Statistical Methodology for Quantitative Linguistics: A Case Study of Learnability and Zipf's Law

MSc Thesis (*Afstudeerscriptie*)

written by

Valentin Vogelmann
(born June 13th 1993 in Freiburg im Breisgau, Germany)

under the supervision of **Dr Willem Zuidema** and **Bas Cornelissen**, and submitted to the Examinations Board in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam*.

Date of the public defense:	**Members of the Thesis Committee:**
March 26th 2020	Prof Dr Ronald de Wolf
	Dr Jelke Bloem
	Dr Ekaterina Shutova (chair)

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

Abstract. Quantitative linguistics is a large and rich field in the study of language that has brought about or played an essential role in debates such as those around Zipf's law or the learnability of language. Using these two topics and their intersection as a case study, we identify in this thesis a fundamental problem of statistical methodology that seems pervasive in quantitative linguistic practice. In essence, the problem can be summarised as a common negligence of the distinction between observed samples of language, that is corpora, and their source distribution, that is the underlying language.

In the first part, we re-derive how upholding the sample-source distinction naturally leads to the problem of statistical estimation and propose and show how to use standard resampling methods to obtain representative and reliable estimates, particularly given the scarcity of resources in linguistics. We use this method to obtain the most reliable estimates of Zipf's law to date and highlight the importance and potential of proper estimation by analysing some of the estimates' properties.

The second contribution consists of the Filtering method, an novel and general adaptation of resampling methods grounded in information theory. This method is intended to facilitate realistic large-scale learnability analyses of the distributional properties of language, in our case of Zipf's law. We derive the Filtering method itself starting again from the sample-source distinction and instantiate it in two exemplary implementations. Subsequently, we validate its usefulness by analysing the sampled corpora in terms of the sampling objectives and the corpora's naturalness and diversity. Given that these objectives seek to weaken Zipf's law, and that this is a difficult objective to achieve, we find relatively high naturalness and diversity of the resulting corpora.

Finally, and bringing the resampling and Filtering method together, we make a proposal for empirically assessing recent advancements in the innateness debate, which analyse the learnability of language via Kolmogorov complexity. The high degree of abstraction makes it difficult to directly evaluate the proposed learning strategies but with the help of resampling and Filtering, and the sample-source distinction more generally, we make a concrete proposal at how this may in fact be achieved.

Acknowledgements. First and foremost, I would like to thank my supervisors, Jelle and Bas, for being inspiring researchers and having been incredibly patient with me. Jelle, I deeply appreciate that you have allowed me to begin researching a topic that I am passionate about, despite the high degree of uncertainty that initially came with it, and for bearing with me and guiding me towards making it concrete. That this thesis manages to make such strong methodological points is owed to a large part to your academic skill and rigour, that I admire and look up to. And Bas, I would like to thank you for being an invaluable source of ideas and motivation. I have and will cherish our discussions and your vocal critique when it was really necessary. The quality of my work, not only stylistic, has highly benefited from your involvement and I am sincerely grateful for the time you have invested in this thesis. I wish you the best of luck for your own academic pursuits and am looking forward to an excellent PhD thesis. I would also like to extend many thanks to the members Jelle's research group who have greatly aided my progress with their feedback and recommendations and in general with a welcoming and nurturing research environment.

At least to me, writing a thesis has proven a very solitary task in which it can be easy to feel lost and alone. I consider myself very lucky to have wonderful and loving friends in Amsterdam and elsewhere who have been incredibly supportive even in times that I have not been able to reciprocate much. You are the ones who make this worthwhile, thank you for enriching my life in so many ways and making me happy.

My final thanks is due to my family, Angelika, Rainer and Linus, without whom none of my accomplishments would have been possible. This thesis is as much yours as it is mine.

Contents

1 Introduction

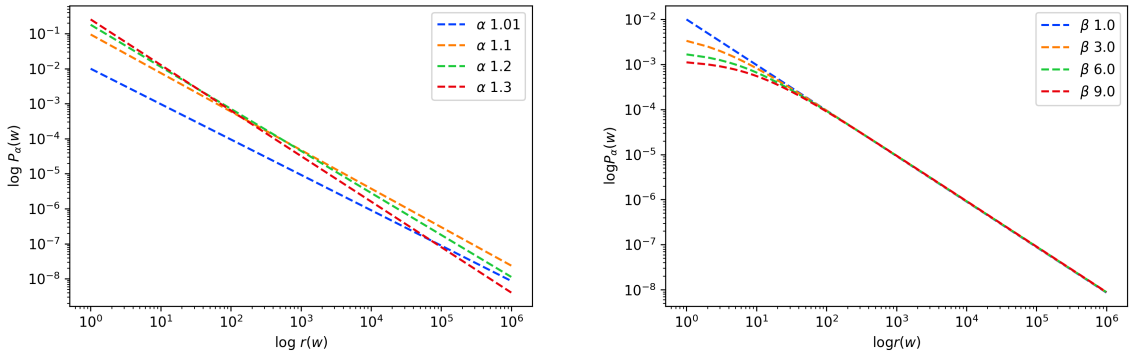
This thesis, being located at the intersection of quantitative linguistics and computational cognitive science, is not concerned with computational models or formal theories, unlike most work in these fields. It consists merely and entirely of methodology, remarks about methodological practice and recommendations for the use of both new and existing methods. Without being too humble, all of this methodology is basic, fundamental even, and that is in fact the point of this thesis.

It is our explicit choice to not introduce further competing theories to fields in which the sheer plethora of these seems to have lead to contentious and seemingly irreconcilable debates (cf. for instance the debates around Zipf's law, the debate about the acquisition of language, the divide between formal and distributional semantics, etc.). Instead, in our opinion, more work should be directed at disentangling and resolving the existing debates rather than cluttering them even further.

Worse yet, as we elucidate in this thesis, there are serious issues in the empirical practice of quantitative linguistics and in the design of analyses of statistical language acquisition. These issues make the debates all the more difficult to resolve since they skew and compromise conclusions drawn from observations and to us make it all the more preferable to focus on methodology instead.

The common theme of the issues, we realise, is negligence or obliviousness to basic statistical concepts, the most fundamental being what we will call the sample-source distinction. Whereas the former, a corpus in the context of linguistics, is observable, the latter, a language, is not and therefore a purely theoretical. Researchers across the language-related fields generally seem to disregard this distinction in their empirical practice and research design and instead seem to equate corpora and language, directly drawing inferences from the former to the latter.

In this thesis, we show that much can be gained from taking the sample-source distinction seriously. On the one hand, it immediately leads to the statistical field of estimation which in turn helps eradicate erroneous and yield reliable empirical observations. These have the power to resolve parts of the very debates about them. On the other hand, a rigorous sample-source distinction suggests and enables new and highly interesting methods in the computational study of language. These can in particular be used for more realistic and detailed analyses of the learnability of language and thus facilitate new conclusions in language acquisition research.



(a) Different values of α.

(b) Different values of β.

Figure 1.1 Illustration of Zipf's law across different values of the parameters α and β. The left plot shows different values of α, with β fixed to 1.0; values were chosen to be in same range as the value we find in natural languages, see Table 2.2. In the right plot, we see different values of β, also chosen to reflect the values in human languages, while α is fixed to 1.01. Notice the log-log scale in both plots.

1.1 Zipf's law

In order to show these advantages of studying methodology, we focus on Zipf's law, the most prominent and well-studied law of quantitative linguistics. Its role in this thesis is that it provides a case study, a prototype in terms of which we can describe and develop our methodological inventions. These, together with the broader stance we are arguing for, go however far beyond Zipf's law and cover all of quantitative linguistics.

Definition

A central element of Zipf's law is the vocabulary of a language, the set of its unique words, denoted $\mathbf{\Sigma}$. We adopt standard terminology and refer to the elements of $\mathbf{\Sigma}$ as (word) types and to the occurrences of types in corpora as tokens. Following empirical evidence of (Blevins, Milin, and Ramscar 2017), we make the important assumption that $\mathbf{\Sigma}$ is unbounded, i.e. it does not have finite cardinality (formally, there does not exist an integer n s.t. $|\mathbf{\Sigma}| \leq n$). We emphasise that this perhaps controversial assumption (although others make it too, e.g. Corominas-Murtra and Solé 2010 or P. M. Vitányi and Chater 2017) has important consequences for the discussions and statistical methods later. Our main reason to assume an unbounded vocabulary is that it leads to a higher degree of generality in our discussions and makes some of the results below less trivial.

Zipf's law describes the distribution over the vocabulary, that is the probability $P(w)$ for each $w \in \Sigma$. Since a word is only ever observed in a context, c, the distribution over the vocabulary can also be seen as the marginal distribution $P(w) =$

$\sum_c P(c)P(w|c)$ and so Zipf's law equivalently describes this marginal. Zipf's law (see Zipf 1949 and Piantadosi 2014) itself is the observation that $P(w)$ is well-described by

$$P_\alpha(w) = \frac{r(w)^{-\alpha}}{\zeta(\alpha)}.$$

Here, α is the law's parameter, and typically found to be close to 1. r is the ranking function, it assigns to each w its probability rank, i.e. is the most probable word has rank 1, and so on. And finally, $\zeta(\alpha) = \sum_{i=1}^{\infty} i^{-\alpha}$ is the Riemann zeta function.

Notice that because of the dependence on r, Zipf's law describes the distribution over words in terms of the relationship between ranks and probabilities of words, the rank-probability relationship, which will be important throughout this thesis. Further, the rank-probability relationship according to Zipf's law is log-linear, i.e. $\log P_\alpha(w) = -\alpha \log r(w) - \log(\zeta(\alpha))$. Both of these facts are important in understanding the common way of plotting of Zipf's law such as in Figure 1.1: $\log\ P_\alpha(w)$ is plotted against $\log\ r(w)$, showing the predicted relationship between ranks and probabilities of the words in $\mathbf{\Sigma}$, and the relationship is linear in log-log space. In Figure 1.1a, we have plotted Zipf's law at different values of its parameter α and clearly, Zipf's law increases in steepness for larger α

Following (Piantadosi 2014), we use Mandelbrot's generalisation of Zipf's law (Mandelbrot 1953). Although we will use the Mandelbrot generalisation throughout this thesis but keep referring to it as Zipf's law. Based on the observation that Zipf's law tends to overestimate the probabilities of the most common types, Mandelbrot introduced a parameter β to correct for this:

$$P_{\alpha,\beta}(w) = \frac{(r(w)+\beta)^{-\alpha}}{\zeta(\alpha,\beta)},$$

where $\zeta(\alpha,\beta) = \sum_{i=1}^{\infty}(i+\beta)^{-\alpha}$ is the Hurwitz zeta function. Notice, that the additional parameter β simply shifts the ranks and thereby decreases the probability of the types with the lowest ranks. Notice also that as the ranks of types grow, the influence of β vanishes. These effects can be seen in Figure 1.1b, where we have plotted $P_{\alpha,\beta}(w)$ for some values of β: Only the head of the distribution, i.e. the lowest ranks, is affected and there, probabilities are lower for higher values of β.

So much for the definition of Zipf's law itself and its mathematical properties. The real interest in Zipf's law lies of course in connecting it to linguistic data, that is corpora. Since this is full of subtleties, some of which are the central points of this thesis, we introduce this properly and in great detail in Section 2.2.

Situated at the intersection of quantitative linguistics and learnability of language, the learnability of Zipf's law is at the core of the topics in this thesis. Despite extensive efforts to uncover the origins and precise nature of Zipf's law in language, very little attention has so far been directed at the law's effect on language. Hence, research on the effects of Zipf's on language acquisition is a young and emerging field (Kurumada, Meylan, and Frank 2013) and indeed there are to date only two studies which address these effects directly.

The first, (Kurumada, Meylan, and Frank 2013), provides an experimental study in conjunction with an artificial language learning paradigm to contrast human word segmentation performance in contexts with uniform word distributions and contexts with Zipfian word distributions. They find that performance is either the same or higher across trials when the word distribution is Zipf's law versus a uniform distribution. Specifically, performance on unknown words is improved, which leads them to hypothesise a "scaffolding effect" of Zipf's law: The very high-frequency words, which Zipf's law gives rise to, serve as anchors for segmentation by surrounding low-frequency and unknown words.

The second study (Hendrickson and Perfors 2019) uses the same experimental paradigm to investigate how Zipf's law affects cross-situational learning of word-meaning pairs. Again contrasting uniform with Zipfian word distributions, they find that human participants achieve higher performance in the context of Zipfian distributions. As they note, this finding is in direct contrast to two computational studies on the same subject, (Blythe, K. Smith, and A. D. Smith 2010) and (Vogt 2012). Both computational experiments found that Zipf's law leads to a degree of sparsity in the low-frequency words which makes it difficult to disambiguate their meanings and thus hampers learning. Without questioning this finding, (Hendrickson and Perfors 2019) identify memory constraints as a potential reason why human performance is heightened, not lowered, by Zipf's law.

Similar to the two computational studies, (Blevins, Milin, and Ramscar 2017) raise the problem of sparsity induced by Zipf's law for the acquisition of morphological inflections classes. Substantiating the enormous degree of sparsity, the observe that, as one increases corpus size, the number of low-frequency words grows at an, importantly, ever-growing rate. This implies that, increasing corpus size does not remedy the problem of sparsity but only increases it, a result of Zipf's law. From this growth-behaviour, they extrapolate that a learner cannot exhaust inflectional classes which therefore must provide a high degree of regularity to allow the learner to generalise.

Taken together, the initial research on the effects of Zipf's law on language acquisition does not unambiguously point in either direction. At least on the two investigated task domains, humans seem to benefit from a Zipfian word distribution, and especially in the case of (Hendrickson and Perfors 2019), this benefit seems to be specific to human memory constraints. On the other hand, it is still unclear how the raised

concerns of sparsity affects learnability of Zipfian word distributions on the whole.

Relevance and Proposed Explanations

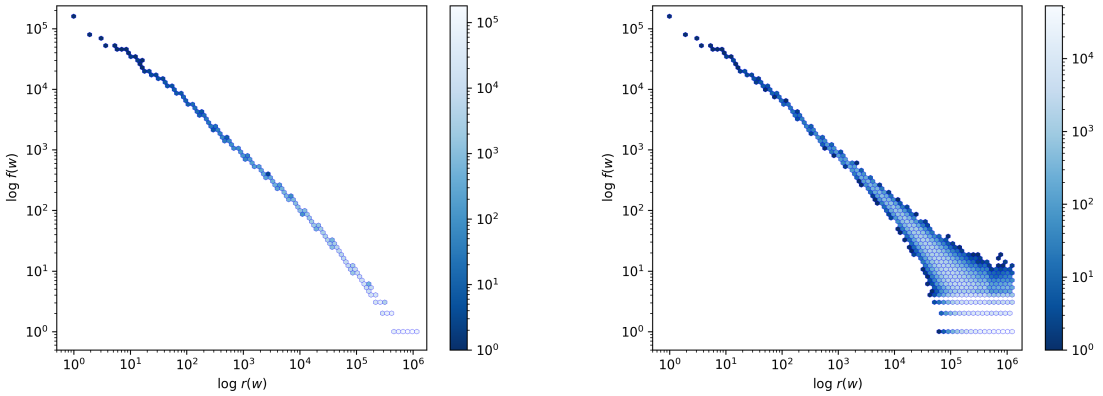
As both (Kurumada, Meylan, and Frank 2013) and (Hendrickson and Perfors 2019) have noted, their studies on learnability of Zipf's law are not only the first to study its effects on learnability but actually the first to study the law's effects on language at all. This is remarkable given the great relevance of Zipf's law in linguistics and other fields, which have attracted a multitude of attempts at explaining it.

Within linguistics, the relevance of Zipf's law arguably stems from what it is about, namely the distribution over the vocabulary, $P(w)$. On the one hand, $P(w)$ can be seen as the distribution over the morphological system of a language (cf. Blevins, Milin, and Ramscar 2017). On the other, $P(w)$ is the distribution over unigrams, i.e. single-word phrases. In this way, $P(w)$ and by extension Zipf's law is the connection between morphology and syntax and a gateway to the combinatorial structure of language. It seems uncontroversial that the combinatorial complexity of language is immense and so it is deeply puzzling that $P(w)$ should follow, if only approximately, a law as simple as Zipf's.

Even adding to this puzzle, Zipf's law has been observed in other, sometimes similarly complex systems, human and natural alike. It has been found in areas of economics, such as income distributions and company sizes (Farmer and Geanakoplos 2008). As extensively discussed in human geography (Arshad, Hu, and Ashraf 2018), Zipf's law governs the distribution over city sizes in a given country and it is commonly encountered in social networks and other aspects of the internet (Adamic and Huberman 2002). Even in biology, where it is observed in the number of species of taxa (Willis and Yule 1922), in geology, for instance in earthquake size (Gutenberg and Richter 1944), and in astrophysics, such as the distribution over solar flares (Lu and Hamilton 1991), Zipf's law is a familiar phenomenon. The extent of Zipf's law suggests that there may be shared properties underlying all these systems, including language, which give rise to the law. This adds to the puzzle because it additionally brings into question which characteristics might make language similar to other systems in the anthroposphere and in nature and if so why such connection should exist.

Since it seems so immediately clear that the existence of Zipf's law in language and its persistence across areas point towards deep characteristics of language, much and in fact most research on Zipf's law has been devoted to deriving it. The sheer amount of the resulting theories has resulted in a long and yet unresolved debate. As (Piantadosi 2014) this is because Zipf's law can indeed be derived from a multitude of different – and often mutually inconsistent – assumptions. We list here only a few those theories in terms of language, see (Piantadosi 2014) for an extensive summary and review.

Several studies have attempted to show that the existence of Zipf's law is uninter-



(a) Commonly used method

(b) Piantadosi's method

Figure 1.2 Rank-frequency relationship in Korean, obtained in (a) from the commonly used method and in (b) from the method proposed by Piantadosi (Piantadosi 2014). Notice that these plots show the frequencies rather than the probabilities of words; these are equivalent in empirical contexts, see Section 2.2 for a detailed discussion.

esting because it even arises in randomly typed texts. This view has, however, been challenged on the ground that texts of natural language are not the outcome of random typing (a discussion can be found in (Ferrer-i-Cancho and Elvevåg 2010)). To a similar conclusion that Zipf's law in language is trivial, (Corominas-Murtra and Solé 2010) have provided a proof according to which Zipf's law may be the only solution for the distribution over the vocabulary of language. The condition for the proof is that the vocabulary expands without bounds over time and the complexity of the language itself stays finite and above 0.

Some accounts of Zipf's law in language have connected it to semantics: For instance, (Manin 2008) manages to reproduce the law by a trade-off between semantic coverage and amount of synonymity in the vocabulary. (Lestrade 2017), on the other hand, shows that Zipf's law arises from an interaction between the sizes of part-of-speech classes and the degrees of vagueness of the word in them. Finally, and famously the theory proposed by Zipf himself (Zipf 1949), is the principle of least effort. In is modern version (Ferrer i Cancho and Solé 2003), this theory predicts Zipf's law from a trade-off in effort between speakers and listeners. Loosely, listeners would prefer a single word in the vocabulary, whereas speakers would prefer a maximally rich vocabulary, and Zipf's law optimises this trade-off.

A Methodological Problem and Piantadosi's Remedy

In this thesis, we explicitly set all issues of explanation and extent of Zipf's law aside and focus on methodology instead. Our starting point is an excellent review of Zipf's law by Piantadosi (Piantadosi 2014), in which he noticed a fundamental problem in

how researchers commonly extract the empirical relationship between the rank and probabilities of words from corpora: Given a corpus of n tokens, C^n, one counts the tokens to establish $P(w)$ for each word type w. Then, one orders all w according to $P(w)$ in descending order and assigns the rank $r(w) = i$ if w is the i-th most probable word. Piantadosi argues that this leads to "spurious regularity" as $r(w)$ is negatively correlated to $P(w)$ by force when obtained this way. Moreover, it does not allow $r(w)$ to vary with respect to $P(w)$ because it is just a deterministic function of the latter. For reference, we have plotted the rank-probability relationship as obtained from this, the common method, in Figure 1.2a. Notice that it describes a single line which is precisely the lack of variance we mean. Moreover, as Piantadosi has remarks, even though this line has some deviation from a straight line, this deviation is uninterpretable because it might just be an artefact of the extraction method.

Based on the request that $r(w)$ and $P(w)$ should be obtained independently, Piantadosi proposes a simple fix: One splits the given corpus C^n in half by randomly assigning its tokens to subcorpora $C_1^{n/2}$ and $C_2^{n/2}$. After that, one computes $r(w)$ and $P(w)$ as before, but now $r(w)$ from $C_1^{n/2}$ and (w) from $C_2^{n/2}$. (Notice that one still needs to compute $P(w)$ from $C_1^{n/2}$ in order to establish $r(w)$ but this is simply discarded.) According to Piantadosi, this allows for variance between $r(w)$ and $P(w)$ and the correlation we find between the two is no longer prescribed by the method of obtaining them. Instead, so Piantadosi, the correlation we do find will be genuine and amenable to interpretation.

The relationship between ranks and probabilities as obtained by Piantadosi's method is plotted in Figure 1.2b. While the heads of the two plots in Figure 1.2 are very similar, the clear difference lies in their tails. The improved method of Piantadosi allows the relationship to deviate from a single line and this is most prominent in the low-frequency word types for which there is high uncertainty about the precise values of $r(w)$ and $P(w)$.

1.2 This Thesis

The Sample-Source Distinction and Subsampling

The first half of this thesis is in fact instigated by Piantadosi's observation that the common methodology to obtain the rank-probability relationship from data is seriously flawed. However, in Chapter 3, we describe why and how even Piantadosi's solution falls short of providing an actual solution. We arrive at this conclusion by first re-assessing the aim of quantitative linguistics in general. Subsequently, we realise that the commonly used, erroneous methodology stems from being oblivious to an essential dichotomy in statistics: the distinction between an observed sample and the theoretical source the sample was drawn from, or the sample-source distinction as we call it.

By extension, the methodological problem in the literature also entails negligence to the problem of estimation, namely that observations are subject to random fluctuations so that inferences from them about the source cannot be made directly. Realising this, we can simply draw from the established estimation methods from statistics and concretely we re-derive and advocate the use of the Subsampling method.

The majority of Chapter 3 is spent on analysing the estimates obtained by the Subsampling method, thereby setting examples of how they can enrich the empirical research on Zipf's law. We close the chapter by generalising the use of the Subsampling method beyond Zipf's law and to other laws of quantitative linguistics with the example of Heap's law. Moreover, since our estimates are arguably the most reliable to date, we re-assess the degree to which language actually conforms to Zipf's law, already in the second half of Chapter 2.

The Filtering Method

Moving on to learnability analyses, the second main topic of this thesis, in Chapter 4 we begin again by clarifying methodological issues in the previous literature. We notice once more that the sample-source distinction is disregarded which leads to invalid alternatives to Zipf's law in the comparative approach to learnability.

Based on the insight which distributions constitute relevant alternatives to Zipf's law in the context of human language, we develop the Filtering method, a novel method for automatically generating data based on the information-theoretic concept of typicality of a corpus with respect to a given distribution. For the particular case of Zipf's law, we describe implementations of two sampling algorithms which instantiate the Filtering method. These special sampling algorithms are required, as we discuss, because of the so-called asymptotic equipartition property, a version of the law of large numbers from information theory.

Because these sampling algorithms have unknown and complex sampling behaviours, we devote the remainder of Chapter to analysing the samples generated by the two filtering algorithms. With these analyses we assess the usefulness of the Filtering method for future research and argue for its success.

Subsampling and Filtering for Learnability Assessments

We briefly return to the original goal of the Filtering method at the end of Chapter 4 and describe how it can be used to facilitate computational learnability studies more realistic and more detailed than those in the previous literature. To generalise this use even further, in the conclusions (Chapter 4) we make our final contribution by detailing how the Subsampling method in connection with the Filtering method can be applied to the debate on general language acquisition. In order to to so, we introduce the innateness debate and summarise recent exciting theoretical work which provides

formal proofs that innate constraint a not a priori necessary for language acquisition. Since they rely on the theoretical concept of Kolmogorov complexity, these proofs are not easily connected to empirical evaluation but, with the help of the Subsampling and Filtering methods, we propose a way to do so. Owing to the weight of the underlying debate and the level of abstraction of the used formal tools, the design the of computational experiments we devise is an apt example of the full potential of the Subsampling and of the Filtering method.

We begin now begin by describing the prerequisite of any empirical study, namely the data we use throughout this thesis.

2 Data & Basic Statistics

The most essential ingredient to empirical research is data. In this Chapter, we lay out and justify our choice of data and describe the pre-processing from raw material into a corpus of language which can be analysed by computational methods. We then ease into the empirical parts of this thesis by showing how to go from the empirical data to Zipf's law and how Zipf's law manifests itself in our specific data set. In the same manner, we describe vocabulary growth as another empirical observation of interest and introduce Heap's law which predicts it. In this way, the current chapter sets the stage for Chapters 3 and 4 in which we develop new methodology and conduct original empirical analyses on Zipf's law.

2.1 Wikipedia as Corpus

Our experiments require collections of text which are both large and available in many different languages. This is a notoriously ambitious requirement but one that Wikipedia can fulfil. Even though Wikipedia is not a perfectly representative linguistic corpus due to its specialised language and partly widespread use of templates and bots for text generation, there are three factors which make it highly convenient for use in our case: (1) Wikipedia is open-source and so are the tools for processing it, (2) Wikipedia authors use mostly the standardised variant of their respective language which makes it comparable cross-linguistically and easy to process computationally and (3) its structure, being a set of independent articles made up of continuous text.

We stress that corpora which are more representative for the language that a learner receives do exist, such as CHILDES (MacWhinney 2014), but are not viable options to our investigations because they are neither multilingual nor large enough.

Languages

We use Wikipedia in seven languages (language codes in parentheses): Esperanto (EO), Finnish (FI), Indonesian (ID), Korean (KO), Norwegian (NO, the Bokmål variant), Turkish (TR) and Vietnamese (VI). The first consideration in choosing this set of languages is that all seven are of similar size and large enough. This is the main reason for excluding English which is too large to be handled straightforwardly.

The second is that our analyses are supposed to hold cross-linguistically, so we

want to cover as much as possible of the variety in the world's human languages. So firstly, none of these seven languages are (closely) genetically related according to scientific consensus – although Esperanto, a constructed language, could be argued to belong to the Indo-European languages together with Norwegian. Esperanto was indeed chosen specifically because it is a constructed language and not the outcome of an evolutionary process. Thus, Esperanto is expected to have high morphological regularity which will have an impact on both Zipf's law and its learnability (cf. Gobbo 2017).

For the other languages, our main focus is morphological variety: Norwegian and Finnish possess fusional morphological characteristics, Korean and Turkish (and to some degree Finnish) have agglutinative morphology and Vietnamese is a language with isolating morphology. Our set of languages thus covers all three of the most general morphological systems. The differences in morphological structure in these systems lead to different conceptions of what constitutes a word and this obviously affects the distributions over words in them. Moreover, the learnability of a language is related to its morphological complexity (Blevins, Milin, and Ramscar 2017), so the learnability of Zipf's law may also differ across morphological systems.

Finally, we choose Indonesian for its status as primarily a lingua franca, i.e. most of its speaker have a different native language). (Ferrer i Cancho and Solé 2001) found that the word distributions of creole languages are much better described by two separate exponents for Zipf's law than just a single one. They explained this finding with the existence of a small and highly productive core vocabulary and a large, mainly unproductive extended vocabulary. Although Indonesian is not a creole, we suspect a similar phenomenon in lingua franca and therefore differences in our findings for Indonesian with respect to the other languages.

The genealogical and typological information we presented here can be found in the World Atlas of Language Structures (WALS, Dryer and Haspelmath 2013).

From Wikipedia to Corpus

A raw Wikipedia is of course not yet suited for computational linguistic analyses, since it contains large numbers of non-linguistic items. For each language, our pipeline from the on-line Wikipedia to a corpus is the following:

1. Download the latest Wikipedia dump from `dumps.wikimedia.org/eowiki/latest` (example for Esperanto (EO)). Such a dump is an XML representation of the entire Wikipedia without media such as images and videos.
2. For each article in the Wikipedia, we extract the main text and remove all XML and list, table and link annotations. For this, we rely on the open-source Python library `WikiExtractor` (Attardi and Fuschetto 2012).
3. Being left with only linguistic data, we clean it by removing all special and meta-

linguistic characters contained in the Unicode blocks `BASIC LATIN`, `LATIN-1`, `ARABIC` and `CJK` (Chinese, Japanese, Korean). We keep essential punctuation characters, such as commas and full stops. This important for sentence and token segmentation to work well, which we perform in the next step.

4. For a set of languages as diverse as ours, there is no unified algorithm for detecting sentence and word boundaries, so monolingual segmenters yield unsatisfactory results and implementations for truly multilingual segmentation are sparse. Fortunately, the multilingual natural language processing library for Python `polyglot` (Al-Rfou, Perozzi, and Skiena 2013) exposes an interface to the Unicode Text Segmentation algorithm (Davis and Iancu 2012) and hence supports sentence and word segmentation. For this algorithm, the Unicode consortium has developed sets of language-specific rules which characters can indicate sentence and word boundaries in which contexts. Thus, while not necessarily state-of-the-art on any particular language, this algorithm outperforms others cross-linguistically.
5. A standard procedure in natural language processing, we lower-case all characters to avoid orthographic variation between tokens of the same type, for example because a token occurred at the beginning of a sentence. Note that this is not unproblematic because it removes orthographic differences between tokens of different types – consider the distinct English words 'polish' and 'Polish'.

The result of this procedure constitutes a corpus for our purposes and this is what we use in the analyses throughout this thesis. Note that although we segment at all levels, we do keep words grouped into sentences and sentences grouped into the original articles of Wikipedia. Therefore, each corpus is a set of articles each of which is in turn a sequence of sentences, each of which is itself a sequence of words. Keeping this structure is important for the methodology we develop and evaluate below.

Basic Quantities

Table 2.1 gives the sizes of the seven Wikipedia corpora after pre-processing in terms of articles, sentences and words (tokens). Notice that, with the exception of Esperanto which has fewer, we length-matched all corpora to have the same number of $50{\cdot}10^6$ tokens in order to increase comparability of cross-linguistic findings. Length-matching was done by simply randomly sampling articles from the original set until the desired number of tokens was reached.

While the average length of articles is rather short, standard deviation is massive, reflecting the typical large amount of stubs and small amount of excellent articles in Wikipedia. Optimally, for our empirical research below, article lengths would be much more uniform but, again, there is not much choice of large multi-lingual corpora besides Wikipedia. The variation in both number of articles and article length across languages is an indicator of varying quality between the Wikipedias, fewer and longer

	EO	FI	ID	KO	NO	TR	VI
articles	$2.45{\cdot}10^5$	$2.94{\cdot}10^5$	$3.31{\cdot}10^5$	$3.4{\cdot}10^5$	$2.81{\cdot}10^5$	$2.82{\cdot}10^5$	$5.02{\cdot}10^5$
(μ len, σ len)	(9.34, 404)	(14, 669)	(9.1, 535)	(9.83, 686)	(10.8, 566)	(12.7, 780)	(4.7, 288)
sentences	$2.29{\cdot}10^6$	$4.1{\cdot}10^6$	$3.02{\cdot}10^6$	$3.35{\cdot}10^6$	$3.04{\cdot}10^6$	$3.58{\cdot}10^6$	$2.36{\cdot}10^6$
(μ len, σ len)	(16.7, 11.9)	(12.2, 6.53)	(16.6, 10.3)	(14.9, 9.45)	(16.4, 8.94)	(14, 10.2)	(21.2, 15.3)
tokens	$38.3{\cdot}10^6$	$50{\cdot}10^6$	$50{\cdot}10^6$	$50{\cdot}10^6$	$50{\cdot}10^6$	$50{\cdot}10^6$	$50{\cdot}10^6$
types	$1.17{\cdot}10^6$	$2.39{\cdot}10^6$	$0.76{\cdot}10^6$	$3.33{\cdot}10^6$	$1.23{\cdot}10^6$	$1.25{\cdot}10^6$	$0.56{\cdot}10^6$
TTR	0.030	0.047	0.015	0.066	0.024	0.024	0.011

Table 2.1 The basic quantitative characteristics of our Wikipedia corpora. The means and standard deviations in rows 2 and 4 (denoted μ and σ) are of the length distributions in the articles and sentences respectively. TTR is the common abbreviation for the type-token ratio (number of types divided by number of tokens).

articles generally indicating higher quality.

Sentence length, on the other hand, is likely to vary across languages not only because of quality but also because of linguistic differences. Vietnamese, an isolating language, has a much lower morpheme-per-word ratio compared to the other languages, leading to a higher number of words per sentence. The effect of isolating morphology can also be seen in the type-token ratio (TTR). Vietnamese displays an extremely low TTR which is owed to its lack of inflectional or derivational morphology. On the opposite, languages with high degrees of inflection like Finnish and Korean have high TTRs because of large numbers of types. In sum, the variation across languages is likely both due to differences in quality and due to differences in linguistic structure but it is difficult to disentangle the contribution of both to that variation. Even though we have made efforts to mitigate this variation by linguistic pre-processing and normalisation, it may influence some of the empirical observations we make in this thesis and something to be kept in mind.

2.2 The Rank-Frequency Relationship & Zipf's Law

Before delving into the main experiments of this thesis, it is useful to first get familiar with the empirical side of Zipf's law. Especially because this thesis is about quantitative linguistic methodology, we go into great detail in introducing how the empirical rank-probability relationship is connected in practice to the theoretical Zipf's law. We show and explain how the empirical rank-probability relationship is represented graphically and then describe maximum likelihood estimation of the parameters of

Zipf's law from that relationship.

