Christian Bentz

Adaptive Languages

Trends in Linguistics Studies and Monographs

Volume 316

Christian Bentz

Adaptive Languages

An Information-Theoretic Account of Linguistic Diversity

DE GRUYTER
MOUTON

ISBN 978-3-11-055758-9
e-ISBN (PDF) 978-3-11-056010-7
e-ISBN (EPUB) 978-3-11-055777-0
ISSN 1861-4302

Library of Congress Control Number: 2018941192.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Printing and binding: CPI books GmbH, Leck

www.degruyter.com

für Dani und Krümel

Preface

This book grew out of my PhD thesis submitted to the University of Cambridge in October 2015. I owe this incredible opportunity to Paula Buttery, the best *Doktormutter* I can imagine. You gave me the freedom to collect the pieces to the puzzle and helped me put them together – when I was puzzled.

I thank Gerhard Jäger, Ted Briscoe, Volker Gast, and an anonymous reviewer for feedback on earlier drafts of this manuscript, as well as Ramon Ferrer-i-Cancho and John Nerbonne for their critical reading of Chapter 4 in particular. The help of Julie Miess and Olena Gainulina at Mouton de Gruyter is also gratefully acknowledged. Moreover, this book would not exist in the current format without critical discussion and collaboration with the following people in alphabetical order: Dimitrios Alikaniotis, Aleksandrs Berdicevskis, Andrew Caines, Morten Christiansen, Michael Cysouw, Ramon Ferrer-i-Cancho, Felix Hill, Gerhard Jäger, Douwe Kiela, Klaus-Peter Konerding, Johanna Nichols, Tatyana Ruzsics, Tanja Samardžić, Annemarie Verkerk, Bodo Winter, and Mei-Shin Wu.

I also greatly profited from discussions at meetings of the EVOLAEMP group: Armin Buch, Johannes Dellert, Marisa Koellner, Roland Mühlenbernd, Taraka Rama, Johannes Wahle, Søren Wichmann, and Igor Yanovich. The *Words and Bones aficionados* Abel Bosman, Monika Doll, Mark Grabowski, Katerina Harvati-Papatheodorou, Lumila Menéndez, Hugo Reyes-Centeno, Yonatan Sahle, Sebastian Scheiffele, and Matthias Urban have created an interdisciplinary working environment for me during the writing of this book.

I would also like to acknowledge the painstaking work of the researchers that built the parallel copora used in the analyses of this book: Phillip Koehn (Europarl Corpus), the *UDHR in Unicode* project, and Thomas Mayer and Michael Cysouw (Parallel Bible Corpus).

Finally, thanks to my family and friends at home in Kuhardt. In particular to my parents and my sister for being role models, and for letting me study such practical and truly important subjects as Philosophy and German Literature. Please keep collecting FAZ articles about unemployed academics for me.

Dani, this book deals with texts written in more than 1000 languages. None of them has evolved the words to express my gratitude to you. It is impossible to quantify how much you helped me with my studies, my PhD, and my life in general. All of it would fall apart without you.

CB
Tübingen, April 2018

DOI 10.1515/9783110560107-201

Contents

Abbreviations

General Terminology

AoFP	Age of First Production
BLUPs	Best Linear Unbiased Predictors
CAS	Complex Adaptive System
FLB	Faculty of Language in a Broad sense
FLN	Faculty of Language in a Narrow sense
FLD	Faculty of Language, Derived components
L1	First Language
L2	Second Language
LD	Lexical Diversity
ML	Maximum Likelihood
REML	Restricted Maximum Likelihood
ModE	Modern English
OE	Old English
POS	Part Of Speech
UG	Universal Grammar

Data Sources

ASJP	Automated Similarity Judgement Program
EPC	European Parliament Corpus
OSC	Open Subtitles Corpus
PBC	Parallel Bible Corpus
UDHR	Universal Declaration of Human Rights
WALS	World Atlas of Language Structures

Glosses

Glosses given for example sentences adhere to the Leipzig glossing rules.

DOI 10.1515/9783110560107-202

List of Tables

DOI 10.1515/9783110560107-203

List of Figures

DOI 10.1515/9783110560107-204

1 Introduction

Words have two sides. The sounds, letters and signs they are shaped of are like an imprint on a coin, a pattern on a physical surface that carries information. Once this materialized information is perceived by a human brain, multiple associations are activated and give rise to meaning. Languages of the world harness this information carrying potential, sending messages from speaker to hearer, from writer to reader, and from signer to perceiver, across spaces of centimetres, metres, kilometres, and around the globe.

Despite this fundamental communicative property of all human languages, they vastly differ in terms of the encoding strategies adopted. Basic vocabularies denoting the world around us differ according to the environment we are surrounded by and the cultural practices we engage in. Complex words are built by adding a wide range of prefixes, suffixes, infixes and circumfixes to word roots, thus modifying the information encoded. Words can be arranged in different orders and with different mutual dependencies in phrases and sentences. Each of the circa 8000 languages in the world represents a unique combination of these strategies of information encoding.

This book focuses on *words* and *word forms* as basic information encoding units. It proposes an account to measure and explain the *lexical diversity* (LD) of languages across the world. Lexical diversity is defined based on the number of unique word types and their token frequencies in written texts. This definition is mainly driven by considerations of computational feasibility and reproducibility of results. However, it can be amended to reflect more fine-grained typological perspectives.

Lexical diversity in this sense is influenced by a host of different factors. For instance, the basic vocabulary used to categorize the world varies across languages. Franz Boas has famously pointed out that the “Eskimo” language distinguishes several types of snow:

> Here we find one word, *aput*, expressing SNOW ON THE GROUND; another one, *qana*, FALLING SNOW; a third one, *piqsirpoq*, DRIFTING SNOW; and a fourth one, *qimuqsuq*, A SNOWDRIFT. (Boas, 1911, p. 26-27)

He argued that fine-grained lexical distinctions “to a certain extent depend upon the chief interests of people”. His claim about specialized vocabulary for snow has been exaggerated in subsequent publications by other researchers and circulated to the general public by media outlets. This has stirred the ongoing debate about the so-called “great Eskimo vocabulary hoax” (Pullum, 1989). The actual number of words for snow in Eskimo-Aleut languages, and other languages alike, will cru-

DOI 10.1515/9783110560107-001

cially depend on how we define “words” and “word roots”. If word forms built by derivational processes are included in the count, the numbers will quickly “snowball” into the hundreds (Martin, 1986). Having said this, a thorough examination of Boas’ fieldwork with the Inuit on Baffin Island by Krupnik and Müller-Wille (2010) puts claims about an alleged hoax into perspective:

> Boas certainly knew *more* than the four Eskimo terms for snow that he cited as an example of the “differences in how the groups of ideas are expressed by specific phonetic groups in different languages” (Boas 1911:25). [...] He could have easily picked several more words for snow from his Baffin Island lexicon, like *axilokoq/aqilokoq* – “softly falling snow,” *mauja* – “soft snow on the ground,” *piegnartoq* – “the snow (which is) good for driving sled,” or from Erdmann’s Labrador dictionary that had more terms, such as *pukak* – “crystaline snow,” *sakketok* – “fresh fallen snow,” *machakit (masak/masayak)* – “wet, mushy snow.” Hence, Martin’s (1986:418) criticism that “Boas makes little distinction among “roots,” “words,” and “independent terms” is a gross misinterpretation, as Boas was very careful not to use any derivative snow terms to illustrate his point. (Krupnik and Müller-Wille, 2010, p. 390-391)

Based on dictionaries in overall ten Eskimo-Aleut languages including varieties of Inuktitut, Kalaallisut, Yupik, and Inupiaq, they further illustrate that the numbers of words for snow range from around 10 to 50 without derivatives, and from around 10 to 100 including derivatives (Krupnik and Müller-Wille, 2010, p. 389). Comparing these numbers with English vocabulary, they conclude:

> If we are to count independent stems only, the diversity of the English snow nomenclature is indeed quite limited, compared not only to the Inuit/Eskimo but also to several other languages, including Indo-European ones, spoken by people having greater exposure to snow and severe winter conditions. [...] This illustrates that some languages with more (or longer) exposure to snow and/or sea ice than English naturally develop detailed and meaningful terminologies for those phenomena that are of practical value to its speakers [...]. (Krupnik and Müller-Wille, 2010, p. 394)

In a quantitative account based on text collections and twitter data, Regier et al. (2016) further corroborate this point, showing that languages spoken in warmer climates tend to use the same word type for both *ice* and *snow*, and thus being lexically underspecified compared to languages spoken in climates where more subtle distinctions are relevant.

Words for snow are certainly the most famous example of how the lexicon of a language reflects subtleties of the environment relevant for the survival and thriving of a human population. Trudgill (2011, p. xx) gives a range of further examples. These include North Saami reindeer herders, whose language exhibits intricate distinctions between *njiŋŋelas* ‘female reindeer’, *čoavjjet* ‘pregnant female reindeer’, *čoavččis* ‘a female reindeer who has lost her calf late in spring, in summer, or as late as autumn’, *stáinnat* ‘female reindeer which never calws’, *čearpmat-eadni* ‘female reindeer which has lost its calf of the same year but is accompanied by the previous year’s calf’, alongside many others (Magga, 2006, p. 26).

In a completely different ecological setting – central Brazilian Amazonia – the speakers of Kayapó have so-called folk-taxonomies of many different species of insects, which are relevant for their everyday life. These folk-taxonomies often match western scientific taxonomies closely. However, on top of mere classification, the terms used are associated with rich knowledge regarding flight patterns, aggressive behaviour, sounds produced, etc. (Posey, 2002, p. 107). The detailed ecological insights of the Kayapó, reflected in their language, is thus also a rich source for science.

Many more examples of this type can be found in book-length treatises of linguistic diversity and its decline, such as Nettle and Romaine (2000) and Evans (2011).[1] In fact, given the rich environmental knowledge reflected in languages, these authors argue that preservation of linguistic diversity is not just of paramount importance for students of language, but also tightly interwoven with the preservation of cultural and biological diversity.

The studies discussed so far mainly focus on the range of word types reflecting subtle differences in the environment. On top of this, Derungs and Samardžić (2017) also illustrate that token frequencies found in writing can reflect the spatial salience of landmarks in the environment. In a large collection of texts about Swiss alpine history, mountains that are locally salient are also featured more prominently in the relevant literature.

Moreover, lexical diversity in the sense of this book is not only driven by basic vocabulary, but also by patterns of word-formation. For instance, many languages encode information about number, gender and case in a multitude of different articles, e.g. German *der, die, das, dem, den, des* or Italian *il, la, lo, i, le, li, gli*, while in English there is only one definite article *the* and in Mandarin Chinese there is none. Regular and irregular morphological marking strategies yield whole batteries of word forms derived from a single lemma.[2] In English, the lemma *go* is inflected according to tense, number, and aspect, thus yielding five verb forms *go*, *goes*, *went*, *gone*, *going*. This is a rather manageable set of word forms per lemma. In German, the set somewhat increases to ten verb forms: *gehen*, *gehe*, *gehst*, *geht*, *ging*, *gingst*, *gingen*, *gingt*, *gegangen*, *gehend*.

[1] See also the online article “When grasshopper means lightning: how ecological knowledge is encoded in endangered languages” by David Stringer at https://medium.com. Last accessed on 04/04/2018.

[2] This term is sometimes used interchangeably with *lexeme* or *citation form*. Here, the term *lemma* is used since this is linked with the computational process of *lemmatization*.

This is still dwarfed by claims about excessive numbers of word forms in other languages. Kibrik (1998, p. 467) calculates that in the Nakh-Daghestanian language Archi a single verb lemma can explode into 1,502,839 forms. This is due to a complex interplay of gender, mood, tense, and number inflections, as well as case-marked participles.[3] However, reminiscent of the words for snow controversy, exact counts of word forms here too depend on typological definitions that are constantly under debate. Therefore, such numbers are necessarily controversial. With regards to Archi, Chumakina (2011, p. 9) reminds us that the tense forms counted by Kibrik are often built periphrastically involving forms of the verb 'be', e.g. *i* (present) and *edi* (past). Following this rationale, constructions with an emphasis in English, as in 'I *did* listen', could also be counted as separate tense forms.

Despite such disagreement in detail, typologists will agree more generally that there are indeed systematic differences between languages. Some use inflection, derivation and compounding productively, while others are almost bare of complex word-formation patterns. Gil (2009, p. 23) gives Riau Indonesian as an extreme example of a language almost entirely lacking grammatical marking. The phrase *ayam makan*, literally 'chicken eat' can denote situations that in English would be variously described as 'the chicken is eating', 'the chickens that were eaten', 'the reason chickens eat', and several others. The lack of more explicit grammatical markers leaves the burden on the hearer to decode the intended meaning by integrating co-textual and contextual information.

There is hence a wide variety of encoding strategies to discover across languages of the world today. Also, encoding strategies are not stable over time. They change in the history and evolution of languages, sometimes quite rapidly. Around 1000 years ago, scribes of Old English still used a whole range of definite articles: *se*, *seo*, *þæt*, *þa*, *þæs*, *þæm*, *þone*, *þære*, *þæra*, *þara*, *þam*, *þy*, *þon*. Thus matching, and even exceeding the article abundance in Modern German. However, within a few hundred years, during the Middle English period, these collapsed into *the* single definite article found in Present Day English.

A fundamental question arising from such patterns of diversity and change is: why do human languages "choose" from a panoply of different encoding strategies, rather than all converging onto an optimal one? And why do they "abandon" certain strategies that seemed viable at an earlier point in time? This book argues that the answer has to be sought within a framework that views languages as *Complex Adaptive Systems* (CAS), shaped by specific learning scenarios, and the preferences of speakers, writers, and signers. This framework hinges upon a

[3] Note that including such deverbal nominal forms in the German count would also further increase the number of "verb forms", e.g. *gehender*, *gehende*, *gehenden*, etc.

clear definition of what we mean by "languages" – a non-trivial issue, surrounded by some of the fiercest debates in modern linguistics.

Chapter 2 sketches the CAS view on language as represented in the literature since the early 1990s. This overview starts with the most basic level of analysis: individuals learning an idiolect, i.e. at the *ontogenetic level*. In a second step, interactions between speakers of different idiolects inevitably lead to the formation of dialects and languages at the *glossogenetic* level. Crucially, this constitutes a definition of "languages" – in the plural – as an *accumulation of linguistic interactions between speakers*. This is contrasted with the traditional definition, focusing on the universal core of "language" in the singular, which is mostly known as Universal Grammar. Furthermore, with reference to the *phylogenetic level*, i.e. the development of language(s) on an evolutionary time scale, the usage-oriented definition opens up new perspectives on the co-evolution of population structure and language structure.

Chapter 3 further clarifies the theoretical framework, more specifically, how language change can be modelled within a CAS account. Since languages are defined on the basis of speaker populations, factors such as population size, percentage of second language learners, and language status start to emerge as important predictors of language structure. These are, in turn, intertwined with historical phenomena, most importantly population drift and dispersal. Drift and dispersal cause the unfolding of proto-languages into branches of genealogical trees, and clusters of geographical grouping. From this perspective, the widely used terminology of "external" and "internal" factors of change is re-examined, and eventually replaced by *descriptive*, *explanatory* and *grouping* factors.

Any statistical modelling of the interaction between population structure and language structure comes with challenging decisions on how to represent languages and the linguistic structures we are interested in more specifically. This book takes a quantitative and information-theoretic approach (Chapter 4). Languages are represented by parallel corpora, that is, the same texts translated into more than 1000 languages. This strategy yields (approximately) constant content in the written material used. Different topics, registers, and styles otherwise confound structural differences between encoding strategies. Based on this parallel text material, the *distribution of word tokens over word types* is here taken as a reflection of lexical diversity, a core information-theoretic property of languages. Lexical diversity lends itself to a large-scale quantitative account, since it does not carry much theoretical "baggage". However, at a minimum, we have to engage with the issue of a coherent wordhood definition, which is a non-trivial issue from a typological point of view.

There is a whole range of LD measures that could, in principle, be used to pin down the exact numbers to work with in further analyses. Some of these are discussed and tested. The measure finally settled on is Shannon entropy (Shannon and Weaver, 1949). Claude Shannon founded modern information theory by quantifying the amount of information that any code can carry. While widely applied in fields such as physics, engineering, and computer science, his findings have not made it into mainstream linguistics. The most important reasons for this neglect appear to be: a) a general dissociation between probabilistic and formal accounts of language structure since the 1950s; b) the refutation of Markov models as models for natural language; and c) the difference between information in an engineering sense and meaning in a linguistic sense. At closer inspection in Chapter 4, none of these reasons discredit information theory as a useful framework to analyse linguistic diversity.

By their very nature languages are codes, and their information encoding capacity is reflected in their entropy. To better understand and quantify this encoding capacity, word entropy estimation methods are tested on a subsample of the parallel texts. They are then applied to calculate word entropies for the full sample of 1833 parallel texts, written in 1217 different languages defined by ISO 639-3 codes. These word entropies per language reflect the range of linguistic encoding strategies to be modelled and explained in this book.

The first level of explanation are *descriptive*, i.e. language “internal” factors. The central question in this context is how information-theoretic properties of texts relate to more well-known linguistic notions such as writing systems, word-formation patterns, as well as registers and styles. Chapter 5 systematically analyses the impact that each of these descriptive factors has on word entropy. To this end, the scripts most widely represented in the texts, e.g. Latin, Greek, Cyrillic, Devanagari, Arabic, and some others are briefly discussed. It is shown that these only minimally influence the range of word forms per text and language. In contrast, word-formation patterns, most prominently inflectional marking, but also derivational morphology, clitics, contractions and compounds, strongly influence word entropy. Also, different registers and styles are represented in the parallel text sample. These emerge as serious confounds when comparing word entropy across different languages.

Going beyond descriptive analyses, one of the core tenets of this book is that linguistic structure can be meaningfully linked to population structure. In fact, within the CAS model, characteristics of speaker populations are the central *explanatory factors* for understanding the rich diversity of languages. In Chapter 6, population size, percentage of second language learners, and language status are introduced as predictors of word entropy. Firstly, the literature on qualitative and quantitative evidence for “external” influence on language structure is reviewed. Secondly, basic correlational metrics illustrate potential links between word entropy and these three explanatory factors.

Furthermore, a third kind of factor, here referred to as *grouping factor*, is assessed in Chapter 7. Grouping pertains to the clustering of languages into families and areas. Due to clustering in these genealogical and geographic dimensions, languages are not to be seen as independent data points. Rather, they evolve as the outcome of deep histories of population contact, migration, drift and dispersal. Levels of clustering are powerful predictors of language structure in general, and word entropy in particular. A direct consequence of this observation is that language families and areas have to be taken into account using more advanced statistical models.

Chapter 8 reports the results of two such models: a *multiple regression model* and a *linear mixed-effects model*. The multiple regression model combines the population level predictors, i.e. population size, second language learner percentage, and language status in a single model. This helps to establish whether each of the predictors contributes independently to explain variation in word entropy, or if there is shared variance between predictors. The linear mixed-effects model, in turn, is a statistical tool to account for non-independence of data points.

The results of these two models corroborate the CAS perspective, namely, that population structure and linguistic structure are not to be seen as separate objects of research, but as deeply interwoven phenomena that co-evolve. Moreover, it is argued that descriptive, explanatory, and grouping factors are conceptually different, which needs to be taken into account when interpreting statistical models. This argument further requires the discussion of several issues relating to the link between the theoretical CAS model, and the actual statistical results (Chapter 9). For example, we need to consider the relationship between *synchronic* data and *diachronic* inference. The parallel texts and the population data derive from databases that give information about languages as they are *now*. However, ultimately we want to infer pathways of change and evolution in the past. The validity of such extrapolation is a central topic addressed in Chapter 9.

Abstracting further away from the details of statistical analyses, there are more general considerations, problems, and caveats linked to the CAS model and the methodology chosen in this book. These are addressed in Chapter 10. For example, a crucial missing link is the psycholinguistic evidence for differential learning strategies that come to be reflected in word entropies of different languages. Which role do so-called “native” speakers play compared to “non-native” speakers, and children compared to adults? Are children really better learners than adults? Another, rather methodological issue, is the reliance on single words as information encoding units. There is psycholinguistic evidence that in language perception and production there are no hard and fast distinctions that would match orthographic practices. Multiword expressions and even whole phrases play an important role too. How does the decision to focus on word entropy hold up against such evidence?

Finally, the results of this study are also related to the growing literature on *language complexity*. If languages adopt different encoding strategies, can they be said to have different complexities? Or do complexities at different levels of information encoding ultimately level off? Or both? These questions finally bring us back to the gulf that has been gaping between “variationists” and “universalists” for the largest part of 20th century linguistics. The universalist stance has dominated our view on language, and researchers have long embarked on the search for the universal core that makes us human. However, after all, their diversity, rather than similarity, might turn out to be *the* unique property of human languages.

To enable further research in this direction as well as critical evaluation of the main analyses reported in this book, the R code and relevant data can be found at my personal website *http://www.christianbentz.de/book.html*. Throughout Chapter 4 to Chapter 10 footnotes give pointers to the code files used to create plots and run statistical analyses.

2 Languages as Adaptive Systems

An emerging branch of linguistics views natural languages as *Complex Adaptive Systems* (CAS). At the core of this proposal lies the idea that language is a communicative tool, used by speakers and signers to encode and transmit information embedded in a rich variety of social contexts. Linguistic norms are permanently under construction, i.e. negotiated via a constant flow of productions and perceptions of language users. Linguistic structure is then the outcome of feedback loops involving interactions between language users (often called agents) and their success or failure. In this most general understanding of language as a CAS, the term "complex" refers to the manifold and intertwined relationships between agents, as well as the messages they are sending, whereas the term "adaptive" expresses that both the agents and the emergent linguistic structures can change diachronically due to their mutual dependence.

In its earliest form, this framework directly mapped generalized schematizations of complex systems to linguistic phenomena. Starting with Gell-Mann (1992, 1994) terminology such as "identification of perceived regularities" and "compression into schemata" was broadly applied to describe processes of language change and evolution. One of the first proposals for a more concrete application to language data is found in the context of research into second language acquisition (SLA). Larsen-Freeman (1997) argues that the process of learning a language reflects principles known from chaos/complexity theory. For instance, the usage of regular and irregular patterns of morphology initially seems chaotic. Only later, when more input has been processed by the learner, stabilization to the system of the target language occurs. Ritt (2004) constitutes an early book-length treatment of the topic. However, instead of focusing on data from language acquisition, he concentrates on phenomena in historical language change, such as Early Middle English vowel lengthening. The "evolutionary linguistics" framework by Croft (2000) reflects a similar spirit, albeit without explicitly mentioning complex systems theory.

More recent work refined these early accounts by formulating mathematical models and exploring their validity via computational simulations, backed by empirical analyses (Blythe and Croft, 2009; Baxter et al., 2009; Briscoe, 1998, 2009, 2003, 2005; Steels, 2000). The CAS model is also the backbone of research into language acquisition (Holland, 2005, 2012) as well as the emergence of language structure in the lab (Cornish et al., 2009; Kirby and Hurford, 2002; Kirby et al., 2007). A volume was dedicated to bringing together CAS-related inquiries in a range of linguistic subfields (Ellis and Collins, 2009), of which Beckner et al. (2009) and Ellis (2013) give an overview.

DOI 10.1515/9783110560107-002

Considering the multitude of linguistic studies within the CAS framework, it is important to keep track of what is exactly meant by each study when referring to “language as a CAS”. Despite differences in detail, there are common characteristics that can be found across varying frameworks:

1. *Adaptation*. Agents individually adapt to the input by forming compressed representations of the data, which are the basis for their own future linguistic behaviour (Gell-Mann, 1994; Holland, 2006, 2012; Beckner et al., 2009). By expressing these compressed representations in their linguistic performance, the agents mutually influence each other in terms of the usage of communicative patterns. The overall communication system, in turn, can be said to adapt to the representations of agents.

2. *Regularities*. Representations are built by extrapolating regularities in the input and by filtering out random variation (Gell-Mann, 1994; Holland, 2006, 2012). These regularities are dubbed *schemata* by Gell-Mann (1994).

3. *Interaction*. Agents interact in complex ways by sending and receiving messages, thus giving rise to multiply intertwined feedback loops (Holland, 2006, 2012; Beckner et al., 2009; Steels, 2000).

4. *Emergence*. Structures in the communication system emerge from complex interactions (Ellis, 2013; Steels, 2000; Holland, 2012; Kretzschmar, 2015). That is, they cannot strictly be reduced to individual interactions, they are not guided by a central control, and they are non-deterministic (Kretzschmar, 2015, p. 19).

Each of these characteristics can be more or less important in specific accounts. For example, Gell-Mann (1994, p.17) focuses on the individual agent as a CAS, whereas in Holland (2006)’s account the collective of agents takes center stage. This distinction already hints towards another theme that emerges across many studies: the CAS theory can be applied to at least three different levels of description (Kirby and Hurford, 2002; Kirby et al., 2007; Christiansen and Kirby, 2003; Christiansen and Chater, 2008; Gell-Mann, 1994):

1. *Ontogeny*, i.e. the development of an individual speaker, learner, agent as a CAS that learns an idiolect from the often messy input throughout their life span.

2. *Glossogeny*, i.e. the level of language communities whose accumulated language production forms dialects and languages.

3. *Phylogeny*, corresponding to the evolution of preadaptations that gave rise to the cognitive capacities enabling humans to use language.

An important characteristic of these levels is that although they can be distinguished in theory – by focussing on either processes of language learning, language change or language evolution – they are not necessarily categorically different by nature. Rather, all three levels are intertwined and mutually dependent. Language *ontogeny* gives rise to language *glossogeny* via transmission of learned patterns across the population and across time (Niyogi and Berwick, 1997). Language *glossogeny* is the "moving target" (Christiansen and Chater, 2008) to which individuals and populations potentially adapt in language *phylogeny*, and which might have left a language-specific trace in the learning capability of humans (Briscoe, 2009, 2003). Finally, closing the circle, the outcome of biological adaptation on the phylogenetic time scale is the universal basis for language learning (ontogeny). In the following, these three levels and their interaction are discussed in more detail.

