>
</tr>
<tr>
<td>French</td>
<td>djamena ouaip rebonjour waouh ahhh ndt deposez antioche paulaye riverdance ils pffftf andro ahh naire iwa adios argh pof alembert</td>
</tr>
<tr>
<td>German</td>
<td>halihallo jaaaa hler nunja siert tigen lasset tive duane nagut achja ich rapie denkste bengt ann uhh pssst genaugenommen nder</td>
</tr>
<tr>
<td>Italian</td>
<td>goditela konishi iume piuttosto poizia ehmm signorinella dipartenza tinkertoj hiruta niene larkspur problema matrimony amram moiti iotta kittle paite oscorp</td>
</tr>
<tr>
<td>Polish</td>
<td>ehm kapler ciyaa niajiepiej rbit fredzia taaa ak taa an danke yep sadzac upi sasz pomyslic te archive niw tyros munin</td>
</tr>
<tr>
<td>Spanish</td>
<td>racias rankeado iah demoteres cest mins erlich lbs although bles khz holaa tante ghz mph ufft prentice btu however nidad</td>
</tr>
<tr>
<td>Portuguese</td>
<td>metai grupo caezinhos khz lidade tenny lsmt tante eev vez clack aquanaut oberst mins btus reconciliacao braddock lft descon although</td>
</tr>
<tr>
<td>Romanian</td>
<td>fives keeling bodin vampira georgette raisin rupeee wigs dba minimizeaza factiunea grasule resultate sottotitoli encounters however tortoise soren notitlee inches</td>
</tr>
<tr>
<td>Swedish</td>
<td>ningar nepp nyliga tjaay podnapii tionen derma terna snyggye prenum pyjamaser mln gaeliska ooops etruskiska ligen raken neeem jupp ivoir</td>
</tr>
</tbody>
</table>

(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. 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
Fig. 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.

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”; for example, Spanish pan (bread). Strings that do not appear in a 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 word lists 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 (Fig. 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).
Fig. 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 subgroups 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).

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 (Fig. 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 frequency and word length (column 4).

Using this tripartite categorization, we can investigate how the somewhat unintuitive correlations in Fig. 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 Fig. 3, right), while in-dictionary words are longer and have higher average information content (blue density plots in Fig. 3, right). The positions of these three different categories make for a strong global correlation, but much weaker per-category correlations as observed in Fig. 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 (Fig. 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 versus 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 on-screen 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 closest ASCII equivalents. For English, this is an inconsequential data processing choice: almost all English words can be encoded losslessly using ASCII character representations.
Fig. 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.

(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, ˙o, ø). But while this mundane data processing choice has minimal implications for English, it is highly consequential in languages that use 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, for example, Spanish sí (approximately English “if”) and si (“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 versus 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 English ($R^2 = .994$ in the 25,000 most frequent words; Fig. 4, left), languages that use a broader inventory of characters in their orthography have substantially more unexplained variance across words (e.g., Czech, $R^2 = .464$; see Fig. 4, center). Further, an examination of the distribution of the differences in average information content estimates in Czech (Fig. 4, right) shows a significant rightward 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 in which word forms 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 et al., 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, 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; for example, vendemelo, 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 word list 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 (a) selecting a set of word forms to analyze in each language; (b) computing word lengths from orthographic representations, and unigram surprisal and average information content for each of these words; and (c) 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/datasetsv2.html (Michel et al., 2011), and OPUS (2013) from 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 & Zwitserlood, 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 1,800 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 (http://github.com/njsmith/zs/).

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. 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. 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.

3.3. Matching word list composition across languages

Here, we propose a control that limits the number of word forms per each lemma in the analysis. To do this, we conduct the analysis over a subset of word forms 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 (IDS, 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 1,300 concepts, of which a subset (800–1,200) are present in a given language. The IDS datasets contain the unmarked form of the
concept; for example, 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 word list. 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. Lexical information content estimates were calculated with the ngrawk2 library, 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 word forms 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) (Fig. 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 (Fig. 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 11 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 2 of the 11 1T languages, English and French, show the
Fig. 5. Language-wide Spearman’s rank correlation coefficients between average information content and word length (gray) 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 Fig. 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. Parentheticals on the x-axis indicate the number of word tokens in the sample.

pattern found by Piantadosi et al. (2011). Five 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 International Dictionary Series. 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
Fig. 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 in Fig. 5.