2.2.1 The Empirical Rank-Frequency Relationship

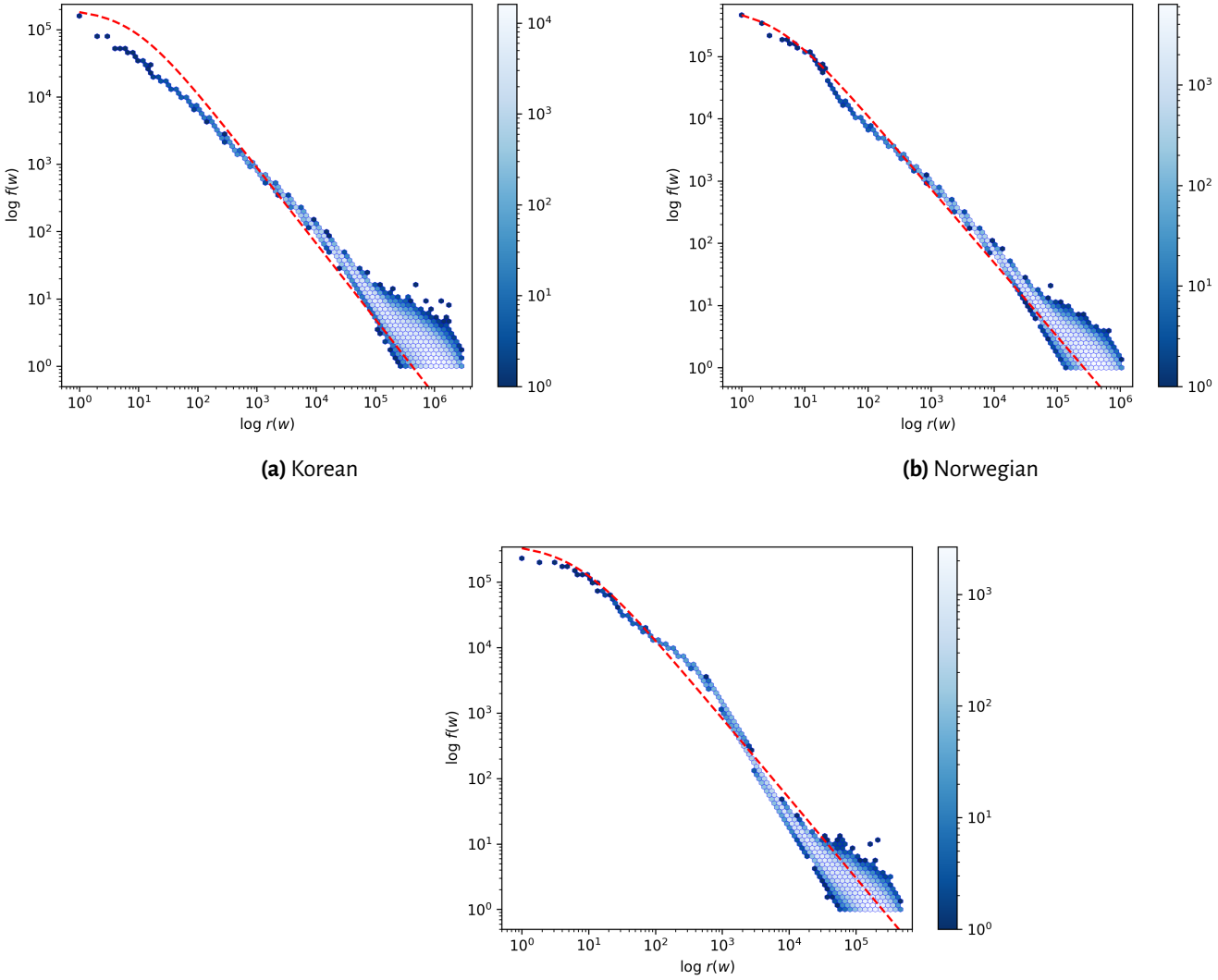
In Figure 2.1, we show Zipf's law obtained from the $50{\cdot}10^6$ tokens of three of the Wikipedia corpora, namely Korean, Norwegian and Vietnamese. We have selected these three language because they are representatives of agglutinative, fusional and isolating morphology, respectively, and therefore cover the spectrum of variation in the word distribution. We will use this subset for the remainder of the chapter for comparability, the plots for the remaining languages can be found at github.com/valevo/Thesis/figures.

In order to construct these graphs, chapterwe have extracted the three relevant observable variables: (1) the set of types (vocabulary), and for each type in that vocabulary and (2) its frequency and (3) its frequency-rank. The rank is a simple transformation of the observed frequency: if a type is the i-th most frequently observed word, its frequency-rank is i. Notice with regard to this transformation that some, and indeed many, types will have the same observed frequency and would be assigned the same rank – we break ties by assigning ranks randomly among these types.

Thus, to be more accurate, Figure 2.1 shows the extracted rank-frequency relationship in the three Wikipedia corpora, rather than Zipf's law which makes a prediction about this relationship – Zipf's law itself is shown as the red dashed line. Although it may seem pedantic, we emphasise this distinction because it is often neglected and a source of confusion between empirical observations and theoretical model of these observations. We will return to and elaborate on the distinction between observation and model in Chapter 4. A second point about accurate methodology we emphasise and strongly insist on is that this relationship is not extracted but estimated since both rank and frequency for each word are in fact estimated. This important statistical distinction concerns one of the main contributions of this thesis and is the central topic of Chapter 3 and we elaborate in great detail there. Notice, importantly, that the relationships in Figure 2.1 are properly estimated which is why they look different from the plots in Figure 1.2 of the Introduction and what readers may have seen in other papers.

In any case, we obtain one two-dimensional data point for each observed type: its rank on the x-axis and its frequency on the y-axis. Rather than plotting every individual data point in Figure 2.1, we follow (Piantadosi 2014) in using a two-dimensional histogram. Each hexagon in the graphs indicates by its shade how many points fall into its area. This technique is aimed at making the plot more robust against visual artefacts and makes areas of high point density visible.

Finally, notice that the y-axis of the graphs shows the log-frequency rather than the log-probability, even though we defined Zipf's law in terms of probability. This is because the probability of a type is not directly observable, only its frequency is. Given



(a) Korean

(b) Norwegian

(c) Vietnamese

Figure 2.1 The rank-frequency relationships in (a) Korean, (b) Norwegian, and (c) Vietnamese. Both scales are log-transformed. The blue hexagons represent two-dimensional bins, the shading (see colour bars on the right of each plot) indicates the number of words which fall into each bin; notice that this shading is also on a log-scale. The dashed red lines correspond to the predictions of Zipf's law with the MLE parameters (see Table 2.2). The predictions are scaled from probabilities to frequencies by multiplying them with the overall number of tokens.

the observed frequency and no prior assumptions, our best estimate of the probability would simply be a scaled version of that frequency. This would only result in a different scaling on the y-axis and therefore not affect the graphs themselves. Because of this direct correspondence between estimated frequency and estimated probability, we use them interchangeably throughout this thesis and so the rank-frequency relationship of words is to be understood as completely equivalent to their rank-probability relationship.

Now, Zipf's law predicts a negative log-linear relationship between the ranks and frequencies of words. And indeed, at first glance, the rank-frequency relationships in Figure 2.1 are highly linear, closely following a relatively straight, downward line. This is especially so considering that many shapes, including pure noise, would have been possible in principle. The high degree of linearity in these graphs is in fact one of the most widely used criteria for positing Zipf's law in empirical observations.

Furthermore, we see that the length of the x-axes and y-axes of the plots are both on the same order of $10{\cdot}10^6$. This fits the more specific prediction of Zipf's law that the parameter α is close to 1. As is predicted by Mandelbrot's extension of Zipf's law (the addition of the parameter β), we see in all languages that the head of the graph curves off slightly. This means that the highest-frequency types are less frequent than would have been predicted by a straight line, i.e. the original Zipf's law.

Looking at the variance, by which we mean deviation from a single line, we observe that while the head of the graph has little variance, the variance increases considerably in the tail. This is not surprising since on the one hand, the tail is inhabited by far more types, allowing for more variance, and on the other hand, low frequencies carry more inherent uncertainty. Although the theoretical Zipf's law predicts a straight line, this variance does not per se invalidate Zipf's law. The reason is once more the mentioned distinction between the observed sample, created by a process which involves randomness, and the theoretical model, or distribution, from which the sample was generated. The distribution itself does not contain any randomness and therefore cannot account for the randomness in the sample which is the source of the variance we observe in the plots. Again, randomness and the sample-distribution distinction is a core topic of Chapter 4.

In summary, from a broad, informal inspection, the rank-frequency relationships look like they may be explainable by Zipf's law. It is already clear, however, that this can only be approximate: Even when disregarding the variance, there are no single straight lines which exactly fit the empirical rank-frequency relationships, especially in the upper third (half). For a more formal and quantifiable analysis we now turn to maximum likelihood estimation.

2.2.2 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a simple and commonly used method for estimating the parameters of a distribution from data. In our case, the parameters of Zipf's law are α and β and the data is a corpus $C^n = (w_1, ..., w_n)$ (the superscript indicates the number of tokens n). MLE works (see e.g. Deluca and Corral 2013) by finding values $\hat{\alpha}$ and $\hat{\beta}$ which satisfy

$$\arg\max_{\alpha,\beta} P_{\alpha,\beta}(C^n) = \arg\max_{\alpha,\beta} \prod_{i=1}^{n} P_{\alpha,\beta}(r(w_i)),$$

where $r(w_i)$ is our estimate of the rank of the word at position i. The equality holds because under Zipf's law, the tokens in a corpus are independent of each other. For numerical stability, one typically minimises $-\log P_{\alpha,\beta}(C^n)$ which leads to the same result. Moreover, since the search for the optimal parameters values is usually intractable, stochastic optimisation is used which implies that the returned parameters are random variables with generally positive variances.

It is worth noting that the MLE is a consistent estimator, i.e. in the limit of sample size $\hat{\alpha}$ converges to α^*, the true parameter value (and similarly for β). That is to say, while the MLE is usually not the most efficient estimator, it is in principle capable of finding the true parameters (see e.g. Moreno-Sánchez, Font-Clos, and Á. Corral 2016). Moreover, MLE also works straightforwardly even for probability distributions with problematic properties such as Zipf's law which has not even a finite mean for $\alpha < 2.0$ (Goldstein, Morris, and Yen 2004). Also, since we do not have any educated guesses on the parameters' prior distributions, the generally more reliable Bayesian maximum a posteriori estimation would simply collapse to MLE. In spite of its simplicity, the MLE is therefore arguably one of the best practical choice in estimating the parameters Zipf's law.

Of course, we do not have to take the parameters returned by the MLE at face value. Instead, we can measure their confidence and quality for which we use the following metrics:

- As mentioned, the optimisation is stochastic and therefore is re-run a number of times. The relative standard error (rel. SE) is the variance of the returned parameters across these runs, normalised to be a percentage. The relative SE we report is given by the `statsmodels` Python package (Seabold and Perktold 2010) which we use to perform the stochastic optimisation. A high SE indicates that highly different optimal parameter values result in high likelihood for the data. This can point towards problematic properties in the likelihood function and will generally lower our confidence in any particular value returned by the MLE.
- McFadden's R^2_{McF} (McFadden 1973) is a pseudo-R^2 measure with an interpre-

	EO	FI	ID	KO	NO	TR	VI
α	1.19	1.14	1.19	1.13	1.19	1.15	1.22
β	1.46	7.44	6.35	9.36	3.48	8.64	6.75
(rel. SE α, rel. SE β)	($6{\cdot}10^{-5}$, $2{\cdot}10^{-3}$)	($4{\cdot}10^{-5}$, 10^{-3})	($5{\cdot}10^{-5}$, 10^{-3})	($3{\cdot}10^{-5}$, $7{\cdot}10^{-4}$)	($5{\cdot}10^{-5}$, 10^{-3})	($5{\cdot}10^{-5}$, $7{\cdot}10^{-3}$)	($7{\cdot}10^{-5}$, $6{\cdot}10^{-3}$)
R^2_{McF}	0.71	0.66	0.71	0.65	0.71	0.67	0.74
rel. BIC	3.44	2.95	3.42	2.81	3.44	3.07	3.78

Table 2.2 Maximum likelihood estimates of the parameters α and β of Zipf's law for our 7 languages. See the main text for the definitions and interpretation of the relative standard error (rel. SE), R^2_{McF} and relative Bayesian information criterion (rel. BIC).

tation similar to the R^2 coefficient of determination: How well does the fitted model, in our case $P_{(\hat{\alpha},\hat{\beta})}$, fit the data in comparison to a null model P_{NULL}? Formally, this is calculated as $R^2_{McF} = 1 - \frac{\log P_{(\hat{\alpha},\hat{\beta})}(C^n)}{\log P_{\text{NULL}}(C^n)}$. Since the likelihood of the corpus C^n under the fitted model is at least as high as under the null model, R^2_{McF} is guaranteed to be in the interval $[0, 1]$ and may hence be interpreted as the percentage to which $P_{(\hat{\alpha},\hat{\beta})}$ provides improved fit over P_{NULL}. The practical range of values of R^2_{McF} is, however, highly dependent on both the model itself and the null model, so much so that there are no general rules for interpretation of specific values of R^2_{McF}.

Choosing an appropriate null model is intricate and for simplicity we set $\alpha = 1$ and $\beta = 0$ to obtain P_{NULL}. Looking at the definition of Zipf's law, this leads to $P_{\text{NULL}} = P_{1,0}(W = w) = \frac{(r(w)+0)^{-1}}{\zeta(1,0)} \propto \frac{1}{r(w)}$, i.e. Zipf's law in which the parameters α and β have cancelled out. Note that $P_{(1,0)}$ corresponds to the flattest and most straight possible Zipf's law. In the actual implementation, an infinite vocabulary requires $\alpha > 1$ and $\beta > 0$, so we set $\alpha = 1 + \epsilon$ and $\beta = 0 + \epsilon$, with ϵ a negligible value. See Section 4.1 for a detailed discussion about the restrictions on the values of α and β and alternatives to Zipf's law in the rank-probability relationship of language.

- The Bayesian information criterion $BIC(P_{\alpha,\beta}) = 2\log(n) - 2\log\left[P_{(\alpha,\beta)}(C^n)\right]$ (Schwarz 1978) has been devised for selection among competing models. Instead of a classical significance test for the MLE parameter values $\hat{\alpha}$ and $\hat{\beta}$, we use the BIC to judge whether whether the data justifies a non-null model (i.e. Zipf's law with the MLE parameters). We do so by calculating $BIC(P_{\hat{\alpha},\hat{\beta}})/BIC(P_{1,0})$, the relative BIC, where we use the null model $P_{\text{NULL}} = P_{(1,0)}$ from above. The higher the relative BIC, the more do the optimal parameters provide a better model than the null model and hence the greater our acceptance for it.

The MLE parameter values $\hat{\alpha}$ and $\hat{\beta}$ for all languages together with their respective

relative SE, R^2_{McF} and relative BIC are given in Table 2.2. The observations about these values we make now hold across languages, language-specific differences are discussed in the next section.

The relative SE is negligible for all languages and both α and β and we therefore regard the MLE as having terminated successfully, i.e. as having found definitive values which maximise the likelihood of the data. The relative SE is higher for $\hat{\beta}$ which hints at the fact that its importance in modelling the data is lower than that of $\hat{\alpha}$, i.e. different values of β lead to relatively similar values of the likelihood of the data.

As is evident from their definitions, R^2_{McF} and the BIC measure similar aspects of the goodness of fit of a model. Their interpretations, however, are quite different. Even without reference values, we find a relatively low R^2_{McF} which indicates that even when the MLE parameters are used, there is substantial variance in the data which Zipf's law cannot model. This is not surprising since, as mentioned, the actual rank-frequency relationship is far more complex than the simple Zipf's law. Therefore, neither the null model nor the MLE parameter model fit the data particularly well, leading to similarly low likelihoods and thus R^2_{McF} to be rather low. At the same time, the relative BIC is well above 2 for all languages, implying that the null model has a BIC at least twice as low as the model with MLE parameter values. As we use the relative BIC in place of a significance test, we regard the fitted model as providing significant fit to the data and therefore as a model of the data that can be termed appropriate. Taking R^2_{McF} and the BIC together, we find that although even a fitted Zipf's law cannot provide very close fit to the data, it is still preferable over the null model, a Zipf's law without parameters. As the result, we retain the MLE parameters as preferable over any other parameter values for Zipf's law and under the premise that Zipf's law altogether may strictly speaking have to be dismissed as the true model of the empirical rank-frequency relationship in language.

In this way, we arrive at a qualitatively similar conclusion as has been argued previously by (Piantadosi 2014) and (E. G. Altmann and Gerlach 2016). Zipf's law cannot precisely fit the rank-frequency relationship of language, as the plots in the previous section make clear as they show systematic deviations and substantial variance. As both previous papers have also argued, this is to be expected, since Zipf's law is merely a statistical model of and not the true underlying source for word use in language. Hence, there are necessarily aspects of the observed word distribution that Zipf's law fails to capture. The use of MLE reveals, however, that the law does fit language well enough to warrant application and interpretation of statistical methods. Specifically, different parameter values lead to differences in fit that can be detected by common goodness-of-fit measures and thus indicate that Zipf's law does capture a significant aspect of the empirical data. Moreover, in an argument related to Occam's Razor, (Piantadosi 2014) comments that the fit of Zipf's law to language is in fact remarkable given the simplicity of the law on the hand and the complexity of language on the other. Given these arguments, we conclude that natural language is what (Piantadosi

2014) calls "Zipfian". That is, the empirical rank-frequency relationship of the words of language is described to a significant extent by Zipf's law but only approximately so. As mentioned in the introduction, this conclusion is in contrast to much of the earlier statistical work on Zipf's law which has sought to unambiguously prove or reject Zipf's law (e.g. Baayen 2002 or Moreno-Sánchez, Font-Clos, and Á. Corral 2016).

In addition to providing a formal tool for assessing the empirical Zipfianness of language, MLE and specifically the MLE parameters of Zipf's law will have a pivotal role in the methodology we develop in Chapter 4. In a way that we will make precise, Zipf's law together with its MLE parameter values can also be used as a stochastic source distribution, one which can be manipulated in data and of which we can measure how closely that data follows it. Both of these uses of the MLE parameter values are only justified if Zipf's law in general manages to provide a reasonably good description of the data. As we have reported and argued, Zipf's law indeed does capture much of the structure in the rank-frequency relationship and therefore can be used as a practical approximate source distribution for it.

2.2.3 Differences Across Languages

So far, we have discussed the empirical rank-frequency relationship as if it was the same in all languages. The reason that we have, and can, is that indeed this relationship is highly similar across languages and to a degree that the entire discussion above holds for all seven languages we have analysed. Great similarity can be clearly seen in Figure 2.1, as well as in the MLE results in Table 2.2, where parameters have similar values and lead to similar fit. This high degree of similarity seems to universally expand to all languages and is a core reason why Zipf's law has received so much attention, see (Piantadosi 2014) for a general review. But, of course, there is also variation in the rank-frequency relationships of different languages and given strong morphological differences between them, this is to be expected. In this section, we give a brief overview of the variation across languages in terms of gradient, curvature and Zipfianness of the rank-frequency relationships. We refer the reader again to github.com/valevo/Thesis/figures for the plots of the rank-frequency relationships of the languages we could not show and discuss here.

A simple and evident example of how the rank-frequency relationships differ are their respective gradients: The relationship is vastly flatter in Korean than in Norwegian and Vietnamese, with the latter being steepest. In all languages, the maximum frequency of a word (the range of the y-axis in the plot of Figure 2.1) is similar, so the difference in gradient is due to different sizes of the relationships' support. This is indeed evidenced by the number of types and TTRs reported in Table 2.1 which show that Korean has by far the largest number of types. Containing a higher number of types makes the rank-frequency relationship flatter because frequency mass needs to be distributed over more items and is not surprising for an agglutinative language such as Korean. We see this also reflected in the MLE values of the parameter

α of Zipf's law (Table 2.2): Korean, together with the other agglutinative languages Finnish, Indonesian and Turkish, corresponds to the lowest values of α, whereas Vietnamese with its isolating morphology leads to a high value of α due to a steep rank-frequency relationship.

Perhaps even more evident but not as easily explained is the fact that the rank-frequency relationships of different languages exhibit different degrees of curvature and in different areas. Generally, curvature seems to occur in one of two areas: Firstly, in the head of the relationship, as strongly exhibited by Norwegian and Vietnamese. It is this curvature which prompted Mandelbrot's inclusion of the parameter β, attempting to correct Zipf's law for it. Secondly, in the middle range, which is the case for Vietnamese. As mentioned, (Ferrer i Cancho and Solé 2001) have characterised curvature in the middle range as a broken power law, i.e. as actually two Zipf's laws with distinct parameters α. They connected this phenomenon to the morphological productivity in language, specifically that of creole languages. The strong curvature of the rank-frequency relationship of Vietnamese could point towards such a broken power law and to a possible connection in morphology between creole and isolating languages such as Vietnamese. On the other extreme, Korean does not show much curvature in either of the two areas and although one may be inclined to conjecture morphology again as the reason, the other agglutinative languages' rank-frequency relationships are not as straight, exhibiting curvature in one of the two areas.

The Zipfianness of the rank-frequency relationships, that is the fit of Zipf's law, turns out to differ along with the relationships' gradient and curvature, as can be seen in the goodness-of-fit measures R^2_{McF} and relative BIC of Table 2.2. Lower gradients lead to lower descriptive power according to the measures because of our choice of the null-model. As mentioned, this null-model corresponds to the flattest and straightest possible rank-frequency relationship, that is the lowest possible values of α and β. Therefore languages with flat relationships, such as the agglutinative ones, lend comparatively low support to their MLE parameter values. Similarly and somewhat counter-intuitively, relationships with low degrees of curvature, such as that of Korean, do not lead to better fit of Zipf's law. Mandelbrot's correction in fact implies that some curvature is expected and if curvature is absent, the correction leads to lower fit.

Generally, deeper investigation is required for more conclusive observations and hypotheses about cross-linguistic differences. Based on the preliminary observations we have made here, such investigation will most likely be fruitful and contribute insight into the nature of the morphological systems which underlie the rank-frequency relationships and hence the different incarnations of Zipf's law in the different languages. For the remained of this thesis, specifically Chapters 3 and 4, we will however ignore language-specific phenomena in the interest of conciseness and keep the discussion general enough to be valid cross-linguistically. The generally high degree of similarity and subtlety of the differences between the rank-relationships of the different languages justify this.

2.3 Vocabulary Growth & Heap's Law

While Zipf's law is arguably the most well-known and deeply studied quantitative law, a number of other phenomena have been observed to also hold across languages and across corpora. One such phenomenon is Heap's law (see e.g. Petersen et al. 2012 or Gerlach and E. G. Altmann 2014). Zipf's law remains the focus of this thesis but Heap's law will serve as an example of how the methodological remarks we make in Chapter 3 are not only applicable to Zipf's law but generalise across quantitative laws. Additionally, in Chapter 4, we will use Heap's law as a control in our experiments. For this reason, we briefly introduce the law itself here and show how well our data is described by it with the help of maximum likelihood estimation (MLE), much akin to the previous section.

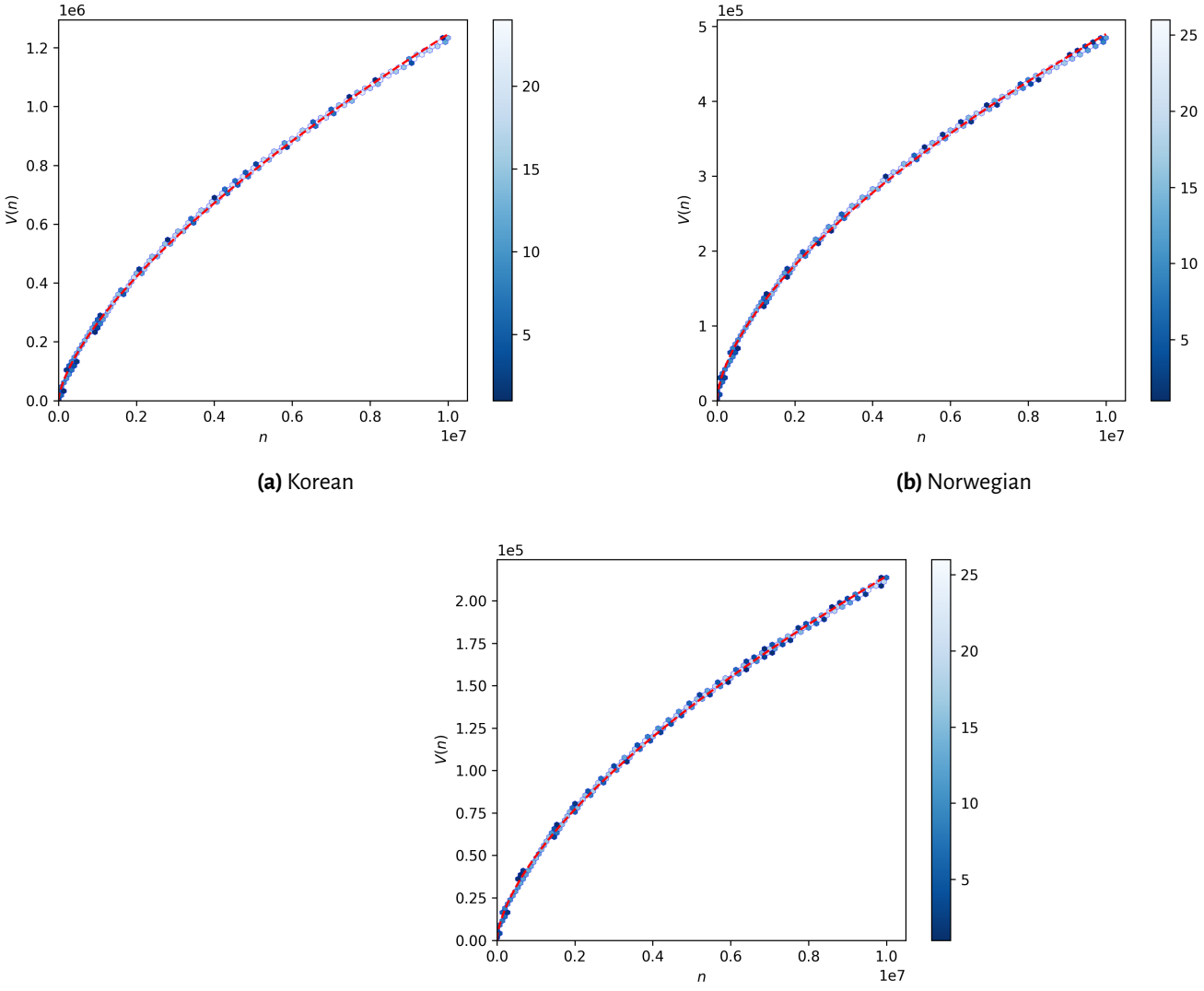
Heap's law describes the vocabulary growth across corpus sizes, i.e. the number of types, $V(n)$, in a corpus C^n. Heap's law is the observation that $V(n)$ is sublinear which is formally stated as:

$$V(n) = \phi * n^{\tau},$$

where $\tau < 1$ and $\phi > 1$. It is in fact trivial that $V(n)$ is sublinear, and therefore $\tau < 1$, because the number of types in a corpus can of course not exceed the number of tokens. What makes Heap's law a meaningful observation is that $V(n)$ behaves precisely like a function of the form $\phi * n^{\tau}$ and not, for instance, like a more complex polynomial or no particular function at all. As is the case with any sublinear function, an important consequence of Heap's law is that the discrepancy between n and $V(n)$ grows with n, even though $V(n)$ does grow to infinity in the limit of n according to Heap's law.

As Figure 2.2 indeed shows, the empirical vocabulary growth shows very little deviation from the simple function given by Heap's law across Korean, Norwegian and Vietnamese. (See github.com/valevo/Thesis/figures for the plots for the remaining languages.) In constructing the graphs in this figure, we stress the same distinction as we did for the rank-frequency relationship (Section 2.2.1): at each n, $V(n)$ is the result of estimation, not calculation. That is, rather than computing $V(i)$ for each $i = 1, ..., n$ from a single corpus C^n, we take a series of independent corpora $C_1, ..., C_n$ and compute each $V(i)$ from corpus C^i. See Section 3.4 for the details of this process.

Just like we did for Zipf's law, we use MLE to determine the optimal parameters $\hat{\phi}$ and $\hat{\tau}$ given our data. There is one caveat with using MLE for Heap's law however: As opposed to Zipf's law, Heap's does not define a probability distribution. Instead, it defines a function and therefore does not inherently assign likelihood to a set of observations. We can still perform MLE but need to make additional assumptions, linking the data to Heap's law by a probability distribution and leading to a generalised linear model (GLM, Nelder and Wedderburn 1972). Specifically, since each individual $V(n)$ is a discrete count, we assume that it is the outcome of a binomial distribution P_{binom}



(a) Korean

(b) Norwegian

(c) Vietnamese

Figure 2.2 Heap's law in (a) Korean, (b) Norwegian, and (c) Vietnamese. The dashed red lines correspond to the predictions of Heap's law with the MLE parameters.

	EO	FI	ID	KO	NO	TR	VI
ϕ	29.3	43.3	35.8	35.9	34.3	62.0	10.7
τ	0.61	0.62	0.56	0.65	0.59	0.56	0.61
(rel. SE ϕ, rel. SE τ)	(0.01, $2{\cdot}10^{-5}$)	($6{\cdot}10^{-3}$, $9{\cdot}10^{-6}$)	($5{\cdot}10^{-3}$, $9{\cdot}10^{-6}$)	($6{\cdot}10^{-3}$, 10^{-5})	($5{\cdot}10^{-3}$, $9{\cdot}10^{-6}$)	(0.01, 10^{-5})	($2{\cdot}10^{-3}$, 10^{-5})
R^2_{McF}	0.99	0.99	1.0	0.99	0.99	1.0	1.0
rel. BIC	1530.0	805.0	3670.0	739.0	1600.0	2380.0	3600.0

Table 2.3 Maximum likelihood estimates of the parameters of Heap's law, ϕ and τ, for our 7 languages together with relative standard error (rel. SE), pseudo-error R^2_{McF} and relative Bayesian information criterion (rel. BIC). See the main text for the null model used in computing R^2_{McF} and BIC.

with mean $\phi * n^\tau$. MLE then operates by finding $\hat{\phi}$ and $\hat{\tau}$ which satisfy:

$$\arg\max_{\phi,\tau} P((C^1, ..., C^n)|\phi, \tau) = \arg\max_{\phi,\tau} P((V(1), ..., V(n))|\phi, \tau)$$

$$= \arg\max_{\phi,\tau} \prod_{i=1}^{n} P_{\text{binom}}(V(i)|\phi, \tau) = \arg\max_{\phi,\tau} \prod_{i=1}^{n} P_{\text{binom}}(V(i); \frac{1}{p} * (\phi * i^\tau), p),$$

where we use $p = 0.5$, the binomial distribution's second parameter, because it leads to maximal variance and therefore to greater numerical stability. Notice that for the second identity we have made use of the fact that under our formulations as outcomes of binomial distributions, the vocabulary sizes $V(1), ..., V(n)$ are independent from each other.

The MLE parameters values are given in Table 2.3 together with their relative standard error, R^2_{McF} and BIC. To calculate R^2_{McF} and the BIC, we again need a null model. We could construct the null model by stripping Heap's law off its parameters, setting $\phi = \tau = 1$ which leads to $V(n) = 1 * n^1 = n$, like we did for the null model for Zipf's law (see Section 2.2.2). But, as it is a linear function, this null model is clearly condemned to grossly overestimate the empirical vocabulary growth at all points. For this reason, we take a different approach to the null model: We simply take the median $m = (V(1), ..., V(n))$ and let this constant be the null model which is then equivalent to Heap's law with parameters $\tau = 0$ and $\phi = m$. This null model is essentially the standard approach in regression modelling and referred to as the intercept.

Partly because the null model is very weak, it is easy for Heap's law with the MLE parameters to be far superior over the null model, leading to exceedingly high R^2_{McF} and the BIC. Even so, as the plots in Figure 2.2 reveal, the prediction by Heap's law (drawn as dashed red lines) indeed fits the empirical vocabulary growth extraordinarily well. In particular, unlike the empirical rank-frequency relationship, the empirical vocabulary growth exhibits almost no variance and is therefore well modelled by a straight line such as that defined by Heap's law. Strong statistical support for Heap's

law as measured by R^2_{McF} and BIC is thus warranted, even if this may be artificially heightened due to the weak null model.

Even comparing the empirical vocabulary growth across languages, it is remarkable how little deviation the graphs show from the simple functional. That is, all languages seem equally well described by Heap's law. At the same time, the vocabulary growths differ significantly in range, visible from both the MLE values of the parameter τ and the y-axes (notices the respective ranges) of the graphs in Figure 2.2 . In particular, Korean displays comparatively very fast vocabulary growth while the same is very slow for Vietnamese. Considering the morphology of both languages, this is not surprising: Korean as an agglutinative language has a high morpheme-per-word ratio which leads to a combinatorially vast space of possible words. The situation is reversed for Vietnamese which, as an isolating language, mainly consists of words of a single morpheme. The chance for a word to recur is thus much higher in Vietnamese than it is in Korean, leading to slower vocabulary growth.

Similarly to the conclusion of the previous section that the languages we have analysed are "Zipfian", the observations in this section lead us to conclude that these languages are also "Heapian". In the sense that their empirical vocabulary growth is approximately but well described by Heap's law. But while there had to be strong emphasis that Zipf's law holds only approximately for language, Heap's law seems to be able to provide almost exact fit.

Chapter Conclusions

In this chapter, we have laid the groundwork for Chapters 3 and 4, in which we present our original empirical work, by introducing and describing the basic quantitative aspects of our data. To these aspects we count the observations about Zipf's and Heap's law which we have described and which our investigations will build upon.

We have motivated and described our corpora, the data we use for our empirical analyses: Extracted and pre-processed text from Wikipedia in seven languages. As the most important features for our work, this is a typologically diverse and large, with $50{\cdot}10^6$ tokens each, set of linguistic data. These corpora are made available for use at `github.com/valevo/Thesis/data`.

While Zipf's law itself, the theoretical model, has been introduced in Chapter 1, we have described in detail in the current chapter how it is connected to empirical observations. For this, we have presented and shown the rank-frequency relationship – empirically equivalent to the theoretical rank-probability relationship – and detailed how it is used to estimate Zipf's law via Maximum Likelihood Estimation (MLE). These concepts have allowed us to establish that our seven languages, or rather their Wikipedias, are indeed Zipfian, i.e. approximately but to a usable extent described by Zipf's law. Establishing this is essential for the remainder of this thesis but in fact also

constitutes a contribution of this thesis. As mentioned, the rank-frequency estimates plotted in Figure 2.1 are arguably the most accurate to date as they are the result of the improved estimation method which is the core of the next chapter. Hence, a re-analysis of the shape of the rank-frequency relationship and the extent of Zipf's law was indeed needed and still is since ours is only preliminary.