2.1 The ontogenetic level: individual speakers

By the early definition of Gell-Mann (1994, p. 25), a CAS identifies *regularities* in previous data, compresses them into *schemata*, and uses these schemata to make *predictions*. The success or failure of these predictions, in turn, feeds back into the viability of the schemata. Gell-Mann (1994, p. 54) suggests that language acquisition is a prime example for the "construction of a schema", in this case, the identification of patterns useful for communication. Ritt (2004, p. 100) elaborates on this model by associating "previous data" with previous communicative behaviour, the sum of "schemata" with the competence state of a speaker/signer, and the predictions with actual language performance or "verbal behaviour". Figure 2.1 is a simplified depiction of this process. In this account, "competence" refers to the entirety of communicative schemata that have worked in the past and that are hence predicted to work in future linguistic interactions with a given probability.

These schemata are derived by identifying the regularities in the input and compressing them. The term "compression" is here used in a rather vague manner, roughly referring to finding regularities. Note that the input here does not necessarily have to be exclusively linguistic. Rather, the input is potentially multi-

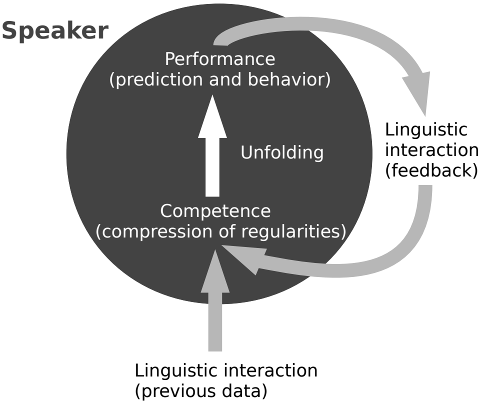

Figure 2.1: Single speaker as a CAS. White arrows and writing represent speaker internal processes and representations, grey arrows denote interactions that cross the boundary between speaker internal and external representation (adopted from Ritt 2004, p. 100).

modal. Visual, auditory, and tactile perceptions are linked to help decide whether the linguistic interaction was successful or not.

At any point in time, the competence acquired through repeated interactions in the past can be "unfolded" as "performance". Repeated linguistic behaviour maximizes the probability of successful communication. The feedback to this linguistic behaviour can then be used to further adjust the competence state, and the feedback loop goes full circle. A first step towards modelling the complex interplay of multiple speakers is to assume two speakers interacting. This simplest possible interaction is illustrated in Figure 2.2.

Crucially, the feedback loop in which speaker *A*'s performance is evaluated involves the competence and performance of speaker *B*. This way, any patterns arising from previous interactions will spread across the speaker population.[1] In consequence, after a given number of interactions the two speakers will converge on common patterns to communicate, i.e. a common language. This is the point of linkage to the next higher level of description, namely, the formation of dialects and whole languages from idiolects, referred to as language *glossogeny*.

[1] Note that Ritt (2004, p. 106) also includes separate feedback loops for speaker *A* and *B* here, probably reflecting the internal reconsideration of schemata. This is somewhat redundant, since any feedback must be linked to the input coming from another speaker.

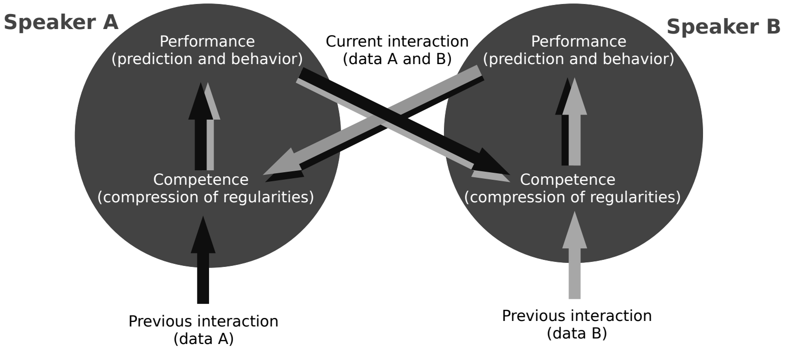
Speaker A
Performance (prediction and behavior)
Competence (compression of regularities)
Previous interaction (data A)
Current interaction (data A and B)
Speaker B
Performance (prediction and behavior)
Competence (compression of regularities)
Previous interaction (data B)

Figure 2.2: Two speakers/signers interacting. Dark blue arrows represent speaker *A*'s previous experience and the pathway by which this is passed on to speaker *B*. Grey arrows denote speaker B's former and current interactions (adopted from Ritt 2004, p. 106).

2.2 The glossogenetic level: formation of dialects and languages

Multiple idiolects, i.e. communicative patterns used by individual speakers/signers, interact to shape a communication system that shares a certain inventory of regularized patterns to encode information. This can be a dialect or a language. The model with two agents in Figure 2.2 is thus extended to a network of multiple agents in Figure 2.3. This is inspired by the utterance selection model in Baxter et al. (2006). In this representation, a set of at least two speakers/signers is linked via arrows representing *linguistic interactions*. Linguistic interactions might consist of single words/signs, phrases, or sentences. Their content and structure is not further defined here. However, they should be seen as utterances that are *successfully* decoded and hence understood by the hearer, otherwise we would include any kind of attempt at communication.[2]

The network thus consists of a set of speakers $\mathcal{S} = \{s_1, s_2, \dots, s_{10}\}$, and a multiset[3] of linguistic interactions $\mathcal{L} = \{l_{1\to5}, l_{2\to3}, \cdots, l_{10\to9}\}$, with the arrows in indices indicating the direction of the interaction. In parallel to the model in

[2] Of course, defining *successful* in the context of communication is again not trivial.

[3] Any linguistic interaction can be repeated a given number of times. We could indicate this by raising it to a power x, e.g. $l^x_{1\to5}$. The linguistic interaction going from s_1 to s_5 is thus repeated x number of times.

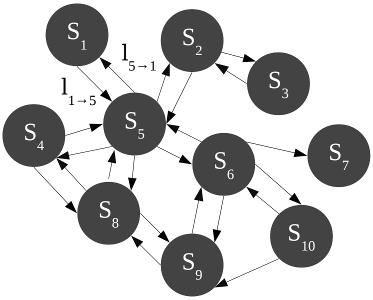

Figure 2.3: Network of multiple speakers/signers. Dark grey circles represent speakers, black arrows indicate linguistic interactions between speakers.

Figure 2.2, linguistic interactions can – but do not have to – happen between each speaker and each other speaker. For example, there is an interaction between s_1 and s_5 ($l_{1\to5}$), and likewise the other way around ($l_{5\to1}$), while there are no linguistic interactions between s_1 and s_2.

This model could be extended to reflect the social status of speakers. For example, if social status is higher for s_1 than for s_5, then the influence of the former's output on the latter's competence might be stronger than the other way around. Baxter et al. (2006, p. 4) and Baxter et al. (2009, p. 275-276) take such differences into account. Another possible variable is the mode of communication, which can vary along different dimensions such as written/spoken, face-to-face/remote, etc. The impact of social status, mode of communication, and further factors could be combined to an "adherence-score", reflecting the degree to which any output of a given speaker will influence the competence of another speaker. For instance, a one-way interaction (e.g. between s_6 and s_7) corresponds to a monologue-like communicative situation as in broadcasting via TV or radio. In this model, the "transparency" of the whole network is reflected by the degree of interconnectedness between the speakers. The number of interactions, their frequency of occurrence, and the adherence-score will determine the network's transparency for change.

Finally, note that in the network given here, the incoming interactions for each speaker are collapsed into a single input arrow, whereas in both Figure 2.1 and Figure 2.2 there are two input arrows: one referring to input via *previous* interactions and the other to input via *current* interactions. In order to represent diachrony in the network model, we need to add the dimension of time (Figure 2.4).

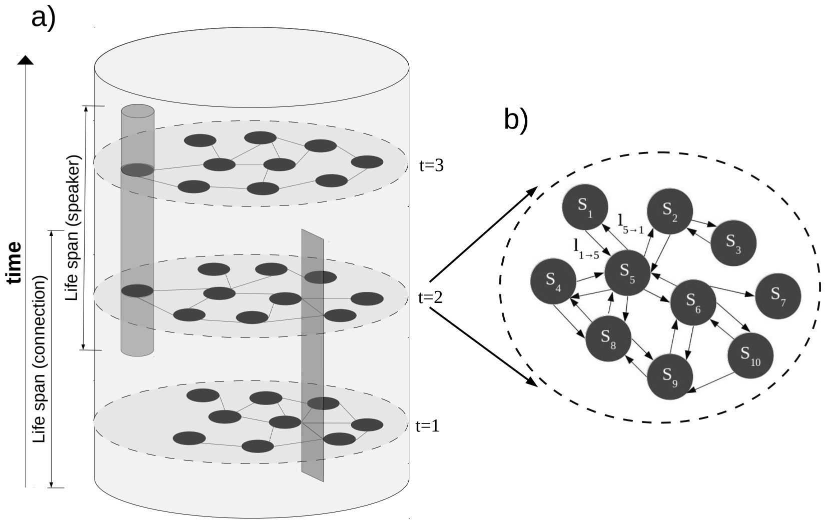

Figure 2.4: Network of multiple speakers/signers across time. a) Dark grey circles represent speakers, black lines indicate interactions between speakers. The vertical dimension represents time and grey ellipses (dashed lines) represent cross-sections of the speaker population at points in time (t=1, t=2, t=3). The transparent grey cylinder stands for the life span of an individual and the transparent grey parallelogram for the life span of an interaction. b) The speaker population and mutual interactions at time t=2.

Though time is a continuous variable, we can imagine a cross-section of the population and its interactions at a specific point in time (e.g. $t = 1, t = 2, t = 3$). This is an idealization of a genuinely gradual and continuous process. In Figure 2.4, the cross-section at time $t = 2$ is depicted as a speaker network paralleling the one in Figure 2.3. Thus, a *speaker population* $\mathcal{S}$ is here defined as the set of n different speakers at time t:

$$\text{Population: } \mathcal{S}(t) = \{s_1(t), s_2(t), \dots, s_n(t)\}. \tag{2.1}$$

The criterion for counting a speaker as part of a population is whether there are regular linguistic interactions with at least one other speaker of that population.[4] Importantly, the "language" at a given point in time is then defined as the multiset

[4] In practice, we might want to define an exact threshold of what to count as "regular" linguistic interactions.

of linguistic interactions across the whole network:

$$\text{Language: } \mathcal{L}(t) = \{l_{1\to 2}(t), \dots, l_{n-1\to n}(t)\}. \tag{2.2}$$

Further assume that we assign a competence state to each i^{th} speaker/signer such that $c_i = \text{comp}(s_i)$. The judgement on whether an utterance is grammatical or ungrammatical is based on this competence state. The competences of all speakers/signers together at time t then constitute a set called *network competence* in the following:

$$\mathcal{NC}(t) = \{c_1(t), c_2(t), \dots, c_n(t)\}. \tag{2.3}$$

Referring back to Gell-Mann (1994, p. 54), the most general definition of competence is related to the storage of regularities and patterns in the mapping from meanings to forms that help speakers to communicate. On the other hand, grammaticality might not be exclusively based on whether the content of an utterance is correctly decoded, i.e. understood. An utterance might be perfectly understandable and still be socially stigmatized as ungrammatical. In this case, communicative failure is due to the social dimension of language usage. In fact, Stockhammer (2014) argues that the history of descriptive and theoretical grammars can only be understood in the light of the social power of grammarians.

Having said this, the definitions given above reflect the *usage-based* view of "language". In this sense, language emerges from the manifold communicative interactions of speakers and signers. As a consequence, competence is shaped by frequency of usage (Bybee, 2006, 2007). In contrast, we could also imagine the language competence c_i of a speaker as a set of modular subcompetences, i.e. $c_i = \{m_{1i}, m_{2i}, \dots, m_{ji}\}$. In this case, "language" can be defined in a *Universal Grammar* (UG) way, namely, as the set of subcompetences shared by every single individual. This is represented here by the intersection of competence sets at a given time:

$$\text{Language: } \mathcal{L}(t) = UG(t) = c_1(t) \cap c_2(t) \cap \dots \cap c_n(t). \tag{2.4}$$

Chomsky (1986) has famously adopted this view on language and termed it "internalized" language, *I-language* for short, which reflects the universal competence of human speakers/signers underlying linguistic performance. He dismissed the *usage-based* view, which he terms *E-language*, short for "externalized" language. He argues that language in the externalized sense is merely an artefact of errors and random variation in performance – genuinely not an interesting subject of study.

2.3 The phylogenetic level: language evolution

Explaining language as a CAS at the phylogenetic level is the most difficult exercise, since it requires linking both ontogenetic changes and glossogenetic changes together on an evolutionary time scale (Kirby and Hurford, 2002). The current proposals in this direction are here subsumed under two broad categories:

1. The *saltational* account. This proposes a sudden *qualitative* change of competence in an individual speaker (or several speakers) – yielding modern human language – which is then perpetuated through the population in the following generations.

2. The *gradual* account. Stepwise or "piecemeal" changes lead towards an altered competence state. This account can be further subdivided by whether it posits discrete stepwise changes leading to a *qualitatively* different competence state (in agreement with the saltational account), or if "graduation" is interpreted as a genuinely continuous process with a *quantitative* rather than *qualitative* difference in outcome. The former is here called *discrete* graduation and the latter *continuous* graduation.

Three different – and often opposing – views emerge from these fundamental distinctions. In particular, saltation on one hand, and continuous change on the other, are logically incompatible. This was coined the *Continuity Paradox* by Bickerton (1981, p. 217) and has caused some of the fiercest debates in 20th century linguistics and cogntive sciences more generally.

2.3.1 The saltational account

The saltational account lends itself to a theory which reduces language to a highly specified core. This specified core might have resulted from a sudden "switch" in competence, sometimes also referred to as mutational "great leap forward", possibly in a single individual (Chomsky, 2005, 2010; Boeckx and Piattelli-Palmarini, 2005). In this context, the computational core of language is speculated to boil down to an operation which allows for linguistic elements to be recursively recombined *ad infinitum*.

Hauser et al. (2002) and Fitch et al. (2005) denote this operation "recursion", and attribute it to the faculty of language in a "narrow sense" (FLN). Fitch (2010b) further elaborates on different types of "recursion", and how these are relevant for the theory of natural language. In the context of the *Strong Minimalist The-*

sis (SMT), Chomsky (2005, 2010) calls this operation "unbounded Merge". The saltational view maintains that recursion, or unbounded Merge, is likely the only human specific capacity in the FLN that sets human language apart from communicative (and non-communicative) competences of other animals, while there might be a multitude of general learning capacities that are shared across species as part of the faculty of language in a "broad sense" (FLB). The position that unbounded Merge is human specific has mainly been adopted by accounts of language evolution associated with the Minimalist Program (Bolhuis et al., 2014; Berwick et al., 2013).

To illustrate this, in Figure 2.5 a), there is an individual (marked in blue)[5] at time step $t = 2$, which evolved the uniquely human language competence c^{hum}. This qualitatively diverges from the "proto-human" competence c^{proto}. As a consequence, the "switching" event is also associated with the split between "proto-language" and "language", for instance, in Piattelli-Palmarini and Uriagereka (2004, pp. 368). The altered competence state is spread throughout the population in subsequent generations.

More precisely, the competence state can be divided into subcompetences, namely, the human specific competence (FLN) and the competence shared with animals (FLB):

$$c^{\text{hum}}(t) = c^{\text{FLN}}(t) + c^{\text{FLB}}(t). \tag{2.5}$$

The assumption of Hauser et al. (2002), and Fitch et al. (2005) is that any competence state before c^{hum}, i.e. before time $t = 2$ in this specific example, exclusively contained (FLB):

$$c^{\text{proto}}(t < 2) = c^{\text{FLB}}(t < 2). \tag{2.6}$$

According to the same authors, the addition of a human specific competence for language, i.e. $c^{\text{FLN}}(t \geq 2)$, might not have had any obvious communicative function and hence not necessarily a selective advantage. From the saltational perspective, the term "language" refers to this universal state of competence shared by all living humans, variously called "I-language", "internal language" or "Universal Grammar". The details are given in Hauser et al. (2002, p. 1570) and Fitch et al. (2005, p. 180). In this framework, language is thus defined as:

$$\text{Language} : UG(t) = c_1(t) \cap c_2(t) \cap \cdots \cap c_n(t) = c^{\text{FLN}}(t). \tag{2.7}$$

[5] The competence state of some individual i at time $t = 2$ should be denoted as $c_i^{\text{hum}}(2)$. This is simplified to c^{hum} in the figure for convenience.

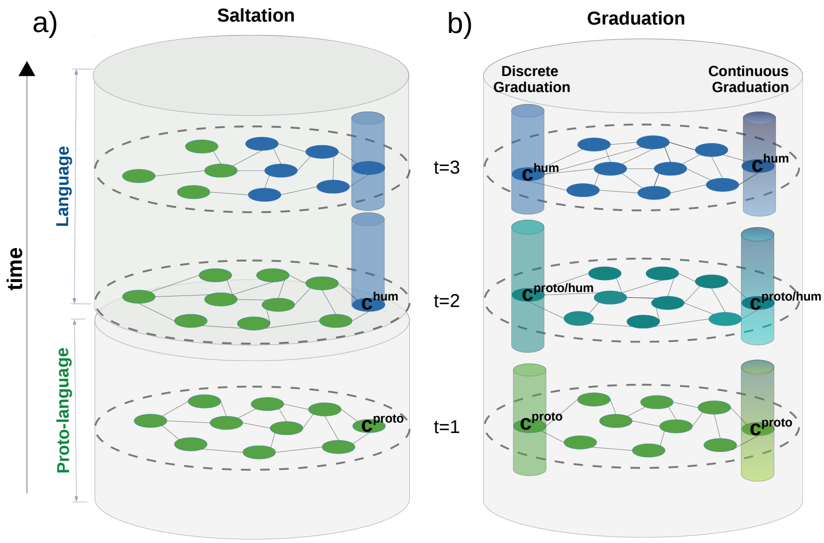

Figure 2.5: Networks of speakers on an evolutionary time scale. a) The saltational account. Green and blue ellipses represent speakers in competence states c^{proto} and c^{hum} respectively (which here correspond to points in time t=1 and t=2). The time period of proto-language is marked in green, the time period of language is marked in blue. b) Two gradual accounts. Both posit a gradual change from proto-language to language (gradual change of green into blue). Competence states change gradually from c^{proto}, over $c^{proto/hum}$, to c^{hum}. There is a further distinction between *discrete graduation* and *continuous graduation* reflecting the difference between small – but discrete – changes in competence states from generation to generation (left) versus gradual changes in competence states across individual life spans (right).

It refers to the universal competence state that spread throughout the whole human population from a given time t onward. Of course, this is rather different from the definition referring to the multiset of linguistic interactions as given in Equation 2.2.

As a result, this model allows to draw a hard and fast distinction between language *competence*, on one hand, and language *performance*, on the other (Chomsky, 1965, p.4). All biological preadaptations to language in the human lineage up to competence state c^{hum} and the evolution of that competence state itself are considered the subject of *language evolution* research, whereas the linguistic interactions between speakers from $t = 2$ onwards are seen as language performance and as such form part of research into *historical language change*. A clear dividing line between proto-language and language follows logically from this dichotomy (Bickerton, 1990, pp. 164).

2.3.2 The gradual account

Opposing the saltational view, the two further accounts depicted in Figure 2.5 b) display gradual changes from competence state c^{proto} to c^{hum}, with intermediate state $c^{\text{proto/hum}}$. Moreover, another distinction is introduced: the left side depicts small – but still discrete – steps of changes in different generations (small cylinders in different colours), while on the right side changes in competence occur even within the life span of an individual (represented by cylinders gradually changing colour). The first type of graduation is here referred to as *discrete graduation*, whereas the second type is referred to as *continuous graduation*.

The discrete account is associated with the Neo-Darwinian process of gradual adaptation of the language faculty (both c^{FLB} and c^{FLN}) by Pinker and Bloom (1990) and Pinker (2003). It has been further corroborated by Pinker and Jackendoff (2005) and Jackendoff and Pinker (2005) in reaction to Hauser et al. (2002) and Fitch et al. (2005). The core tenet of the adaptationist view is that not only the abstract principle of recursion, but several further cognitive components, for example, those relevant for the perception and production of speech, are human and language specific (i.e. part of c^{FLN}). These are assumed to be the outcome of gradual biological adaptation in the human lineage, potentially via the so-called *Baldwin effect*, also referred to as *genetic assimilation*. For a discussion of this effect and computational models illustrating its workings see Briscoe (2003, 2005, 2009). This account is still "discrete" in the sense that it invokes genetic changes, albeit potentially minuscule, which are inherited to later generations via natural selection.

Importantly, this raises the question: what are *competence* states at any given point in time adapting to? The only plausible answer is: language *performance*. Adaptation hinges upon feedback that is necessary to define a fitness function of the trait under selection. Pinker and Bloom (1990, p. 721) realize this when they state that "each intermediate grammar [was] useful to its possessor", and that "every detail of grammatical competence that we wish to ascribe to selection must have conferred a reproductive advantage on its speakers". Note that "usefulness" and "reproductive advantage" are not necessarily opposed to arbitrariness (Briscoe, 2009, p. 13). Namely, arbitrary features of language structure might serve as successful communication "protocols", and thus have a function, independent of other more typical functional pressures such as ease of learning and usage. Broadly speaking, the feedback-loop between competence and performance links language back to a *communicative function*, from which it was dissociated by Hauser et al. (2002) and Fitch et al. (2005). In fact, even a non-functional saltational account might not get around positing some selection pressure towards recursion, or Merge, since sudden mutation might explain how it came into existence, but not how it came to fixation in the human lineage (Briscoe, 2003, p. 303).

Thus, in accounts extending the original proposal by Pinker and Bloom (1990), any competence state can only be selected for or against by means of evaluation of its fitness in communicative performance. This is a far-reaching repercussion of the adaptationist stance. It nullifies the categorical distinction between competence and performance and, by logical extension, also between proto-language and language. Investigations into language evolution and language change are then part of the same overall research programme.

To put it in terms of the definitions given above: the competence state of an individual at a given point in time, e.g. $c_i(t = 2)$, can only be fully understood against the backdrop of language performance (involving the same individual) of former time steps, i.e. the accumulated linguistic interactions $\mathcal{L}_i(t < 2)$. Performance constitutes the basis for the selection of former competence states. In consequence, the network competence at a given point in time is also depending on former performance. Namely, network competence and language in a usage-based sense are intertwined in a continuous feedback-loop along the dimension of time:

$$... \mathcal{L}(t-1) \rightarrow \mathcal{NC}(t-1) \rightarrow \mathcal{L}(t) \rightarrow \mathcal{NC}(t) ... \tag{2.8}$$

Without performance there is no selection and without selection there is no adaptation. If a competence state is the outcome of adaptation, then it cannot be independent of past performance *by logical necessity*. Note that this is independent of whether competence is defined in a FLN or FLB sense. Hawkins (2004, 2014) has outlined the tight link between competence and performance in glossogeny. Assuming an adaptationist account, the same rationale holds in phylogeny as well. Upholding the distinction between competence and performance is probably the prime reason for why Fitch et al. (2005, p. 179) rebut the adaptationist account – for the FLN – stating that: "[...] questions of function are independent of those concerning mechanism".

2.3.3 Cultural evolution: beyond biological transmission

The model depicted under "continuous graduation" on the right of Figure 2.5 b) adds another complication to the picture. Competence states do not only differ from generation to generation as suggested by discrete graduation, but they also change over the course of an individual's life. Competences acquired in this way are potentially handed down to the next generation by means of epigenetic inheritance. The exact workings of this process are currently under debate, and beyond the scope of this book. However, besides genetic inheritance, changes in competence are also handed down to the next generation by means of *cultural inheritance*. Cultural evolution has become a recurrent theme in studies that do not draw a hard and fast distinction between language change and language evolution (Christiansen and Chater, 2008; Christiansen and Kirby, 2003; Chater and Christiansen, 2012; Bentz and Christiansen, 2010). Most prominently, Christiansen and Chater (2008), followed up by Chater and Christiansen (2012), critically review both the saltational and the gradual adaptationist view and argue that both are at odds with the structural diversity found across languages of the world.

According to their criticism, the latest saltational account as proposed by Chomsky (2005, 2010) runs into the problem of underspecifying the mechanism by which linguistic variation arises. In fact, Chomsky (2010, p. 60) himself points out that the variety of languages spoken across the globe is a "violation to the spirit of the strong minimalist thesis". Following the rationale of Figure 2.5 a), he posits that linguistic research should be mainly concerned with the description and analysis of *unbounded Merge*, whereas language diversity might be a epiphenomenon of "externalization" (i.e. performance), which falls outside the domain of minimalism. Following this advice leads to the somewhat paradoxical situation that the core question addressed in the field of linguistics has few – if anything – to do with linguistic diversity, which is, however, part and parcel of linguists' everyday work when documenting and analysing the world's languages. This is most likely the reason why many linguists oppose to the "recursion-only-hypothesis", and argue for a richer FLN that became biologically adapted to deal with "complex communicative propositions" (Pinker and Jackendoff, 2005; Jackendoff and Pinker, 2005).

However, even if the existence of a richer FLN is taken for granted, then there are still three further problems with the adaptationist account (Deacon, 1997; Christiansen and Chater, 2008; Chater and Christiansen, 2012):

1. It presupposes a stable language community, i.e. constant $\mathcal{S}(t)$, over a relatively long period of evolutionary time, in order for structural characteristics of the communication system, i.e. $\mathcal{L}(t)$, to become genetically encoded in the speakers' competences. It is argued in Christiansen and Chater (2008, p. 492-493) that such a long population stasis is hard to reconcile with the facts about human migrations.

2. Languages, in a usage-based sense, are attested to have changed relatively fast in human history. Over a few hundred years they quite commonly diverge into varieties that are not mutually intelligible anymore. Such rapid changes constitute a “moving target” for genetic encoding (Christiansen and Chater 2008, p. 493-494, and Deacon 1997, p. 328). However, for a recent reassessment of this claim based on computational modelling see de Boer and Thompson (2018).