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 toward 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 (Akaike’s Information Criterion and Bayesian Information Criterion, respectively; Table 3). In 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
Table 3
Comparison of mixed-effects models predicting word length for word types in the Intercontinental Dictionary Series (IDS) in three datasets

<table>
<thead>
<tr>
<th></th>
<th>OPUS Model 1</th>
<th>OPUS Model 2</th>
<th>Google 1T Model 1</th>
<th>Google 1T Model 2</th>
<th>Google Books 2012 Model 1</th>
<th>Google Books 2012 Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed effects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average information content</td>
<td>0.251 (0.008)</td>
<td>0.202 (0.007)</td>
<td>0.248 (0.012)</td>
<td>0.348 (0.014)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative log probability</td>
<td>0.395 (0.009)</td>
<td>0.392 (0.009)</td>
<td>0.348 (0.014)</td>
<td>0.348 (0.014)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.021 (0.018)</td>
<td>-0.001 (0.018)</td>
<td>0.033 (0.019)</td>
<td>-0.002 (0.018)</td>
<td>0.004 (0.021)</td>
<td>-0.014 (0.021)</td>
</tr>
<tr>
<td>Random effects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of concepts</td>
<td>1,303</td>
<td>1,303</td>
<td>1,302</td>
<td>1,302</td>
<td>1,292</td>
<td>1,288</td>
</tr>
<tr>
<td>Concepts standard deviation</td>
<td>0.605</td>
<td>0.590</td>
<td>0.625</td>
<td>0.605</td>
<td>0.643</td>
<td>0.657</td>
</tr>
<tr>
<td>No. of languages</td>
<td>12</td>
<td>12</td>
<td>11</td>
<td>11</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Languages standard deviation</td>
<td>$&lt;10^{-10}$</td>
<td>$10^{-10}$</td>
<td>$&lt;10^{-10}$</td>
<td>$10^{-10}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model fits</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log-likelihood</td>
<td>-23,878</td>
<td>-23,411</td>
<td>-20,914</td>
<td>-20,379</td>
<td>-9,976</td>
<td>-9,879</td>
</tr>
<tr>
<td>AIC</td>
<td>47,766</td>
<td>46,833</td>
<td>41,876</td>
<td>40,806</td>
<td>19,962</td>
<td>19,768</td>
</tr>
<tr>
<td>BIC</td>
<td>47,805</td>
<td>46,872</td>
<td>41,837</td>
<td>40,768</td>
<td>19,996</td>
<td>19,803.0</td>
</tr>
</tbody>
</table>

Note: 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. Bold values indicate model with best fit for each dataset.

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 word list 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 10 languages). This quantity is further reduced in preprocessing, once web page content has been filtered out. These datasets also have minimum token counts necessary for inclusion in the dataset; for example, 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) 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. Recomputing average information content on 100 words from English from a 6% subsample of the 1T dataset reveals 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 $10 \times$ (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 such as $n$-grams.

To test how bias in average information content varies as a function of sample size across languages, we again use binomial sampling (Appendix). 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 seven 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 Fig. 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 0.05 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 IDS is that the restricted set of lemmas may not be representative of full lexica, and it may thus obscure patterns that emerge across the entire lexicon. IDS datasets contain a small subset (800–1,200 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, which 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; that is, 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 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 word forms in the language.

5.2.2. Represent words with care

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 word forms 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 word forms, 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 multicharacter 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, 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 con-
sistent use of “normalized” UTF-8 representation, and we encourage researchers to carefully
consider how the dataset was tokenized. Researchers may also consider using learned, gener-
alizable 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 (Kim, Jernite, Sontag, &

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 words in a given analysis. As argued in Section 2.3, the composition of a word list 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 word forms
elicited for a list of concepts such as the IDS (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 analyti-
cal 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. Limit-
ing an analysis to words in a dictionary or using a matched set of words for an analysis
introduce their own methodological drawbacks (see Sections 2.1 and 3.3). Rather, we encour-
age 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” analyti-
cal 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 exper-
imental work, computation-first methods such as corpus analysis allow for researchers to
automate the process of hypothesis testing across a range of models, inputs, and assump-
tions. 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 that may help to elucidate long-standing 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.

Acknowledgments

This material is based on the work supported by the National Science Foundation Graduate Research Fellowship under grant no. DGE-1106400 to SM and NSF grant no. SMA-1228541 to TLG The authors especially thank 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.

Open Research Badges

This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/dj5vr/.

Notes

1. Historically known as “sanity checks,” we encourage the adoption of the term “canary.”
2. The better known Zipf’s Law presented in the same work, concerns the power-law distribution of word frequencies.
3. We do not use the dictionary-based filter to restrict the input to the n-gram models used to compute in-context predictability.
4. We use the 2013 release of the dataset.
5. We allow uppercase forms for German, which capitalizes all nouns by convention.
6. 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).

**References**


Appendix A: 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:

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

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:

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

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; for example, 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 $X^s$ is the sum of all sampled contexts:

$$IC^s = -\sum_i \left(\frac{X_i^s}{X^s}\right) \log\left(\frac{D_i^s}{X_i^s}\right).$$