3 The Sample-Source Distinction and Subsampling

In the previous chapter, we have familiarised ourselves with the linguistic data used in this thesis, as well as Zipf's law and Heap's law in their empirical incarnations. Specifically, we have described how observed ranks and frequencies (or probabilities, recall that these are equivalent in our context) and their relationship relate to the theoretical Zipf's law. With the concepts and tools of the previous chapter at hand, we are now ready to introduce the first major contribution of this theses.

This contribution concerns the remark made when describing how the empirical ranks and frequencies are extracted from linguistic data (Section 2.2.1). It explains as well the dispersion in the plots of the rank-frequency relationship (Figure 2.1) which may be unusual to readers already familiar with Zipf's law. Finally, and most importantly, the contribution of this chapter justifies why we deemed it necessary to reassess the Zipfianness of language when there is of course already an existing body of work on this very question (e.g. E. G. Altmann and Gerlach 2016, Baayen 2002 and notably (Piantadosi 2014); we selected these examples in fact because they constitute otherwise excellent work on the statistics of Zipf's law).

The reason lies in the realisation made by Piantadosi (Piantadosi 2014, summarised in Section 1.1) that essentially the entire body of previous work on Zipf's law has the flaw of committing an error when extracting the empirical rank-frequency relationship from linguistic data. As a result, so the opinion of Piantadosi and indeed also our own, the reported rank-frequency relationships report in previous work are misleading about the precise shape and degree of regularity of the rank-frequency relationship. Therefore, we agree with Piantadosi that a large part of the body of work on Zipf's law is in need of verification, in particular that concerned with assessing Zipfianness. The re-assessment of the previous chapter, while preliminary, is a first step in this direction.

Our contribution in this chapter is to improve upon and significantly generalise the methodological correction Piantadosi proposed. The shortcoming in the way that Piantadosi approaches the problem in the previous literature seems to be that he sees it as a problem of "data visualization" (Piantadosi 2014) but as we strongly object, the problem in the literature goes far beyond that. As we explain in detail in the next section, the problem is one of statistical practice, in particular estimation, and it seems to be rooted in an obliviousness to the distinction between the theoretical and the observed. This, in contrast to Piantadosi whose solution is rather ad hoc and does

not generally solve the problem, allows us to provide an estimation method which is rooted in established statistical theory. As a result, estimates are arguably the most reliable and detailed the can be in practice and the method is versatile enough to be applied to quantitative linguistic laws other than Zipf's law.

Concretely, in the following section, we re-derive the Subsampling method, an instance of the relatively well-known resampling estimation techniques, in the context of quantitative linguistics. Once introduced, we show and analyse the Subsampling method and its properties when applied to the rank-frequency relationship and extend its use to estimating vocabulary growth.

3.1 Estimating Linguistic Quantities

We begin by taking a step back to reflect on the aim of quantitative linguistics in general because, as we shall see, the problems of the common practice in quantitative linguistics stem from neglect of basic statistical insights. Moreover, going to the basics of statistics will lead automatically through solutions to the methodological corrections we propose and analyse in the current chapter. At least for our purposes, quantitative linguistics seeks to extract mathematical quantities from language with the aim to obtain insights into the systematic structure of language (cf. for instance Köhler and G. Altmann 2005, McEnery and Hardie 2011, Fenk-Oczlon and Fenk 1999). Laws in quantitative linguistics, or quantitative laws, are then formal observations about regularities in such quantities and they are usually special because they persist across languages (cf. (E. G. Altmann and Gerlach 2016)). For instance, Heap's law (see Section 2.3) predicts the single quantity vocabulary size, $V(n)$, of a corpus of n tokens. Meanwhile, Zipf's law (see Sections 1.1 and 2.2) predicts the relationship between the two quantities rank, $r(w)$, and probability, $P(w)$, of a word w. What makes them laws is that they hold for most or all corpus sizes n and words w, respectively (although this part of ongoing debates).

Such an understanding of quantitative linguistics leads us to see that its quantities and laws pertain to languages as a whole. This is an important realisation because a language as such is not actually observable. As is common knowledge in and the essential struggle of quantitative (and more generally computational) linguistics, a language is per se a theoretical concept. The only way to observe it is through small and necessarily incomplete windows, namely samples, also called corpora in computational linguistics. This has a profound implication for linguistic quantities: As functions of the language itself, their values are not directly observable either and can also only be observed through necessarily imprecise samples. Thus, just like the language it describes, a quantity such as the rank of a word $r(w)$, is theoretical.

At the same time, a corpus C which we observe leads to an observed value, $r_C(w)$. Henceforth, we let $r(w)$ denote exclusively the theoretical quantity and use the notation $r_C(w)$ to indicate the observed value in corpus C (and similarly for other quanti-

ties). Now, of course, we can expect $r_C(w)$ to vary across corpora C because C itself will vary due to what we shall treat here as randomness. (Whether or not this is the case is certainly debatable (see for instance Kilgarriff 2005) but can be ignored for our discussion.) In either case, this (random) variation in C and hence $r_C(w)$ induces a distribution over the observed values, $P(r_C(w) = r)$. The importance of this distribution lies in the fact that it gives us a connection between the observed quantity $r_C(w)$ and its theoretical counterpart $r(w)$: Namely the distribution's expected value, $r(w) = E[r_C(w)] = \sum_r P(r_C(w) = r) * r$. According to the law of the unconscious statistician (DeGroot and Schervish 2012), the expected value is equal to $\sum_C P_L(C) * r_C(w)$ where the sum is over corpora C and $P_L(C)$ is the probability of C being generated by the theoretical language L.

This sum, albeit the direct connection between the observed and theoretical, is however infeasible, so $r(w)$ stays inaccessible in principle. On the one hand, the sum over all corpora C is infinite and therefore uncomputable and on the other hand, P_L is unknown since the language L itself is, as just described. But this is a well-known problem in statistics and for instance Monte Carlo methods (MC, originally described in Metropolis and Ulam 1949, extensive modern introduction in Rubinstein and Kroese 2016) have been devised exactly for the purpose of approximating sums and expected values. MC approximates the expected value $E[r_C(w)]$ by randomly sampling a set of corpora $S = \{C_1, ..., C_m\}$ and then calculates the mean $\overline{r}(w) = \frac{1}{m} \sum_{i=1}^{n} r_{C_i}(w)$. Since the relative frequency of each corpus C_i in S approaches $P_L(C_i)$ (according to the law of large numbers, see e.g. Wen 1991), $\overline{r}(w)$ converges to $E[r_C(w)]$ as m grows to infinity. Hence, for large but finite m, the mean approximates the expected value and the former is in fact an estimate of the latter. With modern computing power, MC can usually provide good estimates and has the added advantage that it essentially only requires that random sampling of corpora can be done efficiently.

Here is precisely the problem for our context: At least currently, we have no effective reliable source for randomly sampled corpora and on the contrary, corpora are a notoriously sparse resource. The sparsity of resources renders genuine solutions (to which we do count MC) to estimating the expected value $E[r_C(w)]$ and hence the theoretical quantity $r(w)$ an impossibility. So from this point, any method can only be an approximation to a true solution at best and it is good to keep this in mind for the remainder of this thesis and in general. Having said that, we now describe a method for approximating the MC approximation using just a single (large) corpus (notice the double approximation). This method is called Subsampling (first extensively described in (Politis, Romano, and Wolf 1999)) and is a member of the broader class of resampling methods, together with the more commonly-known Bootstrap and Jackknife methods (see e.g. (Simon and Bruce 1991), (Efron and Tibshirani 1986) and (Efron and Stein 1981)).

The only practical difference of Subsampling from MC is the way samples are obtained but this has the important consequence that while MC approximates the true

underlying distribution, the Subsampling method cannot provide the same guarantee. Hence there is no true guarantee that the estimated quantity will converge to the theoretical one and neither are there guaranteed rates at which convergence might happen. At the same time, the Subsampling method can provide estimates of precision of the estimated quantity, such as estimates of variance and confidence intervals (we will discuss variance estimates further in Section 3.3.1). These can, at least, serve as indicators of how much certainty we are to place in the estimated quantity and how well it reflects its theoretical value.

Compared to MC, the essential idea of the Subsampling method is to let a single corpus fulfil the purpose of the sample generating source – a probability distribution – that is used in MC. Under relatively mild conditions (see the following section), and also due to the law of large numbers (Wen 1991), any individual corpus converges to the probability distribution it was generated from (in terms of the relative frequencies of its elements). Therefore, if the corpus used for the Subsampling method is large enough, then it provides a reasonable approximation of the same source that MC would have used to generate samples. Otherwise, the Subsampling method is works in the same way as MC. Formally (Politis, Romano, and Wolf 1999), given an original corpus C^n, one takes a set of random subsamples $S' = \{C_1^k, ..., C_m^k\}$ with n much larger than k (for simplicity, we fix the subsample size but this is not strictly necessary). Each C_i^k is a proper subset of C^n, that is sampled without replacement. S' then has the same function as S from MC and so the estimate for the theoretical quantity $r(w)$ is formed in the same way: $\overline{r}(w) = \frac{1}{m}\sum_{i=1}^{m} r_{C_i^k}(w)$. The mentioned caveat is in formal terms that the quality of approximation of $\overline{r}(w)$ depends on the initial corpus C^n. Although it is worth mentioning that $n \to \infty$ implies that Subsampling becomes equivalent to sampling directly from P_L and therefore to the MC method. Hence, there are large enough n such that the estimates from the Subsampling method are essentially indistinguishable from those of the MC method.

As it becomes important in the methodology we propose in Chapter 4, we emphasise that the Subsampling method samples each C_i^k uniformly from C^n. That is, each element in C^n (see the next section for a discussion about these elements) has equal probability of being drawn. The validity of the Subsampling method hinges on this in order for the sampling distribution to approximate P_L (i.e. for the probability of drawing a subsample C_i^k to approximate $P_L(C_i^k)$) and for the resulting estimates to be unbiased.

Given the Subsampling method for estimating linguistic quantities, we now estimate the rank-probability relationship, i.e. $r(w)$ and $P(w)$ for each word w, in the following way: We take an initial corpus (in our case Wikipedia) C^n and construct two sets of subsamples, $S_1 = \{C_1^k, ..., C_m^k\}$ and $S_2 = \{C_{m+1}^k, ..., C_{2m}^k\}$. Computing $r_{C_i^k}(w)$ from each C_i^k in S^1 and $P_{C_j^k}(w)$ from each C_j^k in S^2, we obtain the averages $\overline{r}(w)$ and $\overline{P}(w)$. These mean values are our estimates for the rank and probability, respectively, of the word w and these are what we relate for the estimate of the rank-probability relationship. Thus, if we speak of the empirical rank-probability relationship, in this

thesis we refer to the relationship between the mean ranks and the mean probabilities of words, since these are the proper estimates according to the Subsampling method. It is this empirical rank-probability relationship we use to assess the Zipfianness of language, and indeed it is what we have used already throughout Chapter 2; the rank-frequency plots in Figure 2.1 shows actually the mean ranks versus the mean frequencies of words. In Chapter 2, and also in this and the next chapter, we use $k = 1{\cdot}10^6$ and $m = 10$ for the parameters of the Subsampling method. We keep them constant across chapters in order to ensure comparability of the results and because these parameters have proven to work well.

Finally, we address the difference between the method Piantadosi proposes in (Piantadosi 2014) and the Subsampling method, which is most easily understood in terms of m, the number of subsamples taken to calculate the average. Piantadosi's method is essentially (not entirely) equivalent to the Subsampling method with $m = 1$, that is just a single subsample is taken for estimating $r(w)$ and a single one for estimating $P(w)$. Of course, then the averages $\overline{r}(w)$ and $\overline{P}(w)$ that are computed for the estimate of the Subsampling method are only over a single value and of course equivalent to that value. An estimate from just a single sample is known as a point estimate (Lehmann and Casella 2006) and it should be clear that its value is heavily influenced by the randomness in drawing the particular subsample (cf. (Bloem et al. 2016)). Because of this randomness, the point is likely to be unreliable, depending on how large the parameters n and k are. In contrast, the Subsampling method, which generally uses $m > 1$, actually approximates the mean of the distribution over values of $r_C(w)$ and $P_C(w)$ and the estimates it produces are therefore less prone to being influenced by randomness.

In order to see how the difference between the point estimate and the estimate by the Subsampling method manifests itself in the estimated rank-frequency relationship, compare Figure 1.2b (point estimate) to Figure 2.1a (Subsampling estimate). The point estimate leads to much higher dispersion in the relationship across words, especially in the relationship's tail, than the Subsampling estimates. This is not surprising because the mentioned randomness inherent in the subsample leads to randomness, that is dispersion, in the estimate. In the Subsampling method, that randomness is marginalised out by taking multiple subsamples, i.e. multiple point estimates, and subsequently their mean value. This is important because the randomness is not actually part of the theoretical value of the quantity we are trying to estimate. Piantadosi claims that his proposed method make the precise structure in the rank-probability relationship and its deviation from Zipf's law interpretable but this is strictly speaking not true. Because his estimates are merely point estimates, any specific shape of the resulting rank-probability relationship could have been caused by the randomness in the subsamples.

As discussed above, even the Subsampling method cannot completely remove the dependence on the observed data (the original corpus C^n) either and with only a single source corpus at hand, generally no method can. However, and herein lies the full

advantage of the Subsampling method over point estimation methods, the set of values from which the average $\overline{r}(w)$ (or $\overline{P}(w)$) is computed approximates and converges to the distribution over values itself (in terms of relative frequencies). Thus, in addition to the estimated value, we can also examine the certainty attached to that value, for instance by examining variance of the distribution and convergence behaviour of estimates. These can reveal the quality of the used data (the corpus C^n) as a representative of the underlying theoretical language. In order to show this potential and to analyse the rank-frequency relationship itself, we devote Sections 3.3.1 and 3.3.2 respectively to analysing the variance and convergence behaviour of the rank-frequency relationship. We begin our empirical investigation, however, by addressing yet another methodological issue and evaluating a proposed remedy in the next section.

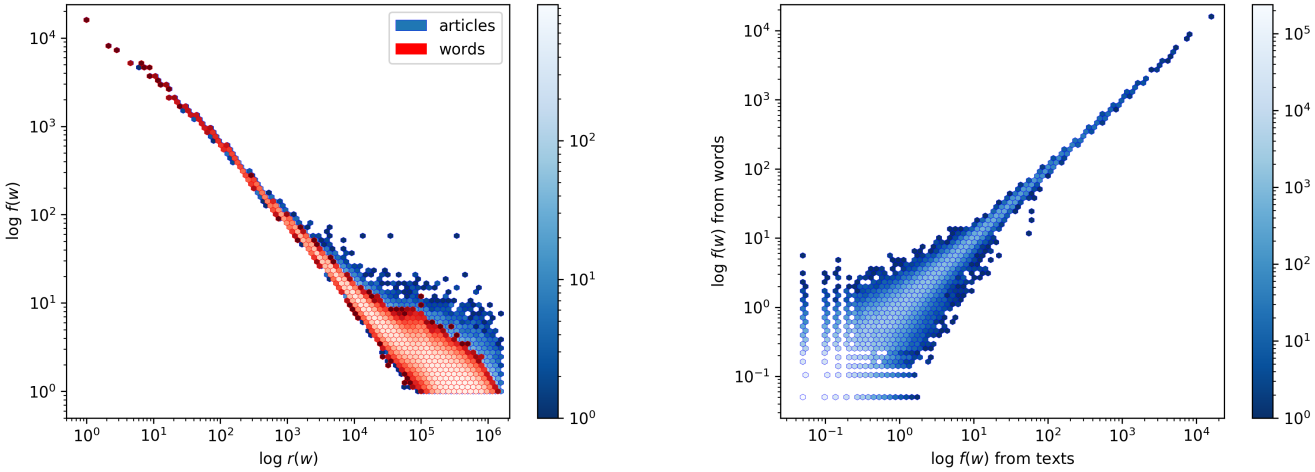
A note on the data used in this chapter: To provide comparability of the plots and results, we investigate the Korean Wikipedia in all of the sections of this chapter. While cross-linguistic comparisons of the results of the Subsampling method applied to the rank-frequency relationship would certainly be instructive about its nature, it is beyond the scope of this thesis. Using seven different (and typologically diverse, see Section 2.1) languages was done mainly done for development purposes, namely to ensure that the Subsampling method produces valid results in all tested languages. The plots and results we report on Korean below can be found for the remaining set of six languages at `github.com/valevo/Thesis/figures`. Indeed, as these show, the results and arguments we provide below are qualitatively the same in all the six languages we used next to Korean. There are, of course, cross-linguistic differences in the details but they are subtle enough for the discussion below to hold across languages.

3.2 Elements Used for Subsampling

Before we can analyse the Subsampling method itself, we need to address one more problem, which pertains to the structure of language. This problem also affects the method proposed in (Piantadosi 2014) but is not addressed there. Although the problem has no general solution, in this section we propose a remedy and analyse how much of the problem it alleviates.

While re-sampling methods, including the Subsampling method, work under quite broad conditions, they break down if there is sequential dependency in the data. This is clearly the case for language, as the words and sentences in a corpus C^n are not independent from each other but exhibit long-range and complex patterns of sequential dependence. It is easy to see why this makes naive subsampling invalid: If we sample subcopora C^k by randomly drawing individual words from C^n, C^k will be syntactically invalid gibberish and not represent the language which generated the original corpus C^n.

In particular, when using Subsampling for estimating the rank-frequency relationship, we expect a strong effect on the low-frequency types: These are characterised by



(a) Text vs Word Rank-Frequency Relationships

(b) Text-Word Correlation of Frequencies

Figure 3.1 (a) Estimates of the mean rank-frequency relationship, between $\overline{r}(w)$ and $\overline{f}(w)$, of Korean. The red bins represent the estimates obtained from sampling on the word level and are superimposed on the estimates from sampling on the text level (i.e. articles in Wikipedia) in blue. (b) The per-word correlation of the estimates of the mean frequency $\overline{f}(w)$ from text-level sampling (x-location) and the same estimates from word-level sampling (y-location). If the estimates were the same for every word, this plot would show as a single diagonal line.

highly clustered and non-uniform occurrence across a given corpus (see (A. Corral et al. 2009) for an investigation into the recurrence statistics of words). Consider for example the word 'Turing' which generally does not occur often but once it does, will have high recurrence in the immediate context, due to its high topicality. Performing Subsampling by sampling individual words destroys such phenomena and we expect it to severely underestimate the variance in the rank-frequency relationship of low-frequency types.

A possible remedy lies in the hierarchical structure of language itself, namely that words are organised into sentences which are in turn organised into texts inside a corpus (in Wikipedia, texts are called articles). As we move up this hierarchy, the sequential dependency of language becomes weaker, since words have the strongest influence on each other's occurrence probabilities in immediate context. In the extreme, it is relatively safe to assume that words in two separate corpora do not influence each other's occurrence probability at all. Hence, rather than sampling subcorpora on the word level, we can sample larger elements, such as sentences or entire texts.

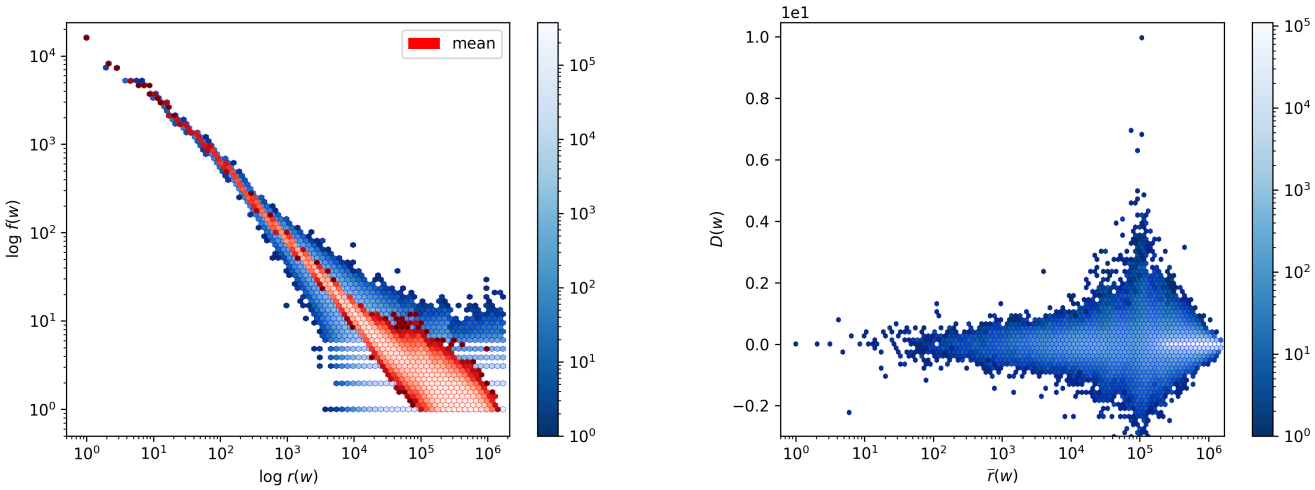
In order to analyse the effects of sequential dependence in language on estimation by Subsampling, we compare the outcomes of random sampling at three different levels in the hierarchy: words, sentences and texts. Concretely, for each of these levels, we randomly sample subcorpora by sampling at that level and then estimate the rank-

frequency relationship from these subsamples as described above. Thus, we obtain three estimates of both $\overline{r}(w)$ and $\overline{P}(w)$, one for each level.

Two of these estimates, namely from the word and from the text level, are plotted in Figure 3.1a, where the word-level estimates are simply superimposed on the text-level ones. The sentence-level estimates are omitted since they would obfuscate the plot and show an intermediate effect between the word- and text-level estimates in any case. As the plot shows, and as we expected, word-level subsamples underestimate the variance in the low-frequency types, as the resulting tail is thinner than that of the estimates from text-level subsamples. Notice that the tail from the text-level subsamples additionally exhibits what could be called outliers, points with very high deviation from the centre of mass. In contrast to the tail, the estimates are virtually the same in the head of the graphs. The reason is likely that high-frequency types are characterised by dense and uniform occurrence patterns. Sampling at the word level evidently reproduces these patterns for the high-frequency words and the estimates from word and text level therefore converge.

Regardless of the fact that sampling at the word level underestimates variance, the two estimates of the rank-frequency relationship apparently still have the same mode across sampling levels, since both histograms of Figure 3.1a have their mass concentrated along approximately the same line. To emphasise this, and to provide a more detailed view, Figure 3.1b shows the word-level estimate of $\overline{f}(w)$ plotted against the text-level estimate of $\overline{f}(w)$. Indeed, this graph is centred around a straight diagonal line, indicating high correlation, which is only weakened by the high variance in the region of low frequencies (lower left corner). Even within this region the mode of values is heavily concentrated on this diagonal as the shading of the bins indicates. Moreover, so the points which do not lie on the diagonal are roughly symmetrically distributed around the diagonal. This plot adds further evidence that sampling on the word-level does not introduce significant bias into the rank-frequency estimates, as compared to text-level sampling, and that the only difference really is underestimated variance.

This observation has an important implication for Zipf's law: As described in Section 2.2.1, Zipf's law predicts a single line, i.e. it cannot predict any of the variance of the rank-frequency relationship. The MLE therefore inherently tries to find the parameters which best fit the mode of the relationship and ignore the variance. This implies that if two relationships have the same mode and only differ in variance, as is the case with the word-level and text-level estimates, MLE will return the same parameters for Zipf's law. In order to confirm this explicitly, we fitted Zipf's law to the rank-frequency estimates obtain from the three sampling levels. Indeed, the MLE parameter values are virtually same, only differing beyond the third decimal place. We even find essentially the same goodness of fit attached to these parameters, as measured by R^2_{McF} and the relative BIC. This is surprising since higher variance leads to reduced significance of the relationship's mode and in turn to lower significance of



(a) Mean vs Full Rank-Frequency Relationship

(b) Fano Factor plot

Figure 3.2 (a) Mean rank-frequency relationship of Korean in red, superimposed on the full set of values from which that mean was obtained in blue. (b) For each word, its mean rank $\overline{r}(w)$ plotted against its Fano factor, $D(w) = \text{cov}_C(r_C(w), f_C(w))/\overline{r}(w)$.

Zipf's law. The text-level estimates would have therefore been expected to lend lower support to Zipf's law than the word-level estimates. The fact that they do not emphasises that the mode of the rank-frequency relationship occupies the large majority of mass.

In sum, even though Subsampling is theoretically invalid given the sequential dependency of language, the results of Subsampling at the different levels of the hierarchy of language show that it only underestimates variance but does not introduce bias into the estimates. Instead, the modes of the estimated relationships coincide to a degree where the results of MLE are indistinguishable. For practical reasons, sampling at the level of texts may not always be possible, as is the case for the method we develop in the next chapter. Therefore and given these observations we conclude that sampling at lower levels, such as the level of sentences, can be a reliable proxy for text-level sampling. Especially if the variance in the tail of the rank-frequency relationship is not of interest. For comparability with the next chapter, all estimates of the rank-frequency relationship in the remainder of this thesis are obtained from sentence-level sampling.

3.3 Estimating Ranks and Frequencies

3.3.1 Variance

Previously – specifically in the previous section and Section 2.2.1 – when speaking of variance in the rank-frequency relationship, we have been referring to the variance of $\overline{r}(w)$ and $\overline{f}(w)$ with respect to each other across words w. As a variance between two random variables, this variance is actually a covariance which we write $\text{cov}_w(\overline{r}(w), \overline{f}(w))$ and use the subscript to emphasise that it is across words w. The covariance is the generalisation of the variance to two dimensions and a measure of linear dependence, specifically it is the un-normalised version of the Pearson correlation coefficient.

Since Zipf's law predicts a log-linear dependence, i.e. a single straight line, for the relationship between $r(w)$ and $f(w)$ across w, the validity of the law depends on $\text{cov}_w(r(w), f(w))$. Briefly put, deviation from a perfect correlation in the logarithmic rank-frequency relationship is deviation from Zipf's law. Since $\overline{r}(w)$ and $\overline{f}(w)$ are our best estimates for and converge to $r(w)$ and $f(w)$, we empirically judge the validity of Zipf's law by $\text{cov}_w(\overline{r}(w), \overline{f}(w))$.

The plots in Figure 2.1 indicate that the covariance is indeed rather high, as dispersion is low and confined to the tail. Notice also that our m and k, the number and sizes of the subsamples, are rather low at respectively 10 and 10^6. The covariance likely further increases with higher values of these parameters of the Subsampling method; we will inspect the behaviour across values of k in the following section. It is worth noting that the covariance in the estimates from Subsampling is especially low in comparison to the point estimates of Piantadosi's method (see Figure 2.1). Piantadosi specifically emphasised the high degree of dispersion and used it as evidence against the validity of Zipf's law. But as we see now, with the more representative estimates for the theoretical rank-frequency relationship, the relationship does show rather high conformity to a single line, that is high covariance, and therefore to Zipf's law. For more definitive conclusion, it should be evaluated in future work whether the rank-frequency relationship assumes perfect correlation in the limits of the Subsampling parameters m and k and equivalently whether $\text{cov}_w(\overline{r}(w), \overline{f}(w))$ can invalidate Zipf's law.

At the same time, there is a second type of variance which, as we will see, entails that the high covariance in the observed mean rank-frequency relationship needs to be taken with a grain of salt. The variance in question is the variance attached to this mean, namely the variance in point estimates across corpora. Again, as we are relating two random variables, this variance is also in fact a covariance, namely $\text{cov}_C(r_C(w), f_C(w))$. In contrast to the covariance across words, the subscript C now indicates that the covariance is across corpora C. Notice that when taking the mean during the Subsamling estimation procedure, it is this covariance across corpora that

we are marginalising over.

Similar to the covariance across words, the covariance across corpora has implications for the validity of Zipf's law, but in a different way. As described in Section 3.1, the mean is taken as the representative of the distribution and as the estimate of the rank-frequency relationship because it converges to the expected value which is in turn equivalent to the theoretical value. But as for any distribution in general, the meaningfulness of the mean as a representative is tied to the variance of that distribution. Simply put, the greater the variance, the less representative the mean is for the distribution, since large variance implies that values other than the mean also have high probability mass. In the extreme case, namely the uniform distribution, the mean has as much probability mass as all other values and hence the mean is not particularly representative of the distribution. So even though we may above have observed high conformity of the mean rank-frequency relationship with Zipf's law, large variance would mean that we assign low significance to that conformity. This is what we assess in the remainder of the current section by analysing the covariance $\text{cov}_C(r_C(w), f_C(w))$.

First, we plot the mean rank-frequency relationship together with the distribution it is computed from in Figure 3.2a. Specifically, we plot $(\overline{r}(w), \overline{f}(w))$ in red; this is the same graph as in Figure 2.1a (except for different values of k). The blue graph which lies beneath it is the full set of ranks $\{r_{C_1^k}(w), ..., r_{C_m^k}(m)\}$ obtained from m subsamples plotted against the full set of frequencies $\{f_{C_{m+1}^k}(w), ..., r_{C_{2m}^k}(m)\}$ obtained from another m subsamples. This blue graph represents a sample from the full distribution over rank-frequency relationship across subcorpora C^k and can equivalently be seen as the union of m point estimates of the rank-frequency relationship. Thus, Figure 3.2a can be seen as the two-dimensional version of the common one-dimensional histogram with the mean; notice that in the two-dimensional case, the frequency mass is indicated by the shading of the bins (see the colour bar). Notice also that the graph of the full distribution has horizontal gaps in the low-frequency region, whereas the graph of the mean does not. This is simply because only natural-numbered values are possible for actual ranks and frequencies in a given corpus, while their means can be real-valued.

Figure 3.2a shows that the high covariance across words in the mean rank-frequency relationship we have observed above is somewhat deceptive: There is substantial dispersion in full distribution which starts from its upper tail and increases substantially in its lower tail. This indicates that for low-frequency words, the precise value of the rank-frequency relationship is essentially due to chance across corpora and decreases the meaningfulness of the mean as a representative of the distribution. Positive variance, although small, can be found even in the high-frequency words which indicates that even there some uncertainty about the precise location of the rank-frequency relationship exists. At first glance and judging from Figure 3.2a, it thus seems that there is high dispersion in the rank-frequency relationship of words across corpora. This is

equivalent to low covariance $\text{cov}_C(r_C(w), f_C(w))$ and, as described, would imply low significance for the precise mean rank-frequency relationship, that is our estimates returned by Subsampling. We now test this conclusion from two further perspectives.

As mentioned, the sample of the full distribution in Figure 3.2a (blue graph) can also been seen as consisting of m point estimates of the rank-frequency relationship. We would like to assess the variability in these individual point estimates but doing so is complicated by the fact that each such point estimate is over the entire vocabulary, i.e. over all words w. Here, Zipf's law can be of help rather than just the object of analysis: As a statistical model, it provides a low-dimensional description of a point estimate, namely in terms of the parameters α and β. Hence, we estimate the parameters separately from each of the m point estimate, obtaining m estimates of Zipf's law. The variability in the estimates of Zipf's law is an efficient, albeit only approximate, indication of the variability in the point estimates and specifically of the variability of the point estimates' modes. Despite the high dispersion we observed above, the MLE parameters show essentially no variation across point estimates, with differences only in the second decimal place (which is also why do not report the MLE results). The same is true for the goodness-of-fit measures, as both R^2_{McF} and relative BIC are virtually the same across all point estimates. This relativises the dispersion we have found above as it implies that the m point estimates of the rank-frequency relationship have similar modes. It also implies, as indicated by the goodness-of-fit measures, that the highly dispersed points in Figure 3.2a carry only very little mass.

Moreover, we would like a more detailed view of $\text{cov}_C(r_C(w), f_C(w))$ in individual words, rather than in the entire vocabulary. In order to achieve this, we treat the words as a sequence of data, akin to a time series, and order them according to $\overline{r}(w)$. Then, we compute the Fano factor (Cox and Lewis 1966) for each word: $D(w) = cov_C(r_C(w), f_C(w))/\overline{r}(w)$, i.e. the covariance divided by the mean. For each time step in a given time series, the Fano factor measures the signal (the mean) to noise (the covariance) ratio of that time step. That is, it indicates whether that time step provided a reliable signal, which is the case if $D(w) < 1$, or not, i.e. $D(w) \geq 1$. For our purpose, the use of the Fano factor stems from the realisation that for words with high $\overline{r}(w)$ the rank-frequency relationship is inherently highly dispersed. These words only have few observations which leads to high uncertainty about the precise value of the rank-frequency relationship. For this reason, in order to investigate the $\text{cov}_C(r_C(w), f_C(w))$ in individual words, we normalise it by the mean $\overline{r}(w)$, as prescribed by the Fano factor.

Figure 3.2b plots for each word its mean rank against its Fano factor, i.e. $\overline{r}(w)$ against $D(w)$. Note that the plots contains negative values because covariance can be negative, the only difference between positive and negative covariance values is the direction of correlation. First, notice that the graph is roughly symmetric along the constant at 0, reflecting that the distribution of values around the mean in Figure 3.2a is also relatively symmetrically distributed. The most important insight from this graph is that by far the most words have a Fano factor of less than or equal to one.

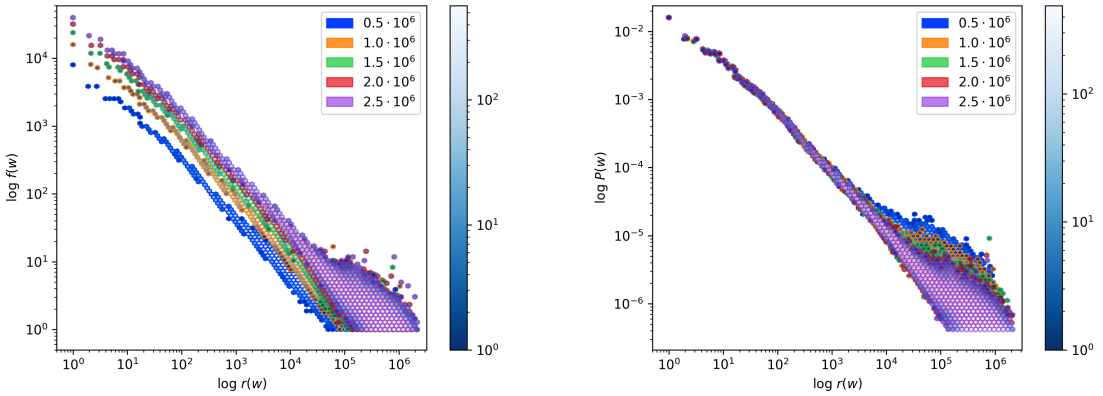
This is a striking observation as it implies that for most words, even the low-frequency ones, the signal provided by the mean rank $\overline{r}(w)$ exceeds the noise that is the covariance across corpora. Hence, even if the low-frequency words exhibit high uncertainty, when accounting for how much is inherent due to low numbers of observation, one finds relatively high reliability. Notice an opposite effect in the middle range of the plot, roughly between mean ranks 10^2 and $5{\cdot}10^4$: Compared to their relatively low dispersion in Figure 3.2a, the Fano factors for these words are relatively high. This reveals that compared to their low inherent uncertainty due to higher numbers of observations, they do not provide very reliable signals. Taken together, Figure 3.2b contributes the important insight that the variance across corpora of low-frequency words is not as high as it seems and in particular that this variance in not excessively high compared to that of more frequent words.