3. There is no *a priori* reason to assume that biological adaptation only targets abstract and “deep” language structures as part of the language faculty, while “surface” characteristics of languages are culturally transmitted and learned (Christiansen and Chater, 2008, p. 494-495). In other words, if adaptation was at work to genetically encode an abstract and arbitrary communication “protocol”, as argued in Pinker and Bloom (1990), then it could just as well work to encode more concrete properties of languages such as the ability to produce clicks or perceive tones. However, so far there is only limited evidence for genetic predispositions that potentially facilitate the production of such particular phonemic features (Dediu and Ladd, 2007; Dediu et al., 2017; Moisik and Dediu, 2017).

Instead of a saltational account with minimal specification of the FLN, or a gradual adaptationist account with a richer FLN, supporters of continuous graduation propose to investigate language structures as an outcome of interactions between domain-general preadaptations to language. This includes aspects of the FLB like increased working memory, word learning abilities, and sequential learning abilities (Christiansen and Chater, 2008, p.508). Amplified by processes of cultural transmission such preadaptations might give rise to the structural diversity of languages we find in the world today (see Beckner et al. 2009; Kirby et al. 2007; Smith and Kirby 2008; Scott-Phillips and Kirby 2010; Kirby and Hurford 2002). From this perspective, language structures are shaped by domain-general learning constraints (i.e. c^{FLB}) rather than “grown” by a genetically hard-wired blueprint (i.e. c^{FLN}).

In consequence, there is no strict discontinuity in the competence states of speakers across time caused by a “great leap forward”, or even by minor genetic changes as in the discrete graduation account. Instead, the shared set of competences are learned from linguistic patterns found in the input at an earlier point in time $t-1$, i.e. linguistic interactions $\mathcal{L}(t-1)$. The cognitive capacities underlying this learning have gradually evolved beyond animal capacities as domain-general competence c^{FLB} and gave rise to complex communication. Hence, while the FLB is universal, in the sense of being shared by all humans, it is not strictly specific to language. In this account, there is no need for a language specific, universal competence state that is shared across all – and only – humans. The FLB is the overlap in competence that makes humans capable of language, while the FLN is empty (Christiansen and Chater, 2015):

$$c^{\mathrm{FLB}} = c_1(t) \cap c_2(t) \cap \cdots \cap c_n(t), \tag{2.9}$$

and

$$UG(t) = c^{\mathrm{FLN}} = \{\}. \tag{2.10}$$

Clearly, this position has not been generally adopted in the literature. Briscoe (2003, 2009), for instance, points out that UG, referred to as the *Language Acquisition Device* LAD, is very likely to be non-empty, if only in the weak sense that it in some way needs to constrain the *possible space* of grammars that could potentially be derived from language input.

Notably, Fitch et al. (2005, p. 201) concede that the actual content of the FLN has to be determined by further empirical research and that it might turn out to be "an empty subset of FLB, with only the integration of mechanisms being uniquely human." In the most recent overview article by Fitch (2018), the neural components relevant for speech are said to be derived from cognitive capacities already present in our "nonlinguistic primate ancestors". These are assigned to the "faculty of language, derived components", abbreviated as FLD. Furthermore, the criteria for assigning a trait to the FLN, namely being human and language specific, are now seen as "too stringent regarding speech" (Fitch, 2018, p. 256). Of course, since the reference is here to "speech" rather than "language", this still leaves open the question of whether the FLN is entirely empty.

2.4 Summary

It is important to keep in mind that the CAS framework, as outlined in the previous sections, is agnostic to the exact definitions of "language" and "grammar". Depending on how competence and performance are defined, it can accommodate everything from the *Strong Minimalist Thesis* (SMT) to usage-based theories of language. In fact, Gell-Mann (1992, p. 15) names both "innately represented principles that constrain grammar", *and* "selection pressures [...] that favour what is adaptive in terms of communication" as factors involved in the evolution of language from a complex systems point of view. Thus, the CAS framework overarches the opposing camps.

Nevertheless, it is fair to say that, over time, the spirit of researchers associating themselves with the CAS framework has tended towards reducing innate constraints to a minimum, and positing instead that language structures largely emerge from cultural transmission (Kirby et al., 2007; Smith and Kirby, 2008; Kirby et al., 2008). The results of Kirby et al. (2008) are a paradigm example of this trend. In their iterated learning experiments, a hallmark of human language – compositionality – emerges from originally random input over several generations of learning and transmission. Importantly, the observation that language structures emerge from repeated cycles of learning and usage might be *in line with* – not *opposing to* – the *Strong Minimalist Thesis*. Chomsky (2010) explicitly states that linguistic diversity is not an *explanandum* of the minimalist program. The core of language in the minimalist sense, i.e. UG, can – by definition – not explain the variety of linguistic encoding strategies found across the world. UG can only delimit the space of "possible languages", any variance within that space is a secondary phenomenon.

It is all the more important to keep in mind which exact definition of terms such as "language" and "grammar" a theory proposes, and which are the implications of these definitions. The definitions given in the previous section can be divided into two categories, according to whether competence and performance are seen as strictly different, or rather intertwined subjects of research. Henceforth, the former is called *minimalist* definition (Min), the latter *usage-based* definition (Use).

$$\mathcal{L}^{\text{Min}}(t) = UG(t) = c^{\text{FLN}} = c_1(t) \cap c_2(t) \cap \cdots \cap c_n(t). \tag{2.11}$$

$$\mathcal{L}^{\text{Use}}(t) = \{l_{1\to 2}(t), \dots, l_{n-1\to n}(t)\}. \tag{2.12}$$

The saltational account is frequently linked to the minimalist definition. "Language" in the minimalist sense equates Universal Grammar, defined as the human and language specific set of overlapping competences in the FLN.

In contrast, the cultural evolution account is often associated with the usage-based definition. "Language" is here defined as an accumulation of linguistic interactions. Importantly, the linguistic competence of a speaker or signer s_i at a certain point in time t, i.e. $c_i(t)$, is the outcome of applying their domain-general learning competence c_i^{FLB} to the input available at an earlier point in time, that is $\mathcal{L}_i(t-1)$. This could be modelled as a "link function" between competence and performance. They are intertwined in a feedback-loop along the dimension of time. This is in line with the so-called *performance-grammar correspondence hypothesis* (Hawkins, 2004, 2014) as well as usage-based approaches to language learning and change (Bybee, 2006, 2007). By its very nature, the usage-based perspective is more likely to inspire research into the diversity of languages, whereas the minimalist account focuses more on language universals.

Interestingly, it is an open question how the discrete adaptationist account can be reconciled with a distinction between competence and performance. On the one hand, it posits a rich FLN as underlying human and language specific competence. This is in line with the minimalist definition, and suggests a strict divide between competence and performance. On the other hand, adaptation requires performance as a basis for selection, i.e. interaction between competence and performance. It is not further specified in Pinker and Bloom (1990), Pinker and Jackendoff (2005), and Jackendoff and Pinker (2005) how both of these views can be logically combined in a single, coherent account.

We have thus laid out different views on what "language" actually is. The overall objective of this book is to start measuring, and ultimately, explaining linguistic diversity across languages of the world. The usage-based definition of language lends itself to this undertaking, more so than the minimalist definition. The minimalist perspective tends to marginalize language variation and change and to focus on universal properties of language. However, whether FLN exists as a non-empty set of subcompetences is not a central question here. Instead, the question in focus is: which aspects of language in a usage-based sense are potentially to be explained by factors external to FLN, and maybe even external to FLB?

Answering these questions requires establishing links between languages and the properties of populations using them as communicative tools. More precisely, it requires investigating how the set of linguistic interactions ($\mathcal{L}(t)$) might change over time due to changes in the population of speakers ($\mathcal{S}(t)$) and their competences ($\mathcal{NC}(t)$). The next chapter discusses how the CAS model can be modified to reflect this link.

3 Language Change and Population Structure

As outlined in the previous chapter, the CAS model – in its *usage-based* formulation – assumes that a speaker/signer population at a given time $\mathcal{S}(t)$ carries network competence $\mathcal{NC}(t)$. This network competence, in turn, underlies the usage of particular linguistic variants and is reflected in language performance, i.e. the entirety of linguistic interactions $\mathcal{L}(t)$. By this definition, if speaker/signer networks were static over time, there would be no – or very limited – change in $\mathcal{L}(t)$.

A notable exception is change driven by random noise, i.e. independent of the language competence underlying an utterance. It is viable to assume that there is quite literally random noise in the transmission of phonetic information, though whether and how this affects language structure more generally is unclear. In fact, there is evidence that spoken language is robust even in the face of noise (Winter, 2014).

Having said that, a core assumption in the following is that linguistic interactions *reflect* network competences. Network competences are likely to change over time for reasons of expansions or reductions of speaker populations, enhancing or cutting back on variation in linguistic interactions respectively. Any single speaker and their linguistic performance can vary along several dimensions, including the pronunciation of words, the usage of base vocabulary, morphological marking, and syntax. To model language change in detail we could consider many different subpopulations (down to the individual speaker and their idiolect) with different network competences. However, to start with, we assume the simplest scenario of variation possible: the existence of two subpopulations A and B, with two differing network competences.

3.1 Populations and languages

Let us assume two subpopulations A and B at time t which are heterogeneous in the sense that they carry differing sets of network competences. In Figure 3.1, these are indicated by different colours, i.e. $\mathcal{NC}^A(t)$ in blue, and $\mathcal{NC}^B(t)$ in red. The proportion of speakers in subpopulation A compared to speakers in subpopulation B can differ over time. Crucially, differences in network competences of subpopulations are reflected in their linguistic output. This is illustrated in the equations on the right of Figure 3.1. The composition of speaker populations at times $t = 1$, $t = 2$, and $t = 3$ is directly reflected in the composition of the respective language in a usage-based sense. Namely, the three languages $\mathcal{L}(t = 1)$,

DOI 10.1515/9783110560107-003

$\mathcal{L}(t = 2)$, and $\mathcal{L}(t = 3)$ differ with regards to the spread of the red and blue features in linguistic interactions. For instance, at time $t = 1$ the red competence is exclusively represented by speaker s_7. Hence, the corresponding red feature is only reflected in their linguistic interactions with other speakers: $l_{7\to3}(t = 1)$ and $l_{7\to10}(t = 1)$. At $t = 2$, however, the new feature is increasingly more widespread in linguistic interactions. Finally, at $t = 3$, it has spread throughout the entire population. Of course, whether we want to call $\mathcal{L}(t = 1)$, $\mathcal{L}(t = 2)$, and $\mathcal{L}(t = 3)$ three separate languages depends on how exactly we define the threshold at which linguistic interactions are not mutually intelligible anymore.

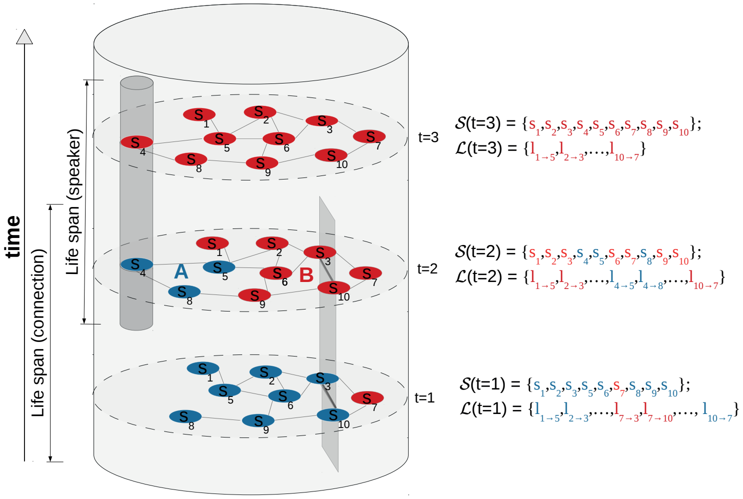
time
Life span (connection)
Life span (speaker)
A
B
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
t=3
t=2
t=1
$\mathcal{S}(t=3) = \{s_1,s_2,s_3,s_4,s_5,s_6,s_7,s_8,s_9,s_{10}\}$;
$\mathcal{L}(t=3) = \{l_{1\to5},l_{2\to3},\ldots,l_{10\to7}\}$
$\mathcal{S}(t=2) = \{s_1,s_2,s_3,s_4,s_5,s_6,s_7,s_8,s_9,s_{10}\}$;
$\mathcal{L}(t=2) = \{l_{1\to5},l_{2\to3},\ldots,l_{4\to5},l_{4\to8},\ldots,l_{10\to7}\}$
$\mathcal{S}(t=1) = \{s_1,s_2,s_3,s_5,s_6,s_7,s_8,s_9,s_{10}\}$;
$\mathcal{L}(t=1) = \{l_{1\to5},l_{2\to3},\ldots,l_{7\to3},l_{7\to10},\ldots, l_{10\to7}\}$

Figure 3.1: Subpopulations *A* and *B*. The cylinder on the left displays proportions of subpopulations (blue and red) changing over time. The equations on the right give definitions of sets of populations and languages at times t=1, t=2, and t=3. It is assumed here that a given competence state of a speaker/signer s_i is reflected in linguistic interactions with other speakers/signers. For convenience, in this figure a linguistic interaction is denoted as $l_{i\to j}$, dropping the exact point in time in parentheses.

Take an actual historical example. Let us assume that at some point in the transition from Old English towards Middle English and finally Modern English a subpopulation A still used the numerous weak and strong plural forms of nouns in the nominative, e.g. *stanas* ‘stones’, *scipu* ‘ships’, *sorga* ‘sorrows’, and *naman* ‘names’, whereas subpopulation B started to use the regular s-plural suffix which applies to the vast majority of English nouns today. Depending on the numerical dominance of speakers of subpopulation A in relation to subpopulation B, either the original or the s-plural forms would have been dominant in the language at that point in time. In other words, if our categorization of language $\mathcal{L}(t)$ into Old, Middle, or Modern English was solely based on nominal plural marking, then our decision would be directly linked to the proportion of regular over irregular forms, and this proportion is a direct reflection of population structure.

In practice, there are many more phonological, morphological and syntactic considerations taken into account in language classification. However, the rationale is similar. There has to be a threshold for our decision on how any given language at time t is classified. Whether this threshold has been reached is directly linked to the prevalence of respective subpopulations, their network competences, and the entirety of linguistic interactions resulting from these. As a consequence, there is a tight link between population structure and language structure in this model.

Ultimately, we want to know the factors that *drove* or *caused* the language to replace one linguistic feature for another. What caused the change of nominal plural morphology from a variety of patterns towards a regular -s pattern in the case of OE and ModE? Within the usage-based CAS account, the answer has to lie in the network competences and the linguistic interactions of subpopulations, that is, in differing pressures of learning and usage. In this context, *neutral change*, i.e. change unrelated to particular selection pressures, has been argued to constitute the null-model – in parallel to biological evolution. Furthermore, the *size of a population*, the percentage of *adult learners* present, and the *status* associated with a language might be causally related to language structures changing over time. Finally, geographical and genealogical phenomena such as *population drift* and *phylogenetic grouping* are intertwined with these effects.

Neutral change

The term "neutral change" denotes a process whereby a particular variant replaces other variants in a population without necessarily conferring an adaptive advantage, i.e. without being directly selected for. This process is sometimes also referred to as "random copying". In a linguistic context, this is the case if "the probability of a language learner adopting any given linguistic variant only depends on the frequency of that variant in the learner's environment" (Kauhanen, 2017, p. 327). If this is the case, then frequency of usage – plus some random noise – is sufficient to explain the rise and fall of linguistic features from phonemes to syntax. Since neutral change in this sense is tightly linked with frequencies of usage, it has been associated with usage-based accounts of language change (Baxter et al., 2006, 2009; Blythe and Croft, 2012). Language-like phenomena that have been argued to arise in this manner include Zipf's law of word frequencies (Reali and Griffiths, 2010) and S-curve shapes characteristic of the frequency increase and decrease of variants in competition (Reali and Griffiths, 2010; Kauhanen, 2017; Yanovich, 2016). Moreover, Kauhanen (2017) argues based on computational simulations that neutral drift should be seen as the default mechanism rather than a last resort to explain language change.

However, the degree to which actual language change phenomena can be explained by neutral mechanisms, versus being driven by social and cognitive selection pressures, is still under debate (see Blythe, 2012; Kauhanen, 2017, for an overview). In a study based on historical corpora of English, Newberry et al. (2017) illustrate how some changes qualify as driven by genuine selection, e.g. irregularization from *sneaked* to *snuck* over the past 150 years, while for others, e.g. *builded* to *built*, neutral drift cannot be rejected as a mechanism. More studies based on empirical data across different languages are necessary to further assess the weighting of neutral drift versus selection.

Also, more theoretical thinking about the basic assumptions of language change in comparison to biological evolution is necessary. In many cases, it is hard to assess how realistic a given mathematical model and its computational implementation is, given the lack of empirical data to evaluate it. For instance, in Kauhanen (2017)'s model, "random noise" is introduced in the form of an innovation parameter. This is necessary for variation to arise, otherwise the same variant would be used consistently throughout a population and change would be impossible. In this context, random noise is conceptualized as speakers choosing a given variant *uniformly at random*, i.e. with equal probability, from a set of variants.

Translated to the history of noun plural formation in English this could mean the following. While the majority of speakers in an early period still used, for instance, the variant *scipu* as the nominative plural of 'ship', some speakers in a later period started to choose randomly between the variant *scipu* and *scip(e)s*. However, this further begs the question where the new variant came from in the first place.

In models of genetic evolution, random mutations can replace one amino acid for another in a substring of DNA, and thus create a new allele. Is the linguistic process of replacing the *-u* marker with an *-(e)s* marker on the stem *scip-* conceptually the same as a random mutation in genetics? Or is there already a cognitive pressure, i.e. selection, reflected in the usage of the *-s* plural? More generally, can there be "true" random noise in the usage of linguistic variants?

Most certainly, these are tough questions to answer. Also, even if we come to the conclusion that biological evolution *is* different from language change in some fundamental regard, this does not undermine the overall usefulness of the above cited studies, since they spearhead the development of methods to rigorously test different models of language change, which are currently lacking. Beyond the process of neutral drift there are further aspects of population structure relevant to language change. These are outlined in the following and detailed in Chapter 6.

Population size

The overall size n of a population of speakers $\mathcal{S}(t) = \{s_1, s_2, ...s_n\}$ is related to the "transparency" for change. Simply speaking, bigger populations require more time and interactions for an innovation to spread throughout the network of competences, everything else being equal. Going back to the example of subpopulations in Figure 3.1, it is conceivable that an innovation like the s-plural in the history of English has a higher probability to spread in a small population than in a bigger population.

Moreover, network competences and hence linguistic interactions in bigger populations might exhibit more variation than in smaller populations – bearing more potential for change. Finally, population size can be linked to another population-related factor: *language contact*. Bigger populations might be those populations that have (by trend) "recruited" more adult learners in the past. The literature on population size as a predictor of language change is discussed further in Chapter 6.

Percentage of adult learners

Imagine that, in Figure 3.1, subpopulation A corresponds to "native" speakers of $\mathcal{L}(t = 1)$, whereas subpopulation B corresponds to "non-native" speakers. Non-native speakers are here conceptualized as adult learners, or more precisely, as learners that have had less exposure to the target language. While at $t = 1$ high-exposure speaker competences (and the corresponding linguistic interactions) are in the majority, at $t = 2$ the low-exposure competences have spread throughout the biggest part of the network. Note that this could either happen by more low-exposure learners coming into the population from outside, or by native speakers adopting low-exposure patterns in their linguistic interactions.

In both cases, it is the competence states of the low-exposure learners that underlie changes in the overall population, and drive the language towards the new states at $t = 3$. The literature and terminology related to such language contact phenomena is outlined in more detail in Chapter 6 and Chapter 10.

Language status

Competences and linguistic interactions could also differ with regards to the social status they are associated with. For example, interactions in A might correspond to dialectal variance in pronunciation, whereas interactions in B might correspond to an evolving national standard such as Received Pronunciation. In this case, the reason for why B competences spread throughout the population at $t = 2$, and $t = 3$ would have to do with the socio-political associations linked to linguistic interactions, not necessarily with communicative efficiency or learning pressures.

3.1.1 Population drift

The effects that different characteristics of speaker/signer populations and their usage preferences have on linguistic structures might further be magnified or diminished by population drift and dispersal. *Population drift* is a concept known from evolutionary biology. When populations of species drift apart (either geographically or by any other means of isolation) the emerging subpopulations will carry only a part of the overall variation found in the source population. A series of such population splits leads to decreasing diversity in consecutive subpopulations. This is called a serial founder effect. The classic notion of linguistic "drift" as coined by Sapir (1921) is rather different at first sight. Sapir's "drift" refers to the tendency – often defying any obvious explanation – of one linguistic form gradually replacing another in a population of speakers. Sapir gives the example of how the question "Who did you see?" unavoidably encroaches on the territory of "Whom did you see?" in the spoken English of his time – whilst being "quite incorrect". While linguistic drift and population drift are conceptually different, Sapir recognized that they are intertwined, namely, when social isolation gives rise to linguistic diversification:

> Now dialects arise not because of the mere fact of individual variation but because two or more groups of individuals have become sufficiently disconnected to drift apart, or independently, instead of together. So long as they keep strictly together, no amount of individual variation would lead to the formation of dialects. In practice, of course, no language can be spread over a vast territory or even over a considerable area without showing dialectic variations, for it is impossible to keep a large population from segregating itself into local groups, the language of each of which tends to drift independently. (Sapir, 1921, p. 73)

Hence, linguistic drift and population drift interact when new languages take shape. To illustrate this, assume that at time $t = 2$ there are two subpopulations A and B that are part of a bigger speaker population in the sense that speakers of A and B are still interconnected via mutual linguistic interactions (see Figure 3.2). After $t = 2$ there is a split that runs through the overall population, i.e. a population drift. This could be a geographic, political, or any other type of isolation that results in reduced linguistic interactions. Remember that our definition of “speaker population” is based on linguistic interactions. Hence, if there are no interactions between two subsets of speakers, then this automatically leads to two separate populations A and B. Arguably, in scenarios of sudden geographic dispersal, it is unrealistic to assume that these two populations speak two different, i.e. mutually unintelligible, languages straight away. Realistically, right after $t = 2$ linguistic interactions are still mutually intelligible between the populations, but communication channels are cut. These potential linguistic interactions are indicated in Figure 3.2 by dashed grey lines.

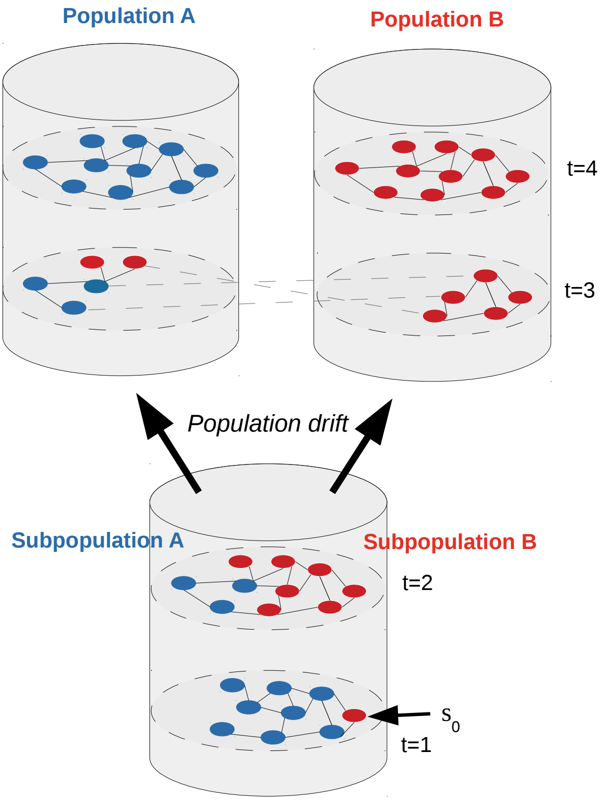

Figure 3.2: Two subpopulations (blue) and (red) drift apart after time t=2. At subsequent times t=3 and t=4 the “branches” develop two different network competences, eventually leading to two separate languages. Potential linguistic interactions between the separated populations are indicated with grey dashed lines. The original innovation in the red branch develops from speaker s_0 onwards.

Crucially, at time $t = 4$ all speakers in population A have come to share the blue feature, while all speakers in population B have come to share the red feature. If these populations are geographically separated, then geographic location predicts the usage of a given feature. This "areal" pattern emerged due to geographic drift, since it isolated speakers with the red feature from speakers with the blue feature. Arguably, however, the population drift by itself is not the ultimate *cause* for the two separate red and blue populations at time $t = 4$. The cause has to be sought a) in speaker/signer s_0 who first adopted the red feature, and b) in the reason(s) for why this feature was spread throughout the original population at time $t = 2$. For example, s_0 could have been an adult learner whose choice of the red feature is explainable by specific learning biases. The feature might then have spread throughout the entire population by means of native speakers adopting it, maybe related to the social status of s_0 or due to the relative ease of usage, etc.

The bottom line is that population drift and the resulting geographic/areal patterns are factors *conceptionally different* from the ones directly relating to the structure of a population at a given point in time, i.e. population size, adult learner percentages, and status. When it comes to explaining the presence or absence of specific features in linguistic interactions these two kinds of factors should be teased appart. Namely, population drift can *magnify* changes that are already under way in subpopulations, but it does not *cause* change by itself.

This is generally true, unless we assume that the "choice" of linguistic variants is genuinely random, that is, unrelated to any systematic biases and selection pressures related to learning and/or usage by speakers/signers. In the case of neutral linguistic features, mere population drift could still give rise to diversification, and might hence be conceptualized as a direct causal factor. For instance, if speakers of Old English "chose" purely randomly from the set of possible plural noun forms, population drift could still causally explain why certain forms became fixated in certain subpopulations and not others. However, as has been hinted at above, it is questionable whether the usage of linguistic forms is ever purely random in this sense.

3.1.2 Language genealogy

Similar considerations are relevant for the phylogenetic dimension. Language families are established on the basis of structural similarities, be it in the phonological, lexical, morphological, or syntactic domain. The distance between languages with regards to structural features is taken as an indication of close, distant or non-existent genealogical relatedness.