The use of Zipf's law and the Fano factor shed a somewhat different light on the covariance $\text{cov}_C(r_C(w), f_C(w))$ than one would be inclined to conclude from only looking at Figure 3.2a: Neither is the covariance particularly excessive, as the Fano factors revealed, nor is there significant variation in the modes of the point estimates across corpora, as indicated by the coinciding MLEs of Zipf's law. Even though there certainly is a significant amount of dispersion in the full distribution over rank-frequency relationships and this needs to be remembered when analysing the mean relationship, dispersion is not as high as may have been expected. We therefore conclude that the mean be seen as a reasonably good representative of the full distribution since it seems to carry the majority of the distribution's mass. Similarly, it may be concluded that Zipf's law, as estimated by the mean, has reasonable significance in the context of the full distribution.

The covariance we have described and analysed in the current section is an important aspect of the Subsampling method: It allows to gauge how dependent the precise values of observations, i.e. estimates, are on the specific data. In light of the randomness that is inherent to any sample, this variance is also to be seen as an inherent, necessary component of any estimated quantity and not to be neglected. But whereas it may seem as a nuisance to deal with at first glance, the covariance across data can also provide arguments to strengthen phenomena in empirical observations. In this vein, we have observed in this section that the rank-frequency relationship across corpora, while displaying some dispersion, is in fact remarkably stable and this strengthens the significance of Zipf's law as an observation about the relationship.

3.3.2 Convergence

In this section we analyse the convergence behaviour the mean rank-frequency relationship across the subsample sizes k and we do so for two reasons. First, we continue the previous sections by adding further insight into how stable and reliable the estimates yielded by the Subsampling method are. And second, we use it as an example



(a) Rank-Frequency Relationships

(b) Rank-Probability Relationships

Figure 3.3 (a) Mean rank-frequency relationships of Korean as estimated on subcorpus sizes $k = 0.5{\cdot}10^6$ through $k = 2.5{\cdot}10^6$. (b) Same plot as (a) but plotting the relationship between mean rank and mean probability rather than mean frequency.

of how properties of the estimated quantities become interpretable and thus meaningfully analysable when estimated properly. In the current section, the property of interest of the rank-frequency relationship is its convergence behaviour, i.e. whether the relationship continues to change in shape until the limit of corpus size or whether such change stagnates beyond a certain size.

As mentioned previously and indeed an important aspect, the Subsampling method alleviates the dependence of the estimates on the specific corpus. The reason is that taking the mean over the rank-frequency relationships of repeatedly sampled corpora effectively marginalises the individual random characteristics of these corpora. In consequence, the precise shape of the estimated rank-frequency relationship is interpretable in the sense of being comparable across corpora and only depends on properties such as corpus size. This is in emphasised contrast to the erroneous, commonly used method of constructing the rank-frequency relationship from a corpus. This method leads to a strong dependence of the relationship's shape on the specific corpus and importantly implies that relationships obtained from different corpora are incomparable.

For this reason, we re-evaluate and in fact challenge previous reports about the changes in the rank-frequency relationship across corpus sizes. We suspect that most of the observed phenomena are artefacts of the misleading estimates and, using proper estimates, resolve the contrary findings that have been made. Using two different books, (Powers 1998) took a number of increasing length prefixes (up to $2{\cdot}10^4$ tokens) of these books. The rank-frequency relationships constructed from each of the prefixes were found to increase in steepness with prefix size. Using the same method of prefixes and similar sizes, (Baayen 2002) replicated this finding in terms of the parameter

α of Zipf's law which was reported to increase with prefix size. Based on this and other observations, (Baayen 2002) in fact questioned the usefulness of the rank-frequency relationship and Zipf's law because it would render comparison across different-sized corpora impossible. Again using the same prefix method but more corpora and larger sizes, (Font-Clos, Boleda, and A. Corral 2013) took a closer look at the behaviour of the rank-frequency relationship across sizes and could not find significant changes. Instead, they reported the relationship to stay quite stable and attributed the previous findings to insufficient corpus sizes. Contrary to this finding, (Moreno-Sánchez, Font-Clos, and Á. Corral 2016) did find changes in the parameter α but this time that α was in negative correlation with corpus size, indicating flatter, not steeper, rank-frequency relationships. (Moreno-Sánchez, Font-Clos, and Á. Corral 2016) is the only study to date that uses a large collection of different-sized texts. Together, the previous studies provide a clearly inconsistent and therefore inconclusive picture of the converge behaviour of the rank-frequency relationship and we suspect that this is caused by the erroneous method of estimating it.

In order to re-evaluate the relationship's convergence behaviour, we inspect both the differences in the mean rank-frequency relationships which emerge across subsample sizes k and the differences in the MLEs of Zipf's law obtained from these mean relationships. We begin with Figure 3.3, in which we have plotted the mean rank-frequency relationships obtained from subcopora of five different sizes, ranging from $0.5{\cdot}10^6$ and $2.5{\cdot}10^6$ tokens. We specifically choose a range of subcorpus sizes around $1{\cdot}10^6$ because the subcorpora in Chapter 4 all have this size. For the same reason, we have used the same size in Sections 3.2 and 3.3.1.

Note that the only difference between Figure 3.3a and 3.3b is the scale of the y-axis. In the latter, frequencies were transformed into probabilities, simply by dividing by the total number of tokens. The relationships in Figure 3.3a move to the upper right corner with increasing corpus sizes for the simple reason that greater corpus sizes lead to higher frequencies. Greater corpus size also leads to higher numbers of observed word types (cf. Heap's law, Section 2.3) and this causes the tails of the rank-probability relationships in Figure 3.3b to increase in length as size increases. Simultaneously, the tails also stretch further down on the y-axis as corpus size increases, since greater vocabulary size implies lower probability mass for the low-frequency types.

Besides these trends, which are inherent to increasing sample sizes, however, the relationships across all sizes exhibit strikingly high similarity. Disregarding minor individual fluctuations, the relationships are centred on virtually the same line and seem to really only grow in the number of types they are defined on. Not even covariance of the relationships across words seems to be substantially higher in smaller samples, as the relationships' tails do not vary greatly in dispersion. Judging from Figure 3.3 it thus seems that the rank-frequency (and equivalently the rank-probability) relationship has converged beyond corpus sizes of $0.5{\cdot}10^6$ and that no major differences in the relationship will arise from further increasing the subcorpus size. Also compare Figure 3.3a to Figure 2.1a, where the rank-frequency relationship is estimated

	$0.5{\cdot}10^6$	$1{\cdot}10^6$	$1.5{\cdot}10^6$	$2{\cdot}10^6$	$2.5{\cdot}10^6$	$10{\cdot}10^6$
α	1.13	1.13	1.13	1.12	1.12	1.13
β	4.57	5.23	5.64	5.93	6.17	9.36
R^2_{McF}	0.63	0.63	0.63	0.63	0.63	0.65
rel. BIC	2.76	2.76	2.76	2.76	2.76	2.81

Table 3.1 Results of performing MLE of Zipf's law, i.e. its parameters τ and ϕ, on the mean rank-frequency relationships obtained from subcorpus sizes $0.5{\cdot}10^6$ through $2.5{\cdot}10^6$. See Section 2.2.2 for the interpretation of R^2_{McF} and relative BIC.

from $10{\cdot}10^6$ tokens, i.e. five times the size of the largest subcorpus used in Figure 3.3a. Even at $10{\cdot}10^6$ tokens, the rank-frequency relationship has essentially the same shape, which strengthens the conclusion that the relationship has converged.

In order to verify this more formally, we turn to MLE of Zipf's law which we perform on the mean rank-frequency relationship for each subcorpus size; the results can be found in Table 3.1. The results of MLE clearly indicate that the mean rank-frequency relationship does not change across subcorpus sizes we use, since the parameter α stays the same across all sizes. The slight decrease in α in corpus sizes $2{\cdot}10^6$ and $2.5{\cdot}10^6$ from 1.13 to 1.12 seems accidental, since the value of α obtained from $10{\cdot}10^6$ tokens is again 1.13. The same holds for the goodness-of-fit measures R^2_{McF} and relative BIC which stay the same across subcorpus sizes and are only slightly higher at $10{\cdot}10^6$ tokens. This indicates that the mean rank-frequency relationship has the same of conformity to Zipf's law across subcorpus sizes. Glaringly, however, β does not converge across corpus sizes but exhibits steady increase. Looking at the rank-frequency relationships themselves in Figure 3.3 it is not clear why β, which controls how strongly the head of the predicted relationship curves off, should increase, since the head of the relationship itself does not increase. This phenomenon will likely require deeper investigation and so we leave it unexplained for now. On the whole, the MLEs of Zipf's law across subcorpus sizes reinforce the finding that there is a high degree of convergence in the rank-frequency relationship at the corpus sizes we use here.

In summary, using much larger corpus sizes than the previous studies and, in contrast to them, proper estimates obtained with the Subsampling method, we find rather strong convergence of the rank-frequency relationship and a remarkable degree of stability of the relationship itself and the resulting MLE of Zipf's law across corpus sizes. In particular, since the rank-frequency relationship converges, so does the corresponding Zipf's law and this indicates that Zipf's law in the underlying theoretical language is well-defined. We emphasise this because it has been questioned in some of the cited previous works (such as Baayen 2002). We also emphasise that our observations re-

produce those of (Font-Clos, Boleda, and A. Corral 2013) and, because of the fact that ours are the most reliable and largely corpus-independent estimates of the rank-frequency relationship, there is high likelihood that the convergence behaviour we report here is the closest to the true one. In the spirit of the methodological concerns we are raising in this chapter, by the true convergence behaviour we mean the way in which the rank-frequency relationship changes across corpus sizes independent from any specific corpus. As explained in the introduction, this is done by averaging over the changes in the rank-frequency relationship across all individual corpora. Here, we have approximated this average by virtue of the Subsampling method.

Besides further emphasising the importance of accurate and reliable estimates and thus providing a further argument for using the Subsampling method, we have investigated the convergence of the rank-frequency relationship because other observations and methods of this thesis depend on it. On the one hand, the strong degree convergence already among corpus sizes of $1{\cdot}10^6$ tokens ensures that our discussion of the precise shape of the rank-frequency relationship in Section 2.2.2 is representative irrespective of the size of the number of tokens used there (namely $50{\cdot}10^6$). The same argument holds true for our discussion of the variance in the previous section. On the other hand, rather high degrees of convergence, as the ones we have observed here, are in fact necessary for the methodology we develop and evaluate in the next chapter. Both for the theoretical discussion of that methodology, which becomes meaningless in the absence of convergence, and for its evaluation, since without convergence the results become difficult to interpret (see in particular Sections 4.3.1 and 4.4.1).

3.4 Estimating Vocabulary Growth

The methodology for estimating quantities from linguistic data which we developed in the current chapter is not confined to the rank-frequency relationship but generalises to all quantities of interest in quantitative linguistics. This is because, as argued in this chapter's first section, any quantity that pertains to the underlying language, rather than the observable samples from it, is subject to the concerns of correct estimation. As far as we are aware, the problem (Piantadosi 2014) originally noticed in the estimation of the rank-frequency relationship is pervasive in the entire field of quantitative linguistics. In particular, to best of our knowledge, the problem also exists in the estimation of vocabulary size in language which is described by Heap's law (see Section 2.3). In order to show how the entire field, not just research around Zipf's law, can benefit from the methodological improvements we are proposing with this chapter, we close the chapter by showing and briefly analysing the Subsampling method applied to estimating vocabulary size.

Recall that for a given number of tokens of a corpus n, Heap's law predicts the number of types, or vocabulary size, $V(n)$ in that corpus. Notice that, unlike Zipf's law which predicts the relationship between two quantities, Heap's law is about the single

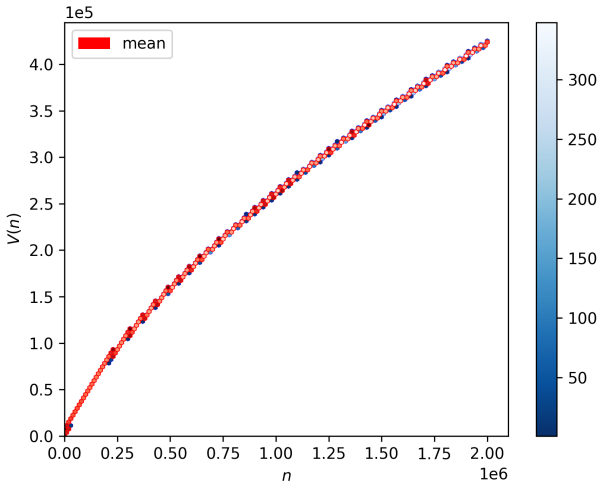


Figure 3.4 Mean vocabulary growth of Korean in red, plotted on top of the full set of values (in blue) from which that mean was computed. Notice that the blue graph is almost entirely hidden by the red, indicating that the mean graph and the graph of all values almost completely coincide.

quantity $V(n)$. At the same time, the actual interest lies in the vocabulary growth, i.e. the sequence of vocabulary sizes $(V(1), ..., V(n))$. The common way (as done for instance by Petersen et al. 2012 and Gerlach and E. G. Altmann 2014) to construct this sequence empirically is to use a corpus C^n and take from it the sequence of all of its prefixes $(C^1, ..., C^{n-1}, C^n)$. Then, one counts the vocabulary size of each such prefix C^i, obtaining $V_{C^i}(i)$ and hence the sequence $(V_{C^1}(1), ..., V_{C^n}(n))$ (notice the use of indices to the function V as described in Section 3.1). Finally, this sequence is used as the estimate of $V(1), ..., V(n)$.

With the discussion of Section 3.1 in mind, it should be clear that this method leads to erroneous and misleading estimates: First, each $V_{C^i}(i)$ is merely a point estimate of $V(i)$ due to the dependence on C^i. As previously, one should use a set of corpora $\{C_1^i, C_m^i\}$, compute $\{V_{C_1^i}(i), ..., V_{C_m^i}(i)\}$ and take the mean $\overline{V}(i)$ of this set. Again, as the influence of the specific corpus is marginalised out by taking the mean, the resulting estimate is a more reliable representative of the theoretical quantity $V(i)$. Second, and yet more problematically, the commonly used method leads to correlation between the estimates of the different vocabulary sizes. Specifically, it forces them to be monotone increasing and the reason is simply that for each i, C^i, used to estimate $V(i)$, is a proper prefix of C^{i+1}, used to estimate $V(i+1)$. The vocabulary size of the prefix C^i must be smaller or equal to that of C^{i+1} and therefore the estimate of $V(i)$ necessarily smaller or equal than the estimate of $V(i+1)$.

Clearly, independent corpora C^i and C^{i+1} should be used to ensure that the estimates of $V(i)$ and $V(i+1)$ are independent. But, as previously, independent corpora are generally not available (in this case, we would require n corpora), and so we again turn to the Subsampling method. That is, we substitute independent corpora with in-

dependently sampled subcorpora of an original single corpus. More formally, given a corpus C^n and $k < n$, we take random subsamples of sizes $1, ..., k$ to construct the sequence of subcorpora $(C^1, ..., C^k)$. Notice that the subcorpora in this sequence, although not independent from C^n, are indeed independent from each other. Given this sequence, we compute the sequence $(V_{C^1}(1), ..., V_{C^k}(k))$ as the point estimate for $(V(1), ..., V(k))$. Of course, the concern from the previous paragraph applies, so we repeat this procedure m times and take the means for the final Subsampling estimates of vocabulary growth.

In fact (and as mentioned there), in Figure 2.2 the estimates of vocabulary growth were already obtained by Subsampling, so $V(n)$ in this plot actually refers to $\overline{V}(n)$. In order to estimate the complete vocabulary growth, we would need to take m subsamples for each number of tokens $i \leq n$, i.e $n * m$ subsamples in total. Even though sampling a subcorpus itself is not too expensive, this number of samples grows too fast. Therefore, we do not estimate $V(i)$ for every $i \leq n$ but instead, leave a constant distance between the i for which we estimate $V(i)$. In Figure 2.2, this distance is $2{\cdot}10^3$. This still provides enough data points for estimating Heap's law but does entail that we miss some of the potential variance of vocabulary growth. Finally, in applying the Subsampling method to the estimation vocabulary growth, the concern of Section 3.2 also holds here, namely that subsampling at the word level is invalid given the sequential structure of language. Therefore, as before, we use sentences as the elements for subsampling rather than individual words.

Because Heap's law is not our main focus, we do not reproduce the full analysis we provided for Zipf's law. In Figure 3.4 we do, however, show the variance of $V_C(n)$ across subcorpora C akin to Figure 3.2a of Section 3.3.1. That is, we plot the mean vocabulary growth $\overline{V}(n)$ together with the set of m values $\{V_{C_1^i}(i), ..., V_{C_m^i}(i)\}$ it was computed from. The resulting graph shows that the variance, or dispersion, from the distribution over values is remarkably low and one might even say that there is essentially no variance. Not even is the variance higher for lower number of tokens, i.e. in the left half of the plot, as one might have expected. The absence of variance therefore does not seem to be related to the fact that our numbers of tokens (up to $2{\cdot}10^6$) are quite large, where a high degree of convergence and therefore low dispersion is expected. It may, however, be related to the gaps we have to leave in estimating vocabulary sizes, mentioned in the previous paragraph (the distance of $2{\cdot}10^3$). Future work should test this but for now, since do not observe any significant variance, there is no detailed analysis of that variance to be conducted.

Similarly to Section 3.3.1, where we have argued that the relatively low variance (i.e. high stability across point estimates) of the rank-frequency relationship, we also conclude here that Heap's law as a description of vocabulary growth is strengthened by the apparent high stability of vocabulary growth across point estimates. This is especially remarkable given that the point estimates of vocabulary size are obtained from independently subsampled corpora; even at large corpus sizes, one might expect individual corpora to exhibit variation in their vocabulary sizes. Thus, and adding to

the observations of 2.3 that the estimated mean vocabulary growth closely follows the predictions of Heap's law, we reaffirm that Heap's law seems to be a rather unambiguously valid description of the vocabulary growth in natural language.

Verifying the validity of quantitative linguistic laws such as Heap's law with proper estimates by means of the Subsampling method is arguably useful in itself and will moreover come in handy in Section 4.4.2, where the strong Heapianness of language will serve to control for the effects of the methods we develop in that chapter (Chapter 4). The actual point of this section has, however, been to exemplify that the concerns of this chapter and the Subsampling method as the proposed way to address them apply not only to ranks and frequencies as linguistic quantities. Instead, they generalise to all of quantitative linguistics, such as vocabulary growth as shown in this section. Not only the discussion on Zipf's law can hence benefit from taking estimation seriously but also that on Heap's and all the other laws in quantitative linguistics.

Furthermore, the example of estimating vocabulary growth shows that the Subsampling method, in addition to being straightforward to apply, is versatile enough to be easily adapted to estimate generally any linguistic quantity of interest. First and foremost, we see this as a reflection of the fact that it is a general and theoretically grounded estimation method, as opposed ad hoc methods such as the one proposed by (Piantadosi 2014) to estimate the rank-frequency relationship. Second, the Subsampling method, and generally the entire family of resampling methods it belongs to, has been devised precisely with the aim of being applicable to large classes of estimation problems without requiring much adaptation (other members of the family of resampling methods are thus aptly called "bootstrapping" and "jackknifing").

Chapter Conclusions

This brings us back to the beginning of the current chapter. We have started with the realisation of (Piantadosi 2014), whose role as the original motivation of this chapter we emphasise, that the rank-frequency relationship is commonly estimated in an erroneous way in the context of Zipf's law. But, as we have argued in Section 3.1 and demonstrated in the previous section, this problem is not specific to the rank-frequency relationship and instead pervades the field of quantitative linguistics. For this reason, and precisely by noticing that the actual issue is deeper than (Piantadosi 2014) asserts, we have presented the Subsampling method – a general and established estimation method – for use in quantitative linguistics. The Subsampling method is able to provide reliable and detailed estimates from just a single corpus. In order to show this aspect and hopefully convince other practitioners of using of the Subsampling method, we have analysed in terms of the variance and convergence behaviour of the rank-frequency relationship in Sections 3.3.1 and 3.3.2. While these are merely preliminary, we hope that they have laid groundwork for similar and more extensive future analyses of linguistic quantities such as the rank-frequency relationship.

In this chapter, we have advocated the use of the Subsampling method for proper and reliable estimation of theoretical quantities from observed corpora. The reasons have done so lie in its simplicity, adaptability and most importantly the scarcity of linguistic resources. However, as mentioned, the Subsampling method is itself only an approximation to true solutions to estimation, such as Monte Carlo methods. Such methods should be reconsidered in future work, because the Subsampling methods has its drawbacks.

First, its ability to provide reliable estimates does hinge on a relative large source corpus, since the size of subsamples should be much lower. However, in some fields of empirical linguistics, such as clinical linguistics, even obtaining corpora of $1{\cdot}10^3$ tokens can be difficult (cf. for instance (Egmond 2018) who studied Zipf's law corpora of aphasic speech of 350 tokens). Although estimation of quantities of the theoretical language is generally difficult with such small numbers of observations, extensions or adaptations to the Subsampling method might help accommodate such cases. Here, for instance Bayesian extensions (such as A. F. Smith and Gelfand 1992 could improve reliability of estimates even at small corpus size by informing them with prior probabilities on estimated values obtained from larger samples.

In part precisely because of its considerable size, and also its availability in many languages, we have used Wikipedia as the corpus for this thesis. Wikipedia, however, is in fact a good example for the second caveat of the Subsampling method. Linguistically, that is in terms of topic and style, Wikipedia is a very homogeneous corpus since it consists entirely of scientific and encyclopedic texts. Recall that the Subsampling method treats the source corpus, i.e. Wikipedia in our case, as an approximation to the underlying language (in the sense of a distribution over corpora). However, it should be clear that Wikipedia is not a good representation of language as a whole which is linguistically far more diverse. Therefore, future work should repeat the analyses we have conducted in this chapter on more diverse corpora. Potential examples are WebCorp (Renouf, Kehoe, and Banerjee 2007) and (Hart 1992) which are similarly large but more diverse.

Altogether, regardless of the advantages and limitations of the Subsampling method itself, the central point of this chapter is not to suggest that the Subsampling method is the only or the superior estimation method. Rather, our aim has been to show that with relatively simple and ready to use techniques, it is possible to overcome erroneous, correlated point estimates in favour of reliable estimates of the full underlying distribution; the Subsampling method is an example of such a technique. Thus, yet more important is the strict distinction between sample and source from which we have started our discussion. Especially in the debate on Zipf's law, where not even its status as a law is clear, proper statistical methodology is in our opinion vital and the sample-source distinction and the issue of estimation are essential basic ingredients thereof.

4 The Filtering Method

In the previous Chapter, we described the Subsampling method for use in quantitative linguistics. It provides a convenient and, given the scarcity of linguistic resources, likely the best way to properly estimate theoretical linguistic quantities from samples of the languages they pertain to. The general strategy of the Subsampling method is to use a single large corpus and to repeatedly sample random subcorpora, or subsamples, from within it. The source corpus thus approximately simulates the stochastic source, that is, the underlying theoretical language from which the subsamples are generated.

As mentioned in Section 3.1, for this strategy to be statistically valid, it is important that subsamples are drawn with uniform probabilities. Specifically, this means that all elements of the source corpus have equal probability of being drawn into a subsample. It ensures that the probability of an element to occur in a subsample is entirely governed by its frequency in the source corpus and therefore approximately its probability in the underlying language. Only in this case is it possible for subsamples to preserve the statistical properties of the source corpus and to converge to the underlying language in the limit of their size. That is, only in this case do subsamples have a high likelihood of being credible and reliable representatives of their source corpus.

Otherwise, that is in the case of non-uniform biased sampling, the statistical properties of subsamples generally become altered with respect to those of the source corpus. As a result, they will with high likelihood not be representative of the source corpus and not converge to the original underlying language, but to a different one which will in general possess different distributional properties. Thus and according to the theory of Subsampling, in the same way that uniform subsamples are representatives the original source corpus and its underlying language, biased subsamples represent a different, hypothetical source corpus with a different underlying language.

In this chapter, we realise the chance that lies in biased subsampling and adopt the the Subsampling method to obtain samples from varying hypothetical sources in which properties of human languages are altered. The sampling biases we use are derived from information theory and we construct two biased sampling algorithms as instantiating examples. The subsamples we obtain from these algorithms enable a comparative approach to studying the effects of the properties that are altered by biased subsampling.

Concretely, we use Zipf's law again as the case study, so the sampling algorithms

we devise are aimed at weakening the Zipfianness of the sampled subcorpora. These will enable a comparative approach to studying learnability, our other topic of interest, and specifically the effect of Zipf's law on the learnability of language. We begin this chapter by reviewing how comparative approaches have been achieved in the previous work on the learnability of Zipf's law and explain why these approaches are problematic. As we shall see, removing the problems of these approaches lead naturally to the information-theoretic approach we take.

4.1 Non-Zipfian Languages

In studying the effects of any quantitative property of language on any aspect of cognitive processing, a convenient strategy is to take a comparative approach. That is, we determine empirically how processing changes when the quantitative property in question is changed. This is in fact a common strategy in cognitive science and has also been employed by the learnability studies (Kurumada, Meylan, and Frank 2013 and Hendrickson and Perfors 2019) summarised in the introductory chapter of this thesis (see Section 1.1).

Concretely, to gauge the effect of Zipf's law on learnability, and as both these studies have done, the approach is to measure learnability in both Zipfian and non-Zipfian languages. If the Zipfian languages turn out to be more learnable, then Zipf's law is concluded to have a positive effect on learnability, and vice versa. Such a comparative approach requires at least two languages, a Zipfian one and a non-Zipfian one, and it should be clear that the choice of both is a key ingredient. As is commonly observed, and as we have re-confirmed in the previous chapters of this thesis, human languages are arguably Zipfian. So human languages are an obvious and the most relevant choice for a Zipfian language in the comparative approach. But herein lies also the problem, namely that all known human-like languages are Zipfian, even when restricted to specific domains (see Piantadosi 2014). Searching for or explicitly constructing a non-Zipfian language is therefore not straightforward and we therefore devote the current section to explore the set of non-Zipfian languages from a theoretical perspective.

One might expect this set to be immensely vast but is not quite as unrestricted as one might think. Recall that Zipf's law describes the relationship between the rank and probability of the words in the vocabulary of a given language. Recall further that a word's rank is actually defined in terms of its probability, i.e. the most probable word has rank 1, the second most probable word has rank 2, and so on. This implies that the relationship between rank and probability is by definition monotonically decreasing (although not strictly). Thus, all languages, including the non-Zipfian ones, trivially have a decreasing rank-probability relationship over their vocabulary.

Being restricted to the set of decreasing functions, the possible rank-probability relationships hence range from uniform (a flat non-decreasing relationship) to one

where the word at rank 1 has probability 1 and all other ranks have probability 0 (the steepest relationship). Zipfian languages lie relatively at the centre between these two extremes, namely where the word at rank 1 has roughly twice as much probability mass as rank 2, roughly three times as much mass as rank 3, and so on. Any other type of rank-probability relationship, both more gradual and steeper, can be termed non-Zipfian. However, usually, and both by (Kurumada, Meylan, and Frank 2013) and by (Hendrickson and Perfors 2019), only more gradual relationships are considered for non-Zipfian languages. In fact, both studies only use non-Zipfian languages with uniform rank-probability relationships, the extreme of the spectrum of relationships more gradual that Zipfian.

Having established the set of possible non-Zipfian rank-probability relationships, it is still underspecified what the set of corresponding languages is. The vocabulary is merely one aspect of a language and languages with vastly different grammatical structure, for instance, can have the same rank-probability relationship. And for any given shape of the relationship, most of the corresponding languages are in fact not relevant for comparison with human languages since their structures and complexities are not comparable to that of human languages. The precise structure and complexity of human languages, however, is unknown and it is therefore not directly possible to construct or even identify non-Zipfian languages that are apt for comparative approaches.

Presumably, this is a reason why both (Kurumada, Meylan, and Frank 2013) and (Hendrickson and Perfors 2019) choose to sidestep the structural complexity of human language and construct their own, artificial languages, both Zipfian and non-Zipfian. This allows them to keep all properties besides the rank-probability relationship equal in the Zipfian and non-Zipfian languages and guarantees comparability. The problem with such artificial languages for our specific context is, however, that they are toy languages with very small vocabularies. Specifically, the vocabularies of the languages in both studies consist of respectively maximally 36 and 28 types. Even though the effects reported by both studies are likely to become even stronger with larger numbers of types, such small vocabularies are clearly far from realistic for natural language. In learning tasks other than the highly specialised ones investigated by these two studies, and especially in holistic language learning, small vocabularies are likely to lead to misleading results.

To be more concrete, our objection mainly stems from the excessively long tail in the rank-probability relationship of natural language, which is in fact also one of a core property of Zipf's law. And this long tail, which consists of the vast amount of (low-probability) words in language, can of course only properly show when vocabulary sizes are also vast. Equivalently, since the long tail is a core property of Zipf's law, it is even questionable to call languages with small vocabularies Zipfian. Simultaneously and importantly, it has been found by (Blevins, Milin, and Ramscar 2017) that the tail's excessive length in natural language has immediate consequences on learnability. As they argue and empirically illustrate, a long tail in the probability distribution over the

vocabulary entails that most word types in any corpus will have an exceedingly low frequency in that corpus. For a learner, this creates a problem of sparsity because most observed word types will provide few or even just a single examples of themselves and their usage. As (Blevins, Milin, and Ramscar 2017) further show, this sparsity problem is not simply countered by increasing corpus size but on the contrary becomes even worse, since the tail is ever-growing. Yet even worse, the authors verify empirically that the ever-growing tail implies that human language should be assumed to have an unbounded vocabulary (cf. Section 1.1). This implies equivalently that the morphological productivity of language has no bounds in the sense that new words can always be created via the language's morphological system. As an immediate consequence, the tail in the rank-probability relationship of the language itself is infinitely long and the learner is not only faced with sparsity issues but even with the impossibility to exhaust the language's vocabulary in its entirety. It is this situation which leads us to the conviction that comparative learnability assessments generally need to respect the vastness and sparsity in the vocabularies natural language. Especially for studying general language learning, the non-Zipfian alternatives hence need to have unbounded vocabularies and therefore unbounded morphology, just like their human and Zipfian counterparts.

The assumption that the vocabulary is, at least in the theoretical language, unbounded leads to a further constraint on the set of possible rank-probability relationships and further restricts the set of monotonically decreasing functions. Notably, this constraint was neglected by (Hendrickson and Perfors 2019) in their discussion of languages with unbounded vocabularies. The constraint stems from the axiom that the probability distributions over the vocabulary must sum to one, $1 = \sum_{w \in W} P(w)$. Clearly, if the vocabulary is unbounded, then this sum has an infinite number of terms. But such an infinite sum does not converge unless $P(w)$ decreases faster than $\frac{1}{r(w)}$ in relationship to $r(w)$ (viz. the harmonic series, which does not converge). Hence, for instance the uniform distribution, which is the alternative discussed by (Hendrickson and Perfors 2019), is not a valid probability distribution if the vocabulary is unbounded. In fact, given this constraint it turns out that Zipf's law (with α close to 1) is the most uniform possible distribution over an infinite vocabulary, i.e. it maximises the entropy over the vocabulary (see the following section). This is a fact which is often overlooked and on the contrary, Zipf's law is seen as a particularly steep distribution for languages' vocabularies (see e.g. (Kurumada, Meylan, and Frank 2013) and (Hendrickson and Perfors 2019)). Thus, in the presence of an unbounded vocabulary, the set of non-Zipfian languages is confined to those with a rank-probability relationship steeper than Zipf's law. As opposed to what was done in previous studies, non-Zipfian languages with relationships less steep than Zipf are an impossibility, at least from a theoretical perspective.

On the other end, in the set of rank-probability relationships steeper than Zipf's law, the vast majority is too steep: Most of these relationships correspond to distributions which assign (effectively) positive probability to only a finite set of words and by our assumption of an unbounded vocabulary these are excluded. A similar and even

stronger conclusion was reached in a mathematical approach by (Corominas-Murtra and Solé 2010) who argue that it may indeed be the case that Zipf's law (with α close to 1) is the only theoretical possibility for the rank-frequency relationship of languages with an unbounded vocabulary. Roughly speaking, they prove that all other, non-Zipfian rank-probability relationships fall in one of two classes: If the relationship is less steep, it leads to an unusable language because the Kolmogorov complexity of its vocabulary becomes infinite. Otherwise, if the relationship is steeper, the vocabulary is effectively bounded and this is again by assumption excluded.

To summarise, this high-level exploration of the set of rank-probability relationship reveals that the theoretical and mathematical perspective leaves rather little room for languages to be non-Zipfian. Instead, given that human languages have unbounded vocabularies, the set of relationships is constrained to monotonically decreasing functions with slope greater but not much greater than $\frac{1}{r(w)}$. Only then can the corresponding probability distribution over the vocabulary be valid and effectively support a potentially infinite set of words. Given the proof of (Corominas-Murtra and Solé 2010), it is moreover possible that the rank-probability relationship is forced to be around $\frac{1}{r(w)}$ and that truly non-Zipfian human-like languages are impossible. Indeed, we have lead the discussion in such detail because this is a problem for the comparative approach and none of the previous learnability studies seem to have realised it. That is, if there are no truly non-Zipfian languages, at least in the theoretical perspective, then the comparative approach to assessing the learnability of Zipfian languages is invalid.

4.2 Information-Theoretic Typicality

The conclusion of the previous section leaves us in a seemingly dire situation: We would like to both take a comparative approach to the learnability of Zipfian languages and respect the vastness and sparsity that is typical to the vocabularies of human languages. But at the same time, we have little to nothing to compare human languages with because the set of non-Zipfian human-like languages is difficult to impossible to enumerate and moreover heavily restricted to the extent that it may even be non-existent.

But all is not lost because two more points need clarification which have also been neglected in the previous work of (Kurumada, Meylan, and Frank 2013) and (Hendrickson and Perfors 2019). First, notice that we spoke of languages being Zipfian or non-Zipfian in the previous section. But of course, and this is closely connected to the topic of Chapter 3 (cf. specifically Section 3.1), a language is merely a theoretical concept and not itself observable; in particular not by a learner who will only ever receive a finite sample from the language. So while it makes sense to speak of the learnability of languages in a theoretical sense, we (and generally any learnability study) are really assessing how learnable finite samples from these languages are. Both because

our assessments can only be carried out on such finite samples and because they are in fact what is relevant because they are what learners observe. Unlike the theoretical languages, of which we have assumed that they are unbounded, samples could very well be non-Zipfian, even uniform. The constraints discussed in the previous section do not hold for them precisely because they are finite. Second, for both languages and the samples from them, being Zipfian or non-Zipfian is not a binary quality. As we have presented in the previous section, the set of rank-probability relationships is the entire set of monotonically decreasing functions over the vocabulary (with further restrictions in the unbounded case). Even though, as discussed, this set is heavily restricted, it is still uncountably infinite. As for an uncountable set, one may hence speak of a distance between rank-probability relationships in the sense that one relationship may be more Zipfian than another. Equivalently, one can then speak of the degree of Zipfianness of a relationship. We make this explicit here because the previous learnability studies have indeed implicitly assumed that a language or sample can only either be Zipfian or non-Zipfian. As we will see, and with the mathematical tools we describe in the following, treating Zipfianness as a matter of degree instead, we can overcome the problem that truly non-Zipfian languages are theoretically questionable alternatives to human languages in the comparative approach.

Amazingly, and in fact the reason for the discussion above, the field of information theory has a concept which formalises exactly this: the degree to which a sample is similar to a given distribution, that is in our case, the Zipfianness of a sample. This concept is known as typicality, itself derived from the typical set, a basic but not often used concept of information theory. In this section, we describe typicality and how it can be applied to searching samples with reduced Zipfianness from human-like languages. As a prerequisite we begin by defining entropy, the core concept of information theory, and deriving the entropy of Zipf's law.

Entropy

The information-theoretic entropy (see Cover and Thomas 2012, the standard reference of information theory) of a distribution P over a vocabulary W, denoted as $H(P)$, is defined as

$$H(P) = \sum_{w \in W} P(w) \log_2 \frac{1}{P(w)} = - \sum_{w \in W} P(w) \log_2 P(w).$$

In words, $H(P)$ is the expected value of the length in bits that it takes to encode the surprisal about encountering a word $w \in W$, where the surprisal is measured by $\frac{1}{P(w)}$. As is intuitive, the surprisal is high if w has low probability according to P and vice versa. Therefore, $H(P) = 0$ if and only if there is a single $w \in W$ which has $P(w) = 1$, as in this case there is no surprisal and P is called deterministic. Conversely, $H(P)$ takes on its maximal value of $\log_2 |W|$ if and only if P is the uniform distribution over W. Hence, $0 \leq H(P) \leq \log_2 |W|$.

For us, of course, the distribution over the vocabulary of interest is Zipf's law. Recall from the introduction (Section 1.1) that its definition is

$$P_{\alpha,\beta}(w) = \frac{(r(w)+\beta)^{-\alpha}}{\zeta(\alpha,\beta)},$$

with α and β the law's parameters and $\zeta(\alpha,\beta) = \sum_{i=1}^{\infty}(i+\beta)^{-\alpha}$ the Hurwitz zeta function. The derivation of the entropy of Zipf's law on an unbounded vocabulary W has, to the best of our knowledge, not been published before. Hence, we derive it here ourselves:

$$\begin{aligned}
H(P_{\alpha,\beta}) &= -\sum_{w\in W} P_{\alpha,\beta}(w)\log_2 P_{\alpha,\beta}(w) \\
&= -\sum_{w\in W} \frac{(r(w)+\beta)^{-\alpha}}{\zeta(\alpha,\beta)} \log_2\Big(\frac{(r(w)+\beta)^{-\alpha}}{\zeta(\alpha,\beta)}\Big) \\
&= -\sum_{w\in W} \left[\frac{(r(w)+\beta)^{-\alpha}}{\zeta(\alpha,\beta)}\log_2((r(w)+\beta)^{-\alpha}) - \frac{(r(w)+\beta)^{-\alpha}}{\zeta(\alpha,\beta)}\log_2(\zeta(\alpha,\beta))\right] \\
&= -\Big(\sum_{w\in W}\frac{-\alpha\log_2(r(w)+\beta)}{(r(w)+\beta)^{\alpha}\zeta(\alpha,\beta)}\Big) + \sum_{w\in W}\frac{(r(w)+\beta)^{-\alpha}}{\zeta(\alpha,\beta)}\log_2(\zeta(\alpha,\beta)) \\
&= \frac{\alpha}{\zeta(\alpha,\beta)}\Big(\sum_{w\in W}\frac{\log_2(r(w)+\beta)}{(r(w)+\beta)^{\alpha}}\Big) + \log_2(\zeta(\alpha,\beta))\frac{\sum_{w\in W}(r(w)+\beta)^{-\alpha}}{\zeta(\alpha,\beta)} \\
&= \frac{\alpha}{\zeta(\alpha,\beta)}\Big(\sum_{w\in W}\frac{\log_2(r(w)+\beta)}{(r(w)+\beta)^{\alpha}}\Big) + \log_2(\zeta(\alpha,\beta)).
\end{aligned}$$

In the computational experiments we describe later in this chapter, we need to actually compute $H(P_{\alpha,\beta})$ for given values of α and β. For this, the term $\sum_{w\in W}\frac{\log_2(r(w)+\beta)}{(r(w)+\beta)^{\alpha}}$ in the final line of the derivation is problematic since it has an infinite number of terms and can therefore not be explicitly computed. Fortunately, an equivalent closed-form expression exists, namely $-\frac{\partial}{\partial\alpha}\zeta(\alpha,\beta))$, the negative partial derivative of the Hurwitz zeta function. Therefore, taking the last line from above:

$$\begin{aligned}
H(P_{\alpha,\beta}) &= \frac{\alpha}{\zeta(\alpha,\beta)}\Big(\sum_{w\in W}\frac{\log_2(r(w)+\beta)}{(r(w)+\beta)^{\alpha}}\Big) + \log_2(\zeta(\alpha,\beta)) \\
&= \frac{-\alpha\cdot(\frac{\partial}{\partial\alpha}\zeta(\alpha,\beta))}{\zeta(\alpha,\beta)} + \log(\zeta(\alpha,\beta)).
\end{aligned}$$

This final form of the entropy of Zipf's law has no direct intuitive interpretation other than that it increases as a function of α and decreases as a function of β. The

important aspect of deriving this form lies in that it allows for efficient computation of $H(P_{\alpha,\beta})$, since it contains no sums and both the Hurwitz zeta function and its partial derivative can be efficiently approximated with modern techniques.

Typicality

Having defined the entropy $H(P)$ and derived an explicit closed-form expression for $H(P_{\alpha,\beta})$, the entropy of Zipf's law, we can now define typicality (see again Cover and Thomas 2012): Given a source distribution P, the typicality function, a, maps a sample S to a real number; this number indicates – in an intuitive sense – how typical it is to obtain S when sampling from P. We hence speak of the typicality of S with respect to P, measured by the typicality function. In our case, a sample is of course a corpus, $C^n = (w_1, ..., w_n)$, i.e. a sequence of n tokens. The formal definition of the typicality of C^n with respect to some source distribution P is

$$a(C^n; P) = H(P) - \frac{1}{n} \log_2 \frac{1}{P(C^n)} = H(P) + \frac{1}{n} \log_2 P(C^n).$$

The shape of $a(C^n; P)$ over all C^n is similar to that of a quantile function (the inverse sigmoid function), so most of the values of a are around 0. Its extremes are $-\infty$ and $H(P)$ since $\frac{1}{n} \log_2 \frac{1}{P(C^n)}$ takes on values between 0 and ∞. Specifically, $a(C^n; P) = -\infty$ if and only if $P(C^n) = 0$, that is if C^n is impossible according to P. On the other hand, $a(C^n; P)$ attains is maximal value of $H(P)$ if and only if $P(C^n) = 1$. Notice, however, that for practically all sources P and corpora C^n, $P(C^n) < 1$, since all but the degenerate probability distributions assign positive probability to more than one element in their domain. A notable exception is the empty corpus, C^0 which has $P(C^0) = 1$ for all sources P and is therefore typical with respect to all P.

Another, and in fact the most interesting case is $a(C^n; P) = 0$ which implies that $H(P) = -\frac{1}{n} \log_2 P(C^n)$. In this case $0 < P(C^n) < 1$ is "just right" and we say that C^n is typical for P. Typical because in this case, P is the most likely among all probability distributions to have generated C^n by random sampling (although that P need not be unique in being the most likely). This aspects becomes clearer by turning to our specific source of interest, Zipf's law i.e. $P_{\alpha,\beta}$, and looking at a special property of it. Namely, the fact that the words in a corpus are independent from each other under Zipf's law. This is formally stated as $P_{\alpha,\beta}(C^n) = \prod_{i=1}^{n} P_{\alpha,\beta}(w_i)$ (cf. Section 2.2.2 where we have also made use of this property). The independence property allows us to rewrite the definition the typicality function a as follows:

$$a(C^n; P_{\alpha,\beta}) = H(P_{\alpha,\beta}) + \frac{1}{n}\log_2 P_{\alpha,\beta}(C^n) \qquad = H(P_{\alpha,\beta}) + \frac{1}{n}\log_2 \prod_{i=1}^{n} P_{\alpha,\beta}(w_i)$$

$$= H(P_{\alpha,\beta}) + \frac{1}{n}\sum_{i=1}^{n}\log_2 P_{\alpha,\beta}(w_i) \qquad = H(P_{\alpha,\beta}) - \frac{1}{n}\sum_{i=1}^{n}\log_2 \frac{1}{P_{\alpha,\beta}(w_i)}.$$

Notice that in the final identity, the second term is in fact an average: The average negative log-probability of the words w_i in the corpus C^n under $P_{\alpha,\beta}$. Notice further that the negative log-probability of the words in the vocabulary is actually also used in the definition of the entropy, $H(P_{\alpha,\beta}$ except only that there, the expectation is taken rather than the average. The fact that this is the only difference between the two terms of the final identity makes it particularly clear why the typicality function a captures typicality in the intuitive sense: The average negative log-probability of the word types in C^n is close to $H(P_{\alpha,\beta})$ (and $a(C^n; P_{\alpha,\beta})$ close to 0) if the types' normalised frequencies are on average close to the respective probabilities assigned by $P_{\alpha,\beta}$. That is, at least on average, the empirical distribution over the vocabulary that arises from C^n (by taking normalised frequencies) mirrors the theoretical distribution $P_{\alpha,\beta}$ and in this case, it is intuitive to speak of C^n to be typical with respect to $P_{\alpha,\beta}$.

A core theorem of statistics and already made use of in Section 3.1, the difference between an average and the corresponding theoretical expectation vanishes in the limit of the size of the sample (known as the weak law of large numbers, see e.g. Gnedenko 2018). Hence, the average negative log-probability converges to the expected negative log-probability and the latter is exactly the definition of entropy. This implies that the typicality function a converges to 0 and importantly means that in the limit of sample size any sample is typical. Convergence here happens because the law of large numbers asserts that the normalised frequencies of the elements in a sample converge to their theoretical probabilities as the sample grows to infinity in size (know as Borel's strong law of large numbers, see e.g. Wen 1991). For this type of convergence to hold, the independence property we have used to rewrite the typicality function a (and which Zipf's law possesses) is crucial. Independence is however not necessary for typicality to converge to 0, and indeed convergence theorems also exist for classes of non-independent sources (see e.g. Algoet and Cover 1988). In general, the fact that typicality converges to 0, regardless of the precise mode of convergence, is known as the Asymptotic Equipartition Property (AEP, (Cover and Thomas 2012)).

Importantly, however, the typicality $a(C^n; P)$ only converges to 0 if P is the true underlying source of C^n, that is only if C^n is really sampled from P. Otherwise, if C^n is sampled from a different source Q, then $a(C^n; P)$ converges in absolute value to $D_{KL}(P \mid\mid Q)$ (Cover and Thomas 2012). D_{KL} is the Kullback-Leibler divergence between P and Q and its interpretation is that of a measure of distance between probability distributions. Thus, as sample size grows, typicality reveals whether a considered source is the true source of that sample or not by converging to non-zero values.

Stronger yet, since it converges to the Kullback-Leibler divergence, typicality even reveals how different from the true source the considered source is.

If it is not in the limit of their size, however, and even if the considered source is the true one, virtually no samples will have typicality exactly 0. This is because of random fluctuations which are almost guaranteed in finite samples and imply that we will in practice never encounter any samples which we would call typical, even if we know their true source. For this reason, we relax which samples we call typical by introducing a tolerance parameter ε. This leads to definition of the typical set, historically the original concept and in fact already an important ingredient of the original theorems which founded information theory (namely Shannon's source-coding theorem, Shannon 2001). The typical set $A_\varepsilon^n(P)$ of a source P is simply the set of all samples which have typicality close to zero, or formally:

$$\mathcal{A}_\varepsilon^n(P) = \{C^n : |a(C^n; P)| \leq \varepsilon\}.$$

Clearly, by increasing ε, more samples C^n become part of the typical set $\mathcal{A}_\varepsilon^n$ but how much ε needs to be increased for a proportional increase of $\mathcal{A}_\varepsilon^n$ depends on the source P. Hence, it is not an easy task to find a ε which reflects in an intuitive or useful way the set of typical samples for a given source P. In the following section, where we need to establish a usable ε for $P_{\alpha,\beta}$, i.e. Zip's law, we will do so by a heuristic which empirically turns out to be workable. We emphasise here, however, that the dependence on a subjectively chosen ε implies that the typical set does not provide a truly objective notion typicality. For this reason, the typical set cannot, for instance, fully replace statistical tests of whether a given distribution is an appropriate source for an observed data set. In the following sections, we will therefore complement the typical set with the statistical tests we have already used in the previous chapters, namely the R^2_{McF} and the BIC (see Section 2.2.2).

Applying Typicality

So how do typicality and the typical set, i.e. the function $a(C^n; P_{\alpha,\beta})$ and the set $\mathcal{A}_\varepsilon^n(P_{\alpha,\beta})$, help overcome the problems we have identified for the comparative approach to the learnability of Zipf's law? Recall that the eventual goal is to determine whether the presence of Zipf's law in corpora of natural language leads to increased or to reduced learnability in comparison to corpora that are non-Zipfian. In the discussion of the previous section, we have established two issues of the previous learnability studies that we are convinced need to be respected: First, the compared Zipfian and non-Zipfian corpora need to both preserve the properties of natural language, specifically the immensely large and sparse vocabularies typical of natural language, for the effects of Zipf's law to properly show. And second, that for a language and its corresponding corpora to be considered non-Zipfian they need not, and in fact cannot, be at an extreme of the distributions over the vocabulary (such as the uniform dis-

tribution as used in the previous studies). Information-theoretic typicality addresses both of these issues by providing an objective and nuanced notion of Zipfianness of a corpus. Importantly, it allows to assess the Zipfianness of any corpus, however large and complex, and with typicality we are not restricted to using corpora of which we by construction know that they are non-Zipfian.

Specifically, and regardless of the issue with determining ε, $\mathcal{A}_\varepsilon^n(P_{\alpha,\beta})$ provides a theoretically well-founded and practical formalisation of whether a given corpus C^n is to be considered Zipfian. Given a value of ε, we simply call C^n Zipfian if C^n is a member of $\mathcal{A}_\varepsilon^n(P_{\alpha,\beta})$ and non-Zipfian otherwise. Importantly, and to be explicit, a corpus may be non-Zipfian according to $\mathcal{A}_\varepsilon^n(P_{\alpha,\beta})$ even if it superficially still resembles Zipf's law. This is in fact the decided advantage of the information-theoretic formalisation over other statistical tests (cf. Section 4.4.1).

Actually, to gauge the effect of Zipfianness on learnability of corpora we need not even use $\mathcal{A}_\varepsilon^n(P_{\alpha,\beta})$ to separate corpora into Zipfian and non-Zipfian ones. Instead, we can directly establish the correlation between the Zipfianness of a corpus C^n, as measured by $a(C^n; P_{\alpha,\beta})$, and the learnability of C^n. Finding that learnability is higher with in corpora of higher Zipfianness (i.e. typicality closer to 0), we would conclude a positive correlation between Zipfianness and learnability and hence that Zipf's law aids learnability. Using the typicality function a directly instead of the typical set $\mathcal{A}_\varepsilon^n$ removes the dependence on ε and moreover yields a more fine-grained perspective. As we see it, using typicality as the correlate with learnability yields the most reliable and detailed representation of the effect of Zipf's law on the learnability of corpora. It is also the most versatile approach, since again, typicality and hence Zipfianness can be measured on any corpus.

4.3 Implementations

At least from a theoretical perspective, actually establishing the correlation between Zipfianness and learnability of corpora is simple: for any size n and any possible corpus C^n of natural language, measure both $a(C^n; P_{\alpha,\beta})$, its typicality, and its learnability (we ignore for now that measuring learnability may not at all be straightforward and discuss this issue in the conclusion of the current chapter, Section 4.4.3). But of course, practically enumerating all corpora of natural language is an impossible task, both because this set has infinite size and because we do not know the underlying language, so we cannot definitely decide whether a given corpus is the outcome of natural language. Even randomly sampling corpora to form Monte Carlo estimates of the correlation is impractical for the same reason that we have no way to (randomly) construct entire corpora.

Now we have arrived at essentially the same situation as the one we described in the first section of the previous chapter (Section 3.1): We require multiple corpora to construct, or rather estimate, the distribution over values of quantitative measures

from these corpora – previously, these measure were ranks and frequencies, now they are typicality and learnability. But, as before, corpora are a scarce resource, so we resorted in the previous chapter to the Subsampling method, which was indeed also the core of that chapter. Thus, since the situation is the same here as it was before, we can re-use the Subsampling method for estimating the correlation between typicality and learnability. That is, given an original large corpus C^n, we take m subsamples of size k, $\{C_1^k, ..., C_m^k\}$. Then, for each C_i^k we measure both its typicality $a(C_i^k; P_{\alpha,\beta})$ and its learnability and the k pairs of values we obtain in this way are our estimate of the correlation.

While in principle a valid approach, the straightforward Subsampling method will unfortunately not work for this case. In order to estimate the full correlation between typicality and learnability we need set of subsamples $\{C_1^k, ..., C_m^k\}$ to cover a large range of typicality values, i.e. the values of $a(C_i^k; P_{\alpha,\beta})$ must be sufficiently different for different i. However, according to the AEP (see previous section), for large subsample sizes k, this will precisely not happen. Instead, as described, the AEP implies that all corpora converge in typicality in the limit of their size and therefore, as k grows, the set of typicality values of the m subsamples will converge to a single value. This single value is the Zipfianness of the underlying corpus C^n and if, as discussed in Chapter 2, C^n is large enough, then it will be highly Zipfian and hence have typicality close to 0. Since the typicality of the randomly sampled subcorpora converges to this value, they converge to being Zipfian as their size grows. Differently put, if n and k are large, then sufficiently non-Zipfian subcorpora (i.e. corpora with typicality significantly different from 0) will practically not occur, rendering the Subsampling method of little use for this case.

As explicitly mentioned in the introduction of the Subsampling method (Section 3.1) and the beginning of the current chapter, the original Subsampling method hinges on uniform sampling to lead to statistically valid estimates. Here, and indeed the central contribution of the current chapter, we develop a way to avoid the AEP by removing uniform sampling and adopting biased sampling schemes. These are intended to actively destroy the AEP and stop subsamples from converging in typicality. With the help of biased sampling, we can overcome the problem of convergence and can then actually still use the Subsampling method to estimate the correlation between Zipfianness and learnability as described above. Thus, the method we propose here could be seen as a special case of the general Subsampling method with the only change in the sampling procedure.

In our case, a biased sampling scheme which avoids the AEP is simultaneously one that favours subcorpora which are atypical with respect to Zipf's law, i.e. less Zipfian subcorpora receive higher probability of being sampled. At the same time, a sampling scheme which achieves this should not beyond necessity disrupt other statistical properties of language besides Zipf's law. That is, a good sampling scheme should yield subcorpora that are less Zipfian but otherwise as natural, i.e. actually typical with respect to the language as a whole, as possible. This is not straightforward

since there are many competing theories of how Zipf's law arises and its interaction with other properties of language is complex (see e.g. Piantadosi 2014 or A. Corral et al. 2009 and cf. Section 1.1). In this section, we propose and describe just two of the many possible biased sampling schemes with the potential to achieve this two-fold goal. In the following section, we evaluate the outcomes of biased sampling, that is we investigate the properties of the sampled subcorpora.

A note on terminology: Even though the algorithms we describe are indeed biased sampling schemes which produce random subcorpora by successively adding sampled elements to an initially empty set. Differently put, these algorithms are constructive sampling schemes. Nonetheless, we will call them "filtering" algorithms and the class of algorithms "Filtering method". We do so for conciseness and because the sampling schemes work by imposing restrictions on which element may be sampled next at a given point in the sampling procedure. We do not mean to imply a connection to statistical algorithms such as particle filters or rejection samplers (both part of the field of Monte Carlo methods, Rubinstein and Kroese 2016 gives a broad overview), although such connections are likely to be worth investigating.

4.3.1 Typicality Filter

The first filtering algorithm we describe is arguably the most general of any potential sampling scheme, since it works in direct reference to the typicality of subcorpora and it is hence that we call it Typicality Filter (TF). In brief, the TF successively samples sentences and keeps track of the typicality of the resulting subcorpus, thus ensuring that the typicality value does not surpass an initially specified value. Before fully describing the algorithm that is the TF, we specify its preliminaries, that is its inputs and parameters.

First, and these inputs are the same as in the general Subsampling method of Chapter 3, we require an initial large corpus C^n from which we will sample subcorpora. For this, we use the same corpora of $50{\cdot}10^6$ tokens as before in Chapters 2 and 3. Moreover, we need to set the size k of these subcorpora. Since we are sampling under a restriction, it is not guaranteed that a subset of C^n which satisfies that restriction actually exists for a given size k. We simply determined k via trial-and-error and found $k = 1{\cdot}10^6$ to be both large enough and to work for all languages that we use. Notice that this is also the size of subsamples we have been using throughout Chapter 3 and indeed we have done so to keep the results there comparable to the results here.

The second set of parameters pertains to typicality and consists of a reference distribution P and a tolerance parameter ε. While the reference distribution P is estimated via MLE, ε needs to be determined experimentally, much like k. In the case of ε, this is however more involved than simple trial-and-error.

As the reference distribution, with respect to which we measure typicality during the sampling procedure, we use of course Zipf's law, $P_{\alpha,\beta}$; what needs to be deter-

mined are the law's parameters α and β. From the theoretical perspective, here is actually the greatest caveat: The point of biased subsampling is to produce samples which do not share some of the same distributional properties of the language they were generated from, in particular we are trying to generate non-Zipfian samples from a Zipfian language. But, to reiterate what we have emphasised above, the distributional properties of the underlying language are not observable and in particular the shape and extent of Zipf's law, together with the closest-fitting parameters α and β, are not observable. These are, however, precisely what we would need for the reference distribution as input to the TF. Although there is, once more, no true solution to this problem, for practicality's sake we can still form a reasonable guess at the values of α and β. Namely, we estimate them via MLE from the corpus C^n, the same corpus from which we will also draw subsamples. These MLE parameter values $\hat{\alpha}$ and $\hat{\beta}$ are the ones we reported in Chapter 2, Table 2.2 and they are indeed the parameter values we use for the reference distribution for the TF. For comparison, and indication of the quality of the MLE parameters as proxies for the parameter values in the theoretical underlying language, refer to Section 3.3.2 (specifically Table 3.1), where we have analysed the convergence behaviour of the rank-frequency relationship.

The TF's parameter ε has the same interpretation and use as in the definition of the typical set: it determines which subsamples are considered Zipfian. That is, given a ε and the reference distribution $P_{\hat{\alpha},\hat{\beta}}$, a subsample C^k is considered non-Zipfian if it has $|a(C^k; P_{\hat{\alpha},\hat{\beta}})| > \varepsilon$ (which is equivalent to $C^k \notin \mathcal{A}_\varepsilon^k(P_{\hat{\alpha},\hat{\beta}})$). And indeed, given a value of ε, the TF returns only randomly sampled subcorpora which are non-Zipfian, i.e. satisfy this condition. Similar to the issue with setting the TF's parameter k, we do however not have a guarantee that subcorpora exist which are non-Zipfian, i.e. are outside of the typical set for a given ε (the lower the value of ε, the lower the likelihood that this is the case). As the consequence, we cannot simply choose any value for ε by must determine it experimentally; here, we take a more principled approach for which we actively exploit the AEP: According to the AEP, and as detailed at the outset of the current section, if a subcorpus is the outcome of uniform subsampling, then it will with high likelihood typical. So if we take any set of uniformly sampled subcorpora $\{C_1^k, ..., C_m^k\}$ then we can assume the vast majority of them to in fact be typical due to the AEP. If we now measure the typicality of each of these subcorpora, $a(C_i^k; P_{\hat{\alpha},\hat{\beta}})$, then we obtain an estimate of the distribution over typicality values of subcorpora which are actually typical to $P_{\hat{\alpha},\hat{\beta}}$. This distribution has a mean μ and a standard deviation σ and we choose the value of ε in terms of μ and σ. Specifically, by using $\varepsilon = \mu \pm \sigma$ we can be relatively certain that the TF does not return any subsamples which should have been considered Zipfian and vice versa. Further, by using $\varepsilon = \mu \pm 2 * \sigma$, we can increase the reduce the set of subcorpora which can be returned by the TF, that is increase the threshold of typicality which subsamples need to surpass to be considered non-Zipfian. In general, we introduce a factor f and use $\varepsilon = \mu \pm f * \sigma$ as the input to the TF; in the following section, we hence speak of f as the TF's parameter, not of ε.

With the parameter values specified, we are ready to describe the Typicality Fil-

procedure TypicalityFilter($C^n, k, P_{\alpha,\beta}, \varepsilon$)
    ntokens $\leftarrow 0$
    i $\leftarrow 1$
    while ntokens $< k$ **do**
        candidate_s $\sim C^n$ ▷ $\sim$ indicates sampling (without replacement)
        candidate_C $\leftarrow (\mathtt{s}_1, ..., \mathtt{s}_i, \mathtt{candidate\_s})$
        if $a(\mathtt{candidate\_C}; P_{\alpha,\beta}) < -\varepsilon$ **then** ▷ use $> \varepsilon$ for the other direction
            $\mathtt{s}_{i+1} \leftarrow \mathtt{candidate\_s}$ ▷ candidate_s is the next sample
            ntokens $\leftarrow$ ntokens $+ |\mathtt{s}_i|$ ▷ $||$ is the length in words
            i $\leftarrow$ i $+ 1$
        end if
    end while
return $C^k = (\mathtt{s}_1, .., \mathtt{s}_i)$ ▷ result is a subcorpus with k tokens and i sentences
end procedure

Algorithm 1 A pseudo-code implementation of the Typicality Filter. See the main text for a detailed high-level description. Notice that the resulting subcorpus C^k has at least k tokens (rather than exactly k), since whole sentences are being added during sampling.

ter itself, its pseudo-code is given in Algorithm 1. To reiterate, the output of the TF is a random subcorpus C^k of the original corpus C^n such that $|a(C^k; P_{\hat{\alpha},\hat{\beta}})| > \varepsilon$, i.e. $C^k \notin \mathcal{A}^n_\varepsilon$. The TF samples C^k by starting with an empty set of sentences and successively adding (randomly drawn) sentences until the desired number of tokens k is reached. At each step, after having successfully sampled i sentences, the algorithm first samples a new candidate sentence s uniformly without replacement from C^n. It then tentatively adds s to the already sampled subcorpus of i sentences, forming the new subcorpus $C = (s_1, ..., s_i, s)$, and measures the typicality of that subcorpus, $a(C; P_{\hat{\alpha},\hat{\beta}})$. Only if $a(C; P_{\hat{\alpha},\hat{\beta}}) < -\varepsilon$ is s accepted as the next sample $i + 1$, otherwise, s and C are discarded. In both cases the algorithm re-starts the procedure by sampling a new sentence s from C^n.

Notice that rather than comparing the absolute value of $a(C; P_{\hat{\alpha},\hat{\beta}})$ to ε (as in the definition of $\mathcal{A}^k_n(P_{\hat{\alpha},\hat{\beta}})$), the TF compares its raw value to $-\varepsilon$. This is because, as mentioned in Section 4.1 and together with the definition of typicality in the previous section, a subcorpus can be atypical in two directions: it can either have too low probability under Zipf's law, leading to a negative typicality value, or too high probability, leading to positive typicality. The definition of the typical set removes this distinction but we make it explicit for the TF. In the given implementation, checking if typicality is lower than $-\varepsilon$, the TF will only return subcorpora which are atypical, i.e. non-Zpfian, because their probability is too low under $P_{\hat{\alpha},\hat{\beta}}$.

Finally, a note on the efficiency of the TF which evidently does not need to make involved computations besides $a(C; P_{\hat{\alpha},\hat{\beta}})$ in each iteration. At the same time, having to compute the typicality from scratch in every iteration would lead to unacceptable complexity because computing $a(C; P_{\hat{\alpha},\hat{\beta}})$ requires iterating over all tokens in the sub-

corpus C. This would lead at least to quadratic complexity of the TF in its parameter k. Fortunately, typicality is additive, or more precisely, the log-probability of a corpus under Zipf's law is, due to the law's independence property. Hence, as a new sentence s is added to a subcorpus C, the typicality of the resulting subcorpus $a(C \cup s; P_{\hat{\alpha},\hat{\beta}})$ can be efficiently updated by adding the log-probability of s to the already calculated log-probability of C. Re-calculating the typicality of the grown subcorpus hence only requires iteration over the added sentence and this makes the individual iterations of the TF very efficient. Even if each iteration has low time complexity, the overall time complexity of the TF can still be prohibitive for very large k (at least larger than $1{\cdot}10^6$). The reason is that sampling a subcorpus C^k from C^n with the TF is essentially equivalent to performing a randomised search over all subcorpora of size k in C^n. Worse yet, because of the AEP, in fact most of the sentences which are drawn during the sampling procedure will have to be discarded. Hence, the number of iterations will most often by far exceed k, but is at least bounded by n.

4.3.2 Speaker Restriction Filter

Note that the Typicality Filter we just described also works for distributions other than Zipf's law because the reference distribution can simply be exchanged for a different one. That is, the Typicality Filter works directly on the abstract concept of typicality and does not make reference to any of the specific properties of Zipf's law. A filtering algorithm can, however, do exactly this in order to sample non-Zipfian subcorpora. The Speaker Restriction Filter (SRF) which we describe in this Section is an example of such a filtering algorithm.

The motivation for the SRF is the following: As (A. Corral et al. 2009) have already noticed, the presence of Zipf's law is connected to the patterns of inter-appearance distance of words in language. While some words have consistently very short inter-appearance distances, that is they usually recur almost immediately after having been uttered, others have thousands of words in between two of their occurrences. It should be clear, that the words with low inter-appearance distances are the high-probability ones and that there can only be few of them. Vice versa, the vast majority of words has high inter-appearance distances, as they are the low-probability words. This exactly mirrors Zipf's law in which there are few high-probability words and many low-probability ones, and in fact, these patterns in inter-appearance distance is clearly one of the reasons that Zipf's law arises.

The SRF works by disrupting exactly this property of language, namely by preventing words from having too short inter-appearance distance. The rationale is that, by obliterating in the sampled subcorpora those phenomena which give rise to Zipf's law, the Zipfianness of these subcorpora will also be affected. Specifically, the SRF imposes a minimum inter-appearance distance on the sampled subcorpus and the words in it are not allow to recur below this distance. This will according to our expectation prevent the high-probability words from accumulating too high frequencies

of occurrence, and in particular from accumulating the excessively high frequencies characteristic for the high-probability words in Zipf's law. Low-probability words, on the other hand, will likely not be affected by the SRF's sampling scheme. These words are characterised by high inter-appearance distances, so enforcing a minimum distance will not affect their frequencies in a subsample returned by the SRF. Thus, we predict that the SRF reduces Zipfianness mainly by lowering the frequencies of the high-probability words but leaving the rest of the distribution over words relatively untouched. Otherwise, however, we cannot make precise predictions about how Zipfian the subcorpora returned by the SRF will be, in particular not terms of their typicality with respect to Zipf's law; this has to be determined empirically and we do so in the next section.

The way the SRF achieves restricted minimal inter-appearance distances is simple: While successively sampling sentences, much alike the Typicality Filter, the SRF blocks a sentence from being sampled next if it contains any of the words which occurred in the previously sampled sentences. This way of working suggests an analogy in terms of speakers which is the reason why we call the SRF so. Namely, we imagine a speaker who utters sentence after sentence. But, in contrast to a typical speaker, this speaker tries to avoid repeating words too often. Thus, if she uses a certain word in the current utterance she will wait a few utterances before using the word again. Since the Speaker Restriction Filter is genuinely equivalent to such speaker behaviour, it describes a cognitively possible, albeit unrealistic, generative process of language. We emphasise that this is in opposition to the Typicality Filter which does produce non-Zipfian subcorpora but has nothing to say about what makes them non-Zipfian.

The speaker's waiting time in the analogy above, we will call history length h and is the SRF's only parameter besides C^n and k (both of which it shares with the Typicality Filter and the general Subsampling method). For a concrete example of the meaning of h, consider the case $h = 1$: A subcorpus C^k, i.e. a sequence of sentences $(s_1, ..., s_i)$, returned by the SRF will satisfy $s_j \cap s_{j+1} = \emptyset$ for all $j < i$. (We abuse notation to let the intersection of two sentences denote the intersection of the words in them.) If $h = 2$, then for all $j < i - 1$ it will be that case that $s_j \cap s_{j+1} \cap s_{j+2} = \emptyset$, i.e. all triples of successive sentences in C^k will have no words in common. And so on for higher h. Once again, the usable values for h need to be determined empirically, since there is once more no guarantee that a subcorpus C^k of the original corpus C^n actually exists which satisfies the constraint for a given h. Just like with the parameter k, we do so by simple trial-and-error.