Figure 3.3 illustrates the hypothetical evolution of speaker populations and linguistic interactions starting with a single ancestral population at $t = 1$, and finally grouping into separate branches (A and B, e.g. families or genera) at time $t = 8$.

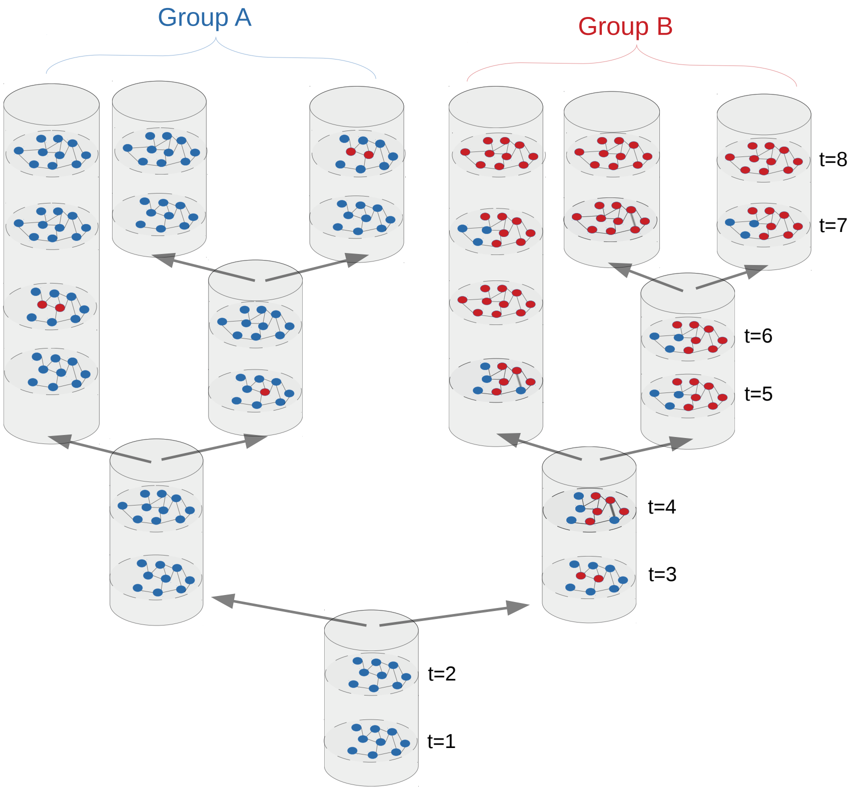

Figure 3.3: Languages grouping into families. Hypothetical phylogenetic tree with branches of two separate phylogenetic groupings which coincide with population splits. Group *A* (left) and group *B* (right) evolved along points in time (t=3 to t=8). The colour of individual speakers/signers represents the presence (red) versus absence (blue) of a particular linguistic feature of interest.

In reality, we often do not know the exact population history as outlined here. Instead, languages between $t = 1$ to $t = 8$ would be represented by fairly restricted subsamples of the actual linguistic interactions, for instance, looking at lists of lexical items or cognates. This is a common method applied in phylogenetic linguistic analyses to build the (presumably) most reliable phylogenetic trees. The structure of a phylogenetic tree derived from lexical material might very well differ from another tree based on other linguistic material. In this particular example, assume it is the distribution of the red and blue features we want to explain. They could reflect, for instance, morphological marking strategies, with red colour reflecting the usage of case markers to indicate *who did what to whom* and other case relations, and blue colour reflecting the absence of such markers.

Now, we might pose the question: *why* do the languages in group A generally not exhibit case marking at time $t = 8$, while the ones in group B do? One answer might refer to the whole group (be it family or genus), simply stating that in the history of group A (from $t = 3$ to $t = 7$) there has never been extensive case marking in the respective languages. Hence, it was *a priori* unlikely to arise in languages at $t = 8$. To give a real world example, we can ask why German has morphological case marking and Mandarin Chinese does not. We could refer to the fact that case marking is typical for languages of the Indo-European family, but not for those of the Sino-Tibetan family. However, this “explanation” is barely scratching the surface. It merely hints at some unknown past processes shaping languages of different families in different fashions.

Just as in the population drift scenario above, a satisfying answer should explain *why* case marking started to be adopted by some speakers in group B at time $t = 3$ and $t = 4$, and how and why it spread in the following generations. Likewise, it is important to know why case marking did not spread in group A at $t = 5$ and $t = 6$, though some speakers started to adopt the strategy. Again, within the CAS framework, the answers to these questions are linked with the competences of speaker populations, and hence with factors such as population size, adult learning, and language status.

To sum up, populations of speakers/signers can differ synchronically and diachronically with regards to their network competences and – as a direct consequence – the encoding strategies they adopt in their linguistic interactions. Thus, variation and change in $\mathcal{L}(t)$ is tightly linked with variation and change in the population $\mathcal{S}(t)$. We have focused here on three factors pertaining to the population of speakers and their network competences: population size, language contact, and language status. These are henceforth referred to as *explanatory* factors. They are primary factors for understanding change in languages over time. Population drift can magnify these effects and yield specific geographical patterns of the distribution of linguistic structures. Similarly, phylogenetic modelling sheds light on how languages relate to each other with regards to the subsample of linguistic interactions we choose to represent them with. However, both geographical and phylogenetic patterns are not genuine explanatory factors. Rather, they are secondary factors of *grouping*.

3.2 A note on “internal” and “external” factors of change

So far we have discussed factors of change related to the population of speakers. This raises the question how the CAS model relates to the classical distinction between language “internal” and “external” factors of change. This dichotomy is pervasive in studies of language evolution and historical language change, and crosses the boundaries of opposing views. It is ubiquitous in frameworks related to sociolinguistics (Croft, 2000; Jones and Esch, 2002; Jones and Singh, 2005), genetic linguistics (Thomason and Kaufman, 1988), as well as principles-and-parameters (P&P) theory alike (Briscoe, 2000b,a; Clark and Roberts, 1993; Lightfoot, 1979; Pintzuk et al., 2000; Yang, 2000). However, definitions of “internal” and “external” factors vary considerably across these frameworks. They can be interpreted in at least two ways.

Structuralist interpretation

Within accounts influenced by structuralism languages are viewed as closed systems of intertwined structures at different levels: phonology, morphology, and syntax. In this view, linguistic systems inherit structural features from their proto-languages and follow a “natural” or “normal” path of change, according to general principles such as *assimilation*, *analogical extension* and *analogical levelling* (Jones and Singh, 2005, p. 18-19). A typical example is the nominal plural marking system of Old English as outlined above. While the plural markers of nouns in OE could differ according to declension classes (e.g. *stānas* ‘stones’, *scipu* ‘ships’), in Modern English the plural -s is analogically extended to the vast majority of nouns. The structuralist interpretation of such “internal” language change is a reordering of markedness patterns in a closed language system (Thomason and Kaufman, 1988, p. 22). Having said this, language “external” factors then refer to *language contact*, i.e. the impact of child bilingualism or adult second language learning on a target language. For more extensive discussions of the structuralist take on “internal” and “external” factors of change see Thomason and Kaufman (1988, p. 1-12), Jones and Singh (2005, p. 1-29), and Lightfoot (1979, p. 381).

Generative interpretation

Generative accounts of language change tend to focus on synchronic language acquisition of native speakers, though there are extensions to diachrony (Briscoe, 2000a; Clark and Roberts, 1993; Lightfoot, 1979; Niyogi and Berwick, 1997; Roberts and Roussou, 2003; Roberts, 2007; Yang, 2000). In the generative context, "internal" refers to the innate blueprint of UG that limits the space of possible grammars (Yang, 2000, p. 232). The internal force in the context of P&P accounts is UG, an innate, universal and presumably fixed set of parameters. This is spelled out most clearly in Clark and Roberts (1993, p. 340), but see also contributions in Biberauer et al. (2010) for more recent elaborations within the Minimalist framework. Since the input varies depending on the language a child learns, parameters are set in acquisition to adjust for these particularities. The "external" dimension then refers to language input during acquisition.

Beyond "internal" and "external" factors

It has been argued recently that both of these accounts are somewhat limited for modelling and explaining language change. As pointed out by Jones and Singh (2005, p. 25-26), the structuralist perspective on language change lacks explanatory power. Although the descriptive tools of this framework are elaborate and elicit *what* is happening in great detail, they are generally silent on *why* it is happening. Structuralists might refer to earlier stages of a language as an explanation for the loss of current features, but this just dilutes from the important question of *why* specific features started to erode in the first place. Likewise, holding system internal pressures responsible for changes falls short of accounting for the triggering events. Jones and Singh (2005, p. 26) discuss these shortcomings more explicitely referring to the *Great Vowel Shift* in Middle English.

The problem with the P&P view, on the other hand, is that its notion of "internal" factor is too broad. Namely, the internal factor UG only requires a grammar to be learnable or "possible". As a matter of fact, all the grammars of languages across the world are learnable and "possible" in this sense. By definition, all natural languages are within the scope of UG. Hence, UG as an internal factor is not predictive when it comes to changes in specific languages, and is therefore only marginally relevant if we want to model how languages evolve over time. Again, this seems to be in line with the newest facet of the minimalist program as outlined in Chomsky (2010).

Moreover, the notion "external" in the P&P sense refers to linguistic input for the next generation of learners. Variation in the input will cause the learners to select different viable grammars (Roberts and Roussou, 2003; Roberts and Holmberg, 2010; Yang, 2000). That is, learners are exposed to a random, limited sample of input sentences from which they can analytically derive a possible grammar. This mechanism will lead to convergence with adult grammars in most cases, but also produce new grammars in a few non-converging cases. These cases inevitably give rise to language change (Roberts and Holmberg, 2010, p. 53-54). Note, however, that this framework does not attempt to explain in more detail where the variation in the input for the next generation of learners stems from. Interestingly, there are P&P accounts asserting that variation can be triggered by migration, phonological erosion, and linguistic innovation (see Lightfoot, 1979, p. 381; Niyogi and Berwick, 1997, p. 717; and Yang, 2000, p. 237), but neither the mechanisms nor the outcome of such genuinely "external" pressures are modelled in more detail.

In conclusion, the definition of "internal" factors in P&P accounts is too broad to predict changing encoding strategies and the definition of "external" factors does not include a detailed account of how language contact and other sociolinguistic pressures can trigger change.

Finally, researchers working within both structuralist and generative frameworks often automatically assume that an explanation has to be either "internal" or "external". However, Farrar and Jones (2002, p. 3) argue that this *either-or* mentality is bound to be misleading when dealing with complex phenomena such as language change. Rather than imposing strict distinctions upon gradual and intertwined phenomena, they propose to highlight the interplay of various factors instead (Farrar and Jones, 2002, p. 8).

In line with these considerations, a strict distinction between language "internal" and language "external" factors is not backed by the usage-based variant of the CAS model. A strict distinction can only be upheld if we assume that there is a universal competence state independent of linguistic interactions – as argued by the saltational account in Section 2.3.1. In contrast, within the usage-based account linguistic interactions and competence states – and hence "internal" and "external" factors – are intertwined and change in parallel over time.

3.3 Description, grouping, and explanation

According to the arguments given in the previous section, the "internal" vs. "external" distinction is not of central importance here. Instead, it is necessary to tease apart so-called *descriptive*, *explanatory* and *grouping* factors.

Descriptive factors

Descriptive factors are such that help describe the entirety of linguistic interactions of a speaker population, or a subsample of these. Essentially, this is any information – phonological, morphological, syntactic or otherwise – we can gather about a language at a certain point in time $\mathcal{L}(t)$. Note that depending on which descriptive features of that language we are interested in, other descriptive features might function as predictors. For example, the loss of phonological distinctions, e.g. between a bilabial nasal [m] and an alveolar nasal [n], can lead to – and hence seemingly "explain" – the loss of morphological distinctions. However, such an "internal" account is not the ultimate explanation in the usage-based CAS model. In this example, it only shifts the focus from a question about morphology to a question about phonology, but does not provide a causal answer to either.

Explanatory factors
In contrast, *explanatory* factors are tightly linked to speaker/signer populations and their competences. A change in the population, and hence the network competences, directly predicts and explains changes in linguistic interactions. In the usage-based CAS model, there is little or no change from $\mathcal{L}(t = 1)$ to $\mathcal{L}(t = 2)$ without a change in the network competences from $\mathcal{NC}(t = 1)$ to $\mathcal{NC}(t = 2)$. The network competences, in turn, are a reflection of the speaker/signer populations $\mathcal{S}(t = 1)$ and $\mathcal{S}(t = 2)$. Explanation in this sense is the ultimate goal of linguistic analysis.

Grouping factors
Finally, if changes in competences and linguistic interactions happen within a certain branch of a phylogenetic tree at time $t = 1$, then the "offspring" languages $\mathcal{L}(t > 1)$ of this branch are more likely to carry the changes. Likewise, if they happen before a population drift, this drift might magnify the differences in subpopulations and lead to new languages distinguished by the changed features in question. Hence, phylogenetic and geographical grouping patterns are predictive with regards to linguistic features, but not explanatory by themselves. They constitute a third kind of factor, here referred to as *grouping* factor.

3.4 Summary

In conclusion, there is a range of characteristics at the population level that are directly linked with variation in language competences, and the linguistic interactions emanating from them. These include – but are not limited to – the size of populations, language contact via subpopulations of adult learners, and social status. This is not an exhaustive list. It could be extended to other factors, such as the degree to which languages are learned systematically in schools, or whether a written variant of the spoken language exists. These too might explain the presence or absence of structural features in linguistic interactions at a certain point in time. The number of predictors included in a statistical model is often only limited by the data available. On top of such explanatory factors, there are grouping factors relating to geographical patterning or family relationships. These can be predictive, but have to be linked to explanatory factors to constitute causal explanations.

Applying this usage-based CAS model to actual languages, we want to explain why a language A uses certain linguistic features to encode information, whereas language B does not. To this end, we first need to define the linguistic feature(s) we are interested in and consider all the descriptive factors relevant to their analysis. Second, we need to link these descriptive factors to explanatory and grouping factors in a quantitative model. The following chapter focuses on the first step. Namely, it discusses a core information-theoretic property of languages – their *lexical diversity* – as the linguistic feature to be cross-linguistically modelled and explained.

4 Lexical Diversity across Languages of the World

4.1 Introduction

As pointed out in the introduction to this book, languages across the world are astonishingly diverse. They use a potpourri of phonemes, morphemes, and syntactic structures to encode information. Modelling and explaining this diversity quickly becomes an infeasible endeavour. To reduce the complexity of the task, we have to decide which of the manifold linguistic aspects we are interested in more specifically. This decision will also determine our choice of data to represent languages.

If we are interested in variation in lexica, we can use cross-linguistic information about basic word lists, for instance, those collected in the *Automated Similarity Judgement Program* (ASJP) (Wichmann et al., 2013). If we are interested in structural features such as phoneme inventories, morphological marking strategies or word orders, we can use databases such as the *World Atlas of Language Structures* (WALS) (Dryer and Haspelmath, 2013). However, in the first case we reduce languages to a short list (40-100 words) of base vocabulary and in the second case we reduce languages to specific phonological, morphological or syntactic features defined by typologists.

On top of this, collecting information about structural features across hundreds of languages requires definitions of abstract linguistic categories such as *phoneme*, *morphological marker* and *subject*, *verb*, *object*. How many phonemes exist for any given language? What exactly qualifies as an accusative case marker cross-linguistically? How do we define the concept of subjecthood? These are theory-heavy decisions that need to be taken into account. Any further analyses of the data will carry this theoretical baggage.

Information theory offers us a more basic, quantitative, production-based, and less theory-driven perspective on the core properties of languages. Shannon (1948) considered the purely probabilistic aspects of the English alphabet to be informative as to how the encoding of information works in this language. He developed a measure, the *entropy*, to capture the information encoding potential of “events”. These events could be phones in a speech stream, letters, words, symbols, or sentences in writing, or signs in the case of sign languages. Shannon’s theory is agnostic to the exact information encoding units chosen – as long as they can be measured. His proposal essentially boils down to asking: how much surprisal, uncertainty, or *choice* is associated with the units we use? Simply put, if we repeat the same unit over and over again, i.e. with high frequency, then there is not much surprisal in our code, and hence not much information. Whereas if

DOI 10.1515/9783110560107-004

we use a multitude of different units with low frequencies, then we can transmit more information. In written language production in particular, the information encoding potential is closely related to the number different characters, words, phrases, etc.

The account outlined in this book focuses on the *word* as basic information encoding unit. The abundance of word forms or *word types* in relation to their *token* frequencies will henceforth be referred to as *lexical diversity* (LD). Lexical diversity is a core information-theoretic property of a given text, corpus and – by extension – a language. The range of lexical diversities we find across languages of the world is part of the puzzling diversity of information encoding strategies. Lexical diversity can be calculated directly from written language production, and does not come with much theoretical baggage. In the remainder of this book, lexical diversity across languages of the world is the phenomenon to be measured and explained. Estimating lexical diversity per language, and measuring differences in lexical diversity between languages is the central topic of this chapter.

4.1.1 Sampling from languages: parallel corpora

Remember that we defined the language of a population of size n as the multiset of *linguistic interactions* within that population at time t, i.e. $\mathscr{L}(t)$. To calculate the lexical diversity for an entire language we would need to capture the totality of linguistic interactions between speakers of that population at time t. Moreover, to compare two languages A and B we would have to capture $\mathscr{L}(t)^A$ and $\mathscr{L}(t)^B$. Clearly, this is downright impossible. However, we can sample from languages using written and/or spoken corpora to approximate the complete set of linguistic interactions and their lexical diversities. There are two major issues here:

1. Any subsample might not be representative of the whole language. This issue is typically addressed in corpus linguistics by using *balanced corpora*, i.e. corpora that represent different registers and styles.

2. The *content* of texts chosen to represent a language can be a major confound. Different contents might yield different LDs and hence blur the actual differences between LDs for languages. Hence, we need to find samples for different languages that (at least roughly) encode the same information, i.e. *parallel corpora*.

We face a trade-off when we try to address both of these issues at the same time: If we use maximally representative and balanced corpora, such as the *British Na-*

tional Corpus (BNC) for English, it is impossible to find a corpus that is exactly parallel in another language. On the other hand, if we use use parallel corpora to make the lexical diversity comparison across languages more meaningful, we risk not representing the respective languages to a satisfying degree.

In this study, the strategy is to use parallel corpora of different registers to address both issues to some extent. This approach has been proposed within the framework of quantitative typology (Cysouw and Wälchli, 2007; Dahl, 2007). To assess how strongly corpus composition biases our results, a part of the analyses involve estimations of systematic differences between registers.

The parallel corpora used in this study are the *Universal Declaration of Human Rights* (UDHR),[1] the *Parallel Bible Corpus* (PBC) (Mayer and Cysouw, 2014), and the *Europarl Parallel Corpus* (EPC) (Koehn, 2005). The UDHR comprises a collection of more than 400 parallel translations, though not all of these are fully converted into unicode. They represent more than 300 different languages (unique ISO 639-3 codes). The UDHR is a short legal text of 30 articles and the average number of word tokens across the texts is ca. 2000. The PBC is a collection of parallel translations of the Bible. It comprises around 1500 texts of more than 1000 languages.[2] The average number of tokens per text is in the ten thousands. Note that this corpus is not perfectly parallel since particular verses might not be represented in a given translation. The EPC is a collection of transcripts of discussions in the European Parliament in 21 European languages. The average number of tokens is around 7 million.

4.2 Theoretical preliminaries: type and token frequencies

Types and tokens

Any measure of variation in lexical diversity of texts, corpora and languages has to be based on the distinction between *word types* and *word tokens*. Since we work with written language and want to automatically process it, we need to assume a technical definition. A *word type* is here defined as a unique string of unicode characters (lower case) delimited by non-alphanumeric characters (e.g. white spaces and punctuation). A *word token* is then defined as any recurring instance of a specific word type. For example, the first sentence of the first article of the *Universal Declaration of Human Rights* reads:

[1] http://www.unicode.org/udhr/

[2] As of June 2016.

(1) English (eng, UDHR 01)[3]
All human beings are born free and equal in dignity and rights.

The set of word types (in lower case) for this sentence is

$$\mathcal{T} = \{\text{all,human,beings,are,born,free,and,equal,in,dignity,rights}\}. \tag{4.1}$$

Hence, the number of word types in this sentence is 11, but the number of word tokens is 12, since *and* occurs twice. Note that word types here include different word forms of the same lemma, such that *right* and *rights* are two separate word types. In theory, this definition of word types and tokens is straightforward, and underlies a good portion of corpus and computational linguistic studies. In practice, however, there are some caveats even to this simple definition. Grefenstette and Tapanainen (1994) have pointed out that tokenization is a non-trivial issue, especially for very big corpora. There is a whole range of decisions that need to be taken on the status of words and word delimiters. When analysing a wide range of different languages, different writing systems, and different scripts such issues become even more apparent.

For example, characters such as the apostrophe (’) can be ambiguous. In some languages it denotes clitics and contractions as in English *John’s* or *she’s*, representing *John=GEN.SG* and *she is* respectively, while in other languages, it denotes phonemic distinctions such as ejectives or glottal stops. For instance, in the UDHR translated into Cuzco Quechua (quz), we encounter the words *k’iri* ‘wound’, *p’unchay* ‘day’, and *sut’in* ‘truth’. In combination with the preceding consonants, the apostrophes here denote velar, bilabial and dental ejectives. Furthermore, in adaptations of Latin scripts used to write Mesoamerican languages, e.g. of the Otomanguean family, the apostrophe can represent a glottal stop. We find plenty of examples such as *xa’bi’* ‘man/human’, *che’n* ‘of’, *co’* ‘who/what’ (Ruegsegger and Ruegsegger, 1955) in the Miahuatlán Zapotec translation of the Bible. In English, genitive clitics and contractions might – or might not – be teased apart and analysed as separate word types, depending on our linguistic analysis. In Chapter 5, it is assessed how much difference such decisions make for word type distributions in English and German. In Cuzco Quechua, Miahuatlán Zapotec, and other languages with ejectives and glottal stops, apostrophes should clearly not be part of non-alphanumeric characters on which to split word types.

Another problematic example is numerical tone marking, which is widespread in Latin scripts used to write tone languages of Mesoamerica, e.g. Otomanguean languages of the Chinantecan branch (Skinner and Skinner, 2000). For instance,

[3] Throughout this book, example sentences are accompanied by a line giving the language name, its ISO 639-3 code, the acronym of the corpus, and a line identifier.

in the Bible translation into Usila Chinantec (cuc), the name *Abraham* is written *A*^{3}*brang*^{23}. The numbers here indicate that the first vowel is pronounced with constant mid-level pitch, while the second vowel has a falling pitch contour. Importantly, tone numbers are by default included in the set of non-alphanumeric unicode characters. Hence, a standard tokenization algorithm splits on tone numbers and will yield *A brang* as two separate word types.

Fortunately, typologically aware corpora such as the Parallel Bible Corpus are specifically curated to use white spaces in an informative manner. Namely, they delimit word forms independent of punctuation, as in the examples below.

(2) English, Darby version (eng, PBC 01001005)
And God called the light Day , and the darkness he called Night .

(3) Cuzco Quechua (quz, PBC 01001005)
Hinaspan Diosqa k'anchayta “ P'unchay ” nisqata suticharqan , laqhayaqtataq “ Tuta ” nisqata suticharqan .

(4) Usila Chinantec (cuc, PBC 40001001)
I^{4}*la*^{3} *ti*^{2}*ton*^{3} *la*^{4}*jang*^{34} *sa*^{1}*jeun*^{3} *quian*^{1} *Jesucristo a*^{3}*lang*^{43} *jon*^{43}*tyie*^{1}
A^{3}*brang*^{23} *jian*^{3} *Da*^{3}*vei*^{23} *.*

(5) Miahuatlán Zapotec (zam, PBC 40001001)
Loo libr ndxè' nda' cuent cón che'n rye mèn co' ngòc xudgool che'n Jesucrist co' nde bin David co' ngòc rey póla .

Note that there are white spaces between characters and punctuation, but not between apostrophes and tone markings that are part of a word type. This simplifies the problem of tokenization considerably. Adding white spaces is also an important practice in scripts that delimit words by other means. The first sentence of the UDHR in Amharic (amh), written in the Ge'ez script, is given as an example here.

(6) Amharic (amh, UDHR 01)
Original:
የሰው፡ልጅ፡ሁሉ፡ሲወለድ፡ነጻና፡በክብርና፡በመብትም፡እኩልነት፡ያለው፡ነው።
White spaces added:
የሰው ፡ ልጅ ፡ ሁሉ ፡ ሲወለድ ፡ ነጻና ፡ በክብርና ፡ በመብትም ፡ እኩልነት ፡ ያለው ፡ ነው ።

In the traditional Ge'ez script, words are delimited by a “colon” (፡), rather than a visual space in between the characters. These need to be added separately to enable automated tokenization. Yet other writing systems and scripts do not give word boundary indications at all. A well-known example is Mandarin Chinese.

(7) Mandarin Chinese (cmn, UDHR 01)
人人生而自由, 在尊严和权利上一律平等.

The Chinese script is often referred to as a logography, i.e. a writing system representing each word by a separate logogram. However, at close inspection it is arguably more of a “phonetically imprecise syllabary” (Mair, 1996, p. 201). This means it encompasses up to sixty thousand so-called *sinograms* that can sometimes represent semantic radicals such as 足 *zú* ‘foot’, but are most often (namely 81% of the time) a combination of phonetic and semantic subcomponents (Mair, 1996, p. 201).

There is some rudimentary punctuation (a comma and a full stop) employed in the Chinese example, but there are no strict and coherent rules of when to use them. Thus, translations into Mandarin and Cantonese (cmn, yue), as well as some other languages, including Japanese (jpn), Mon Khmer (khm), and Burmese (mya) do not lend themselves to simple automated tokenization. There is work under way to semi-automatically add white spaces between words for these scripts too. However, at this point, these languages are not included in the sample.