Notice in these examples that the order of sentences is important, so different reorderings of the sentence of a subcorpus returned by the SRF may not be valid outputs anymore. This is in emphasised contrast to the Typicality Filter which may return any ordering of the same set of sentences. Yet more importantly this is also in contrast to Zipf's law which, due to its independence property, cannot distinguish between different orderings of a corpus. For this reason, there exist subcorpora with the same distribution over their words (and therefore the same typicality with respect to Zipf's

```
procedure SpeakerRestrictionFilter(C^n, k, h)
    s_1 ~ C^n                          ▷ ~ indicates sampling without replacement
    ntokens ← |s_1|                    ▷ | | signifies length (here, number of words)
    i ← 2
    hist ←Queue(s_1)
    forbidden ← ∅ ∪ s_1                ▷ the union of sentences is the union of their words
    while ntokens < k do
        candidate_s ~ C^n
        if candidate_s ∩ forbidden = ∅ then
            s_i ← candidate_s          ▷ candidate_s is the next sample
            ntokens ← ntokens + |s_i|
            i ← i + 1
            hist.push(s_i)
            if |hist| > h then
                hist.pop()
            end if
            forbidden ← ⋃_{s∈hist} s
        end if
    end while
return C^k = (s_1, .., s_i)            ▷ result is a subcorpus with k tokens and i sentences
end procedure
```

Algorithm 2 The pseudo-code implementation of the Speaker Restriction Filter.

law) of which one is a valid output of the SRF while the other is not.

The practical implementation of the SRF is not as simple as that of the Typicality Filter, since the SRF needs to keep track of unions and intersections of words while sampling, see Algorithm 2. Its implementation does, however, admit a simple high-level description: Given a corpus C^n to randomly draw sentences from, a target subcorpus size k and a history length h, the SRF starts by sampling a single sentence s_1 uniformly from C^n. After having already successfully sampled i sentences, a newly sampled sentence is accepted as sample $i+1$ only if none of the words in it occur in any of the h previously sampled sentences. This is repeated until the sampled sentences amount to at least k tokens.

In closing the description of the SRF, we outline how it can be adapted into the opposite effect, namely into working on the low-probability words of the vocabulary. The analogous realisation is that there are so many low-probability words, and the rank-frequency relationship's tail correspondingly as heavy because speakers introduce new words at a very high rate during discourse. By disallowing such behaviour, that is by putting a bound on this rate, the relationship's tail can be reduced. To be more precise, one may restrict the number of sentences which introduce new words, that is introduce a parameter v which controls how many sentences need to be sampled before a new word may be introduced into discourse. Setting $v = 1$ would imply

that every next sentence to be sampled can introduce new words, $v = 2$ that only every second sentence can introduce new words, and so on. Already at $v = 2$, half of the sentences in the resulting subsample would not have introduced any new words and the size of its vocabulary heavily reduced due to lower numbers words with low frequencies. At the same time, the frequencies of high-probability words would be left largely unchanged, as these could continue to recur across sentences at their original high rates. Thus, adapting the SRF in this way would affect the Zipfianness of resulting subcorpora in a similar, albeit opposite, way.

4.4 Results

The algorithms of the Typicality Filter (TF) and the Speaker Restriction Filter (SRF), Algorithms 1 and 2, can be quite directly practically implemented as such, except for minor efficiency improvements. So we have done and created large numbers of subsamples, for each filtering algorithm, from our Wikipedia corpora. The current section is devoted to analysing those subsamples and empirically validating the filtering methodology.

Rather than putting filtering to use for learnability analyses straight away, we analyse here the subsamples themselves because we need to ensure that they have the properties we expect, as laid out in the first part of the current Chapter. The analysis carried out in this section is hence mainly one about the usefulness of the filtering methodology for learnability and other assessments. Regardless of the soundness of the derivation of the filtering algorithms, the usefulness of filtered subsamples is by no means given: The sampling biases of both the TF and the SRF are too complex to be easily derived analytically. This entails that, besides the explicit restrictions we impose during sampling in these filtering algorithms, we cannot make precise informed predictions about the properties of filtered subsamples. The effects of the sampling biases of the TF and the SRF therefore need be analysed by empirical means. We focus here on three of the core aspects of the subsamples' properties but emphasise that an empirical investigation is necessarily non-exhaustive.

As mentioned in the algorithms' descriptions, the values of some of their parameters need to be determined by experimentation, that is, whether subsamples with those values can indeed be found by the sampling algorithms. These parameters are the subsample size k, for both algorithms, the factor f to compute $\varepsilon = \mu + f * \sigma$ for the TF and the history length h for the SRF. In order to ensure comparability, we used the same value of k for both algorithms and all seven languages. While slightly higher values could have been possible, $k = 1{\cdot}10^6$, or 2% of the original corpus C^n that was sampled from, is large enough for our purposes and leads to reasonable run-times of the filtering algorithms. Notice that the same sample size $k = 1{\cdot}10^6$ was also used in Sections 3.2, 3.3.1 and 3.3.2 of the previous Chapter. This was indeed done so that the findings about uniformly sampled, typical subcorpora from the previous Chapter can

f	2	6	10	14	18	22	26
EO	•	•	•	•	•	•	•
FI	•	•	•	•	•	•	
ID	•	•	•	•	•		
KO	•	•	•	•			
NO	•	•	•	•	•	•	•
TR	•						
VI	•	•	•	•	•		

h	2	4	8	16	32	64	81
EO	•	•	•	•	•		
FI	•	•	•	•	•	•	•
ID	•	•	•	•	•	•	•
KO	•	•	•	•	•	•	•
NO	•	•	•	•	•		
TR	•	•	•	•	•	•	•
VI	•	•	•	•	•		

Table 4.1 Per-language indication of the values of the parameter f of the Typicality Filter (left table) and the parameter h of the Speaker Restriction Filter (right table) used. See Section 2.1 for the meanings of the language codes. A bullet point (•) indicates that the respective filtering algorithm managed to find a set of $m = 10$ subcorpora for the corresponding combination of language and parameter value.

be compared to those about filtered subcorpora in the current Chapter.

For the respective parameters f and h of the TF and SRF we used a range of values rather than just a single one. Both for a more detailed perspective on the effects of the filtering algorithms and because we are eventually interested in subsamples from the entire spectrum of typicality, as explained at the outset of Section 4.3. The used ranges of f and h can be found in Table 4.1; the specific values were, rather arbitrarily, chosen to create an evenly spaced spectrum. The respective maximum values are the highest values of these ranges for which subsamples could be found after a fixed amount of run-time of the filtering algorithms. Notice the partly strong differences between languages: Whereas Esperanto and Norwegian allow for factors f of up to 26, for Korean the TF could not find subsamples for factors beyond 14; we consider the case of Turkish an outlier which we will not investigate for now. The pattern is reversed for the history length h of the SRF, where Esperanto, Norwegian and Vietnamese are the only languages which did not permit history lengths of up to 81.

It is conceivable that some patterns in the maximum parameter values are tied to language-specific properties, such as the morphological differences discussed in 2.1. However, as also discussed there, it is not clear to what extent these patterns are tied to Wikipedia rather than the languages themselves which is why we refrain from interpretation of the patterns for now. Either way, it is promising to see that all languages collectively allow relatively high parameter values and that each language allows a high value for at least one of the filtering algorithms.

In the same vein as Chapter 3, we take multiple subsamples for each type of subsampling (uniform, TF and SRF) and each value of the parameters f and h in order to obtain an estimate of the sampling distribution rather than just a point estimate. Like in Chapter 3 we use $m = 10$, i.e. take 10 subsamples per sampling type and parameter value. Thus, to be explicit, three sampling types, seven values for each parameter and

10 subsamples each yields 150 subcorpora for each language.

As Table 4.1 shows, Finnish admitted the highest parameter values for both the TF and the SRF. For this reason, all properties of the subcorpora resulting from the two filtering algorithms are investigated using Finnish. Just like in the previous chapter (see the final paragraph of Section 3.1), it would be highly interesting, and most likely be instructive about the nature of both the filtering algorithms and the rank-frequency relationships across languages, to analyse all languages and carry out cross-linguistic comparisons. But again, such analyses would go beyond the scope of this thesis, and so we focus on verifying the general properties of the filtering algorithms. We do, however, provide all results given below (plots, graphs and numerical values) at `github.com/valevo/Thesis/figures`. As these results show, the results we present in terms of Finnish generally also hold for the other languages, although some trends are more or less extreme.

4.4.1 Assessing Zipfianness

The first and most obvious assessment to be done on the outcomes of the TF and the SRF is how Zipfian they are. Here, we assess the Zipfianness of the resulting subcorpora in two ways: First, their typicality because this the formal measure of conformity of a sample to a theoretical distribution of our choice and arguably one of the most objective. Bear in mind, as described in Section 4.3.1, that typicality of the subcorpora is measured with respect to the parameters of Zipf's law found in the source corpus C^n, $P_{\hat{\alpha},\hat{\beta}}$. Hence, if a subcorpus is atypical with respect to $P_{\hat{\alpha},\hat{\beta}}$, this could also imply that the subcorpus conforms to Zipf's law with different parameters α and β. Second, we assess the shapes of the rank-frequency relationship in the filtered subcorpora, as we have already done in Section 2.2. Although such an assessment will be less rigorous than in terms of typicality, it is likely closer to the intuitive notion of Zipfianness because it is difficult to anticipate the shape of rank-frequency relationships of atypical subcorpora.

These two ways of assessment also correspond closely to the dichotomy between the TF and the SRF: We know how typical the outcomes of the TF will be with respect to $P_{\hat{\alpha},\hat{\beta}}$ because it is an explicit ingredient of the TF's definition. What we can however not predict precisely is what shapes of the rank-frequency relationship we will find in these outcomes. The situation is reversed for the SRF, whose sampling restriction directly affects the shape of the relationship in the sampled subcorpora. But here we do not know in advance how typical these subcorpora will be with respect to Zipf's law.

We analyse the typicality of the subcorpora in reference to Table 4.2 which is constructed by first computing the typicality $a(C_i^k; P_{\hat{\alpha},\hat{\beta}})$ for each subcorpus of the set of subcorpora $\{C_1^k, ..., C_m^k\}$ obtained from each sampling type and each parameter value, leading to the sets $\{a(C_1^k; P_{\hat{\alpha},\hat{\beta}}), ..., a(C_m^k; P_{\hat{\alpha},\hat{\beta}})\}$. From each such set, we then compute

	UNIF	TF							SRF						
		2	6	10	14	18	22	26	2	4	8	16	32	64	81
μ	6.55	6.54	6.41	6.20	5.97	5.70	5.40	–	6.48	6.44	6.37	6.29	6.19	6.08	6.04
σ	.002	.002	.0	.0	.0	.001	.0	–	.001	.001	.001	.001	.001	.001	.001

Table 4.2 Mean μ and corresponding standard deviation σ of the distributions over typicality $a(C_i^k; P_{\hat{\alpha},\hat{\beta}})$ of individual subsamples C_i^k. For each sampling type, one of uniform (UNIF), Typicality Filter (TF) and Speaker Restriction Filter (SRF), and each value of the filtering parameters f and h, respectively, the distribution over typicality values is constructed from m subsamples. Values for the TF with f are not available, see Table 4.1.

its mean μ and standard deviation σ which correspond to the two rows of Table 4.2. Notice that σ is negligible for all sampling types and parameters value which is why we did not additionally plot the histograms. In the case of uniform subsamples, σ is this low because the asymptotic equipartition property (see Section 4.2) has lead the typicality of uniformly sampled subcorpora to largely converge to the mean μ. The same is most likely also true for the filtered subsamples which have converged to their respective means. Even though these are the outcome of biased subsampling, their typicality will also converge in the limit and the low collectively low values of σ indicate a high degree of convergence at our subsample size of $k = 1{\cdot}10^6$. Thus, while a low standard deviation σ is per se not surprising, it makes it all the more remarkable that filtered subcorpora with filtering parameter values as high as $f = 22$ and $h = 81$, respectively, exist.

As expected, higher values of the parameter f of the TF indeed lead to lower mean typicality of the resulting subcorpora. Even though this was clear from the algorithm's design, empirically verification is still necessary due to caveats in estimation. The lowest achieved typicality value is 5.4, which corresponds to 80% of the typicality of uniform subsamples, and which is not a drastic reduction in typicality in absolute terms. In relative terms, however, the high degree of convergence of due to the asymptotic equipartition property implies that corpora with typicality value 5.4 have essentially 0 probability under uniform sampling, which leads to typicality values strongly concentrated around 6.55. Since it is this probability which determines the significance of the reduction in typicality, at for our purposes, already the reduction of typicality at $f = 6$ can be termed significant. Additionally, notice that increases of the parameter f lead to steady, linear decreases in typicality which underlines the significance of the reductions.

This is in contrast to the decrease typicality values across parameter values of the SRF: Even though there is steady decrease, the parameter values need to increase quadratically, the reason for which lies in the interpretation of the parameter h as the history length. Generally, the SRF leads to a less pronounced reduction in typicality as compared to the TF. Seeing as it was not given that the SRF would result would even result in noticeable reductions of typicality, it does however achieve remarkable

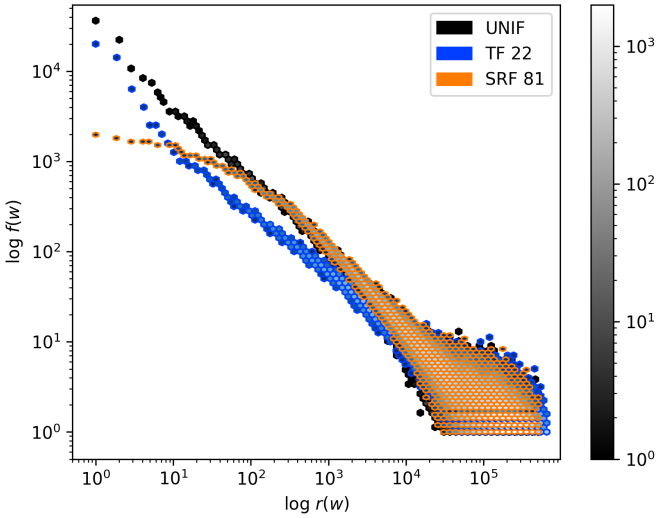
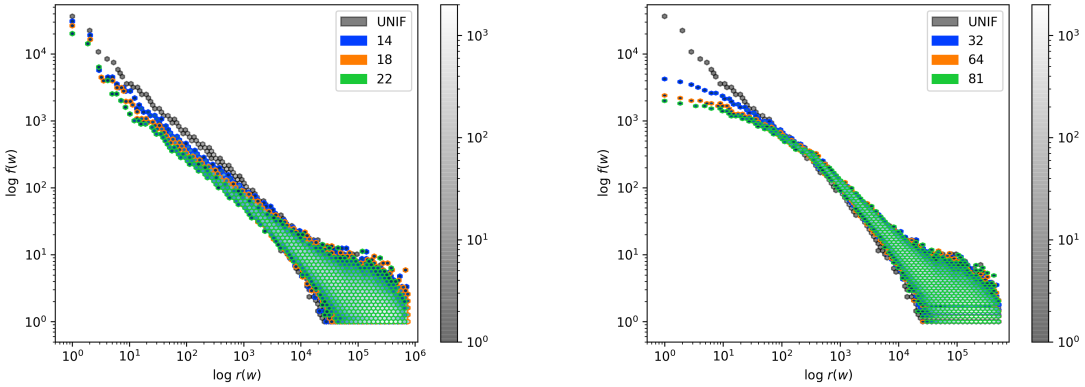


Figure 4.1 Mean rank-frequency relationships, $\overline{r}(w)$ and $\overline{f}(w)$, obtained from uniform subsampling ("UNIF", green), the Typicality Filter with parameter $f = 22$ ("TF 22", blue) and the Speaker Restriction Filter with parameter $h = 81$ ("SRF 81", orange).

effects. With the same reasoning as above, the reductions can be deemed significant, even at $h = 2$.

Thus, Table 4.2 confirms that the primary goal of filtering has been achieved, namely a way to sample subcorpora of lowered typicality. Again, since sampling subcorpora with typicality values as low as those of the filtered subcopora with uniform sampling is practically impossible, the reductions by both filtering algorithms have significance. Specifically, their significance should suffice to convince a hypothetical learner that the filtered subcorpora are not outcomes of the original Zipfian language, since that assigns probability close to 0 to them. In addition, the steady and roughly linear decrease in typicality across the values of the parameters of the filtering algorithms shows that typicality can be quite precisely controlled. This enables the original motivation for filtering, namely mapping out the spectrum corpora in terms of their typicality, as the filtering algorithms allow to sample subcorpora of arbitrary typicality within the achievable bounds.

Even though typicality provides an exact and objective measure, the values of Table 4.2 are too abstract for an intuitive understanding of how the filtering algorithms affect Zipfianness. For this it will be more revealing to analyse the resulting rank-frequency relationships themselves and we begin with a direct comparison between the three types of sampling: uniform, TF and SRF. The mean rank-frequency relationships resulting from each sampling type are shown in Figure 4.1. The relationship obtained from uniform sampling has the shape familiar from Chapters 2 and 3 and can informally be termed Zipfian. Notice that the shape resulting from the TF could



(a) Typicality Filter

(b) Speaker Restriction Filter

Figure 4.2 (a) Mean rank-frequency relationships obtained from the Typicality Filter with parameter values $f = \{14, 18, 22\}$ (see the legend of the plot). (b) Mean rank-frequency relationships resulting from the Speaker Restriction Filter with parameter values $h = \{32, 64, 81\}$. Mean relationship from uniform subsampling (labelled UNIF) also plotted in both (a) and (b) for comparison.

also be characterised as Zipfian, albeit less so because of the downward curvature in the graph's upper body. Indeed, the two graphs are mainly distinguished by their steepness, with the TF leading to a more uniform relationship.

Presumably, this is also the reason why the TF achieves subcorpora with lower typicality than the SRF. It leads to more uniform word frequencies in the entire graph, not just the head as is the case with the SRF. The body of the relationship contains more words and more probability mass, leading to a genuinely more uniform word distribution. Thus, although the effect of the SRF is visually more salient, it is confined to the head and affects less of the overall probability mass. At the same time, one could argue from a graphical point of view that the SRF leads to less Zipfian subcorpora because of its strong deviation from the straight line predicted by Zipf's law. Notice, finally, that the tails which result from the three types of sampling are virtually the same. This confirms our theoretical images of the TF and the SRF, since neither was designed to affect the relationship's tail and indeed neither empirically does.

Notice that we used the highest values of the respective parameters f and h of the TF and SRF in Figure 4.1. In order to contextualise these, and to see how the rank-frequency relationship evolves across different parameter values, we constructed Figures 4.2a and 4.2b. With the mean relationship from uniform subsamples as the reference in both plots, we plot the mean rank-frequency relationships extracted from the outcomes of filtering with the three highest parameter values.

Similar to the mean typicality values in Table 4.2, while a linear increase in f leads to a linear increase of the effect of the TF on the relationship, h needs to increase quadratically for the SRF's effect on the relationship to increase linearly. This is visible in the relative difference between the shapes at $h = 64$ and $h = 81$ which is not

	UNIF	TF							SRF						
		2	6	10	14	18	22	26	2	4	8	16	32	64	81
α	1.15	1.15	1.14	1.13	1.12	1.12	1.11	–	1.14	1.14	1.14	1.13	1.13	1.13	1.12
β	4.55	4.55	4.63	4.73	4.82	4.98	5.2	–	5.0	5.2	5.4	5.59	5.76	5.85	5.89
R^2_{McF}	.66	.66	.65	.64	.63	.62	.60	–	.66	.65	.65	.65	.64	.64	.63
rel. BIC	2.99	2.98	2.92	2.83	2.74	2.64	2.56	–	2.95	2.93	2.9	2.87	2.82	2.78	2.76

Table 4.3 Results of performing MLE of Zipf's law, i.e. its parameters α and β, on the mean rank-frequency relationship obtained from each sampling type and parameter value of the filtering algorithm. UNIF stands for uniform subsampling, TF for the Typicality Filter and SRF for the Speaker Restriction Filter. The values below TF and SRF are the values of parameters f and h of the Typicality Filter and the Speaker Restriction Filter, respectively. See Section 2.2.2 for the interpretation of R^2_{McF} and relative BIC.

as strong as the difference between $h = 32$ and $h = 64$. Besides this, we observe clear trends in both plots: Across values of f, the TF successively leads to more uniform distributions over words and hence to a more flat rank-frequency relationship. The downward curvature in the relationships' body becomes more pronounced while its tail stays the same across all of f's values. The same is case for the SRF where growing parameter values apparently only affect the head of the relationships. The head curves off more and more until the point where the most frequent words have approximately equal mean frequency with $h = 81$.

Finally, for a more formal assessment of the Zipfianness of the rank-frequency relationships resulting from the TF and the SRF, we perform MLE of the parameters α and β of Zipf's law. Specifically, we perform MLE on each mean rank-frequency relationship obtained from each sampling type and each parameter value the results of which are given in Table 4.3. With regards to the parameters α and β themselves, we find that neither of them seems to converge across values of the parameters f and h of the filtering algorithms. The steepness parameter α steadily decreases and although the decrease is slight, α is an exponent which implies that even small changes lead to large differences. The decrease is slightly stronger in the case of the TF which mirrors the stronger decrease in typicality values. β, on the other hand, steadily increases which was strongly expected for the SRF because β controls the curvature in the head of the distribution. It is not clear, however, why β increases across the values of the parameter of the TF, since the curvature does not increase.

More importantly for an assessment of the fit of Zipf's law to these rank-frequency relationships, the goodness-of-fit measures R^2_{McF} and relative BIC (see Section 2.2.2) show a steady decrease, indicating that both the fit of and statistical support for Zipf's law weakens as the parameter values of the filtering algorithms increase. We emphasise that these goodness-of-fit measures are relative to the each of the respective MLE parameters and not to $P_{\hat{\alpha},\hat{\beta}}$ obtained from C^n. This is an important to notice since if even the best-fitting parameters lead to a worse fit, then this implies that the rank-

frequency relationships really are less Zipfian. Hence, even though the outcomes of the TF visually still appear quite Zipfian in the sense of conforming to a straight line, the decreasing values of the measures R^2_{McF} and relative BIC reveal that the TF really does decrease Zipfianness and not only typicality with respect to $P_{\hat{\alpha},\hat{\beta}}$.

On the other hand, the decrease in goodness-of-fit is not quite as strong as expected and the R^2_{McF} and relative BIC in fact still lend support to Zipf's law, even for the highest values of f and h. This implies that the filtering algorithms do not completely eliminate Zipf's law from the subsamples' rank-frequency relationships which is an unattainable goal in any case. It further points towards the dichotomy between goodness-of-fit measures and typicality. Whereas the former will not be drastically lower as long as there are traces of Zipf's law in the rank-frequency relationships of filtered subsamples, the latter can still be strongly reduced. With reduced values, typicality will indicate that a distribution other than Zipf's law should be favoured as the hypothesised source, even if Zipf's law is still a reasonably good source.

Whether or not typicality is sufficiently reduced in the outcomes of the TF and the SRF will depend on the specific application. However, as argued, the reductions we reported in Table 4.2 are significant enough to convince a hypothetical learner that the filtered subcorpora are not the outcome of the Zipf's law found in the source corpus C^n. This conjecture is strengthened by the fact that we used the mean ranks and frequencies $\overline{r}(w)$ and $\overline{f}(w)$ for our analysis of the rank-frequency relationships in the filtered subcorpora, akin to the methodology described in Chapter 3. Just like the means from uniformly sampled subcorpora, these means also approximate their theoretical values. The degree of deviation of the filtered mean rank-frequency relationships from the original ones we observed in Figures 4.1 and 4.2 and verified by MLE (Table **??**) indicate that the underlying theoretical languages really are less Zipfian.

4.4.2 Assessing Normality

As mentioned in Section 4.1, besides being less Zipfian, the samples used for learnability analyses of Zipf's law should also preserve as many of the statistical properties of the original language as possible. This is important, since differences in learnability may otherwise arise trivially and it would be difficult to isolate the influence of Zipf's on learnability.

At the same time, we have kept these statistical properties vague and the simple reason is that they are complex and largely unknown. For if they were known, then language could be explicitly and precisely statistically modelled, but the current state of the field of language modelling shows clearly that this is not the case. Hence, there can as of now be no definitive statements as to the degree to which a corpus represents the statistical properties of the entire underlying language it was sampled from. Notice, however, that if we knew the underlying language, then we could measure this degree precisely, namely by the core tool of this Chapter: typicality. As described in Section 4.2, a corpus is typical with respect to a distribution precisely if it mirrors that

	UNIF	TF							SRF						
		2	6	10	14	18	22	26	2	4	8	16	32	64	81
ϕ	9.62	–	9.75	–	9.19	–	7.83	–	–	10.0	–	10.6	–	11.5	–
τ	0.72	–	0.73	–	0.75	–	0.786	–	–	0.72	–	0.72	–	0.72	–
R^2_{McF}	0.99	–	0.99	–	0.99	–	0.99	–	–	0.99	–	0.99	–	0.99	–
rel. BIC	1600	–	1420	–	727	–	278	–	–	1610	–	1560	–	1480	–

Table 4.4 Results of performing MLE of Heap's law, i.e. its parameters τ and ϕ, on the mean vocabulary growth obtained from each sampling type and parameter value of the filtering algorithm. See Section 2.2.2 for the interpretation of R^2_{McF} and relative BIC. Missing values are omitted because estimation of vocabulary growth is expensive.

distribution or equivalently has the same statistical properties.

That said, we still make a non-exhaustive attempt at judging the typicality of the filtered subcorpora with respect to language as a whole, not just Zipf's law, and thus how "normal" they are. Specifically, we investigate some of the known and empirically testable statistical properties of language in the filtered subcopora in comparison to those same properties in uniformly sampled subcorpora of the same size. The higher the degree of similarity between the filtered and the uniformly subsampled in terms of these specific properties, the higher the likelihood that the properties are similar in the underlying languages. At the same time, some differences are to be expected, as changing the rank-frequency relationship of words will likely lead to changes in properties of language that are entangled with the word distribution. In a way that will become precise below, we need to therefore distinguish between such differences and those differences which, if found, would arise from the filtering algorithms themselves and would be seen as undesirable.

Vocabulary Growth

In deciding which statistical properties to compare, a convenient avenue is to look at the other known quantitative regularities of language, such as Heap's law which we have introduced in Chapter 2 for this purpose. The advantage here is that we know what shape the growth of the vocabulary size $V(n)$ should have in samples of language, namely one which can be described by $V(n) \approx \tau * n^{\phi}$ for some parameters τ and ϕ. This was confirmed in Section 2.3, where the MLEs of Heap's law showed near-perfect fit, as indicated by the goodness-of-fit measures R^2_{McF} and relative BIC. As Table 4.4, which was constructed in the same way as Table 2.3, shows the filtered subcorpora have the same near-perfect fit, that is their respective vocabulary growths are as well described by the equation of Heap's law. Hence, the filtered subcorpora, or rather their underlying theoretical languages, are as Heapian with respect to their

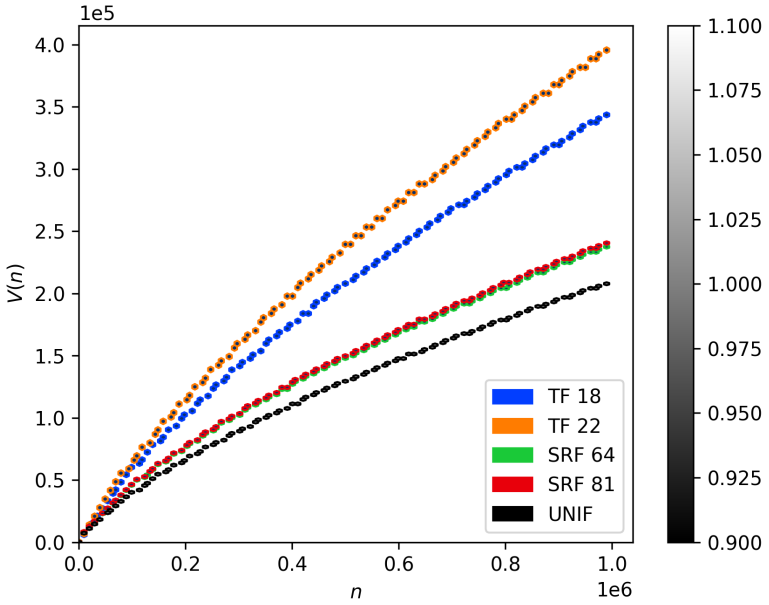


Figure 4.3 Mean vocabulary growth obtained from uniform sampling (UNIF), Typicality Filter (TF) with parameter values $f = \{18, 22\}$ and Speaker Restriction Filter (SRF) with parameter values $h = \{64, 81\}$ (see label in plot).

MLE parameters as the original language they were sampled from.

At the same time, looking at the actual vocabulary growth, we find some differences. This is reflected in the MLE parameters τ and ϕ in Table 4.4 as well as Figure 4.3, which displays the mean vocabulary growths themselves. Most strikingly, the TF leads to increased vocabulary growth and two about twice the vocabulary size at $1{\cdot}10^6$ tokens in comparison to uniform subsamples. As for the SRF, it looks as though it leads to increased vocabulary growth but the MLE parameters reveal that this increase is negligible. In fact, even the increase in vocabulary growth due to the TF is not as substantial as it may seem, and specifically it is still on the same order as that of the other sampling types. Using the MLE parameters, we can derive that one would need to increase sample size to about $1{\cdot}10^{18}$ for the order of vocabulary size to be different between the TF and uniform sampling.

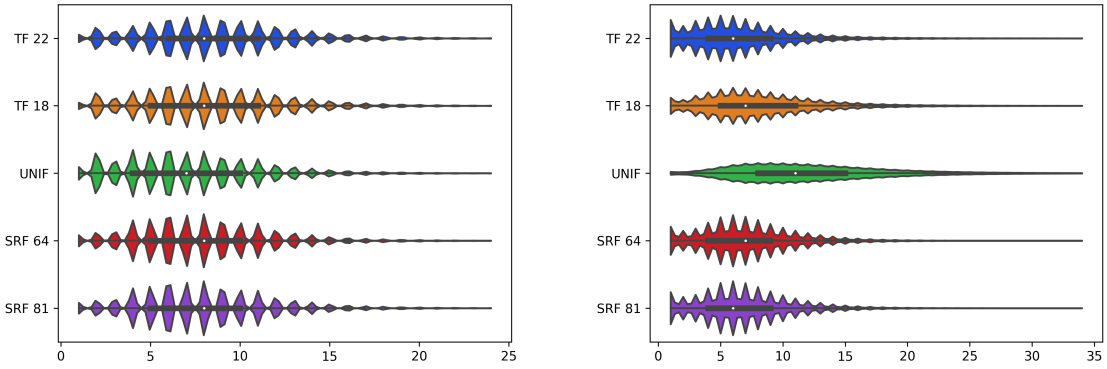
Finding exactly the same rates of vocabulary growth in the different types of sampling would have been surprising in any case: The distribution over words, which is changed as the outcome of the filtering algorithms, is surely connected to vocabulary growth since the latter governs the probability that a word is sampled. The TF and Speaker Restriction force more uniform word distributions which implies that low-probability words gain probability mass and therefore occurrence probability which implies that the vocabulary grows faster. In this way, it can even be argued that changes in vocabulary growth are a direct outcome of reducing Zipfianness, not just

a side-effect. Thus, even if increased vocabulary growth would rather trivially imply reduced learnability of the filtered subsamples, this increase is itself a direct consequence of reduced Zipfianness and sill pertains to the learnability of Zipf's law.

Length Distributions

A second statistical property of corpora and their underlying languages we turn are the distributions over the lengths of their constituents and we will consider both word and sentence length. On the one hand, these distributions provide controls for the normality of the filtered corpora that are easy to compute and to examine. On the other hand, we have also chosen them for their interpretations and applications in the literature: Treating letters and words as symbols in sequences and without assigning them any prescribed meaning, we enter again the realm of information theory and speak of the amount of information a word carries. According to information theory, the information content of a word depends on its probability of occurrence as well as the degree to which it is tied to specific contexts (Piantadosi, Tily, and Gibson 2011). Words with high probability and low specificity carry low information content and vice-versa. Importantly for our context, taking the information contents of all words in corpus together is then an indicator of how much information the corpus as a whole carries. Since the information content of a word is costly to compute, and can only be approximated in any case, we turn to a surrogate: its length. Word length has traditionally been viewed as correlating negatively with frequency (originally put forward by Zipf, Zipf 1932, and known as Zipf's law of Abbreviation, see also Bentz and Ferrer Cancho 2016). But recently evidence and arguments have been put forward that the length of a word correlates more closely with its information content (Mahowald et al. 2013). We therefore stipulate that a difference in word length distribution between corpora points towards a difference in their information content and emphasise the importance this gives to the word length distribution as a simple control.

We compare the distributions by a simple inspection and for this purpose we have plotted them in Figure 4.4a. In order to construct the distribution for each sampling type and filtering parameter value, we have pooled the individual distribution in each of the m subsamples which correspond to that type and parameter value. That is, the length distributions in Figure 4.4a display estimates of the word length distributions the underlying languages rather than of particular subsamples. Upon visual inspection, all distributions are highly similar, having the majority of their similarly spread out mainly between word lengths 5 to 11. Even though a Kruskal-Wallis test for equal means and a Levene's test for equal variance are below significance level, this is unsurprising given our large data size and that the processes which generated the length distributions really are different. As we are only interested in verifying similar distributions, we are inclined from Figure 4.4a to conclude that high similarity is indeed the case. To the extent that the collective information content of the words in a corpus adequately represents the information content of the corpus itself, this finding indicates that filtering leads to subcorpora with comparable amounts of information content.



(a) Word Length Distribution

(b) Sentence Length Distribution

Figure 4.4 (a) Distribution over word lengths obtained from uniform sampling (UNIF), Typicality Filter (TF) with parameter values $f = \{18, 22\}$ and Speaker Restriction Filter (SRF) with parameter values $h = \{64, 81\}$ (see labels). Each distribution is formed by pooling the individual distributions from the m subsamples into one. (b) Same as (a), but taking the distributions over sentence lengths.