Overall, it was outlined in this section that there are systematic ways of dealing with problems relating to the diversity of scripts. As a consequence, the orthographic definition of “wordhood” is often taken as a given in corpus and computational linguistic studies. Nevertheless, it is controversial from a linguistically more informed point of view.

The indeterminacy of words

What is a “word” in the first place? From a psycholinguistic and typological point of view, the mere existence of words as a cross-linguistically coherent category is questionable. Haspelmath (2011), for instance, points out that there are at least ten different ways of defining words, based on phonotactic, morphological or syntactic criteria. He argues that none of them are “necessary and sufficient” by themselves and, what is even worse, no combination of these yields a definition that unequivocally matches common orthographic practices. For example, the bigram *car park* in English is probably just as cognitively entrenched as *Parkplatz* in German (literally meaning ‘parking spot’) and probably does not bear more of a pause in speech, but it is written with a white space in English, and in German without. Likewise, the German infinitive marker *zu* ‘to’ is sometimes separated from verbs, as in *zu gehen* ‘to go’, while it is integrated in the verb form when preceded by a particle, as in *wegzugehen* ‘to go away’ (Haspelmath, 2011, p. 36). Such orthographic decision seem arbitrary.

The problem of a coherent wordhood definition becomes particularly apparent when considering polysynthetic languages, which are traditionally defined as combining a range of different stems and affixes into complex word forms. Consider the following example from Mapudungun (arn), a polysynthetic language of Southern Chile (example from Salas 2006 as cited in Bickel and Zúñiga 2017).

(8) Mapudungun (arn)

pepi-rume-küme-wentro-nge-tu-rke-i-ngu.
can-very-good-man-be-TEL-REP-IND-3DU

"Both of them were able to turn into very rich (lit. good) men, they say."

From the point of view of linguistic analysis, it can be argued that a whole sentence is here construed as a single complex word form by incorporating the noun phrase *rume-küne-wentro* 'very good man' into a construction involving a modal marker *pepi* 'can', the copula *nge* 'to be', as well as further morphological markers. However, as pointed out by Bickel and Zúñiga (2017), speakers of Mapudungun might actually write something akin to *pepi rume küme wentru ngeturkeingu*, due to influence by Spanish spelling conventions.

Along similar lines, Wray (2014) suggests that spelling conventions using white spaces are rather the cause for – than the effect of – our intuition that words are elementary units of language. Hence, orthography could be the main reason for our intuition that words exist as meaningful units of language. In this view, rather than being caused by a universal psycholinguistic bias to think in word units, white spaces are more of an illusion introduced by the somewhat arbitrary orthographic practices of Western scholars.

With regards to natural language processing, the bias to think about sentences as strings of clearly defined words has potentially led to many inefficiencies. Grefenstette (2010) makes a first attempt to estimate the number of multiword concepts used in the English language, and argues that these, rather than single words, will be the basic units of future natural language processing systems.

In the context of modelling processes of language learning, Baayen et al. (2016) present a computational model of auditory comprehension that does not explicitly try to segment speech input into neat word units, but is trained on whole utterances and finds statistical regularities in continuous input along a moving window. This "comprehension without segmentation" model is shown to exhibit sensitivity to statistical patterns comparable to infants in artificial language studies.

Thus, there are good reasons to be critical of orthographic wordhood definitions. Advanced typological, computational and psycholinguistic analyses are important to further understand the basic units of information encoding in natural languages and their acquisition. For example, Stoll et al. (2017) analyse striking differences in the distribution of verb forms in English and Chintang, a polysynthetic language of Nepal. This study is based on a typologically refined definition of a word type (more specifically verb form type) and it harnesses data from longitudinal acquisition corpora. This allows a much finer-grained assessment of linguistic differences than analyses based on orthography.

However, there is also evidence that some of the theoretical writing on the indeterminacy of words might be overly pessimistic. A recent study on the *informativeness of linguistic unit boundaries* (Geertzen et al., 2016) applies standard compression algorithms to parallel texts of the European Parliament Corpus in English (Indo-European), Finnish, Estonian, and Hungarian (all Uralic) and investigates which unit boundaries in written language are the most important from an information-theoretic point of view. It illustrates that word boundaries (i.e. white spaces in written corpora), rather than morpheme boundaries or sentence boundaries, are most informative. In other words, white spaces are most efficient for finding and compressing regular patterns in written language. Arguably, the range of typologically different languages is small in this study, and particularly the analyses comparing the informativeness of morpheme boundaries and word boundaries are limited to English, as there are currently no high-quality morphological analysers or large manually tagged corpora available for other languages. Still, Geertzen et al. (2016) is a first important step to vindicate the usage of words in written corpora as basic information encoding units. In fact, building on this and further recent studies, Blevins (2016) develops a new framework for morphological analysis based on information-theoretic considerations.

Word frequency distributions

Given our definition of word types and word tokens we can define lexical diversity more precisely as the *distribution of word tokens over word types (given constant content of a message)*. In other words, lexical diversity is the distribution of word token frequencies per word types.

To formalize this, let $\mathcal{W} = \{w_1, w_2, w_3, ...w_V\}$ be the set of word types, where V is the vocabulary size, i.e. the number of different word types. For this set of word types there is a distribution of token frequencies $F = (f_1, f_2, f_3, ...f_V)$ which reflects the frequency count of each word type (in a given corpus) such that for example $f_1 = \text{freq}(w_1)$. Note that the term “frequency” here means *frequency count* – without normalization. For example, the “frequency” of the word type *the* in the UDHR is 121. Given this definition, the overall number of tokens N for as set of types is therefore

$$N = \sum_{i=1}^{V} f_i. \tag{4.2}$$

Clearly, there are differences in how word frequencies are distributed over types in different sentences, texts and in whole corpora. An example of two differing word frequency distributions, namely a *uniform* distribution of equal frequencies and a *non-uniform* distribution of varying frequencies, can be seen in Figure 4.1.

In linguistic examples, the ranks (x-axis) of these frequency distributions correspond to word types, and the frequencies on the y-axis to the number of tokens per word type. To get a better overview of the rank/frequency profile, we can follow Zipf (1932, 1935, 1949) and rank the distribution of token-counts (i.e. the distribution F) from highest to lowest. The actual non-uniform ($F_{\text{non-uni}}$) and uniform (F_{uni}) token frequency distributions chosen for illustration in Figure 4.1 are:

$$\begin{aligned} F_{\text{non-uni}} &= (45, 20, 15, 10, 5, 1, 1, 1, 1, 1) \\ F_{\text{uni}} &= (10, 10, 10, 10, 10, 10, 10, 10, 10, 10) \end{aligned} \tag{4.3}$$

These are directly reflected on the y-axis of Figure 4.1 a) and logarithmically transformed in Figure 4.1 b). The log-transformation is a convention to make the full distribution visible, even for long-tailed distributions with thousands of ranks.

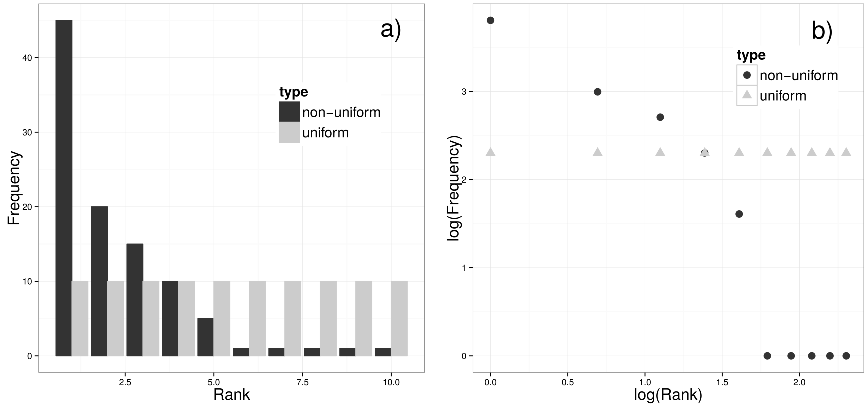

Figure 4.1: Uniform and non-uniform frequency distributions. An example of visually comparing a uniform (grey) to a non-uniform (black) frequency distribution. a) Illustrates the frequencies of the two distributions ranked from highest to lowest. b) Illustrates the log frequencies and log ranks for the uniform (grey triangles) and the non-uniform distribution (black dots).

The purely numerical examples in Figure 4.1 and Equations 4.3 illustrate how frequency distributions can differ. While the non-uniform distribution is skewed towards the y-axis (i.e. higher frequencies), the uniform distribution is an example of skewness towards the x-axis (i.e. low frequencies). In linguistic examples the “shape” or “skewness” of a word frequency distribution reflects its lexical diversity. To compare lexical diversities we need measures to capture these distributional differences.

Measuring lexical diversity

There is a wide range of lexical diversity measures in quantitative and applied linguistics (see Baayen, 2001; Jarvis, 2002; McCarthy and Jarvis, 2007, 2010; Mitchell, 2015; Tweedie and Baayen, 1998, for an overview). For example, Mitchell (2015) reports a total of 50 different models based on the so-called *type-token ratio* (TTR). The TTR is simply the number of word types divided by the overall number of word tokens, i.e. in our notation

$$TTR = \frac{V}{\sum_{i=1}^{V} f_i} = \frac{V}{N}. \tag{4.4}$$

Taking the ratio of word types to word tokens is one of the simplest and crudest ways to approximate the “shape” of a word frequency distribution. However, it has been argued in Tweedie and Baayen (1998) that TTRs strongly depend on the size of the language sample (i.e. the overall number of tokens), and that this is an undesirable property of a lexical diversity measure. As a consequence, a series of modified measures was developed to render LDs independent of sample sizes (see Jarvis, 2002; McCarthy and Jarvis, 2007, 2010).

These LD measures can be roughly divided into two categories: *parametric* and *non-parametric* measures. *Parametric* measures require an underlying model – e.g. the Zipf-Mandelbrot model (Mandelbrot, 1953) – that needs to be fitted to the empirical ditributions. In contrast, non-parametric methods do not assume an underlying model. Furthermore, the non-parametric measures might be divided into ones that are based on TTR, and those which are not. In principle, any of these LD measures could be used to approximate the “shape” of a word frequency distribution, and to calculate the difference between distributions (i.e. ΔLD). However, there are advantages and disadvantages to each of them.

For example, when it comes to fitting parametric models, it is difficult (though possible) to determine the right balance between underfitting by using the most parsimonious model (few parameters) and overfitting by using models with many parameters. Hence, the results of simple parametric curve fitting procedures are

Table 4.1: Differences in lexical diversities between uniform and non-uniform distributions as depicted in Figure 4.1.

Measure	non-uniform	uniform	ΔLD	Type
ZM α	8.67	NA	NA	parametric
ZM β	12.45	NA	NA	
HD-D	7.04	9.97	2.93	
Shannon H	2.27	3.32	1.05	non-parametric
Yule's K	2680	900	1780	
TTR	0.10	0.10	0	non-parametric (TTR-based)
MSTTR	0.17	0.10	0.07	
MATTR	0.16	0.19	0.03	
Herdan's C	0.50	0.50	0	
Guiraud's R	1.00	1.00	0	
CTTR	0.71	0.71	0	
Dugast's U	4.00	4.00	0	
Summer's S	0	0	0	
Maas index	0.50	0.50	0	
MTLD	2.20	2.04	0.16	

Note: details about the LD measures used here (except for Zipf-Mandelbrot's α and β, and Shannon entropy H) can be found in Michalke (2014).

an easy target for criticism. Altmann and Gerlach (2016) discuss these issues more extensively and give potential remedies.

Table 4.1 gives values for some of the most well-known LD measures applied to the uniform and non-uniform distributions (F_{uni}, $F_{\text{non-uni}}$) from above (Equation 4.3). Most of these values were calculated using the *koRpus* package in *R* (Michalke, 2014). ZM parameters are estimated by using the *likelihood* (Murphy, 2013) package, and Shannon entropy (Shannon and Weaver, 1949; Shannon, 1951) by using the *entropy* (Hausser and Strimmer, 2014) package respectively.

Importantly, only the ΔLD values of the measures marked in grey reflect the difference of the uniform and non-uniform distributions. For all the other measures ΔLD is 0. This is because measures such as *Herdan's C*, *Guiraud's R*, *CTTR* etc. are all based on TTRs, and TTRs are insensitive to the differences in the distributions given in Equations 4.3. To see this, note that the number of types is 10, and the total number of tokens is 100 for both the uniform and the non-uniform distributions. Hence, the TTRs are

$$TTR_{\text{uni}} = TTR_{\text{non-uni}} = 10/100 = 0.1. \tag{4.5}$$

Clearly, in more realistic distributions derived from actual texts, it is very unlikely that the TTRs are exactly the same. However, the point about uniform and non-uniform distributions illustrates that simple TTRs and measures based on them are insensitive to the exact distributions of word frequencies. All that matters for TTRs, by definition, is the total number of tokens and the overall number of types.

In contrast, measures such as ZM parameters and Shannon entropy reflect the "shape" of a distribution of word frequencies in more detail. Though a problem with ZM parameters in this specific example is that the algorithm gives NAs for the uniform distribution. Nevertheless, for word frequency distributions of real texts estimation of ZM parameters is a viable method. It has been applied to measure changes in frequency distributions over historical time (Bentz et al., 2014; Koplenig, 2015; Chand et al., 2017), in language learning (Baixeries et al., 2013), and across different languages (Bentz et al., 2015).

There are further potential LD measures listed, such as *HD-D*, *Yule's K*, *MSTTR*, *MATTR* and *MTLD*. Each of these reflects the difference between the uniform and non-uniform distributions (though for MSTTR and MATTR this difference is minor). Hence, all of these could, in principle, be suitable to measure differences in frequency distributions across languages. Note, though, that these were specifically developed with the objective of being constant across text sizes, a property that requires specific modifications of the original TTRs, and makes these measures less straightforwardly interpretable.

Finally, Shannon entropy comes with the advantage of being a non-parametric measure, not requiring curve fitting based on a hypothesized underlying model. Moreover, it is interpretable within the framework of standard information theory, it is directly linked to theories of complex systems, and it is widely applied and tested in many different contexts. Taking into account all these considerations, Shannon entropy is arguably one of the most convenient measures for LD. It is discussed in more detail below.

4.3 Word entropy

In *A mathematical theory of communication*, Claude Shannon (1948) laid out his account to measure the information potential of a symbol system transmitted via a serial channel and thereby founded modern day information theory. The same paper was republished a year later with minor changes – and an introduction by Warren Weaver – as *The mathematical theory of communication* (Shannon and Weaver, 1949). *Entropy* is a core concept in this work. It is a fundamental property of symbol strings based on a given symbolic repertoire. In natural languages, this could be phonemes, morphemes, words, phrases, or any other definable unit of information encoding. The theoretical entropy of a set of possible "events", better conceptualized here as word types, and their probabilities $p_1, p_2, ... p_n$ is defined as (Shannon and Weaver, 1949, p. 50):

$$H(p_1, p_2, ... p_n) = -K \sum_{i=1}^{n} p_i \ \log \ p_i, \tag{4.6}$$

where K is a positive constant determining the unit of measurement, and n is the number of different types. The default is $K = 1$, and the logarithm taken to the base 2, yielding *bits* of information. The mathematical notation used to define entropy can vary. Another standard way is to define a discrete random variable X with "alphabet" (i.e. finite set of values) denoted by $\mathcal{X}$. The probability mass function is $p(x) = Pr\{X = x\}, x \in \mathcal{X}$. This gives entropy equivalently as (Cover and Thomas, 2006, p. 13-14):

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \ \log \ p(x). \tag{4.7}$$

The logarithm is by default to base 2. The Shannon entropy in Equation 4.6 and 4.7 is often circumscribed by the term "uncertainty". Shannon himself asked:

> Can we find a measure of how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome? (Shannon and Weaver, 1949, p.49)

Hence, "choice" and "uncertainty" are two sides of the same coin. We choose symbols to transmit information, and we create uncertainty to be deciphered by the recipient. When we talk about language, the term "choice" resonates better with our intuition that we try to transmit ideas to the hearer and thereby create more certainty, rather than uncertainty. However, the tight link between "choice" and "uncertainty" reflects a fundamental information-theoretic trade-off in natural languages and communication systems more generally: more choice enables the creation of more elaborate messages, but also increases uncertainty, and hence impedes rapid interpretation. For an extended discussion of information-theoretic trade-offs in natural language see Ferrer-i-Cancho (2017b).

To illustrate the intuition behind entropy, take the example of uniform and non-uniform distributions from above: the uniform distribution assigns equal probability to each "event", i.e. word type, whereas the non-uniform distribution is skewed in terms of word type probabilities. To put it differently, if we choose a word token at random from a text with uniform word type probabilities, e.g.

$$P_{uni} = \left(\frac{10}{100}, \frac{10}{100}, \frac{10}{100}, \frac{10}{100}, \frac{10}{100}, \frac{10}{100}, \frac{10}{100}, \frac{10}{100}, \frac{10}{100}, \frac{10}{100}\right), \tag{4.8}$$

then our best guess about which word type we will draw gives us a uniform $10/100$, i.e. $1/10$ chance to be right. However, if we draw from the non-uniform distribution, e.g.

$$P_{\mathit{non\text{-}uni}} = \left(\frac{45}{100}, \frac{20}{100}, \frac{15}{100}, \frac{10}{100}, \frac{5}{100}, \frac{1}{100}, \frac{1}{100}, \frac{1}{100}, \frac{1}{100}, \frac{1}{100}\right), \tag{4.9}$$

then our best guess is $45/100$, and the second best guess $20/100$, etc. Hence, in the uniform distribution, there is more “choice” giving rise to more “uncertainty”. Shannon realized that a measure of the information encoding potential would have to involve the individual probabilities of events, and it would have to be maximal for uniform distributions. These and further criteria led him to the formulation of entropy in Equation 4.6. In fact, the entropy of the uniform distribution in our example is

$$\begin{aligned} H_{\mathit{uni}} = -1 & \left(\frac{10}{100}\log_2\frac{10}{100} + \frac{10}{100}\log_2\frac{10}{100} + \ldots \right. \\ & \left. + \frac{10}{100}\log_2\frac{10}{100}\right) \sim 3.3 \text{ bits/event}, \end{aligned} \tag{4.10}$$

and for the non-uniform distribution it is

$$\begin{aligned} H_{\mathit{non\text{-}uni}} = -1 & \left(\frac{45}{100}\log_2\frac{45}{100} + \frac{20}{100}\log_2\frac{20}{100} + \ldots \right. \\ & \left. + \frac{1}{100}\log_2\frac{1}{100}\right) \sim 1.6 \text{ bits/event}. \end{aligned} \tag{4.11}$$

Thus, with about 3.3 bits per word type there is more choice/uncertainty in the uniform distribution than in the non-uniform distribution with approximately 1.6 bits per word type. Considering word types as “events”, entropy can be used as an index for lexical diversity. Namely, languages with more different word types of lower probabilities have higher lexical diversities and higher word entropies. This can be seen in parallel to the entropy index for biodiversity (Jost, 2006; Chao and Shen, 2003), where the distributions of species’ frequencies are compared across different habitats.

Before we further delve into particularities of word entropy estimation, a more general question needs to be addressed: is it even “meaningful” to count tokens and types and apply information-theoretic measures to natural language data?

4.3.1 Information, meaning, and natural language

Since Shannon’s seminal paper, information theory has found applications in a wide range of scientific disciplines including physics, engineering, biology, computer science, and many others. Right from the start, it was appealing to consider the mathematical theory of *communication* also a mathematical theory of *natural language*. In fact, one of the first real-world applications of entropy – proposed by Shannon (1951) himself – captured the uncertainty in characters of written English. Several studies have attempted to refine his approach, and approximate the entropy of written English (Brown et al., 1992; Schürmann and Grassberger, 1996; Kontoyiannis et al., 1998; Gao et al., 2008), as well as other languages (Behr et al., 2003; Takahira et al., 2016) with highest possible precision.

Entropic measures more generally have found a range of applications at different levels of language structure, starting with phonemes (Borgwaldt et al., 2004, 2005), and morphemes (Moscoso del Prado Martín et al., 2004; Milin et al., 2009; Ackerman and Malouf, 2013), extending to words (Montemurro and Zanette, 2011, 2016; Bentz et al., 2015, 2017b; Koplenig et al., 2017), and finally sentences (Fenk and Fenk, 1980; Fenk-Oczlon, 2001; Hale, 2001, 2016; Jaeger and Levy, 2006; Levy, 2008; Jaeger, 2010). Entropic measures are often a natural first choice to precisely measure "complexity" at different levels of language structure (Juola, 1998, 2008; Bane, 2008; Ehret and Szmrecsanyi, 2016a; Ehret, 2016; Ehret and Szmrecsanyi, 2016b; Bentz et al., 2016; Hale, 2016) – a concept that is otherwise often used in a rather intuitive and vague manner. The same and further studies have also investigated complexity trade-offs between different levels of structure (Juola, 2008; Moscoso del Prado, 2011; Montemurro and Zanette, 2011; Futrell et al., 2015; Montemurro and Zanette, 2016; Ehret and Szmrecsanyi, 2016a; Ehret, 2016; Koplenig et al., 2017).

In line with the proposal advocated in this book, earlier corpus-based studies have also harnessed entropy-related metrics as a reflection of differences in word frequency distributions (Bochkarev et al., 2014; Altmann et al., 2017; Bentz et al., 2017b). More specific linguistic phenomena can also be investigated within an information-theoretic framework. For example, the mutual intelligibility of related languages (Moberg et al., 2006), the functionality of gender paradigms and pronominal adjectives (Dye et al., 2017a,b), the efficiency of naming systems (Dye et al., 2016), kinship and color terminology (Regier et al., 2015), as well as the variation in color naming across languages of the world (Gibson et al., 2017). It has further been suggested that entropy-profiles can help to distinguish writing systems from symbolic systems more generally (Rao et al., 2009; Rao, 2010; Rao et al., 2010). However, the usefulness of this approach has been called into question and is debatable (Sproat, 2014).

In the area of quantitative linguistics, information-theoretic explanations further our understanding of linguistic laws, such as Zipf's law of abbreviation (Piantadosi et al., 2011; Mahowald et al., 2013; Ferrer-i-Cancho et al., 2015; Bentz and Ferrer-i-Cancho, 2016; Kanwal et al., 2017) and Zipf's law for word frequencies (Ferrer-i-Cancho and Solé, 2003; Ferrer-i-Cancho, 2005). In fact, the principle of compression – a centerpiece of modern information theory – emerges as the principle underlying quantitative linguistic laws in general (Ferrer-i-Cancho, 2017b), not only at the level of characters and words, but extending to dependency lengths and word orders (Ferrer-i-Cancho, 2017c).

All these theoretical considerations and practical applications illustrate the relevance of information theory to natural language. This is certainly not an exhaustive list of relevant studies. Geertzen et al. (2016) give further examples and an historical overview. Nonetheless, it is fair to say that information theory as a framework has not become part of mainstream linguistics. In fact, most of the authors of the above cited studies are not linguists by training, but computer scientists, physicists, and cognitive scientists. In all likelihood, there are three main reasons for this disconnection: First, the shift away from distributional and probabilistic models of language structure since the 1950s. Second, Noam Chomsky's criticism of Markov processes as models of natural language. Third, the often purported dissociation between "information" and "meaning". All three are briefly discussed in turn.

Reason one: formal vs. probabilistic accounts

In *Syntactic Structures*, Chomsky (1957, p. 15) famously contrasts two sentences: "colorless green ideas sleep furiously", and "furiously sleep ideas green colorless". These examples are construed to illustrate that grammaticality is neither a function of *meaning* nor *probability of occurrence*. With regards to the latter, Chomsky remarks that both sentences might very well have the same frequency of occurrence in the "corpus" of linguistic experiences of an English speaker – namely zero. If grammaticality was a direct function of frequency, we would expect the same grammaticality judgements for both sentences, but English speakers are more likely to consider the first sentence grammatical and the second ungrammatical. In the course of the second half of 20th century linguistics, this and similar examples were raised to disconnect syntactic analyses from corpus linguistic considerations. In an attempt to overcome this gulf, Pereira (2000, p. 1242) points out that formal syntactic and empirical accounts are fully compatible. The problem of assigning probabilities to unseen linguistic "events" can be tackled by estimation methods that predate Chomsky's colourless green ideas. In fact, using a bigram model trained on newspaper text Pereira (2000, p. 1245) estimates that the probability of ever encountering the sentence "colorless green ideas sleep furiously" is 2×10^5 times higher than for "furiously sleep ideas green colourless".

However, the relationship between formal syntax and usage-based accounts is still ambiguous. Though there is no denying that "statistical reasoning" could be a "potentially significant area of research" (Chomsky, 2011, p. 270), pure modelling-based approaches that train and test on empirical data are still seen as "dramatic failures" and "obviously absurd" when it comes to scientific explanation of human language (Chomsky, 2011, p. 266).

Reason two: non-Markovian languages

A second important criticism is levelled more directly at information theory. In a technical paper on *Three models for the description of language* Chomsky (1956) cites Shannon and Weaver (1949) as entertaining "language as a particularly simple type of information source, namely, a finite-state Markov process." In the extended discussion of *Syntactic Structures*, Chomsky further explains:

> To complete this elementary communication theoretic model for language, we assign a probability to each transition from state to state. We can then calculate the "uncertainty" associated with each state and we can define the "information content" of the language as the average uncertainty, weighted by the probability of being in the associated states. (Chomsky, 1957, p. 20)

In the following paragraph, he concedes that the information-theoretic conception of language is an "extremely powerful and general one". However, the famous bottom line of this discussion is that the English language – and by extension any other natural language – cannot be modelled by *any* finite-state Markov process. A simple example of a non-Markovian language is a discrete set of symbols $\{a, b\}$ which are recombined according to a "grammar" that allows only sentences of the form $a^n b^n$, where n is potentially infinite. This language generates "sentences" of the form *ab*, *aabb*, *aaabbb*, etc. Chomsky (1956) argues that there is no finite-state Markov process which would produce this – and only this – set of sentences. Namely, there is a set of dependencies of size n, which the generating process needs to keep track of. As n goes to infinity, a finite-state Markov process reaches the limit of its capacity to capture the dependencies. For Chomsky, it is the single most important feature of natural language that such an infinite potential for short and long-range dependencies between elements exists, though this is not provable based on empirical data, since sentences are always finite. Having said this, the limitations of finite-state Markov processes directly bear on the problem of modelling natural languages. These are important considerations with repercussions on automata theory.