A rationale similar to that of word lengths can and has been applied to sentence length, although here more pandering to the intuition that if a sentence contains more words, then it is also more likely to convey a higher amount of information. The distribution over sentence lengths has found common application in stylometry and especially in authorship attribution, since it has proven sensitive enough across corpora to even reveal the author of a given corpus. Hence, again besides as a simple control, differences in sentence length distribution can point towards deeper differences such as information content. Figure 4.4b, which shows the sentence length distributions, was constructed in exactly the same way as Figure 4.4a. Unlike the word distribution, it is obvious that both filtering algorithms alter the distribution over sentence lengths. Whereas the distribution that arises from uniform subsampling has high variance and has significant mass up to sentence lengths of 20, the distributions from the filtering algorithms show strong biases for sentence lengths below 10.

At least for SRF, this observation has a simple possible explanation: The restriction during sampling is that a newly sampled must not contain any words of the h previously sampled sentences and shorter sentences have a higher chance of satisfying this restriction. This would lead to a bias for short sentences of the form we observe in Figure 4.4b. Since the TF shows essentially the same bias, there could, however, also be a reason connected to Zipfianness itself. We leave this unexplored for now but note that it has consequences for the filtered subcorpora. According to the given rationale, shorter sentences contain less information and the substantially decreased sentence lengths in filtered subcorpora indicate that they have lowered information contents. This may have considerable implications for learnability analyses, since it is a basic result form information theory that one way for data to be highly learnable is if it has low information content. Yet more concretely, short sentences are likely to be easier to

learn to any learner because they are less likely to be syntactically complex than their long counterparts.

Even though a bias for short sentences can be explained by the filtering algorithms' design, it is somewhat curious to find such great differences in sentence length distribution between filtering and uniform sampling, especially given no such bias in the word length distribution. Figure 4.4b reveals another factor which may have enabled such strong biases for shortness: Notice that the filtering algorithms' distributions have relatively large mass on sentence length 1 and generally a considerable mass on sentence lengths below 4. But, of course depending on the language, sentences of lengths below 4 should be rare because at such short lengths, they can barely convey any information. Finding this is likely an artefact of Wikipedia which, as an encyclopedia, contains a high number of lists, tables and enumerations, the elements of which are often parsed into separate sentences. Using a corpus other than Wikipedia or removing the shortest sentences before subsampling could hence remedy the filtering algorithms' biases.

Lexical Diversity

Besides information theoretic information content, it is of course the semantics of language which has a large influence on its communicative value. In verifying that the filtering algorithms do not impair the communicative value of the sampled subcorpora, we would like to also verify that the semantics of the subcorpora are similar. While semantics is not necessarily inherently a statistical property of language, the two are intricately connected and this has been formalised by distributional semantics (which dates back much further but is extensively discussed in Sahlgren 2006). In this framework, semantic similarity is based on statistical patterns of co-occurrence and measured in high-dimensional so-called semantic spaces into which the elements of language are projected. What we mean by similar semantics can therefore be stated in terms of the semantic space which indeed captures statistical properties of language: If the semantic spaces which arise from different corpora are, in rather vague terms for now, similar in their granularity and dimensionality, then corresponding corpora could be characterised as having similar degrees of semantic expressivity.

Unfortunately, constructing and dealing with semantic space is an involved task and beyond the scope of this thesis. So we once more turn to surrogates, which we find in measures of lexical diversity which have seen a wide range of applications across linguistic fields (for instance in clinical linguistics, e.g. Watkins et al. 1995, in analysing language learning, e.g. Malvern et al. 2004 and Foster and Tavakoli 2009, in sociolinguistics, e.g. Bradac and Wisegarver 1984, or in computational linguistics, e.g. Cybulska and Vossen 2014). Lexical diversity is commonly meant to capture the richness and variability of word usage in a given corpus and generally the converse of repetition and redundancy. Although it is a very crude and oversimplifying surrogate for the semantic expressivity of a corpus, frequent use of varied and specialised

words in that corpus would generally lead to the corresponding semantic space being densely populated and hence highly expressive.

Among the most commonly used measures for lexical diversity is the type-token ratio (TTR) of a corpus, simply dividing the size of its vocabulary by the number of its tokens. Besides being overly simple, the problem with the TTR is that it would not add anything to our discussion because of its close relation to vocabulary growth and Heap's law; notice that Heap's law describes how the TTR evolves across number of tokens. Hence, the similarities between the filtered and the uniformly subsampled corpora we have seen analysing vocabulary growth would be exactly the same when looking at their TTR. Requiring measures of lexical diversity which are not equivalent to the TTR, we choose the more sophisticated MTLD and HD-D. (P. M. McCarthy and Jarvis 2010)

HD-D, itself an analogue of voc-D but more efficient to compute, measures lexical diversity in terms how evenly frequency mass is distributed and among the words in a given corpus C^k. Formally, for each word w in C^k, it computes the probability p_w that w occurs in a randomly drawn set of words of a fixed size. This probability is given by the hypergeometric distribution and depends on $f_{C^k}(w)$, the frequency of w in C^k. HD-D is then simply the sum $\sum_w p_w$, i.e. the total probability that any of the word types in C^k occur in the randomly drawn fixed-size set of words. Thus it captures the intuition that lexical diversity should be high if the corpus C^k contains many words with high frequency and low if only few words occur frequently.

MTLD takes a different approach and measures lexical diversity by the mean length of subsequences of C^k which are above a certain fixed TTR. That is, MTLD attests high lexical diversity for C^k if there are many long sequences in C^k that have high TTR. Notice that, even though it is defined in terms of TTR, MTLD is relatively independent of the overall TRR of C^k since MTLD also depends on how evenly word types are spread across C^k. Because both HD-D and MTLD contain parameters in their definitions whose values are arbitrarily chosen and which influence their magnitude, their absolute values on given corpora are not meaningful and are therefore only to be interpreted comparatively.

Just as we have done previously we compute HD-D and MTLD for all m individual subcorpora for each sampling type and filtering parameter value and compare the resulting distributions. The means and variance of these distributions, respectively for HD-D and MTLD, are given in Table 4.5 and since the variances turned out to be negligible compared to their corresponding means, we refrain from plotting the distributions.

Generally, both HD-D and MTLD clearly indicate that filtered subsampling leads to more lexically diverse corpora and the values of both measures increase with the filtering algorithms' parameter values. Where HD-D and MTLD differ is in the ordering between filtering algorithms: Both the TF and the SRF show the same increase in HD-D across their respective parameter values and in both cases the increase is very

	UNIF	TF							SRF						
		2	6	10	14	18	22	26	2	4	8	16	32	64	81
μ HD-D	0.82	–	0.84	–	0.89	–	0.93	–	–	0.85	–	0.88	–	0.93	–
σ HD-D	0.0	–	0.01	–	0.01	–	0.0	–	–	0.01	–	0.01	–	0.01	–
μ MTLD	3090	–	4540	–	12757	–	39827	–	–	4239	–	6024	–	9944	–
σ MTLD	101	–	30	–	696	–	264	–	–	66	–	120	–	170	–

Table 4.5 Means μ and standard deviations σ of the respective distributions over HD-D and MTLD values. For each sampling type, uniform (UNIF), Typicality Filter (TF) and Speaker Restriction Filter (SRF), and each respective filtering parameter value, numbers below, the distribution over lexical diversity values is constructed from m subsamples. Missing values were omitted because calculation of both HD-D and MTLD is expensive.

moderate. As measured by MTLD, the Speaker Restriction leads to similarly moderate increase of lexical diversity but the situation is quite different for the TF. In terms of MTLD, the lexical diversity of the outcomes of the TF explodes, showing higher than linear growth across values of the parameter f.

Being defined in terms of the frequencies of the individual word types in a given corpus, HD-D in fact assigns lexical diversity in relation to the rank-frequency relationship of that corpus. The more evenly frequency mass is distributed among word types, i.e. the flatter the relationship, the higher HD-D. Due to the use of the hypergeometric distribution, HD-D is however a non-linear function of the flatness of the rank-frequency relationship and can therefore not be directly predicted from the relationship's shape. Apparently, the changes in the rank-frequency relationship that the TF and the SRF lead to (see Figure 4.1) have similar consequences for the value assigned by HD-D.

The fact that this is different when measured by MTLD is likely to be understood in connection to the differences in vocabulary growth we have observed above, see Figure 4.3. There, as well as here, does the TF lead to much more marked increase. Even though MTLD is not a direct function of the vocabulary size ((P. M. McCarthy and Jarvis 2010) report a correlation of 0.3), the increases in MTLD values are consistent with the vocabulary growths. Hence, although it is curious that the TF leads to such an explosion in MTLD values, we can find a likely explanation in terms of vocabulary size. Seeing as the MTLD is a function of the lengths of token sequences which surpass a certain fixed TTR, it is apparent that most of the sentences in the subcorpora sampled by the TF surpass that TTR individually. This entails that the individual sentences have uncommonly high lexical diversity and are likely highly complex sentences.

Once again, it is reasonable to expect that increased lexical diversity is generally also an inherent outcome of the reduced Zipfianness, i.e. more uniform rank-frequency

relationships, in the filtered subcorpora. HD-D measures lexical diversity in terms of how uniformly frequency is distributed across word types in a given corpus which is indeed how the filtering algorithms achieve lowered Zipfianness. As for MTLD, and as we have observed above, the filtering algorithms lead to increased vocabulary growth which implies higher TTR which in turn increases the chance for long token sequences with individually high TTR. Increase of MTLD in the filtered subsamples is therefore an outcome just as expected as increased vocabulary growth.

In sum, moderate increase in lexical diversity due to filtered subsampling is an expectable side-effect of lowered Zipfianness. The SRF indeed exhibits such moderate increase with respect to both HD-D and MTLD. On the other hand, the explosion in MTLD by the TF shows that it increases lexical diversity beyond what lowered Zipfianness predicts. Apparently, for the TF to sample subcorpora with more uniform word distributions, requires it to sample lexically highly diverse sentences. As it seems, only such sentences are sufficiently atypical under Zipf's law. As sentences with extreme lexical diversity are individually also atypical under language as a whole, this dampens the typicality of the resulting subcorpora. Of course, it is impossible for the TF to sample less Zipfian subcorpora without affecting the individual sentences in them. But optimally, it would do so by sampling sentences in which lexical diversity is uniformly moderately increased.

Summarising our investigation into the normality of filtered subcorpora, we have looked at their vocabulary growth, length distributions and lexical diversity and compared them to those found in uniformly sampled corpora. Returning to the rationale at the outset of this section, we argued that high similarity between filtered and uniformly sampled subcorpora in terms of these properties can be seen as evidence that filtered subcorpora retain a high degree of typicality under the language they were sampled from.

We have observed that filtering leads to evident changes in all studied properties, except the distribution over word lengths, and in all cases, except the sentence length distributions, there is a clear trend of the filtering algorithms leading to more extreme values of these properties. That is, higher values of the algorithms' parameters f and h lead to increased vocabulary growth and lexical diversity. Because of the observed differences and, even stronger, such trends, it seems necessary to conclude that the filtered corpora are not "normal", typical samples from the original language.

It is worth noting that the observed trends are consistent with decreasing Zipfianness across parameter values of the filtering algorithms and particular decreasing typicality under Zipf's law (see 4.2). This elucidates that fact that the objective of filtering is a min-max problem, namely that, as mentioned, typicality under Zipf's law is to be minimised while typicality under the original language is to be maximised, that is left unchanged with respect to uniform sampling. As mentioned throughout this Section, however, samples with minimal Zipfianness and maximal typicality under the source language are impossible to attain. The simple reason being that the distribution over words in the original language is both highly Zipfian and integral

component. As mentioned at the outset, we therefore need to shift from requiring no or little difference in statistical properties between uniformly sampled and filtered subcorpora to requiring that the magnitude of the difference is low enough and that these differences do not distort the statistical properties' basic shape.

With this relaxed requirement for normality of the filtered subcorpora, the observations of this section become more positive: We find that although vocabulary growth rates increase as a result of filtering, they stay on the same order and more importantly, Heapianness, i.e. the adequacy of Heap's law as a description of vocabulary growth, does not decrease. Similarly, we find clear but moderate increase in lexical diversity as measured by HD-D. As both vocabulary growth and lexical diversity can be expected to grow with more uniform word distributions, we are thus inclined that the observed changes are effects of the changed rank-frequency relationship itself. This, together with the fact that the distribution over word lengths, which is not obviously tied to the relationship's shape, remained largely unchanged leads us to attest relatively high normality, that is typicality, to the filtered subcorpora.

Additionally, the way in which we construct atypical samples, namely by filtered subsampling, has the pronounced, albeit trivial advantage that a high degree of normality is guaranteed: Since the filtered subcorpora are actual subsets of the original source corpus, that is they consist entirely of real sentences, it is given that they are internally valid and therefore typical. For instance, all filtered subcorpora contain only syntactically valid sentences and these are obviously more normal than syntactically invalid ones. This is also the first reason why the measures of normality we have probed in this Section are principally about information content rather than information structure; the second reason being that from a theoretical perspective, it is mainly information content which determines learnability.

At the same time, as the sentence length distributions and the use of MTLD have revealed, those sentences which do end up in filtered subcopora are largely exceptionally short and complex in terms of their lexical constituents. It is not clear why reduced Zipfianness should result in shortened sentences and similarly, the observed increase in MTLD, equivalent to an explosion, is beyond the expected effects of the changed rank-frequency relationship. These outcomes are thus created by the filtering algorithms themselves and hamper the typicality of the filtered subcorpora. As it seems, both filtering algorithms only manage to find corpora which satisfy the respective constraints by sampling excessively short and complex sentences. But, and this is the essential reason why such a bias is undesirable and hampers normality, it is certainly conceivable that corpora exist which consist of sentences with lengths and lexical complexities typical for the original language while still having reduced Zipfianness. Whether subcorpora of Wikipedia can fulfil this requirement, or whether it is made impossible by the sampling restrictions of the TF and the SRF will remain open for now.

4.4.3 Assessing Sample Diversity

As mentioned in the descriptions of the TF and SRF (see Section 4.3.1), both algorithms are equivalent to a randomised search or subcorpora which fulfil the given constraints. We have also mentioned that the constraints are too complex to predict whether such subcorpora even exist but as we have seen, they do, at least for our source corpora and the parameter values given in Table 4.1. But even knowing that such subcorpora exist, we cannot trust that they are sufficiently dissimilar, that is have low enough overlap in terms of the sentences they consist of. Besides being an indicator of the success of the filtering methodology as a whole, sufficiently low overlap of the subsamples is an important prerequisite for the Subsampling method, since otherwise variance estimates are likely to be misleading. Hence, this is what we ensure in the current Section.

Specifically, we treat each subcorpus as a multi-set of sentences and define overlap between subcorpora as multi-set similarity. To measure multi-set similarity we use the multi-set generalisation of the standard Jaccard similarity index J. The Jaccard similarity index of two sets A and B is simply defined as the size of their intersection divided by the size of their union, $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ (originally studied in Jaccard 1901). A basic inequality in set theory, $|A \cap B| \leq |A \cup B|$ and therefore J takes on values between 0 and 1 and can be regarded as a percentage. As a measure of similarity, $J(A,B) = 0$ is equivalent to A and B being disjoint, whereas $J(A,B) = 1$ if and only if $A = B$. The multi-set generalisation we use is defined as (see e.g. Kosub 2019)

$$J(A,B) = \frac{\sum_{x \in \mathcal{X}} \min(f_A(x), f_B(x))}{\sum_{x \in \mathcal{X}} \max(f_A(x), f_B(x))},$$

where $f_A(x)$ denotes the frequency of x in A and $\mathcal{X}$ is the universe that the elements in A and B are drawn from, in our case the set of sentences in the source corpus C^n. Although the multi-set generalisation may seem quite different, it reduces to the original definition of the Jaccard similarity index if A and B are sets rather than multi-sets, and importantly, the generalised definition still has the same properties.

Recall that for each sampling type and each parameter value of the filtering algorithms, we have m subsamples, $\{C_1^k, ..., C_m^k\}$. We measure the Jaccard similarity index on all pairs in this set of subsamples of which there are $\binom{m}{2}$ and since we use $m = 10$ this results in 45 pairs. The similarity values of all these pairs taken together induce a distribution, one for each sampling type and each filtering parameter and we compare these distributions across them. Since, as has been the case before, the distributions have negligible standard deviations, we give their mean values in Table 4.6 rather than plotting them.

As expected, uniform subsampling leads to subcorpora whose average Jaccard similarity indices are close to 0. That this is to be highly expected can be seen in the number of different subcorpora there are: If the source corpus C^n contains ℓ sentences

	UNIF	TF							SRF						
		2	6	10	14	18	22	26	2	4	8	16	32	64	81
sentence μ	.01	.01	.02	.02	.04	.09	.27	–	.02	.03	.03	.05	.07	.1	.12
sentence σ	.0	.0	.0	.0	.0	.01	.01	–	.0	.0	.0	.0	.01	.01	.0
token μ	.63	0.63	.61	.57	.54	.52	.51	–	.63	.63	.63	.62	.62	.62	.63
token σ	.01	.01	.01	.0	.01	.01	.0	–	.01	.01	.01	.0	.01	.0	.0

Table 4.6 Jaccard similarity $J(C_i^k, C_j^k)$ of all pairs C_i^k and C_j^k of every set of subsamples of each sampling algorithm and each algorithm parameter value: uniform (UNIF), Typicality Filter (TF) and Speaker Restriction Filter (SRF). Reported values are means (μ) and standard deviation σ of all pairwise similarities in a set of subsamples. Similarities computed on two linguistic levels, namely sentences and tokens.

from which the subcorpora are drawn and each subcorpus contains on average l sentences, then there are approximately $\binom{\ell}{l}$ possible distinct subcorpora. On our case, see Table 2.1, $\binom{\ell}{l}$ is on the order of at least $10^{100\,000}$ and each individual subcorpus has a probability on the order of $1/10^{100\,000}$ of being drawn. Even if many of this large set of subcorpora have rather high similarity, the probability of drawing high-similarity subcorpora is minuscule.

Overall, it can be positively stated that also the filtered corpora have rather low overlap with generally under 10% overlap of sentences. So apparently the filtering algorithms do manage to find genuinely different subcorpora in their randomised search. Even though these are much higher than the similarities among uniformly sampled subcorpora, keep in mind that there are far fewer subcorpora which fulfil the constraints imposed by the filtering algorithms and hence their a priori chance of having high similarity is increased.

This can be seen in the clear trend of subcorpora showing higher similarities as the filtering algorithms' parameter values increase. In the extreme case of the TF with $f = 22$, the similarity is as high as 0.27, that is the resulting subcorpora share 27% of their sentences on average. Although one could still speak of different subcorpora at such high overlap, this does reflect the fact that the amount of atypical material, that is sentences, in the source corpus is naturally limited. As the filtering restriction becomes more rigid, such material will hence included across subcorpora by the TF.

Increased overlap of the subcorpora, however much, is potentially problematic because for any quantity, such as rank or frequency, whose mean value we compute from a set of subcorpora, that mean will likely be skewed and its variance underestimated. At the same time, increased overlap on the sentence level does not imply increased overlap on all levels, as the mean token-level similarities, also given in Table 4.6, reveal. These were constructed in the same way as the sentence-level similiarities, except only that subcorpora were treated as multi-sets of tokens (rather than sentences). The fact that token level similarity is lower, if affected at all, due to filtered subsampling

can remedy the problems which come with high sentence-level overlap.

In fact, the TF with $f = 22$ leads to the lowest token-level similarity, which evidently generally is an effect of the TF but not of the SRF, a contrast which is seemingly rooted in the SRF only affecting the head of the word distribution. Meanwhile, for the TF, increased sentence-level overlap together with decreased token-level overlap can only mean that those sentences which are not shared between two subcorpora are highly dissimilar in terms of the tokens they contain.

This perhaps odd-seeming behaviour of the TF could actually be revealing a feature of Zipfian language: Samples from Zipfian language will generally have the high word-level similarity that we found in uniform subsamples from Wikipedia, about 63% or nearly two thirds. Zipf's law is itself the source for this phenomenon, since in any sample, the frequent word types will take up most of the token mass and across samples it is always the same types which are frequent. Lowered Zipfianness, as for instance due to the TF, implies reduced frequency of the frequent word types and mass shifted towards the low-frequency ones. The low-frequency types, however, are largely not shared across samples, leading to lowered overlap on the token level.

Thus, similarly to the properties we analysed in Section 4.4.2, the changes in overlap between subcorpora which arise from filtered sampling are not necessarily specific biases of the TF and SRF but can be argued to be intrinsic to lowered Zipfianness. Indeed, both (Kurumada, Meylan, and Frank 2013) and (Hendrickson and Perfors 2019) have argued that high degrees of token-level overlap between samples could be one of the most essential effects of Zipfianness on learnability of language. For any hypothetical learner, high overlap implies that upon being given a new sample the learner has already encountered the majority of that sample's lexical material in previous samples. This alleviates the sparsity issue mentioned in Section 4.1, despite not solving it, by situating new, previously unencountered lexical items in a sample in a context of majorly known items.

Chapter Conclusions

The starting point of the current chapter has been the re-evaluation of the previous studies of the learnability of Zipf's law (Kurumada, Meylan, and Frank 2013 and Hendrickson and Perfors 2019), specifically the data these studies used. Their approach, like ours, is a comparative one, that is it attempts to uncover the effect of Zipf's law by comparing Zipfian to non-Zipfian corpora. We have argued that the corpora used for comparison must preserve the defining properties of language and especially Zipfian language, most importantly the heavy and infinite tail of the word distribution. Respecting this requirement leads automatically also to respecting the distinction between source and sample, since the requirement is on the former but a learner observes the latter; the source-sample is also the most basic aspect that connects the current to the previous chapter. The source-sample distinction together with the need for a

formalisation of Zipfianness then leads quite naturally to the information-theoretic concept of typicality. In this formal framework, Zipfianness is importantly not per se a binary (i.e. Zipfian vs. non-Zipfian) but a continuous notion (i.e. more vs. less Zipfian). In sum, as we are convinced and in contrast to the previous studies, the data for the comparative approach to the effects of Zipf's law should be a set corpora which are as natural as possible and which represent varying degrees of Zipfianness. From the variation in this set, learnability trends may be estimated and these reveal the effect of Zipf's law on learnability.

With this revised request for data in mind, we have developed and evaluated two sampling algorithms (the Typicality Filter (TF) and the Speaker Restriction Filter (SRF)) which yield subsamples of an original natural corpus that are reduced in typicality with respect to Zipf's law. These algorithms correspond to non-uniform sampling schemes and explicitly introduce biases in the produced subsamples. Biases are needed because of the asymptotic equipartition property (AEP) which ensures that large subsamples of a Zipfian corpus are with high likelihood Zipfian themselves. Because the sampling algorithms' biases are too complex to be investigated analytically, we then devoted the remainder of this chapter to an empirical investigation of the properties of the produced subsamples. Because they are the most important aspects for our use in learnability analysis, we focused on analysing their Zipfianness, their naturalness (or normality) in terms of distributional properties other than the rank-frequency relationship and finally their diversity among each other.

Generally, we find that the filtered subsamples both produced by the TF and the SRF meet our requirements of relatively natural and less Zipfian corpora. The subsamples of both algorithms are by all used measures significantly less Zipfian than their uniformly subsampled counterparts. In particular, the filtered subsamples would convince a learner of not having been generated by a Zipfian language, since they are not in the typical set for Zipf's law. Just as importantly, the TF and the SRF do not seem to disrupt the naturalness of generated subsamples beyond what can be argued to be tied to Zipfianness, at least in terms of distributional properties such as lengths distributions of words or lexical diversity. This conclusion is emphatically only preliminary since we studied a few individual distributional aspects and more extensive and holistic tests are required to properly determine to what extent the naturalness of filtered subsamples is impaired. Finally, we have successfully verified that the filtered subsamples are sufficiently diverse subsets of the the original source corpus, that is their pairwise overlap is not too high. Significantly reduced Zipfianness, relatively high degrees of naturalness and low overlap in the sets of filtered subsamples lead us to conclude that the filtering algorithms we have devised, the TF and the SRF, and the general methodology of filtering indeed has the potential to produce valuable resources for the comparative approach to the effects of Zipf's law.

As the only caveat of the filtering algorithms, our empirical assessments have revealed a rather strong trade-off between Zipfianness of the filtered subsamples on one side and naturalness and diversity on the other: The more rigid the sampling bias (i.e.

the less Zipfian the outcomes), the the less natural and less diverse the set of filtered subsamples. This is a rather unsurprising trade-off and the theoretical perspective of the initial section of this chapter suggests that subsamples of extremely reduced Zipfianness should not even exist. However, it does limit how much we can practically reduce the Zipfianness of subsamples without negatively affecting naturalness and diversity too much. This trade-off also complicates the use of the filtering algorithms, as will probably need to be balanced experimentally for each individual use case.

Like the Subsampling itself, and even more so, the filtering methodology (which is derived from the Subsampling method) hinges on the availability of very large source corpora to sample from. Simply put, the larger the source corpus, the higher the chance that it contains large numbers of large and atypical subsamples. This could be seen as a limitation of the filtering methodology but we choose to take a different perspective: We see the availability of massive amounts of linguistic data nowadays a fundamental enabling factor of the filtering methodology. In fact, high effectiveness with which we are able to diminish Zipf's law, let alone that this is at all possible in practice, is a remarkable and even somewhat surprising feat. At the same time, to achieve this we have not required heavy statistical or computational machinery but merely a large enough corpus. We hope that, with this example, others will be encouraged to make similarly daring attempts at exploiting the power of massive corpora.

Analysing the Learnability of Zipf's Law

In closing the current chapter, we sketch in detail how the filtering method is used for the goal we started from: assessing the effect of Zipf's law on learnability. As should be clear by now, this assessment will be in terms of a correlation, namely the correlation between the typicality of a subcorpus with respect to Zipf's law and the measured learnability of that subcorpus. The contribution of the filtering method in constructing this correlation is to provide us with subcorpora of varying degrees of typicality.

Concretely, as we have seen in Section 4.4.1, for any typicality value t (within certain bounds), the filtering algorithms allow us to sample a subcorpus C^k with $a(C^k; P_{\alpha,\beta}) \approx t$. Given C^k, we measure its learnability, to obtain learnability value l, and we then form the correlation with t and l. By repeating this procedure with every typicality value t, i.e. by drawing multiple subsamples of the same typicality, we obtain the Subsampling estimate over the entire distribution over the correlation between typicality and learnability. (Note that taking only one subsample per value of t would lead to a point estimate of the correlation, cf. Section 3.1.) In this way, we can not only assess the directionality and strength of the correlation but additionally also the certainty we should place in it. In the same way, in comparison to the previous studies (Kurumada, Meylan, and Frank 2013 and Hendrickson and Perfors 2019) we obtain a far more detailed perspective on the learnability of Zipf's law.

At the same time, in contrast to (Kurumada, Meylan, and Frank 2013) and (Hen-

drickson and Perfors 2019), our methodology does not allow for human experiments to measure learnability, for many reasons including that humans cannot realistically be exposed to corpora of the sizes we propose to use. Hence, learnability must be measured with computational methods and these can of course only approximate human performance. Recent advancements in natural language processing and cognitive modelling have, however, shown that close approximations are possible. The learnability of a corpus as a whole can be captured by (neural) language models and it has been shown in a recent study (Gulordava et al. 2018) that these have evolved beyond only modelling conditional probabilities of short n-grams. Instead, neural models of the long short-term memory (LSTM) family seem to capable of capturing long-distance syntactic phenomena which resemble human processing performance. Moreover, the specific task of word segmentation learning investigated by (Kurumada, Meylan, and Frank 2013) has been thoroughly analysed in the field of cognitive modelling. Here, Bayesian models (Frank et al. 2010) have managed to very closely match human learning performance. With this state of the art, which keeps developing fast, computational learning models can probably provide proxies for human learning which are good enough to assess the correlation with Zipfianness and hence the effect of Zipf's law on learnability.

5 Conclusions

We begin our conclusions by summarising the contributions we have made in this thesis to quantitative linguistics and analyses into the learnability of Zipf's law. Subsequently, we actually make another final contribution: we show how the filtering methodology can be applied beyond the learnability of Zipf's law to the learnability of language in general. Here, we focus specifically on recent advances which use Kolmogorov complexity to challenge the negative perspective on learnability of language in the tradition of the work by Chomsky. Finally, we close by sketching the bigger picture and broader uses of the Subsampling and Filtering methods: First, we name some of the most immediate and salient contexts besides learnability to which both methods have the potential to contribute insights. Then, we

5.1 Contributions

Data

As the first contribution (see Section 2.1) we have published a cleaned and segmented version of Wikipedia as a corpus in seven languages. The corpora of the individual languages are both massive ($50{\cdot}10^6$ tokens each, except Esperanto which has around $38{\cdot}10^6$) and linguistically highly diverse: covering 6 of the world's language families – only Esperanto and Norwegian could be argued to both belong to the Indo-European family – and all types of morphological systems observable in human languages. Although numerous corpora based on Wikipedia have been published before (e.g. Schenkel, Suchanek, and Kasneci 2007 and Singh et al. 2012), our is, to the best of our knowledge, the first open-source one with these characteristics.

The Sample-Source Distinction and Subsampling

The second contribution of this thesis consists of the use of the Subsampling method and more generally the sample-source distinction as a methodological issue in quantitative linguistics. In Section 3.1, based on the work of (Piantadosi 2014), we have identified a fundamental problem in statistical practice in quantitative linguistics and provided a solution. Namely, researchers seem to be trying to use values of quantities in observed corpora to draw inference about the values of the same quantities in the underlying, theoretical languages. This is invalid and can be characterised as

neglecting of the distinction between sample and source, i.e. the distinction between observed corpus and underlying language. The problem it leads to in practice is that researchers simply compute the values of quantities from corpora rather than properly estimating them.

To address proper estimation, we have proposed to use the Subsampling method, a well-founded and established estimation method from statistics, to estimate linguistic quantities. The Subsampling method is straightforward to apply and is especially apt in the face of scarce of linguistic resources. Because the Subsampling method is invalid in the presence of sequential dependencies (as is the case in natural language), we have proposed in Section 3.2 to simply sample elements higher in the syntactic hierarchy of language, namely sentences rather than tokens. As we have empirically verified in the same section, this remedy seems to work well.

In Section 2.2, using estimates from the Subsampling method, we have preformed an assessment of the Zipfianness of the seven languages in our corpus. Although preliminary, this assessment is based on the most reliable and detailed estimates of the rank-frequency relationship to date. Even though such assessments have been carried out before with great detail and using sophisticated statistical machinery (e.g. Baayen 2002 or Moreno-Sánchez, Font-Clos, and Á. Corral 2016), ours is the first based on properly estimated rank-frequency relationships.

To further exemplify the use and advantages of the Subsampling method and provide first analyses of the properties of the properly estimated rank-frequency relationship, we have dedicated Sections 3.3.1 and 3.3.2 to investigations into the variance and convergence behaviour of the rank-frequency across corpora.

Concluding Chapter 3, in Section 3.4, we have generalised the methodological issue of estimation and our solution, namely the Subsampling method, to another area of quantitative linguistics: vocabulary growth and Heap's law which describes it.

The Filtering Method

Moving on to the study of the learnability of Zipf's law in Chapter 4, we have begun our work (Section 4.1) by clarifying a theoretical issue in the methodology of the previous studies of (Kurumada, Meylan, and Frank 2013) and (Hendrickson and Perfors 2019): We have elucidated that uniform distributions are irrelevant alternatives for comparison to Zipfian languages and argued that, due to the unbounded vocabularies of human languages, not many distributions over the vocabulary other than Zipf's law are in fact valid alternatives. In this way and in the vein of the sample-source distinction, we have provided a theoretical discussion of the possible alternatives to Zipf's law which are required for a comparative approach to studying the effects of the law.

Further (see Section 4.2), based on the insight that investigated alternatives should be less Zipfian languages (rather than non-Zipfian ones) and that sample-source dis-

tinction must again be respected, we have presented information-theoretic typicality as an objective formalisation of the Zipfianness of a sample. The concept of the typical set, from which typicality is derived, is mainly a tool for proofs within information theory, so a part of this contribution also consists of bringing the typical set to practical use.

As a way to avoid the asymptotic equipartition property and to be able to study corpora that are atypical under Zipf's law, in Section 4.3 we have developed two biased sampling algorithms for producing subsamples and have thereby initiated what we call the Filtering method and what can be seen as an extension to the Subsampling method. One of these algorithms, the Typicality Filter, has its sampling bias defined directly in terms of typicality and is therefore applicable to any probability distribution, not only Zipf's law. In order to verify the practical usability of the filtering algorithms, we have generated a large set of subsamples from the seven Wikipedias that cover a range of parameter values of both algorithms. As we have found in Section 4.4.1, the range of parameter values indeed leads to stark reduction in typicality with respect to Zipf's law and Zipfianness of the filtered subsamples. Additionally (Section 4.4.2), the normality of the resulting subsamples, at least in terms of the measured distributional characteristics, is relatively high. Diversity among the subsamples, i.e. pairwise overlap of sentences and tokens, has also been found to be low in Section 4.4.3 These two findings, together with the reduced Zipfianness of the filtered subsamples, have lead us to conclude that the Filtering method is generally successful in achieve its goal and the Filtering method is thus a useful contribution for application in learnability studies.

The final contribution of this thesis will be presented now; due to time-constraints it is merely speculative but we are convinced that, if carried out, the investigations we propose will be fruitful. Below, we will discuss how the methodological remarks and contributions of this thesis, namely the Subsampling and the Filtering methods, can be applied to the prominent debate of the general learnability of language.

5.2 Complexity and the General Learnability of Language

Unlike the learnability of Zipf's law, the learnability of language in general has been a prominent and long-standing puzzle of cognitive science, in particular the question how children could acquire their native language. Originating with Chomsky (Chomsky 2014), a major stance on this question has been that children must possess an innate, biologically evolved language faculty (e.g. Baker and J. J. McCarthy 1981 and Hornstein and Lightfoot 1985). Within this stance, a major theory is Universal Grammar (Chomsky 1980) which describes the language faculty to a considerable degree of specificity. The stance itself has mainly been based on poverty of the stimulus arguments (first mentioned also by Chomsky 1980) which hold that children's linguistic experiences do not provide sufficient information to precisely demarcate the

language they are trying to learn (Fodor and Crowther 2002 or Pinker 2013 among the many more). The claims of innateness have lead to a long and contentious debate in which its critics have questioned the high degree of specificity of innateness that is derived from poverty of the stimulus (Zuidema 2003, Christiansen and Chater 2008 or Clark and Lappin 2010 to name just a few).

An important formal basis for poverty of the stimulus have been seminal proofs by Gold (Gold 1967). In these, he showed that the class of context-free formal languages is not learnable based on finite samples. Both the theoretical learning framework introduced by Gold, identification in the limit, and the hypothesis that human languages are at least context-free (Pullum and Gazdar 1982) are widely accepted. Therefore, Gold's proof have been interpreted to imply that entirely cognition general learning, i.e. with no innate knowledge, of their native language by children is impossible.

Cognition-General Language Learning is Possible

In recent work, Chater and Vitányi (P. M. Vitányi and Chater 2017) have succeeded in providing a proof with the same framework of identification in the limit that cognition-general learning can indeed be possible. Concretely, they prove that a learner can in the limit precisely identify the language which generated the stream of observed linguistic material and require essentially only the assumption that the language is computable. In the interest of brevity, we need to omit the details of the proof but give at least an intuition how this result is possible, despite being in opposition to Gold's and hence perhaps surprising.

The key innovation of Chater and Vitányi's proof (and in contrast to Gold's) is a shift in the formal representation of language: Rather than turning to formal language theory and characterising a language by a formal grammar, Chater and Vitányi characterise it by a probability distribution over all possible utterances in the language (details of the definition and argument in Chater and P. Vitányi 2007 and P. M. Vitányi and Chater 2017). Notice that the different representation implies that the grammaticality of an utterance is no longer a binary but a graded concept, namely the probability the language assigns to the utterance.

Even though this property arguably makes probability distributions more general than formal grammars and might hence be expected to complicate learning even further, have a crucial property: the unity axiom, that is that all valid distribution functions have the same total mass. This allows the learner to make inferences about unobserved utterances, namely that the systematic absence of an utterance can serve as evidence that it has low grammaticality in the language to be learned. Formal grammars, in contrast, allow no such inferences since an actually grammatical utterance could have simply been absent by chance and this is indeed the basis for Gold's proof.

The formal tool Chater and Vitányi use in their proof to formalise this notion and to make such inferences possible is the theory of Kolmogorov complexity (Li and P.

Vitányi 2008 is the standard reference). Briefly put, by approximating the Kolmogorov complexity of the observed utterances, the learner can assess how typical they are with respect to a considered probability distribution. (The Kolmogorov complexity notion of typicality is indeed a generalisation of the information-theoretic typicality used in Chapter 4, see P. D. Grünwald and P. M. Vitányi 2003.) If a considered distribution is not the true source of the observed stream of utterances, this will eventually reveal itself via typicality until only the true source remains. Importantly, this happens within finite time and therefore a finite number of observed utterance and hence identification in the limit is possible.

This proof only requires that the considered probability distribution are computable and elements of a recursively enumerable set. To the extent that these assumptions are valid, the proof shows that innate knowledge about the target language is not strictly necessary and even that entirely cognition-general learning can succeed on language. However, as Chater and Vitányi emphasise themselves, the proof gives not indication about how realistic such learning is in practice, where children are faced with very precise constraints on time and computational resources. Moreover, the learning algorithm given in the proof bears no relevance for human cognitive learning. Hence, even though it remarkably opens up the possibility of cognition-general learning, the proof yields nothing more than the possibility.

Language Learning by Simplicity

In their earlier work (Chater and P. Vitányi 2007), Chater and Vitányi have in fact also addressed realistic and practically relevant cognition-general learning. This is based on the simplicity principle, which has been applied successfully across domains in cognitive science, and which results in a learner whose only learning bias is simplicity. Being the intuitive opposite of complexity, the simplicity is in terms of Kolmogorov complexity, namely as it inverse which is a probability distribution called the universal distribution. The simplicity-based learner is thus also defined in terms of Kolmogorov complexity, like the learner in the proof above.

Unlike that learner, however, the simplicity-based learner makes no attempt at identifying the true source of the observed utterances and instead has a fixed strategy for predicting them: Given a sequence of utterances $s_1, ..., s_n$, the learner predicts the continuation s_{n+1} such that the Kolmogorov complexity of $s_1, ..., s_n, s_{n+1}$ is minimised. Accordingly, learning performance is not measured in terms of successful identification but in terms of prediction error on s_{n+1}. As Solomonoff (Solomonoff 1964a and Solomonoff 1964b), the original inventor of the universal distribution, proved, the prediction error decreases faster than $1/n$ in the length n of the sequence. Moreover, the total error on the entire sequence is upper bounded by the Kolmogorov complexity of the true source which generated the sequence. In formal learning theory, these performance guarantees are seen as sufficient for learning to be in principle successful (for instance in the PAC learning framework Valiant 1984).

Notice that the performance of the simplicity-based learner depends only on the Kolmogorov complexity of the language which produces the sequence of utterances. Considering the simplicity-based learner, the learnability of a language is therefore entirely equivalent to its inverse Kolmogorov complexity – languages of higher complexity are less learnable by leading to higher prediction error. To be explicit, this constitutes the link between Kolmogorov complexity and the learnability of language and shows why it is insightful to investigate both complexity and learnability.

Now, the different view on the possible learning strategies for language acquisition put forward by Chater and Vitányi also implies a different perspective on language itself: In order to reduce the load on the learner, innateness in the form of Universal Grammar puts hard and specific restrictions on which of the vast set of all languages are potential human languages. When considering a cognition-general learner, on the other hand, there are in principle no such constraints. Instead, a language establishes itself as a potential human language by being learnable enough, that is by possessing an inherent degree of learnability high enough for the time and resource constraints faced by children. Thus, in order to assess the possibility of cognition-general learning for humans, one must investigate the general learnability of their languages. By general learnability, we mean objective and learner-independent quantifications of learnability, such as based on Kolmogorov complexity. If the Kolmogorov complexity of human languages should be revealed to be low enough, then simplicity-based learning becomes a real alternative to hypotheses based on innateness such as Universal Grammar.

But then problem remains how low is low enough and, as far as we aware, there is no direct solution to this problem. At the same time, we argue here (Christiansen and Chater 2008 provide a similar argument) that the language of simplicity-based learners, i.e. with no innate constraints such as Universal Grammar, should be particularly learnable. That is, due to the inverse relationship between learnability and Kolmogorov complexity, such language should have lower complexity than other hypothetical languages. The reasoning is that simplicity-based learners have no constraints or biases for choosing their own language other than favouring languages which are simple (and, of course, fulfil the communicative functions that is required from them. According to this reasoning, finding that human language has lower Kolmogorov complexity than its alternatives constitutes evidence that humans are simplicity-based learners. Notice that this approach is in its essence a comparative approach just like the one taken by the learnability studies of (Kurumada, Meylan, and Frank 2013) and (Hendrickson and Perfors 2019) and what we have discussed in Section 4.1.

Practical Learnability Assessments

That said, there are difficult practical problems in studying the Kolmogorov complexity, or learnability, of human language: Kolmogorov complexity is uncomputable (Li and P. Vitányi 2008), making precise measurements impossible and the absolute val-

ues of its approximations useless due to error constants of unknown magnitudes. At the same time, multiple practical approximations have been developed and are well-studied. An exceedingly simple and widely used example is the compression rate achieved by common compression algorithms (e.g. Cilibrasi and P. M. Vitányi 2005). Simplicity-based learning is in practice directly instantiated by Minimum-Description Length (MDL, see Rissanen 1983 and P. M. Vitányi and Li 2000) and has seen application in machine learning (e.g. P. Grünwald 1995) but is rather complicated to implement. Common to all practical methods is that they are bound to overestimate complexity, i.e. underestimate learnability.

In a first investigation into the possibility of simplicity-based learning, Chater and Vitányi (together with Hsu, Hsu, Chater, and P. M. Vitányi 2011) scaled the issues of measuring Kolmogorov complexity down by using MDL. To further simplify the problem, they focused on specific syntactic constraints of English, such as "I enjoy going to Italy" vs. *"I enjoy to go to Italy" (the latter is ungrammatical). The relevance of constraints such as this is that they have been used to explicitly support poverty of the stimulus arguments. It was found that the learnability of the constraints, as measured by MDL, is correlated with the confidence of adult speakers in their respective grammaticality. Such correlation would not be expected if grammaticality was governed by Universal Grammar. On the opposite, grammaticality seems corresponds mainly to probability of occurrence, and therefore learnability in MDL, as evidenced by the found correlation. This is evidence that there are no additional learning biases and that humans are simplicity-based learners.

In close connection to the argument of Section 4.1, namely that learnability studies should consider linguistic observations or corpora which show the full morphological unboundedness of human language, we argue here that approaches such as that of (Hsu, Chater, and P. M. Vitányi 2011) can only be preliminary. We believe that it is the learnability of language as whole, rather than individual aspects of it, because of the morphological and syntactic unboundedness of language. Following (Blevins, Milin, and Ramscar 2017), we are convinced that the sparsity that the unboundedness of language leads to has a significant effect on the learnability and complexity of language. An effect that will be missed by only studying individual linguistic phenomena.

Learnability Assessments via Subsampling and Filtering

But measuring the Kolmogorov complexity, or learnability, of language is inherently impossible because language itself cannot be observed. This problem brings us back to the core discussion of Chapter 3 (see Section 3.1), where we have already noted the same problem in the context of accessing the rank-frequency relationship of language, a theoretical concept. So just as before, we can use the Subsampling method to overcome this problem. Concretely, we may take any preferred way of approximating Kolmogorov complexity or learnability, for our example we will assume approximating Kolmogorov complexity by the compression method (Cilibrasi and P. M. Vitányi

2005). We then estimate the compression rate of the underlying language (n.b. not of individual corpora) by the Subsampling method, that is we take m subsamples from a large corpus, such as Wikipedia, and measuring the compression rate of each subsample. As before, the final estimate of the compression rate of the language as a whole is then simply the average of the m measured compression rates.

As stated above, the Kolmogorov complexity (or learnability) of language should be assessed in a comparative approach. Such an approach, just like before in Section 4.1, of course requires hypothetical alternatives to human language. But in the vast space of all possible human-like languages, which alternatives should we choose? For simplicity, we can actually simply use the same alternatives that we have created in the course of Chapter 4, namely languages that are less Zipfian than human language. That is, to generate such alternatives we can simply return to the Filtering method and we produce a set of filtered subcorpora that have reduced typicality with respect to Zipf's law. Just like uniformly sampled represent the human language, this set of filtered subcorpora will represent a non-human language, namely one that is less Zipfian than human language. We repeat this with different parameters of the filtering algorithms in order to obtain different degrees of typicality and thus a set of subcorpora which represent multiple non-human languages, each with a reduced but different degree of Zipfianness.

With a set of subcorpora S representing human language (i.e. uniformly subsampled) and a number of sets of subcorpora $\{T_1, ..., T_n\}$ each representing a different non-human language (i.e. sampled with a filtering algorithm), we are ready for the comparative assessment of the learnability of human language. Concretely and similar to Chapter 4, Section 4.4.3, we will correlate estimated "humanness" to estimated Kolmogorov complexity for this assessment. Humanness will be defined in terms of the typicality under Zipf's law and relative to the mean typicality of the set S (i.e. the Zipfianness of human language). So for each set T_i of non-human subcorpora, we let the humanness of the language represented by T_i be the mean typicality of S minus the mean typicality of T_i. Formally,

$$\mu(\{a(C; P_{\alpha,\beta}) \mid C \in S\}) - \mu(\{a(C; P_{\alpha,\beta}) \mid C \in T_i\}).$$

As desired, the set S, which represents human language, has humanness 0. The humanness of T_i is further from 0 than the humanness of T_j if T_i has lower mean typicality under Zipf's law than T_j. In this way, we can quantify the distance between the hypothetical languages, represented by filtered subcorpora, and human language.

As already described, the Kolmogorov complexity of language is estimated by the Subsampling method, so the estimated Kolmogorov complexity is the average compression rate of the subcorpora in S. In the same way, the mean compression rates of the subcorpora in a set T_i of non-human subcorpora constitutes the estimate of the Kolmogorov complexity of the underlying hypothetical non-human language. So finally, we construct the correlation between the humanness and the Kolmogorov complexity of languages by, for each set of subcorpora computing its humanness and its

Kolmogorov complexity estimate.

This correlation, as we have argued above, will provide evidence of whether or not humans are simplicity-based learners: If humans are indeed simplicity-based learners, then we expect a humanness of 0 (i.e. no distance from human language) to correspond to the lowest Kolmogorov complexity and Kolmogorov complexity to increase as humanness moves further away from 0. Finding no such correlation would be evidence against the hypothesis. Note, however, that some non-human languages might have lower Kolmogorov complexity than human language without this being evidence against the hypothesis: For any language to fulfil the communicative functions that human language does, a certain amount of complexity is required. Therefore, even a simplicity-based learner would not choose a language with too low Kolmogorov complexity since that language would simply not serve the purpose of communication.

5.3 Final Remarks

The focus on Zipf's law and learnability we have taken in this thesis have been because both topics provide highly relevant case studies. At the same time, the methodological concerns of this thesis go beyond them and carry over to most of quantitative and cognitive linguistics. Correspondingly, the Subsampling and Filtering methods are general and versatile enough to be applied to other domains and problems. Therefore, before closing remarks, we sketch the some of the most immediate strands of possible future work.

Other Uses of the Subsampling and Filtering Methods

One of the most immediate uses of the proper estimates obtained by the Subsampling method is to weed out the plethora of models and explanations of Zipf's law. As remarked by (Piantadosi 2014), Zipf's law can be derived from many (even mutually inconsistent) assumptions, so the ability of a proposed model to reproduce Zipf's law is weak evidence for its adequacy and this makes it difficult to decide between them. At the same time, the relatively high dispersion from the straight line predicted by Zipf's law that we have observed in the proper estimates (see Sections 2.2 and 3.3.1) will be difficult for some models to reproduce. The degree to which a model is able to reproduce that dispersion is then an indicator for its adequacy and conversely, if a model fails to predict any dispersion then it can be ruled out as an explanation for Zipf's law in language.

Going into more detail, and as already noticed by (Piantadosi 2014), the covariance of the rank-frequency relationship across words, $\mathrm{cov}_w(r(w), f(w))$ (see Section 3.2), exhibits significant structure. This structure deserves extensive investigation in itself, as it can provide a gateway into the morphological processes of language. It

can, moreover, also inform the debate about the origin and correct model of Zipf's law. Presumably, many of the proposed models are again not capable of predicting this structure and their failure to do so automatically eliminates them as correct models of why human language is Zipfian. In this way, and without even entering the debate itself, the proper estimates from the Subsampling method could help resolve the long and contentious debate about the origins of Zipf's law.

Both the Subsampling method and the Filtering method can also be used to investigate the effects of Zipf's law beyond learnability. For instance, an interesting avenue would be the interaction between Zipf's law and the semantics of language, in particular the semantic space of language. At least two accounts of the origin of Zipf's law in language have linked the law to semantic properties of language: (Manin 2008) has proposed that the law may arise as the solution to a trade-off between maximising semantic coverage of the vocabulary and minimising the combined amount of synonymity of the words in the vocabulary. (Lestrade 2017) has reproduced Zipf's law from an interaction between the sizes of part-of-speech classes in the vocabulary and the vagueness of the words in it, emphasising that neither alone is enough to result in Zipf's law.

First, by use of the Subsampling method, one could again test the validity of these accounts related to semantics, namely their ability to reproduce the structured dispersion of the rank-frequency relationship in language. Second, and yet more interestingly, the outcomes of the Filtering method, namely subcorpora atypical with respect to Zipf's law, could be investigated in terms of their semantic space: Measuring the degree synonymity of the words in them, one could directly test whether avoidance of synonymity is a necessary condition for Zipf's law and thereby evaluate the proposal of (Manin 2008). The same approach, measuring sizes of part-of-speech classes and vagueness instead, could similarly test the proposal of (Lestrade 2017). More generally yet, measuring the semantic properties of filtered subcorpora would indicate to which degree they are tied to the Zipfianness of language in the same way as the comparative approach to learnability.

Another, and perhaps the most obvious, area of future work consists of applying the Subsampling and the Filtering methods to other laws of quantitative linguistics. We have already initiated this line of work in this thesis with the example of Heap's law, see Section 3.4, which should additionally be subjected to the Filtering method. Other notable laws which come to mind include: The law of abbreviation (e.g. Bentz and Ferrer Cancho 2016) which states that frequency (or information content) of a word is in an inverse relationship to its length. Or the recently discovered law that the mutual information of characters in a text decreases as a power-law function of their distance (Lin and Tegmark 2016). In both cases, researchers have been oblivious to proper estimation and because both laws arguably have relations to learnability, applying our methods to them will likely lead to new insights.

Notice that none of the laws we have listed now are actually defined in terms of probability distributions, unlike Zipf's law. They are, however, still amenable to the

Typicality Filter, the most general filtering algorithm: As we have showed in Section 2.3, Heap's law can be turned in a distribution (in that Section we assumed a binomial distribution) that assigns probabilities to corpora and can hence be submitted to the Typicality Filter. By using similar assumptions, other laws can also be forced to assign probabilities to corpora. This highlights the generality of the Filtering method as a method to research any of the observations in quantitative linguistics.

The Big Picture

The common theme of this thesis, and in our opinion its most important point, is that the sample-source distinction must be given more attention in quantitative linguistics and neighbouring fields. Although it might be considered common sense knowledge that linguistic quantities are subject to the same random fluctuations as the corpora they are observed in, the field of quantitative linguistics generally seems to be unaware of the consequences this has for statistical practice. Especially researchers who work at the (very exciting) boundaries of empirical and theoretical science, constantly verifying theoretical models with observed data, must not be oblivious to this distinction. In contrast, one of the crucial effects of being aware of the sample-source distinction is that the problem of estimation, i.e. of approximating the values of theoretical quantities with their observed counterparts, is an unavoidable problem. As we have shown in this thesis, methods for reliable estimation such as the Subsampling method have long existed, are straightforward to apply and computationally efficient.

The Subsampling method in turn is the precursor for the Filtering method which we have introduced in this thesis. In fact, proper estimation has made the Filtering method at all possible: the random fluctuations in the observed quantities are the key to the filtering algorithms which exploit these to find atypical subsamples. Erroneous estimation, which is pervasive in the literature, hides these fluctuations and hence renders the Filtering method impossible, and even unlikely to conceive of in the first place. The example of the Filtering methodology thus further strengthens the importance of proper estimation, as the latter can even give rise to new and useful methodology that would otherwise not be considered.

In closing, we emphasise that all of the work provided in this thesis is the result of basic but also easy-to-understand considerations and insights, all of which are moreover well-known in statistics. That is, to develop the new perspective and the new findings about Zipf's law we have presented in this thesis, we have not required any heavy formal or statistical machinery and have instead relied exclusively on basic techniques from statistics and information theory. Yet at the same time, we see the outcome of this thesis as a radically different perspective on Zipf's law (and quantitative linguistics in general) and as one that can profoundly influence discussions of its origins in and effects on language. This shows how much can be gained by simply reconsidering the methodological basis of empirical practice. We hope that these considerations find the attention they deserve and that they inspire future research of similar kinds.

Bibliography

Adamic, Lada A and Bernardo A Huberman (2002). "Zipf's law and the Internet." In: *Glottometrics* 3.1, pp. 143–150.

Algoet, Paul H and Thomas M Cover (1988). "A sandwich proof of the Shannon-McMillan-Breiman theorem". In: *The annals of probability*, pp. 899–909.

Altmann, Eduardo G and Martin Gerlach (2016). "Statistical laws in linguistics". In: *Creativity and Universality in Language*. Springer, pp. 7–26.

Arshad, Sidra, Shougeng Hu, and Badar Nadeem Ashraf (2018). "Zipf's law and city size distribution: A survey of the literature and future research agenda". In: *Physica A: Statistical Mechanics and its Applications* 492, pp. 75–92.

Attardi, Giuseppe and Antonio Fuschetto (2012). *WikiExtractor 2.2*.

Baayen, R Harald (2002). *Word frequency distributions*. Vol. 18. Springer Science & Business Media.

Baker, Carl Lee and John J McCarthy (1981). "The logical problem of language acquisition". In:

Bentz, Chris and Ramon Ferrer Cancho (2016). "Zipf's law of abbreviation as a language universal". In: *Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics*. University of Tübingen, pp. 1–4.

Blevins, James P, Petar Milin, and Michael Ramscar (2017). "The Zipfian paradigm cell filling problem". In: *Perspectives on Morphological Organization*. Brill, pp. 139–158.

Bloem, P et al. (2016). "Single sample statistics: Exercises in learning from just one example". In:

Blythe, Richard A, Kenny Smith, and Andrew DM Smith (2010). "Learning times for large lexicons through cross-situational learning". In: *Cognitive Science* 34.4, pp. 620–642.

Bradac, James J and Randall Wisegarver (1984). "Ascribed status, lexical diversity, and accent: Determinants of perceived status, solidarity, and control of speech style". In: *Journal of Language and Social Psychology* 3.4, pp. 239–255.

Chater, Nick and Paul Vitányi (2007). "'Ideal learning'of natural language: Positive results about learning from positive evidence". In: *Journal of Mathematical psychology* 51.3, pp. 135–163.

Chomsky, Noam (1980). "Rules and representations". In: *Behavioral and brain sciences* 3.1, pp. 1–15.

— (2014). *Aspects of the Theory of Syntax*. Vol. 11.

Christiansen, Morten H and Nick Chater (2008). "Language as shaped by the brain". In: *Behavioral and brain sciences* 31.5, pp. 489–509.

Cilibrasi, Rudi and Paul MB Vitányi (2005). "Clustering by compression". In: *IEEE Transactions on Information theory* 51.4, pp. 1523–1545.
Clark, Alexander and Shalom Lappin (2010). *Linguistic Nativism and the Poverty of the Stimulus*. John Wiley & Sons.
Corominas-Murtra, Bernat and Ricard V Solé (2010). "Universality of Zipf's law". In: *Physical Review E* 82.1, p. 011102.
Corral, Alvaro et al. (2009). "Universal complex structures in written language". In: *arXiv preprint arXiv:0901.2924*.
Cover, Thomas M and Joy A Thomas (2012). *Elements of information theory*. John Wiley & Sons.
Cox, David Roxbee and P.A.W Lewis (1966). "The statistical analysis of series of events". In: *Monographs on Applied Probability and Statistics*.
Cybulska, Agata and Piek Vossen (2014). "Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution." In: *LREC*, pp. 4545–4552.
Davis, Mark and L Iancu (2012). "Unicode text segmentation". In: *Unicode Standard Annex* 29.
DeGroot, Morris H and Mark J Schervish (2012). *Probability and statistics*. Pearson Education.
Deluca, Anna and Corral (2013). "Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions". In: *Acta Geophysica* 61.6, pp. 1351–1394.
Dryer, Matthew S. and Martin Haspelmath, eds. (2013). *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://wals.info/.
Efron, Bradley and Charles Stein (1981). "The jackknife estimate of variance". In: *The Annals of Statistics*, pp. 586–596.
Efron, Bradley and Robert Tibshirani (1986). "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy". In: *Statistical science*, pp. 54–75.
Egmond, Marolein van (2018). *Zipf's law in aphasic speech: An investigation of word frequency distributions*. Vol. 513. LOT.
Farmer, J Doyne and John Geanakoplos (2008). "Power laws in economics and elsewhere". In: *Santa Fe Institute*.
Fenk-Oczlon, Gertraud and August Fenk (1999). "Cognition, quantitative linguistics, and systemic typology". In: *Linguistic Typology* 3.2, pp. 151–177.
Ferrer i Cancho, Ramon and Ricard V Solé (2001). "Two regimes in the frequency of words and the origins of complex Lexicons: Zipf's law revisited". In: *Journal of Quantitative Linguistics* 8.3, pp. 165–173.
— (2003). "Least effort and the origins of scaling in human language". In: *Proceedings of the National Academy of Sciences* 100.3, pp. 788–791.
Ferrer-i-Cancho, Ramon and Brita Elvevåg (2010). "Random texts do not exhibit the real Zipf's law-like rank distribution". In: *PLoS One* 5.3.
Fodor, Janet Dean and Carrie Crowther (2002). "Understanding stimulus poverty arguments". In: *The linguistic review* 18.1-2, pp. 105–145.
Font-Clos, Francesc, Gemma Boleda, and Alvaro Corral (2013). "A scaling law beyond Zipf's law and its relation to Heaps' law". In: *New Journal of Physics* 15.9, p. 093033.

Foster, Pauline and Parvaneh Tavakoli (2009). "Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity". In: *Language learning* 59.4, pp. 866–896.
Frank, Michael C et al. (2010). "Modeling human performance in statistical word segmentation". In: *Cognition* 117.2, pp. 107–125.
Gerlach, Martin and Eduardo G Altmann (2014). "Scaling laws and fluctuations in the statistics of word frequencies". In: *New Journal of Physics* 16.11, p. 113010.
Gnedenko, Boris V (2018). *Theory of probability*. Routledge.
Gobbo, Federico (2017). "Are planned languages less complex than natural languages?" In: *Language Sciences* 60, pp. 36–52.
Gold, E Mark (1967). "Language identification in the limit". In: *Information and control* 10.5, pp. 447–474.
Goldstein, Michel L, Steven A Morris, and Gary G Yen (2004). "Problems with fitting to the power-law distribution". In: *The European Physical Journal B-Condensed Matter and Complex Systems* 41.2, pp. 255–258.
Grünwald, Peter (1995). "A minimum description length approach to grammar inference". In: *International Joint Conference on Artificial Intelligence*. Springer, pp. 203–216.
Grünwald, Peter D and Paul MB Vitányi (2003). "Kolmogorov complexity and information theory. With an interpretation in terms of questions and answers". In: *Journal of Logic, Language and Information* 12.4, pp. 497–529.
Gulordava, Kristina et al. (2018). "Colorless green recurrent networks dream hierarchically". In: *arXiv preprint arXiv:1803.11138*.
Gutenberg, Beno and Charles F Richter (1944). "Frequency of earthquakes in California". In: *Bulletin of the Seismological Society of America* 34.4, pp. 185–188.
Hart, Michael (1992). "The history and philosophy of Project Gutenberg". In: *Project Gutenberg* 3, pp. 1–11.
Hendrickson, Andrew T and Amy Perfors (2019). "Cross-situational learning in a Zipfian environment". In: *Cognition* 189, pp. 11–22.
Hornstein, Norbert and David Lightfoot (1985). "Explanation in linguistics. The logical problem of language acquisition". In:
Hsu, Anne S, Nick Chater, and Paul Vitányi (2013). "Language Learning From Positive Evidence, Reconsidered: A Simplicity-Based Approach". In: *Topics in cognitive science* 5.1, pp. 35–55.
Hsu, Anne S, Nick Chater, and Paul MB Vitányi (2011). "The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis". In: *Cognition* 120.3, pp. 380–390.
Jaccard, Paul (1901). "Étude comparative de la distribution florale dans une portion des Alpes et des Jura". In: *Bull Soc Vaudoise Sci Nat* 37, pp. 547–579.
Kilgarriff, Adam (2005). "Language is never, ever, ever, random". In: *Corpus linguistics and linguistic theory* 1.2, pp. 263–276.
Köhler, Reinhard and Gabriel Altmann (2005). "Aims and methods of quantitative linguistics". In: *Problems of Quantitative Linguistics*, pp. 12–42.

Kosub, Sven (2019). "A note on the triangle inequality for the Jaccard distance". In: *Pattern Recognition Letters* 120, pp. 36–38.
Kurumada, Chigusa, Stephan C Meylan, and Michael C Frank (2013). "Zipfian frequency distributions facilitate word segmentation in context". In: *Cognition* 127.3, pp. 439–453.
Lehmann, Erich L and George Casella (2006). *Theory of point estimation*. Springer Science & Business Media.
Lestrade, Sander (2017). "Unzipping Zipf's law". In: *PloS one* 12.8, e0181987.
Li, Ming and Paul Vitányi (2008). *An introduction to Kolmogorov complexity and its applications*. Vol. 3. Springer.
Lin, Henry W and Max Tegmark (2016). "Critical behavior from deep dynamics: a hidden dimension in natural language". In: *arXiv preprint arXiv:1606.06737*.
Lu, Edward T and Russell J Hamilton (1991). "Avalanches and the distribution of solar flares". In: *The astrophysical journal* 380, pp. L89–L92.
MacWhinney, Brian (2014). *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.
Mahowald, Kyle et al. (2013). "Info/information theory: Speakers choose shorter words in predictive contexts". In: *Cognition* 126.2, pp. 313–318.
Malvern, David et al. (2004). *Lexical diversity and language development*. Springer.
Mandelbrot, Benoit (1953). "An informational theory of the statistical structure of language". In: *Communication theory* 84, pp. 486–502.
Manin, Dmitrii Y (2008). "Zipf's law and avoidance of excessive synonymy". In: *Cognitive Science* 32.7, pp. 1075–1098.
McCarthy, Philip M and Scott Jarvis (2010). "MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment". In: *Behavior research methods* 42.2, pp. 381–392.
McEnery, Tony and Andrew Hardie (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
McFadden, Daniel et al. (1973). "Conditional logit analysis of qualitative choice behavior". In: *Frontiers in econometrics*, pp. 104–142.
Metropolis, Nicholas and Stanislaw Ulam (1949). "The monte carlo method". In: *Journal of the American statistical association* 44.247, pp. 335–341.
Moreno-Sánchez, Isabel, Francesc Font-Clos, and Álvaro Corral (2016). "Large-scale analysis of Zipf's law in English texts". In: *PloS one* 11.1, e0147073.
Nelder, John Ashworth and Robert WM Wedderburn (1972). "Generalized linear models". In: *Journal of the Royal Statistical Society: Series A (General)* 135.3, pp. 370–384.
Petersen, Alexander M et al. (2012). "Languages cool as they expand: Allometric scaling and the decreasing need for new words". In: *Scientific reports* 2, p. 943.
Piantadosi, Steven T (2014). "Zipf's word frequency law in natural language: A critical review and future directions". In: *Psychonomic bulletin & review* 21.5, pp. 1112–1130.
Piantadosi, Steven T, Harry Tily, and Edward Gibson (2011). "Word lengths are optimized for efficient communication". In: *Proceedings of the National Academy of Sciences* 108.9, pp. 3526–3529.

Pinker, Steven (2013). *Learnability and cognition: The acquisition of argument structure*. MIT press.
Politis, Dimitris N, Joseph P Romano, and Michael Wolf (1999). *Subsampling*. Springer Science & Business Media.
Powers, David MW (1998). "Applications and explanations of Zipf's law". In: *Proceedings of the joint conferences on new methods in language processing and computational natural language learning*. Association for Computational Linguistics, pp. 151–160.
Pullum, Geoffrey K and Gerald Gazdar (1982). "Natural languages and context-free languages". In: *Linguistics and Philosophy* 4.4, pp. 471–504.
Renouf, Antoinette, Andrew Kehoe, and Jayeeta Banerjee (2007). "WebCorp: an integrated system for web text search". In: *Corpus linguistics and the web*. Brill Rodopi, pp. 47–67.
Al-Rfou, Rami, Bryan Perozzi, and Steven Skiena (2013). "Polyglot: Distributed word representations for multilingual nlp". In: *arXiv preprint arXiv:1307.1662*. URL: https://polyglot.readthedocs.io.
Rissanen, Jorma (1983). "A universal prior for integers and estimation by minimum description length". In: *The Annals of statistics*, pp. 416–431.
Rubinstein, Reuven Y and Dirk P Kroese (2016). *Simulation and the Monte Carlo method*. Vol. 10. John Wiley & Sons.
Sahlgren, Magnus (2006). "The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces". Doctoral dissertation.
Schenkel, Ralf, Fabian Suchanek, and Gjergji Kasneci (2007). "YAWN: A semantically annotated Wikipedia XML corpus". In: *Datenbanksysteme in Business, Technologie und Web (BTW 2007)–12. Fachtagung des GI-Fachbereichs" Datenbanken und Informationssysteme"(DBIS)*.
Schwarz, Gideon et al. (1978). "Estimating the dimension of a model". In: *The annals of statistics* 6.2, pp. 461–464.
Seabold, Skipper and Josef Perktold (2010). "Statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*.
Shannon, Claude Elwood (2001). "A mathematical theory of communication". In: *ACM SIGMOBILE mobile computing and communications review* 5.1, pp. 3–55.
Simon, Julian L and Peter Bruce (1991). "Resampling: A tool for everyday statistical work". In: *Chance* 4.1, pp. 22–32.
Singh, Sameer et al. (2012). "Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia". In: *University of Massachusetts, Amherst, Tech. Rep. UM-CS-2012* 15.
Smith, Adrian FM and Alan E Gelfand (1992). "Bayesian statistics without tears: a sampling–resampling perspective". In: *The American Statistician* 46.2, pp. 84–88.
Solomonoff, Ray J (1964a). "A formal theory of inductive inference. Part I". In: *Information and control* 7.1, pp. 1–22.
— (1964b). "A formal theory of inductive inference. Part II". In: *Information and control* 7.2, pp. 224–254.

Valiant, Leslie G (1984). "A theory of the learnable". In: *Communications of the ACM* 27.11, pp. 1134–1142.
Vitányi, Paul MB and Nick Chater (2017). "Identification of probabilities". In: *Journal of mathematical psychology* 76, pp. 13–24.
Vitányi, Paul MB and Ming Li (2000). "Minimum description length induction, Bayesianism, and Kolmogorov complexity". In: *IEEE Transactions on information theory* 46.2, pp. 446–464.
Vogt, Paul (2012). "Exploring the robustness of cross-situational learning under Zipfian distributions". In: *Cognitive Science* 36.4, pp. 726–739.
Watkins, Ruth V et al. (1995). "Measuring children's lexical diversity: Differentiating typical and impaired language learners". In: *Journal of Speech, Language, and Hearing Research* 38.6, pp. 1349–1355.
Wen, Liu (1991). "An analytic technique to prove Borel's strong law of large numbers". In: *The American mathematical monthly* 98.2, pp. 146–148.
Willis, John C and G Udny Yule (1922). *Some statistics of evolution and geographical distribution in plants and animals, and their significance*.
Zipf, George Kingsley (1932). "Selected studies of the principle of relative frequency in language". In:
— (1949). "Human behavior and the principle of least effort." In:
Zuidema, Willem H (2003). "How the poverty of the stimulus solves the poverty of the stimulus". In: *Advances in neural information processing systems*, pp. 51–58.