The way the discussion is framed in Chomsky (1957), however, also suggests to the reader that *any* information-theoretic account of natural language is flawed to start with. This inference is a fallacy. The relevance of information theory for natural language does not stand and fall with the applicability of finite-state Markov processes. Estimating the probabilities of linguistic events (e.g. characters or words) is possible via a wide range of methods, which might or might not make strong assumptions about the underlying generative process. It is true that Shannon and Weaver (1949, p. 45) construe Markov processes as a discrete source underlying strings of symbols. This is an unequivocal example to illustrate how the entropy of a source is calculated given a known set of probabilities. However, in subsequent years, other methods have climbed up the Chomsky hierarchy harnessing, for instance, phrase structure grammars to estimate word probabilities (see Hale, 2016 for a discussion). Alternatively, experimental data elicited from human subjects (predicting the next character or word given a preceding text) can be used to the same effect. This method constitutes a theory-neutral estimation purely based on empirical observation. It is promoted by Shannon (1951, p. 50) himself as being "more sensitive" and taking account of "long range statistics, influences extending over phrases, sentences, etc." Clearly, Shannon expressed his awareness of the limitations of finite-state Markov models. These limitations, however, do not justify to discount information theory as a useful tool to investigate properties of natural languages.

Reason three: information and meaning

The third reason why information theory has not been accepted as a general framework for natural language is summarized in a recent overview on *The evolution of language* by Tecumseh Fitch (2010a). He describes Shannon's encoder-decoder model, and then states:

> [...] as many critics have since noted, and as Shannon was well aware, this model is not appropriate as a model of human language because "information" in Shannon's technical sense is not equivalent to "meaning" in any sense. (Fitch, 2010a, p. 132)

Presumably, this refers to a statement right at the beginning of *A mathematical theory of communication*, where Shannon lays out the main aim of his study:

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from *a set of possible* messages. (Shannon and Weaver, 1949, p. 31)

Shannon's statement about the irrelevance of semantics to the engineering problem might suggest that the inverse statement also holds, namely, that the engineering problem is irrelevant to semantics. However, Weaver's more general introduction to Shannon's technical article further clarifies:

> The word *information*, in this theory, is used in a special sense that must not be confused with its ordinary usage. In particular, *information* must not be confused with meaning. In fact, two messages, one of which is heavily loaded with meaning and the other which is pure nonsense, can be exactly equivalent, from the present viewpoint, as regards information. It is this, undoubtedly, that Shannon means when he says that "the semantic aspects of communication are irrelevant to the engineering aspects." But this does not mean that the engineering aspects are necessarily irrelevant to the semantic aspects. (Shannon and Weaver, 1949, p. 8)

This passage, in turn, refers back to a distinction laid out earlier in Weaver's introduction where he formulates three levels relevant to the "communication problem", and the questions associated with these (Shannon and Weaver, 1949, p. 4):

- Level A. How accurately can the symbols of communication be transmitted? (The technical problem.)
- Level B. How precisely do the transmitted symbols convey the desired meaning? (The semantic problem.)
- Level C. How effectively does the received meaning affect conduct in the desired way? (The effectiveness problem.)

Level A is concerned with the purely technical problem of how to transmit information through a serial, noisy channel. In this sense, it is comparable to the linguistic field of phonetics. Level B is the extension of the technical problem to the study of the association between codes and the concepts they refer to, i.e. semantics. Level C, or "the effectiveness" problem, is nowadays referred to as pragmatics, i.e. the question of "how to do things with words", anticipating Austin and Searle's speech act theory.

Shannon certainly focused on Level A, the most fundamental level, which the other levels build upon. Since an overarching theory to capture the information encoding potential of symbol repertoires was lacking at the time, Shannon gave paramount importance to this endeavour. However, the findings at Level A are not irrelevant to Levels B and C. Rather, a theory of information has important repercussions on semantics and pragmatics. Levels B and C can only harness "signal accuracies" that are in accord with the technical limitations of a code at level A. Thus, Weaver concludes that "a theory of Level A is, at least to a significant degree, also a theory of Levels B and C" (Shannon and Weaver, 1949, p. 6). In other words, the information encoding potential of natural languages underlies their meaning and usage.

Almost sixty years later, Ferrer-i-Cancho and Díaz-Guilera (2007) apply standard information theory to support Weaver's intuition. They consider a set of signals $\mathcal{S} = \{s_1, \dots, s_i, \dots, s_n\}$ and a set of stimuli $\mathcal{R} = \{r_1, \dots, r_j, \dots, r_m\}$, i.e. the "physical or conceptual entities" that Shannon refers to in the quoted passage above. Signals and stimuli have to be mapped onto each other in a communicative setting. This is often the setup in artificial language learning studies, where a set of signals has to be mapped onto meanings in a meaning space $\mathcal{M}$, e.g. Smith et al. (2003, p. 545). Based on three simple information-theoretic preconditions, Ferrer-i-Cancho and Díaz-Guilera (2007, p. 13) show that the entropy of the signal set $H(S)$ is an upper bound on the mutual information $I(S, R)$ (called *Shannon information transfer* in this particular paper), such that

$$I(S, R) \leqslant H(S). \tag{4.12}$$

In other words, the mutual information – or information transferred – between signals and meanings has to be smaller or equal to the entropy of the signal set. Thus, the entropy of a signal set is the upper bound on the "expressivity" of that set (see also Ferrer-i-Cancho, 2017b, Ferrer-i-Cancho, 2017a, and Bentz et al., 2017a).

To see this, assume a language L, consisting only of two sets of words, one for denoting objects $\mathcal{W}_{\text{obj}} = \{\text{car, box}\}$ and the other for colours $\mathcal{W}_{\text{col}} = \{\text{yellow, red, blue}\}$. Further assume that the grammar of L allows only the recombination of one colour term with one object in the order adjective followed by noun as a grammatical sentence. Thus, the set of potential sentences (signals) is

$$\mathcal{S} = \{\text{yellow car, yellow box, red car, red box, blue car, blue box}\}. \tag{4.13}$$

If each signal is equally probable, this yields the maximum entropy given as

$$H(L) = -\sum_{i=1}^{6} \frac{1}{6} \log_2 \frac{1}{6} = \log_2(6) \sim 2.6 \text{ bits/signal}. \tag{4.14}$$

These 2.6 bits/signal are a fundamental constraint on the "expressivity" of language L. It is the upper bound on the information regarding concepts in the "real world" we can *unequivocally* transfer using L. As pointed out by Weaver, the actual form of the signals can vary. We can construe an infinite number of signal sets equivalent to $\mathcal{S}$ in terms of their information encoding capacity. Also, the signal-to-object mapping can be arbitrary. The signal *yellow car* could refer to an actual yellow car, or a black cat or any other concept. The crucial information-theoretic property of L – reflected by its entropy – is independent of the exact linguistic realization of the signals and the signal-to-meaning mapping. In fact, the entropy is the same for any set of six signals of equal probability.

Thus, while Shannon's technical definition of "information" is not equivalent to meaning, it is also not irrelevant to theories of meaning either. Namely, non-zero entropy is a *necessary* but not *sufficient* condition for meaning. For similar points about information-theoretic interpretations of natural language see also

Deacon (2010). Of course, when humans communicate “encoding” and “decoding” of meaning can go far beyond simple associations between signal and referent, and instead require rich knowledge of the potential communicative intentions of a signaller depending on particular real-world contexts. To capture and explain even the most complex patterns of human communication, Scott-Phillips (2015) proposes that we have to go beyond a simple code model. Namely, taking into account the expression and recognition of communicative intensions in a so-called *ostensive-inferential* model. However, complex pragmatic inferences of this kind go beyond the scope of this book.

To sum up, examining closely the writing of Shannon and Weaver, we find that none of the three reasons given for disregarding information theory as a useful tool to study natural language really holds. Firstly, Pereira (2000) shows that the divide between formal language theory and probabilistic language modelling is artificial and can be overcome. Secondly, early results on the generative capacity of finite-state Markov processes, while important for automata theory, are not to be interpreted as disqualifying information-theoretic models of natural language. Thirdly, it is true that “information” in Shannon’s sense is not equivalent to “meaning”. Instead, information is a more basic property of a set of signals and a fundamental prerequisite for the mapping of forms to meanings. Without understanding information, we cannot understand meaning.

It is an irony of 20th century linguistics that Shannon’s theory of information, though explicitly linked to semantics, was deemed irrelevant by linguists, while Chomsky’s formal syntax, though explicitly dissociated from semantics, was adopted as the default theory of natural language.

4.3.2 Estimation methods

Apart from general considerations about the usefulness of information theory for linguistic inquiry, there are further, more technical issues. Namely, it cannot be taken for granted that token frequency counts in any given text or corpus directly correspond to probabilities of words. This is pointed out, for instance, by Pereira (2000). Particularly, there are two conceptual problems:

Problem 1: Representativeness
Token frequency counts can differ for reasons of text size, content, style, register, etc. This begs the question of *representativeness*. Can normalized token counts be taken as reliable estimations for the probability of word types in the “whole” language?

Problem 2: Non-independence

In natural languages, words are not *independent and identically distributed* (i.i.d) events. Due to co-occurrence patterns as well as short and long-range correlations the probabilites of words dependent on their *co-text* and *context*. *Co-text* is here defined as the word tokens preceding a particular word token of interest.[4] *Context*, on the other hand, is more generally any additional information (gesture, prosody, general world knowledge) that might reduce uncertainty in a string of words. Both co-text and context play an important role in natural languages.

The first problem is addressed by advanced methods which alleviate the estimation bias by taking potentially unseen events into account. A range of such estimators is discussed in Appendix 12 and tested in the following sections. Moreover, variation due to different text types is assessed in Chapter 5. The second problem is mathematically involved. Several studies have ventured to estimate the entropy of words taking conditioning by the preceding co-text into account (Montemurro and Zanette, 2011, 2016; Koplenig et al., 2017; Bentz et al., 2017a). However, the exact conditions for reliable entropy estimation, and its applicability to running text, are an active field of research. For further discussion see, for instance, Bentz et al. (2017a), Dębowski (2016), and Dębowski (2017).

However, in the current study, the focus is on lexical diversity, and hence the entropy of words independent of their co-text and context. Henceforth, this is called *unigram word entropy* – or *unigram entropy* for short.

Advanced entropy estimators

It is difficult – if not impossible – to estimate the "true" or "actual" unigram word entropy of a whole language. Remember that entropy is defined with reference to the probabilities of events $p(x)$. If we assume words as events, we need to estimate probabilities for each i^{th} word type, i.e. $p(w_i)$. A crucial caveat is that even *if* we could capture the complete set of interactions between speakers of a population at time t, i.e. $\mathscr{L}(t)$, we would still not capture the "productive potential" of the language beyond the finite set of linguistic interactions.

Theoretically speaking, it is always possible to expand the vocabulary of a language by recombining word types to compounds, by adding productive morphology to roots, or by creating neologisms. This is related to Humboldt's principle to "make infinite use of finite means" that is prominent in syntax (Chomsky, 1965,

[4] It could also refer to word tokens following a given word token. When restricted to preceding tokens, the co-text is sometimes called "prefix".

p. 8). The consequence is that even for massive corpora like the British National Corpus (BNC) the vocabulary of word types keeps increasing with the number of tokens (Baroni, 2009; Baayen, 2001). This suggests that using corpora – however extensive – we never sample the full set of *potential* word types of a language, and are hence prone to assign zero probability to word types that might actually have non-zero probabilities.

However, whether the exact probability of an event is ever *known* in an absolute sense is a matter of philosophical debate. For practical purposes it can either be defined a priori or estimated based on empirical data. Hence, in practice, the question is not so much: *what is the exact unigram word entropy of a language?*, but rather: *how precisely do we have to approximate it to make a meaningful cross-linguistic comparison possible*?

To estimate the entropy as given in Equation 4.7 with highest possible precision, the critical part is to get good approximations of the probabilities of word types $p(w_i)$. Estimated probabilities are henceforth denoted as $\hat{p}(w_i)$ and estimated entropies correspondingly $\hat{H}$. The so-called plug-in or maximum likelihood (ML) method simply uses frequency counts in a text sample, i.e.

$$\hat{p}(w_i)^{ML} = \frac{f_i}{\sum_{i=1}^{V} f_i}. \tag{4.15}$$

Note that the denominator here is equivalent to what we defined earlier as N (Equation 4.4), which is the total number of tokens of a text sample with V word types. For a given text T, plugging Equation 4.15 into Equation 4.7 yields

$$\hat{H}(T)^{ML} = -\sum_{i=1}^{V} \hat{p}(w_i)^{ML} \log_2(\hat{p}(w_i)^{ML}), \tag{4.16}$$

This is the unigram word entropy based on the ML method for estimating word type frequencies. This method yields reliable results for situations where $N >> V$, i.e. the number of tokens is much bigger than the number of types (Hausser and Strimmer, 2009, p. 1470). In other words, it is reliable for a small ratio of word types to word tokens V/N. Since in natural language this ratio is typically large for small texts and only decreases with N (Baayen, 2001; Baroni, 2009), unigram entropy estimation tends to be unreliable for small texts.

Since Shannon's original work in the 1950s, a range of entropy estimators has been proposed to overcome such biases. Nine of these are used in the following. They are discussed in more detail in Appendix 12, and are accessible via the R package *entropy* (Hausser and Strimmer, 2014) and an implementation in Python.[5] This sample includes the "naive" ML estimator defined above alongside

[5] https://gist.github.com/shhong/1021654/

Table 4.2: Entropy values attained for uniform and non-uniform distributions of Figure 4.1 by using nine different estimators.

Estimator	**Abbreviation**	$\hat{\mathbf{H}}_{\textbf{non-uni}}$	$\hat{\mathbf{H}}_{\textbf{uni}}$
Maximum likelihood	ML	2.27	3.32
Miller-Meadow	MM	2.34	3.39
Jeffreys	Jeff	2.38	3.32
Laplace	Lap	2.47	3.32
Schürmann-Grassberger	SG	2.3	3.32
Minimax	minmax	2.47	3.32
Chao-Shen	CS	2.43	3.32
Nemenman-Shafee-Bialek	NSB	2.3	3.31
James-Stein shrinkage	shrink	2.37	3.32
		mean=2.37	**mean=3.33**
		SD=0.08	**SD=0.02**

some of the newest – and demonstrably less biased – estimators such as the Nemenman, Shafee and Bialek (NSB) estimator (Nemenman et al., 2002). Hausser and Strimmer (2009) and Nemenman et al. (2002) argue that the statistically most advanced method is the NSB estimator. However, based on analyses with generated data, Hausser and Strimmer (2009) illustrate that the so-called Chao-Shen (CS) estimator and the James-Stein shrinkage (shrink) estimator yield equally reliable results. Moreover, the James-Stein shrinkage estimator is computationally more efficient (by a factor of one thousand) than the NSB estimator. The mathematical details of estimators are relegated to Appendix 12, but it is shown in the following that the choice of estimator turns out to be a minor issue.

To get an overview, the estimators are summarized in Table 4.2 with estimated entropy values for the uniform and non-uniform distributions of earlier examples. The mean of estimated values for the non-uniform distribution is $\mu = 2.37$, with a standard deviation of $\sigma = 0.08$. The mean for the uniform distribution is $\mu = 3.33$, with a standard deviation of $\sigma = 0.02$. Given these different entropy estimators and their results, there are three questions relevant from a practical point of view:

1. Which is the best estimator even for small texts?
2. How “small” is small? That is, what is the minimum text size we need for stable entropy estimation?
3. To what extent can estimations by different methods be used interchangeably?

The first two questions are addressed in the following section on *entropy and text size*, while the third question is addressed in the section on *correlations between extimators*.

Entropy and text size

The first and second question above refer to the relationship between the size of a text (i.e. number of tokens N) and the precision of the estimated entropy values. To investigate the behavior of estimated entropies with growing text size, the *European Parallel Corpus* (EPC) is used. To reduce processing cost, only the first 100,000 tokens per language are considered, text sizes are increased in steps of 1000 tokens (rather than one token at a time).[6] Figure 4.2 illustrates the results of this analysis for English, and Figure 4.3 for all 21 languages.

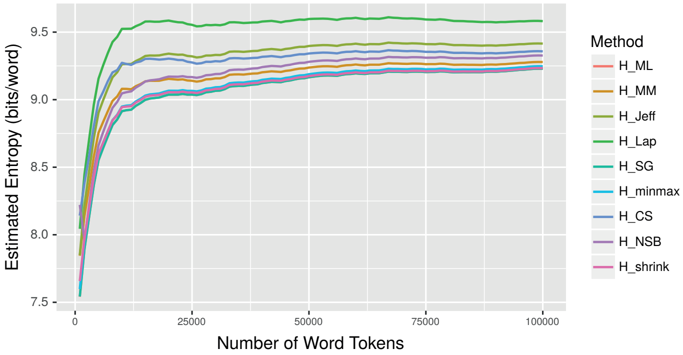

Figure 4.2: Estimated entropy as a function of text size for English. Estimated entropies (y-axis) with growing text size N (x-axis) in the English version of the EPC.

For English, the estimated entropies of all estimators "stabilize" at around 25000 tokens. They still appear to rise gently through to 100,000 tokens (see discussion below). The values ordered from highest to lowest at 100,000 tokens are given in Table 4.3.

Note that Bayesian estimations with Laplace and Jeffreys priors (H_Jeff and H_Lap) somewhat overestimate the entropy compared to other methods. They employ uniform priors for unseen word frequencies, and the estimation starts

[6] Rcode/Chapter4/entropyTextSize.R

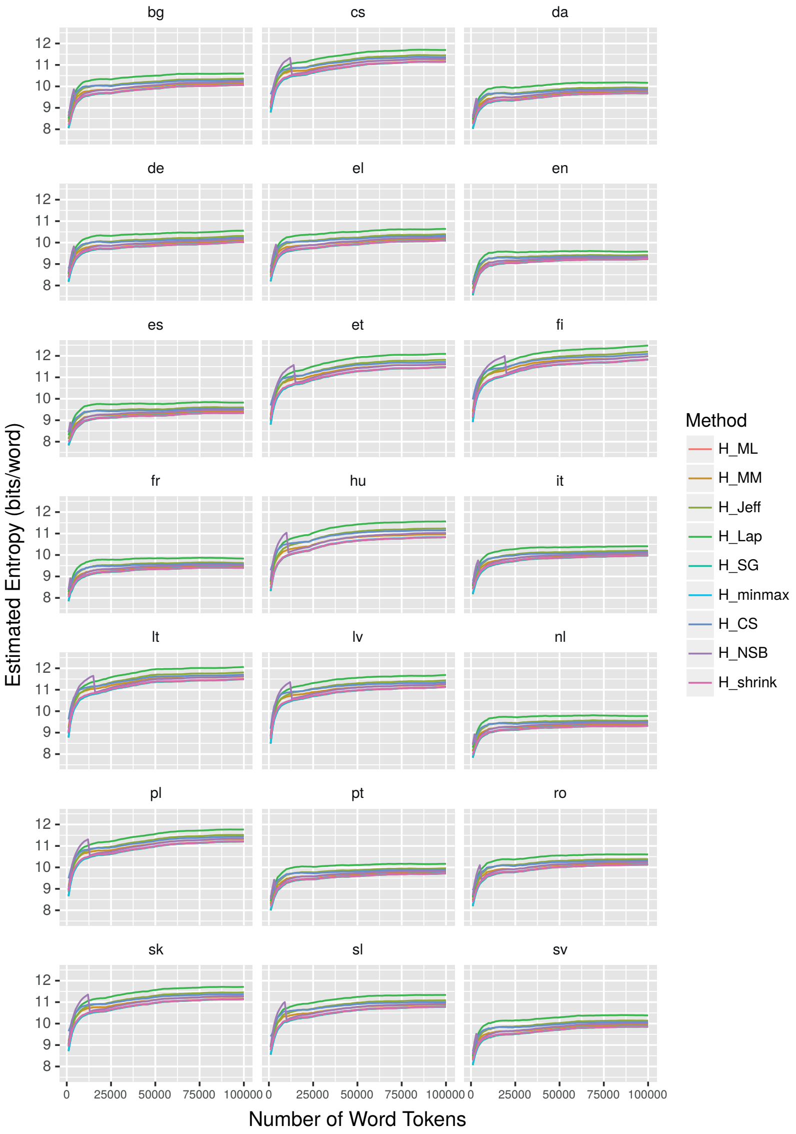

Figure 4.3: Entropy estimations and text sizes for 21 languages of the EPC. The language codes are: Bulgarian (bg), Czech (cs), Danish (da), German (de), Greek (el), English (en), Spanish (es), Estonian (et), Finnish (fi), French (fr), Hungarian (hu), Italian (it), Lithuanian (lt), Latvian (lv), Dutch (nl), Polish (pl), Portuguese (pt), Romanian (ro), Slovak (sk), Slovene (sl), Swedish (sv). Colours indicate different entropy estimators.

Table 4.3: Entropy values given by nine different estimation methods for the English EPC at 100,000 tokens.

Estimator	Abbreviation	$\hat{H}$
Laplace	H_Lap	9.58
Jeffreys	H_Jeff	9.42
Chao-Shen	H_CS	9.36
Nemenman-Shafee-Bialek	H_NSB	9.32
Miller-Meadow	H_MM	9.28
Minimax	H_minmax	9.25
Maximum likelihood	H_ML	9.23
Schürmann-Grassberger	H_SG	9.23
James-Stein shrinkage	H_shrink	9.23
		mean=9.32
		SD=0.12

with an overly high entropy estimate. This has also been pointed out by Nemenman et al. (2002, p. 4). Likewise, Hausser and Strimmer (2009, p.1475) generate language-like data based on a Zipf-distribution in their “scenario 4” and show the same overestimation bias for Jeffreys and Laplace priors. The Chao-Shen, NSB, and Miller-Meadow estimators are to be found in the middle range. The ML, Schürmann-Grassberger, James-Stein shrinkage, and Minimax estimators yield the lowest estimations, and are almost indistinguishable in Figure 4.2. By and large, these results hold across all 21 languages (Figure 4.3).

Looking closely at the plot for English and the cross-linguistic plot, we notice that even at the point where entropy measures seem stable, they still display a slightly positive increase with text size. This is likely related to the observation pointed out earlier, namely, that vocabulary growth in languages is never coming to a halt. Importantly, this assessment of entropy estimators and their performance with growing text size is based on a particular concept of “stabilization” as defined in Bentz et al. (2017a). It refers to the process whereby the standard deviation of entropy estimations (for a given set of consecutive numbers of tokens) falls below a certain threshold. In this regard, the behaviour of different unigram estimation methods is very similar (see Figure A2 in Bentz et al., 2017a). This is different from studies such as Hausser and Strimmer (2009), where data is generated based on pre-specified models, meaning that the actual entropy is known, and the error in estimation can be calculated precisely. As outlined above, this is not possible for natural language data. We do not know the actual entropy. However, the fact that for English all entropy estimators stabilize on roughly the same value around 9.32 (with a standard deviation of 0.12), is encouraging.

With regards to the first question above, we can say that, in terms of stabilization, all estimators are virtually equivalent. The answer to the second question is that stabilization sets in around 25000 tokens across the 21 languages of the EPC. In Bentz et al. (2017a) stabilization properties were also tested across a sample of 32 typologically diverse languages – with the same result.

Correlations between estimators

The third question can be interpreted as referring to the correlation between different entropy estimators: if we choose one entropy estimator over the others, then how much can this choice influence our results? To assess this, entropies are estimated with all nine estimators for all three parallel corpora.[7]

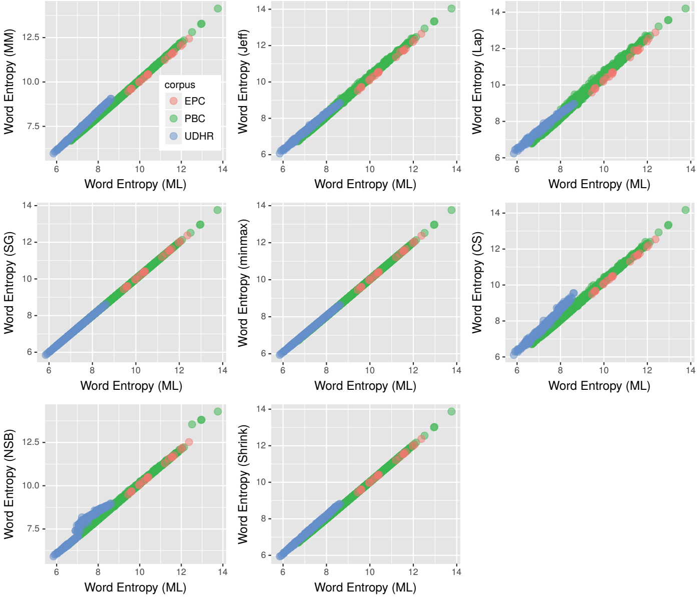

Figure 4.4: Correlations between values of different entropy estimators. The x-axis gives values of ML estimated entropies for the full texts of the EPC, PBC, and UDHR. The y-axes give values for the other entropy estimators. Individual points represent texts.

[7] Rcode/Chapter4/entropyEstimation.R

Table 4.4: Information about parallel corpora used.

Corpus	Register	Overall Size*	Text Size*	Texts	Languages
UDHR	Legal	314,000	1000	314	288
PBC	Religious	75 mio.	50,000	1498	1114
EPC	Written speeches/Discussions	circa 21 mio.	circa 1 mio.	21	21
	Total	**circa 96 mio.**	–	**1833**	**1217**

*in number of tokens

The entropy values per estimator, text, and parallel corpus are then plotted against the entropy values of the ML estimator, used as a baseline. The result can be seen in Figure 4.4.

Overall, the Pearson correlations between values given by different estimators are remarkably strong. In fact, for both the EPC and the PBC the correlations we find for all 36 possible combinations of entropy estimators range between 0.99 and 1 (see Bentz et al., 2017a Table A1 and A2). While the exact values given by each estimator might differ, the relative ordering of values is virtually the same. For the UDHR texts, the correlations range between 0.95 and 1. This slightly wider range is most likely due to small text size. Remember from the analyses in the previous section that a few thousand tokens is not enough for entropy estimations to stabilize and this leaves more room for variation in the values. However, even for the UDHR texts the values of the most "naive" entropy estimator (ML) compared to the values of the NSB estimator give us a Pearson correlation of 0.96. In conclusion, for the purpose of the current study, namely comparing the entropies across different texts and languages, the choice of estimator is a minor issue – though it is advisable to consistently apply the same estimator.

Another notable pattern in Figure 4.4 is that the red dots of the EPC are shifted further to the upper right corner than for texts of the PBC, and these are in turn shifted further up than the dots of the UDHR. EPC texts have, on average, higher unigram word entropies than texts of the PBC and the UDHR. Indeed, we expect this from the entropy growth curves in Figure 4.3. Since entropies keep growing with text size, we expect bigger texts to have higher entropies, everything else being equal.

4.4 Word entropies for 1217 languages

Given the results on text size dependence of unigram entropies in Section 4.3.2 and the well-known dependence of estimation bias on text/sample size, it is advisable to use texts of the same size to ensure comparability. Since the average text sizes vastly differ between the three corpora, three different default sizes are used here, namely, 1000 tokens for the UDHR, 50,000 tokens for the PBC and circa 1 million tokens for the EPC. These criteria exclude some texts for being too small. The overall corpus sample then contains 1833 texts representing 1217 different languages (identified by ISO 639-3 codes). This working sample is summarized in Table 4.4. Additionally, Figure 4.5 gives world maps with locations of the respective languages by corpus.[8]

Sample balance

According to Glottolog 2.7 (Hammarström et al., 2016),[9] the languages represented in the current parallel text sample belong to 106 top-level families.[10] According to the AUTOTYP classification (Nichols et al., 2013), they belong to 145 stocks, i.e. macro-families for which there is some linguistic evidence.[11]

Figure 4.6 illustrates the sample balance with regards to the Glottolog 2.7 classification. Dark grey bars reflect percentages of languages classified as belonging to a given family in the overall Glottolog sample of close to 8000 languages. In comparison, light grey bars reflect percentages of languages belonging to a given family in the parallel text sample. Only top-level families with more than five languages are illustrated in this plot, that is, 30 out of the original 106. The way to interpret this plot is as follows. If the percentage of languages in the corpus sample belonging to a given family (light grey bar) is higher than the percentage of languages in the overall Glottolog sample belonging to the same family (dark grey bar), then this particular language family is over-represented in the corpus sample. For example, around 15% of languages in the corpus sample belong to the Indo-European family. However, only ca. 7.5% of the languages in Glottolog belong to this family. Hence, Indo-European languages are clearly over-represented in the corpus sample. We see that most top-level families are over-represented in the corpus sample, while Sino-Tibetan languages are under-represented. For the two biggest families, Atlantic-Congo and Austronesian, the representation is approximately balanced.

The general trend towards over-representation of big families is due to the existence of many small families, and particularly isolates, which are not repre-

[8] Rcode/Chapter4/entropyWorldMapCorpora.R

[9] Accessed on 10/06/2016

[10] Note that “NA”, i.e. not classifiable, is counted here as well.

[11] Rcode/Chapter4/entropySimpleStats.R

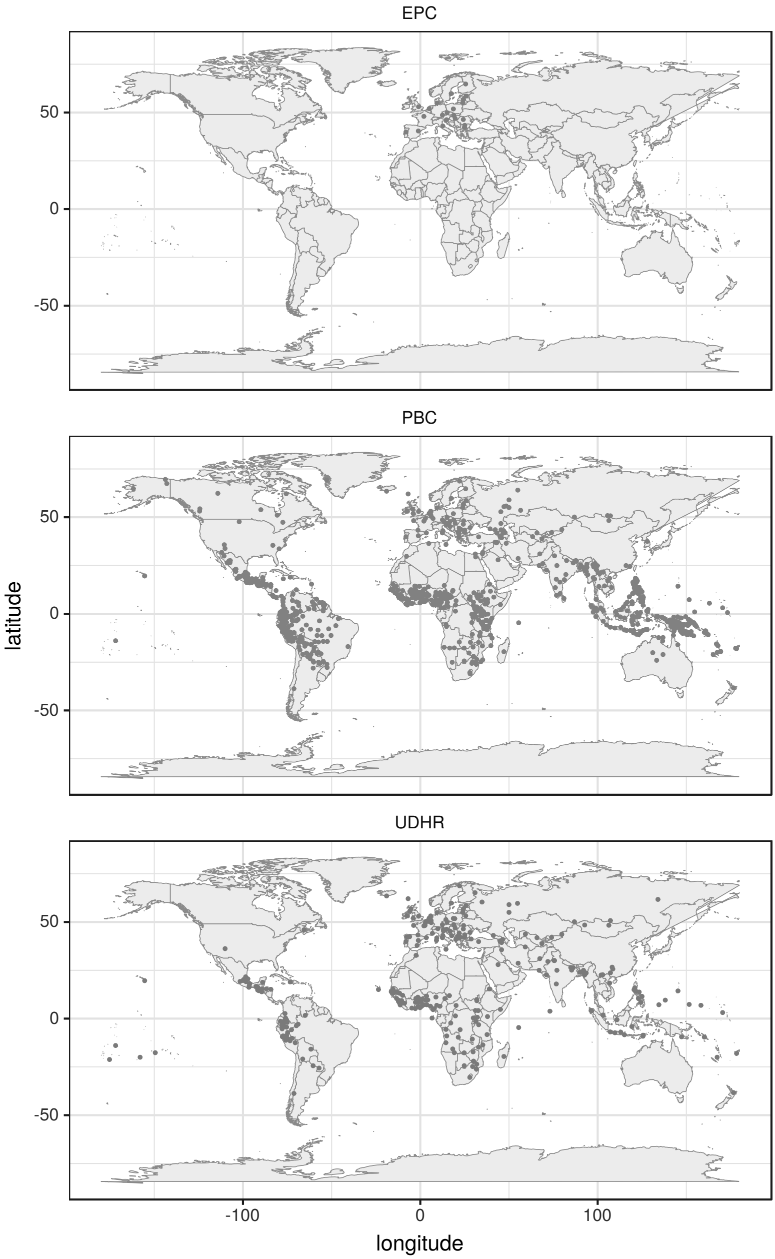

Figure 4.5: World maps with overall 1217 languages represented in the corpus samples. Latitude and longitude information is taken from Glottolog 2.7 (Hammarström et al., 2016).

sented in the parallel text sample at all. This corroborates Bickel (2013)'s point that samples of languages used for typological investigations are likely to be biased towards large families, and we might miss a considerable amount of the variance represented by small families and isolates. He proposes methods related to the "Family Bias Theory" to overcome this issue in future studies.

The sample balance for geographic macro-areas is illustrated in Figure 4.7. The representation of Africa and Papunesia are approximately balanced, Eurasia and Australia are under-represented, and South America and North America are over-represented. The over-representation of the Americas is likely related to more extensive missionary work in this area of the world that resulted in more Bible translations.

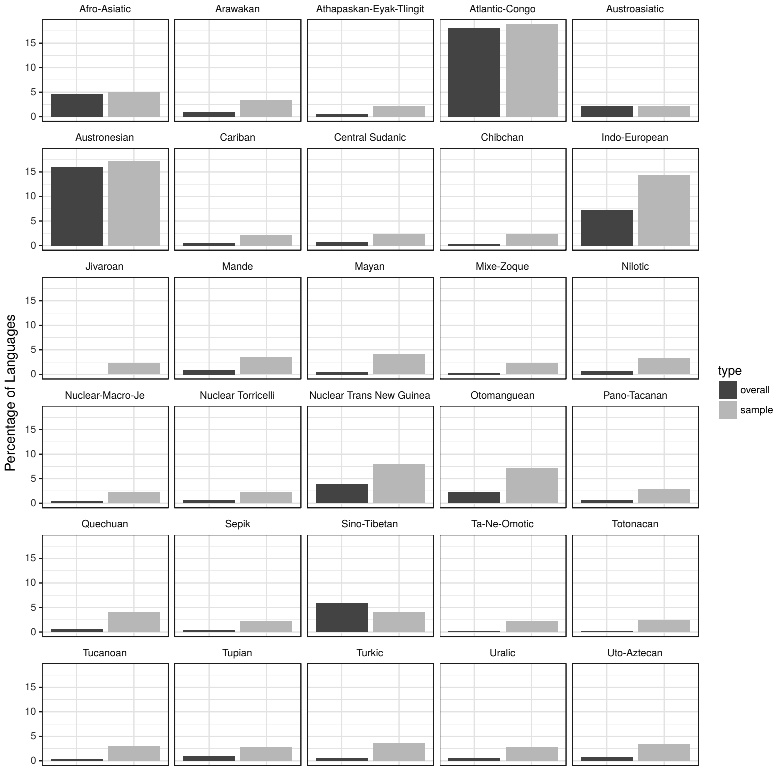

Figure 4.6: Percentages of languages represented per family. "Overall" (dark grey) refers to the percentages of languages belonging to a given family in the overall Glottolog sample of close to 8000 languages. "Sample" (light grey) refers to the percentages of languages belonging to a family in the sample of parallel texts.

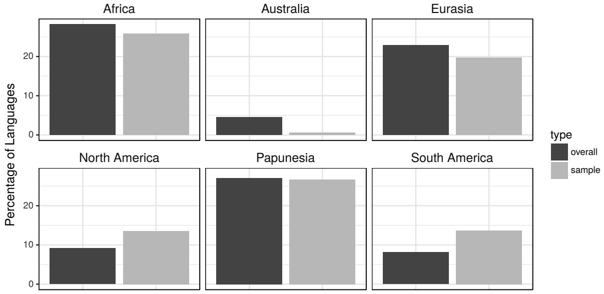

Figure 4.7: Percentages of languages represented per geographic macro-area. "Overall" (dark grey) refers to the percentages of languages assigned to a given area in the overall Glottolog sample of close to 8000 languages. "Sample" (light grey) refers to the percentages of languages assigned to a given area in the sample of parallel texts.

Results across 1217 languages

For all texts of the corpus sample, unigram entropies are calculated using the nine different estimators discussed in Section 4.3.2.[12] In the PBC and UDHR, some languages are represented by several texts. For example, there are 31 different Bible translation for English (eng). To get a single entropy value per language, the mean value across all texts with a common ISO-639-3 code is taken.[13] Furthermore, instead of running analyses for the outcomes of all nine entropy estimators, the James-Stein shrinkage entropy is chosen as a representative value. Henceforth, when reference is made to estimated unigram entropy $\hat{H}$, then the James-Stein shrinkage entropy $\hat{H}^{\text{shrink}}$ is meant.

To facilitate comparison despite different text sizes in the three corpora, James-Stein shrinkage entropies are centred and scaled by corpus c to yield z-scores.[14] The resulting entropy is called $\hat{H}^{scaled}$ (*H_scaled* in plots) and is derived as

$$\hat{H}^{\text{scaled}} = \frac{\hat{H}^{\text{shrink}} - \mu_c}{\sigma_c}, \tag{4.17}$$

where μ_c and σ_c are the mean and the standard deviations of $\hat{H}$ values per corpus. Given these transformations, we can now have a first look at unigram word

[12] Rcode/Chapter4/entropyEstimation.R

[13] Rcode/Chapter4/entropyAggregation.R

[14] Rcode/Chapter4/entropyScaling.R

entropies found across languages of the world. Figure 4.8 a) is a density plot of scaled unigram entropy values across all 1217 languages of the three parallel corpora.[15] It appears that lexical diversities of languages around the world follow a unimodal distribution with a slight right skew, i.e. towards higher values. Potential information-theoretic causes for this right skew are further discussed in Bentz et al. (2017a). The full range of $\hat{H}^{\text{scaled}}$ values goes from –2.35 (low lexical diversity) to 3.65 (high lexical diversity), which corresponds to a range from around 6 bits/word to around 13 bits/word in terms of unscaled unigram entropy.

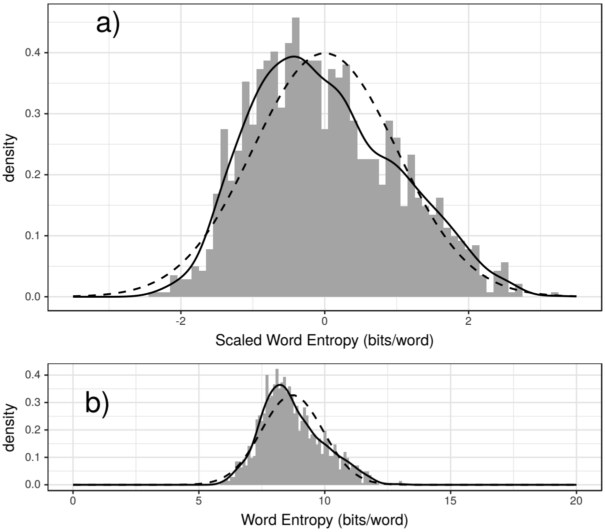

Figure 4.8: Entropy distribution across 1217 languages. Density plot of scaled (a) and unscaled (b) unigram entropy values. In a) the density histogram is overlaid with a density line and the theoretical normal distribution (dashed). In b) the full range of hypothetical entropy values (unscaled) is illustrated. Note that density values do not reflect probabilities (which would have to sum up to 1), since the width of bars is smaller than one.

Figure 4.8 b) compares the range of unscaled $\hat{H}$ values to the overall range of potential entropy values. The minimum entropy is 0, i.e. the entropy of a hypothetical language that would use the same exact word type all the time. The theoretical maximum entropy is harder to estimate. The maximum entropy is represented here by a hypothetical language for which all word types have the same probability. It is unclear what the overall number of word types in such a language might

[15] Rcode/Chapter4/entropyDensity.R

be. As an approximation, the maximum number of word tokens is taken, i.e. 1 million. This yields a maximum entropy of: $\log_2(1000000) = 19.93$. Natural languages (6 to 13 bits/word) only span around 37% of the possible range (0 to 20 bits/word).

Some of the extremely high lexical diversity outliers (in scaled values) are Eskimo-Aleut languages such as Kalaallisut (kal), Eastern Canadian Inuktitut (ike), Northwest Alaskan Iñupiatun (esk), as well as the Abkhaz-Adyghe language Adyghe (ady), and ancient Hebrew (hbo). Some of the low lexical diversity outliers include the Otomanguean language Cuixtla-Xitla Zapotec (zam), the Austronesian language Tahitian (tah), and the Creole language Sango (sag).[16] To visually illustrate the difference between high and low unigram entropy languages, the first paragraph of the UDHR for Kalaallisut and Tahitian, alongside the English version, are given below.

There are some obvious visual difference when we eyeball these paragraphs. To start with, the Kalaallisut translation uses 16 word tokens, English 30, and Tahitian 40 to encode the essentially same content. The reasons for why languages differ in terms of unigram entropies are discussed in the following chapters.

(9) Kalaallisut (kal, UDHR 01)
Inuit tamarmik inunngorput nammineersinnaassuseqarlutik assigiimmillu ataqqinassuseqarlutillu pisinnaatitaaffeqarlutik . Solaqassusermik tarnillu nalunngissusianik pilersugaapput , imminnullu iliorfigeqatigiittariaqaraluarput qatanngutigiittut peqatigiinnerup anersaavani .

(10) English (eng, UDHR 01)
All human beings are born free and equal in dignity and rights . They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood .

(11) Tahitian (tah, UDHR 01)
E fanauhia te tā'āto'ara'a o te ta'ata-tupu ma te ti'amā e te ti'amanara'a 'aifaito . Ua 'ī te mana'o pa'ari e i te manava e ma te 'a'au taea'e 'oia ta ratou ha'a i rotopū ia ratou iho , e ti'a ai ;

4.5 Summary

This chapter introduced the concept of lexical diversity as the distribution of *word tokens* over *word types* in a text. It can be estimated across different texts, corpora,

[16] Assigned to the Atlantic-Congo top-level family by Glottolog 2.7.

and ultimately languages by sampling from the set of interactions of the population of speakers. In order to hold the content of interactions constant, parallel texts are used. Several measures of lexical diversity were presented, and their advantages and disadvantages discussed. In the context of the current study, the unigram entropy emerges as a promising measure.

A more general discussion addressed some classical objections against using information-theoretic models to investigate natural languages. These include arguments that probability is unrelated to grammaticality, that languages are not finite-state Markov processes, and that "information" is not to be confused with "meaning". While these caveats need to be kept in mind, they do not strictly dismiss information theory as a useful tool to investigate diversity in natural languages. Two rather practical problems surrounding unigram entropy estimation were also pointed out: the infinite potential to create new word types, and the non-independence of words in co-text and context. Different methods of entropy estimation were tested to alleviate in particular the first problem.

What we can take away from analyses of text sizes and estimated entropy values is that the number of tokens N does indeed play a role, especially if texts have less than 25000 tokens. Beyond this text size, increasing N has only a minor impact. Moreover, analyses of correlations between entropy values given by different estimation methods suggest that there is only minimal bias introduced by the idiosyncrasies of estimators.

Finally, the overall parallel text sample representing 1217 languages was discussed, and its typological balance assessed using information from Glottolog 2.7 and the AUTOTYP database. Unigram word entropies were estimated for all texts and languages of this sample. They display a right-skewed unimodal density distribution, which covers only around 37% of the potential unigram entropy spectrum. While the spectrum realized by natural languages across the world is relatively narrow, there are still remarkable differences between languages. Some of the extreme high and low entropy outliers were pointed out. Such systematic differences in lexical diversity are in need of explanation. The following chapters aim to further elucidate this diversity.

5 Descriptive Factors: Language "Internal" Effects

In the previous chapter, the entropy of a distribution of word tokens over word types was established as a measure of lexical diversity. Higher numbers of word types and lower token frequencies in a parallel text (i.e. at near constant content) correspond to longer tailed distributions, and hence higher entropy. In this chapter, the scaled unigram word entropy $\hat{H}^{\text{scale}}$ is used to measure differences in word frequency distributions across all the texts and languages covered in the UDHR, EPC, and PBC. To explain differences in lexical diversities, a range of descriptive, "internal" factors are considered. These include the *script* a language is written in, productive *word-formation* patterns, and the *register and style* of texts used.

A crucial question to answer is: *why* do some languages have higher lexical diversities than others? And how can we predict where on the spectrum a specific language falls? The *why*-question can be answered in two distinct ways: either by a description of the linguistic properties of the texts and languages themselves, or with reference to explanatory factors relating to the properties of the speaker population. The focus here is on the first kind of factor.

5.1 Writing systems and scripts

Languages across the world use a panoply of different writing systems. These are traditionally categorized into *logographies* (graphemes representing words), *syllabaries* (graphemes representing syllables), and *alphabets* (graphemes representing phonemes) (Daniels and Bright, 1996, p. 8). However, the actual *scripts* that languages are written in (e.g. Latin, Cyrillic, Devanagari, etc.) are very rarely purely logographic, syllabic or alphabetic, but rather a mixture. For example, in so-called *alphasyllabaries* graphemes represent either syllables or individual phonemes.

To get script information for the parallel corpora used here, the dataset of estimated unigram entropies per language is merged with the *Ethnologue* dataset (Lewis et al., 2013). This reduces the sample to 1408 texts and 1204 languages. In this sample, we find languages written in 27 different scripts overall. Table 5.1 gives an overview of the most common scripts (and combinations of scripts) of languages in the three corpora.[1] The distribution of scripts is highly skewed with 91% of the texts written in Latin and only 9% written in any of the 26 other scripts.

[1] file: Rcode/Chapter5/entropyWritingSystems.R

DOI 10.1515/9783110560107-005

Table 5.1: Information on scripts represented in the three parallel corpora. For some languages there can be text versions using different scripts, e.g. the Uighur (uig) Bible has an Arabic and a Latin version.

Script	Texts	Percentage
Latin	1280	90.91%
Cyrillic	42	2.98%
Devanagari	15	1.07%
Arabic	14	0.99%
Geez	7	0.50%
Cyrillic/Latin	7	0.50%
Thai	4	0.28%
Unified Canadian Aboriginal Syllabics	3	0.21%
Modern Greek	3	0.21%
Hebrew	3	0.21%
Bengali	3	0.21%
Syriac	2	0.14%
Gurmukhi	2	0.14%
Gujarati	2	0.14%
Georgian	2	0.14%
Geez/Latin	2	0.14%
Armenian	2	0.14%
Arabic/Latin	2	0.14%
Vai	1	0.07%
Thaana	1	0.07%
Telugu	1	0.07%
Tamil	1	0.07%
Sinhala	1	0.07%
Oriya	1	0.07%
Greek	1	0.07%
Coptic	1	0.07%
Burmese	1	0.07%
Others	4	1.58%

Some of the most frequently used scripts, including Latin, Cyrillic, Devanagari, and Arabic, are discussed in turn.

Latin (Romance) scripts

Latin scripts are based on the Roman alphabet of 26 letters (5 vowels and 21 consonants), but they can differ vastly with regards to the number of special diacritics added (e.g in Vietnamese, Slavic languages, African languages, and Mesoamerican languages). For instance, the first sentence of the UDHR, i.e. "all human beings are born free and equal in dignity and rights”, in Vietnamese, Serbian and Yoruba reads:

(12) Vietnamese (vie, UDHR 01)
Tất cả mọi người sinh ra đều được tự do và bình đẳng về nhân phẩm và quyền.

(13) Serbian (srp, UDHR 01)
Sva ljudska bića rađaju se slobodna i jednaka u dostojanstvu i pravima.

(14) Yoruba (yor, UDHR 01)
Gbogbo ènìyàn ni a bí ní òmìnira; iyì àti ẹ̀tọ́ kọ̀ọ̀kan sì dọ́gba.

In these examples, diacritics are used to expand the common Latin alphabet to accommodate for phonemic particularities, e.g. for tone, as in the case of the low tone in Vietnamese *và* ‘and’, and in Yoruba *ènìyàn* ‘human’. Note, however, that the same special character or diacritic does not always correspond to the same IPA phoneme in different scripts. For instance, the Vietnamese *đ* in *đẳng* ‘rank/grade/-class’ corresponds to [d] (Dình Hoà, 1996, p. 649), whereas in Serbian *rađaju* ‘born’ it corresponds to [dʒ] (Feldman and Barac-Cikoja, 1996, p. 771).

Generally speaking, modifications of Latin scripts accommodate for specific phonemes that could otherwise not clearly be identified in the written form of the language. Hence, they reflect phonemic properties of a language, rather than just being an idiosyncrasy of the script itself.

Greek script

A script similar to Latin is the Greek alphabet. It consists of 24 letters (7 vowels and 17 consonants) with the letter sigma [s] taking different shapes depending on whether it occurs in word initial and medial position (σ) or word finally (ς).

(15) Greek (ell, polytonic, UDHR 01)
Ὅλοι οἱ ἄνθρωποι γεννιοῦνται ἐλεύθεροι καὶ ἴσοι στὴν ἀξιοπρέπεια καὶ τὰ δικαιώματα.

(16) Greek (ell, monotonic, UDHR 01)
'Ολοι οι άνθρωποι γεννιούνται ελεύθεροι και ίσοι στην αξιοπρέπεια και τα δικαιώματα.

Moreover, there are diacritics indicating rough (῾) and smooth (’) breathing, as well as acute (e.g. ύ), grave (e.g. ὶ), and circumflex (e.g. ῦ) pitch accents and syllable stress as in άνθρωποι *anthropoi* ‘humans’, where the stress is on the first syllable. The inverted apostrophe (῾) on top of a vowel indicates that aspiration [h] is added, e.g. Ὅλοι ‘all’ would be transcribed as *holoi*. Up to the 1980s all of these diacritics were in widespread use in printing, i.e. represented by the so-called *polytonic* variety of the Greek script. Since then, the *monotonic* variety is on the rise. It only marks an acute accent on any stressed syllable (Threatte, 1996, p. 277).

Cyrillic script

The Cyrillic script, illustrated here with Russian, is based on an alphabet of 33 letters, again similar to the Latin script. It is augmented by diacritics that can represent differences in meaning. For example, the first word *Все* [fs^{j}ɛ] ‘all’ has a plural meaning as in ‘all humans’. It can be modified to *Всё* [fs^{j}ɔ] ‘every’ with a singular meaning as in ‘every human’ (Cubberley, 1996, p. 353).

(17) Russian (rus, UDHR 01)
Все люди рождаются свободными и равными в своем достоинстве и правах.

Devanagari script

Devanagari scripts, as in the Hindi example in (18), are alphasyllabaries (abugidas), where a consonant-vowel combination (called akṣara) defines the basic unit of writing. It consists of a character representing the consonant, plus an obligatory diacritic representing the vowel (Bright, 1996, p. 385). For example, the verb है *hai* ‘is/are’ (last word in the example sentence), is a combination of the consonant ह [ɦ] and the diphthong ऐ [ɛ]. The spaces in between complexes of akṣaras are, in classical writing, not necessarily systematic with regards to word type boundaries. However, in modern writing, punctuation (। and ।।) and spaces follow the more systematic usage of Latin scripts (Bright, 1996, p. 386).

(18) Hindi (hin, UDHR 01)
सभी मनुष्यों को गौरव और अधिकारों के मामले में जन्मजात स्वतन्त्रता
और समानता प्राप्त है ।

Arabic script

The Arabic script is composed of 28 letters that historically represent consonants only (Bauer, 1996, p. 561), an example of a so-called *abjad*. Though each letter represents exactly one consonant, there can be different graphic realizations according to where in a word that consonant occurs (initial, medial, final). Normally, letters within a word are joined together in cursive script, but there are some exceptions with letters such as the alif ا, which cannot be joined to the following consonant. Hence, there can be minimal spaces within words as in الناس *al-nas* ‘people’ (read from right to left). We need to take into consideration that here – and generally in Arabic writing – the definite article ال *al* ‘the’ is prefixed to the noun, and is hence not a separate word type as in English. Generally speaking, the Arabic script also delimits word types by white spaces, though lexical units that are represented by only one letter are joined to the following word type (Bauer, 1996, p. 559). Note that Arabic script has to be read from right to left. Hence, the sentence in (19) starts with the verb يولد ‘born’, which is typical for Arabic varieties of the VSO type.

(19) Standard Arabic (arb, UDHR 01)
يولد جميع الناس أحرارا متساوين في الكرامة و الحقوق

Korean (Hankul) script

The case of the Korean *Hankul* script is somewhat different from the other scripts discussed here. It was invented in the 15th century as a phonemically based alphabet (King, 1996, p. 219). In this writing system, words are composed of syllables that are in turn composed of graphemes corresponding to phonemes. For example, in (20) the word 인간은, transcribed into Latin as *inkan-eun* and meaning ‘humans’, consists of three syllables. The first syllable 인 (*in*) is composed of a zero marker ㅇ (indicating that the syllable does not start with a consonant), plus the vowel ㅣ [i], and the consonant ㄴ [n]. The second syllable 간 (*kan*) is composed of ㄱ [k], ㅏ [a] and ㄴ [n]. The third syllable 은 (*eun*) is composed of ㅇ (zero marker), ㅡ [eu] and ㄴ [n] again. Notably, there are spaces in between complexes of syllables that correspond to word types.

(20) Korean, Hankul (kor, UDHR 01)
모든 인간은 태어날때부터 자유로 우며 그 존엄과 권리에 있어 동등하다.

5.1.1 Scripts and unigram entropies

Considering the subtle differences in scripts, we might ask whether these have an impact on word type distributions and hence the unigram entropies we are measuring in parallel texts. In Figure 5.1, the entropies of languages are binned by the most common scripts: Latin, Devanagari, Arabic, Cyrillic, and Ge’ez. The Ge’ez script was discussed in Chapter 4. These five script types cover circa 97% of the parallel texts in the corpus sample. All other scripts are represented by less than five different languages and are excluded in this analysis. Languages written in Latin tend to have the lowest scaled unigram entropy values ($\mu = -0.1, \sigma = 0.96$), followed by Devanagari ($\mu = 0.48, \sigma = 0.69$), and Arabic scripts ($\mu = 0.77, \sigma = 0.76$), while languages written in Cyrillic ($\mu = 1.24, \sigma = 0.64$) and Ge’ez ($\mu = 1.37, \sigma = 0.45$) have the highest average entropy values. Some of these differences are statistically significant according to a Wilcoxon rank sum test. This is the case for Latin and Devanagari ($p < 0.05$), while for Cyrillic and Ge'ez there is no significant difference ($p > 0.05$).[2]

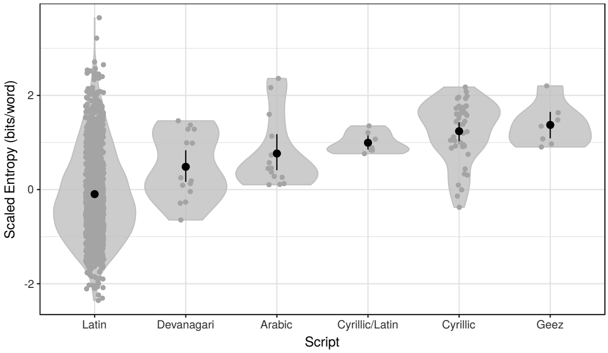

Figure 5.1: Violin plots of scaled unigram entropies per script. Scaled unigram entropies (y-axis) are categorized by script type, i.e. Latin, Devanagari, Arabic, Cyrillic/Latin, Cyrillic, and Ge'ez. Scripts represented by less than five data points are excluded. Black dots indicate mean values with confidence intervals. Light grey violins outline symmetric density distributions of entropic values. Individual data points are plotted in grey, with jitter added for better visibility.

Importantly, this is not to say that different scripts are the *cause* for different average entropies. Other properties of the languages involved, e.g. morphological marking strategies, can likewise account for the differences. A more telling test is to measure the unigram entropy of the same language written in different scripts. In fact, for some languages of the UDHR and PBC this is possible. The Greek (ell) UDHR is documented in both a *polytonic* and *monotonic* version. The UDHR in Azerbaijani (azj), Bosnian (bos), Serbian (srp), and Uzbek (uzn) is available in Latin and Cyrillic. The UHDR in Uyghur (uig) is available in Latin and Arabic, etc.

Unigram entropy values for two different scripts used in the same language are illustrated in Figure 5.2. Here, unscaled shrinkage entropies are reported, since scaled entropies are based on average values across different translations of the same text, and hence would not display the deviation between translations.
The biggest difference we find for the Korean (kor) Bible translations in Latin and Hankul scripts, amounting to $\Delta\hat{H} = 0.36$. This corresponds to a 3.2% unigram entropy change. There is also some visible difference between the Ge'ez and Latin translations into Gamo (gmv) ($\Delta\hat{H} = 0.08$) as well as between the Arabic and

[2] file: Rcode/Chapter5/entropyWritingSystems.R

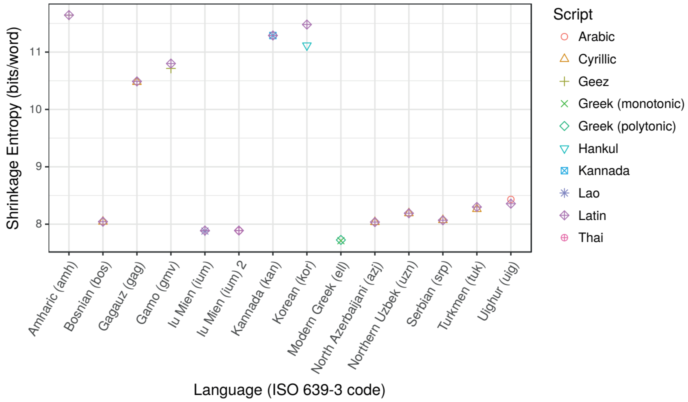

Figure 5.2: Same language written in different scripts. Shrinkage (unscaled) unigram entropy values (y-axes) for languages that are transcribed into two different scripts indicated by colours and shapes.

Latin Bible in Uyghur (uig) ($\Delta\hat{H} = 0.07$). All other differences are below 0.05, i.e. less than 1% unigram entropy change. The monotonic and polytonic Greek versions display a difference of $\Delta\hat{H} = 0.02$. The Latin and Cyrillic versions in Azerbaijani (azj), Bosnian (bos), and Uzbek (uzn) are virtually indistinguishable in terms of entropy values and so are the Kannada translations in its traditional script compared to Latin. The discrepancy in the Greek scripts might be due to the fact that polytonic scripts use more diacritics. Additional diacritics in the polytonic version might generate two separate word types where in the monotonic script we find only one. This can (slightly) increase the lexical diversity of a text written in the polytonic script.

However, the differences observed here can also be due to other factors, such as translation style, and usage of vocabulary of a different breadth. Hence, these differences have to be seen as the *maximum potential difference* caused by scripts. Across the board, these have low values.

To conclude this section we can state that while the architecture of scripts can be vastly different across languages of the world, this difference is only marginally reflected in unigram entropy values. As long as scripts delimit character strings by white spaces and non-alphanumeric characters, they are suitable for automated tokenization. Within the subset of scripts that delimit words by white spaces, the entropies and hence lexical diversities are affected only minimally by the choice of script. This is the case at least for the ones tested here.

5.2 Word-formation

Another descriptive, language “internal” candidate shaping word frequency distributions is productive word-formation. The strategies that languages adopt to create complex word types differ widely across areas and families of the world. One of the most well-known scales runs from the *analytic* to the *synthetic*, or even *polysynthetic* type of language (see Greenberg, 1960, for a historical overview of these terms).

The idea behind this typological distinction is to categorize languages according to the number of morphemes per word they allow (Aikhenvald, 2007, p. 5), and in turn, the average number of words found in sentences. Theoretically, a language with the minimum possible number of morphemes per word – one morpheme for each word – is considered purely analytic, whereas a language with multi-morpheme word types is considered synthetic, though neither an exact average number, nor the maximum number of morphemes required for the synthetic type is clearly defined in the literature.

5.2.1 Analytic, synthetic and polysynthetic languages

Take, for instance, the English phrase *I will go*. We find three independent morphemes, corresponding to three separate word types. They indicate first person, future tense, and the type of action in three separate units. This represents an analytic strategy of encoding information. In contrast, the Italian *andr-ò* encodes the same information by means of using two morphemes. Namely, *andr-* as the root of *andare* ‘to go’ and *ò* as an inflectional marker of first person and future tense. Hence, in relation to English, Italian represents a more synthetic strategy. The synthetic type is taken to its extreme by *polysynthetic* languages with a whole range of prefixes and suffixes surrounding the root morpheme, sometimes even allowing for nouns to become part of verbs, a phenomenon called *noun incorporation* (Aikhenvald, 2007, p. 5).

The difference between analytic, synthetic, and polysynthetic languages is illustrated in examples (21), (22), and (23). The first gives a verse of the Hawaiian (Austronesian) Bible and the latter two a similar verse in the Turkish (Turkic) and Iñupiatun (Eskimo-Aleut) Bibles.[3] Hawaiian is often referred to as an extremely analytic language. The Hawaiian verb *olelo* ‘say’ is here used in its infinite form, while the perfective aspect of completed action is indicated by the particle *ua*.

[3] Not all verses are given for each language in the PBC.

The infinite form can be combined with further particles to indicate imperfective aspect as in *e olelo ana* ‘was saying/ will say’, present (continuing) as in *ke olelo nei* ‘saying’, imperative as in *e olelo* ‘say’, and negative imperative as in *mai olelo* ‘don’t say’ (Pukui and Elbert, 1975, p. 228-229). In this regard, Hawaiian is even more analytic than English, which uses different word types in these contexts in a more synthetic strategy (*said*, *saying*, *say*).

(21) Hawaiian (haw, PBC 41006018)

A	*ua*	*olelo*	*aku*	*o*	*Ioane*	*ia ia*	[...]
Then	PERF	say	to	SUBJ	Johan	he.DAT	[...]

“Then Johan said to him [...]”

(22) Turkish (tur, PBC 41006004)

Ýsa	*da*	*on-lar-a*	[...]	*de-di*
Jesus	also	3P-PL-DAT	[...]	say-3SG.PERF

“Jesus also said to them [...]”

(23) Iñupiatun (esk, PBC 41006004)

Aglaan	*Jesus-ŋum*	*itna-ġ-ni-ġai*	[...]
But	Jesus-ERG	this-say-report-3S.to.3PL	

“But Jesus said to them (it is reported) [...]”

Contrast this with the Turkish example. While the proper name *Ýsa* and the adverb *da* are construed as separate entities here as well, the personal pronoun *onlara* ‘them’ carries information about person, number and case, and the verb form *dedi* ‘said’ consists of the verb root *de* ‘say’ and an allomorph of the perfective marker *ti*. While for this particular example, the Turkish “syntheticity” is comparable to English, with *them* and *said* carrying similar grammatical information, there are examples where Turkish takes suffixation to its extremes, as in the following complex word form given by Göksel and Kerslake (2005, p. 65).

(24) Turkish (tur)

Ev	*-ler*	*-imiz*	*-de*	*-ymiş*	*-ler.*
home	-PL	-1PL.POSS	-LOC	-EV.COP	-3PL

“Apparently they are/were at our homes.”

Still, Turkish is generally not considered a polysynthetic language, since it tends to build even its complex words on single roots, e.g. *ev* ‘house/home’, while polysynthesis is commonly defined as involving several roots in a single complex word form. However, there are several criteria defining polysynthetic languages (Aikhenvald, 2007, p. 5-6) and there is no general agreement on a single criterion.

In the Iñupiatun example, the finite verb form *itnaġniġai* consists of a complex verb stem *itna-q-* 'to say this' (MacLean, 2012, p. 174). By itself *itna* means 'this', and *-q-* means 'say'. The verb root thus merges with the demonstrative into a verb stem meaning 'to say this' (MacLean, 2012, p. x). Further, the suffix *-ni* 'to report' indicates reported speech and the suffix *ġai* clarifies who said what to whom, namely 'he/she/it to them' (Lanz, 2010, p. 83). Iñupiatun, and other Eskimo-Aleut languages, are often given as typical examples of polysynthesis.

Hence, while the highly analytic Hawaiian uses separate and independent word types to encode information about who is saying what to whom, the same information is encoded in two synthetic word types in Turkish and a single word type in the polysynthetic Iñupiatun. Clearly, such differences in word-formation change the distributions of word tokens over word types. In Hawaiian, the same particles *ua*, *e*, *ana*, *kei*, *mai*, as well as the infinite verb forms occur over and over again, namely, whenever a specific tense is used. In Turkish, on the other hand, there is a panoply of different word types built on verb roots such as *de* 'say', and in Iñupiatun there are many different word types consisting of a stem like *itnaq-* and inflectional modifications. Hence, texts written in Turkish and Iñupiatun have (under the assumption of constant content) a longer tailed distribution of word tokens over word types and higher lexical diversity than the rather analytic Hawaiian. This is part of the reason why the Hawaiian Bible translation has a scaled unigram entropy of $\hat{H}^{\text{scaled}} = -1.81$, compared to $\hat{H}^{\text{scaled}} = -0.38$ for English, and $\hat{H}^{\text{scaled}} = 2.08$ for Turkish. Iñupiatun is among the highest entropy languages with $\hat{H}^{\text{scaled}} = 3.22$.

Thus, the theoretical considerations about analytic, synthetic, and polysynthetic languages are reflected in unigram entropies. These observations do not only hold for verb and noun forms, but apply in principle to any other part of speech that is part of productive word-formation. Also, we do not have to look into extreme examples like Hawaiian and Iñupiatun to see the differences. Even for closely related languages like English and German subtle nuances in inflectional marking strategies accumulate and lead to clear deviations in word frequency distributions. To illustrate this effect, the Zipf distributions of token frequencies for word types in English and German are plotted in Figure 5.3.[4] Token frequencies are taken from a corpus built in Bentz et al. (2017b).

Across different parts of speech (pronouns, definite articles, nouns and verbs) German is more synthetic in the sense of having more different word types encod-

[4] file: Rcode/Chapter5/entropyMorphologyOverview.R

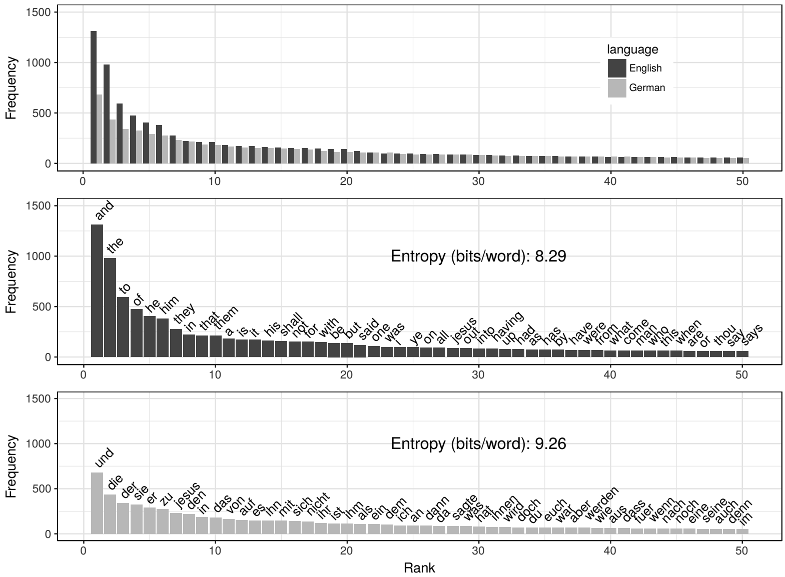

Figure 5.3: English and German word frequency distributions. Example distributions of the first 50 ranks based on Bentz et al. (2017b). Unigram entropies are unscaled shrinkage values for the whole distributions. The panels give English and German token frequencies side by side (top), as well as token frequencies for English (middle), and token frequencies for German (bottom) with the respective word types written above each bar.

ing grammatical information by bound morphemes than English. Consider the range of definite and indefinite articles. While English uses only two word types *the* and *a* with high frequencies, German uses *der*, *die*, *das*, *dem*, *den*, *des* as well as *ein*, *eine*, *einem*, *einen* with lower frequencies respectively. This will lead to a longer tailed, more uniform distribution for German and hence (at least in accumulation) to higher unigram entropies.

The idea that synthetic and analytic encoding strategies are reflected in word frequency distributions is certainly not new, but goes back to George Kingsley Zipf, who manually counted and analysed the patterns of word frequencies in Latin, Chinese and English (Zipf, 1932), as well as Old English, French, Hebrew, Plains Cree and others (Zipf, 1949, 1935). His analyses suggested that it is possible to measure the “degree of inflection” in what Zipf called “positional” (i.e. analytic) and “inflected” (i.e. synthetic) languages. Based on modern corpora and computational tools several researchers have confirmed that such a syntheticity index is possible (Bentz et al., 2014, 2015, 2017b; Baroni, 2009; Popescu et al., 2009, 2010; Ha et al., 2006).

These quantitative accounts of inflectional typology illustrate that the opposition between analyticity and syntheticity is rather a scale than a binary choice – just as the distinction between writing systems such as logographies, syllabaries and alphabets is hardly ever an absolute one. Individual languages can range anywhere on the scale from extremely analytic to extremely synthetic (or polysynthetic), and they can change their position over time. A prominent example is the history of English, which became more analytic from Old English towards Modern English (Szmrecsanyi, 2012; Bentz et al., 2014). Similar variation is also attested synchronically between different varieties of English (Szmrecsanyi, 2009).

5.2.2 Isolating, agglutinative and fusional languages

Another set of distinctions, overlapping with the analyticity/syntheticity scale, is the cline from *isolating* to *agglutinative*, and finally *fusional* languages (Aikhenvald, 2007; Pereltsvaig, 2012; Dixon, 1994). The perfectly analytic type and the isolating type are equivalent, since both exclusively use independent, i.e. unbound, morphemes as word types (Aikhenvald, 2007, p. 3; Dixon, 1994, p. 182).

Languages of the synthetic type, on the other hand, can be further subcategorized into *agglutinative* and *fusional* languages. Agglutinative languages have clear morpheme boundaries with multiple – but distinct – morphemes "glued" together, while fusional languages "fuse" morphemes of different grammatical functions into a single morpheme. Recently Bickel and Nichols (2007) have elaborated that this classic cline confuses three separate dimensions of morphological marking which they call *exponence*, *fusion* and *flexivity*.

Consider the English noun *man*. We can illustrate the agglutinative type with its Hungarian equivalent *ember*, and the fusional type with the German equivalent *Mann*. In Hungarian, *ember* is pluralized with the inflection *-ek,* i.e. *ember-ek* 'man-PL', and marked for accusative case with the inflection *-et*, i.e. *ember-ek-et* 'man-PL-ACC' (Aikhenvald, 2007, p. 4). This is called 'concatenative' morphology on the fusion scale by Bickel and Nichols (2007). In German, the plural of *Mann* is formed by adding the suffix *-er* (plus Umlaut) to the root, i.e. *Männ-er* 'man-PL'. However, the same word type is used in a nominative *and* in an accusative context, i.e. *Männ-er* 'man-PL.NOM/ACC'. So it can be argued that the inflection *-er* "fuses" both plural and accusative meaning together in German, instead of having separate inflections "glued" together as in Hungarian. However, Bickel and Nichols (2007) further clarify that usage of different markers for different grammatical purposes relates to the grammatical dimension of "exponence", with monoexponential (Hungarian) and polyexponential (German) being the two basic types. The degree of phonological fusion, on the other hand, is construed on a separate scale from isolating to concatenative (Hungarian), and finally nonlinear, i.e. nonconcatenative morphology.

Again, there is generally no hard and fast distinction between purely isolating, purely agglutinative, and purely fusional languages – or any of the values on the exponence, flexivity and fusion scale for that matter. Aikhenvald (2007) gives some languages as standard examples for the classic categories, i.e. Vietnamese and Chinese as isolating, Hungarian and Turkish as agglutinative, and Latin and Russian as fusional languages, but there is generally some variation even within the same language. Note that the German dative plural of *Mann* is *Männ-er-n*, where the *-n* suffix marks dative case. Hence, the *-er* suffix can be said to be polyexponential for representing nominative and accusative plural, while the *-n* suffix is monoexponential for indicating the dative plural only.

For the approach advocated here – namely measuring unigram entropy – these typologically fine-grained distinctions are currently out of reach. Unigram entropy estimation does not capture whether the plural of the concept *man* is encoded by adding regular morphology as in Hungarian *ember-ek*, by change of the stem vowel as in English *men*, or a combination of these as in German *Männ-er*.[5] What matters is only the range of different word types that languages use to encode the same information. This is not a limitation of information-theoretic accounts in general. They can be adjusted to take into account word internal structure as well (Koplenig et al., 2017).

To sum up, it is to be expected that inflectional morphology, and word-formation patterns in general, will have various effects on the distributions of word types in a language and by extension also on unigram entropies. In the following, this impact is measured more precisely across several languages, thus clarifying *how much* variance word-formation introduces to unigram entropies. Several metrics to measure unigram entropy differences and cross-linguistic variance are discussed in turn.

5.2.3 Entropy share and cross-linguistic entropy variance

Assume a corpus A is manipulated, for example, by removing certain word types, by neutralizing inflections, or by splitting all compound words, thus yielding a modified corpus B. To measure how much impact this manipulation has on unigram entropies, we can estimate them before (i.e. $\hat{H}_{\mathrm{A}}$) and after (i.e. $\hat{H}_{\mathrm{B}}$) manip-

[5] However, it seems likely that languages categorized as agglutinative generally have less syncretism in their paradigms than fusional languages and hence have higher entropies on average.

ulation. The unigram entropy difference between the original and manipulated corpus is then

$$\Delta \hat{H}_{\mathrm{A}\rightarrow\mathrm{B}} = \hat{H}_{\mathrm{A}} - \hat{H}_{\mathrm{B}}. \tag{5.1}$$

This is an index of how much information was removed (or added) to corpus A by means of a given manipulation. Furthermore, we can normalize this difference by dividing it by the original unigram entropy of corpus A. This is here called the *entropy share* of a given manipulation, defined as

$$\hat{S}_{\mathrm{A}\rightarrow\mathrm{B}} = \frac{\Delta \hat{H}_{\mathrm{A}\rightarrow\mathrm{B}}}{\hat{H}_{\mathrm{A}}}. \tag{5.2}$$

In other words, this is the percentage of unigram entropy change as we manipulate the corpus. It can be conceptualized as the "share" that the information encoding feature, which was manipulated, has in the overall unigram entropy of a given corpus. This can tell us more about how a given language uses different encoding strategies to transmit information.

Another important question is how much of the unigram entropy difference that we find across languages is due to differences in particular encoding features. For a given set of languages (i.e. corpora) A, we can estimate the unigram entropy variance before and after manipulation, and again normalize it to get the percentage of *explained variance* (EV) of the encoding feature that was manipulated:

$$\hat{EV}_{\mathrm{A}\rightarrow\mathrm{B}} = \frac{\mathrm{Var}(\hat{H}_{\mathrm{A}}) - \mathrm{Var}(\hat{H}_{\mathrm{B}})}{\mathrm{Var}(\hat{H}_{\mathrm{A}})}. \tag{5.3}$$

The variance is here calculated over all unigram entropies of a given set of languages. $\hat{EV}_{\mathrm{A}\rightarrow\mathrm{B}}$ measures the contribution of an encoding feature to the cross-linguistic difference in unigram entropies.

In the following, productive processes to form new words are subcategorized into *inflectional morphology* (Section 5.2.4), *derivational morphology*, *clitics/contractions*, as well as *compounds* (Section 5.2.5). Finally, we also look at the information encoding potential of *tones* (Section 5.2.6). In some cases, these analyses require automatic and manual modifications of corpora which are not available for the current corpus sample. Therefore, some of the relevant material is taken from an earlier study by Bentz et al. (2017b). For this older corpus, differences in numbers of tokens are not a concern. Hence, the unscaled unigram entropy is used throughout these sections.

5.2.4 Inflectional morphology

To analyse systematic differences between inflected and non-inflected parallel corpora, state-of-the-art lemmatization tools are employed to automatically neutralize inflections. In English, for instance, inflections are neutralized for different parts of speech, mainly regular and irregular verbs (e.g. *decides/decide/decided* → *decide*, *sings/sang/sung* → *sing*) and nouns (e.g. *noses* → *nose*, *children* → *child*). An example of word types ranked according to their token frequencies for an English corpus, compared to the lemmatized version of it, is given in Figure 5.4.[6]

While the three most frequent word types (*and*, *the*, *of*) are not affected by lemmatization, we already see a difference in the fourth rank: the pronoun *he* is replaced by the lemmatized copula *be*. This is because token frequencies that were distributed over different inflectional variants in the original corpus (*is*, *are*, *was*, *were*, *being*) are now “accumulated” in the lemma *be*. As a general trend, frequencies of former inflectional variants of verbs (e.g. *say*, *have*) and nouns (e.g. *man*) accumulate in the respective lemmas. As a consequence, these are “pushed” further up the ranks. This leads to a shift of the word frequency distribution towards the y-axis, i.e. towards higher frequencies and hence to lower unigram entropies. In English, this is a subtle effect. The unigram entropy in this example drops from 8.29 bits/word to 7.94 bits/word, which is around 4%. The entropy share will increase for languages where inflectional marking is more pronounced.

To measure the size of this effect across languages, a combination of the full UDHR and the full PBC is used. The range of languages included is limited by the availability of lemmatization tools. We arrive at parallel corpora of 12000 to 17000 tokens for 19 different languages (see Table 5.2 for details). The word tokens of these corpora are lemmatized by using the BTagger (Gesmundo and Samardžić, 2012) and TreeTagger (Schmid, 1994, 1995). Both the BTagger and the TreeTagger first assign a part-of-speech tag (POS tag) to each word token, and then neutralize it to the most likely lemma. For example, given the English word token *rights* the BTagger outputs: *rights/Nc/right*. This is the original token, the POS tag for common noun and the respective lemma.

Of course, automated processing inevitably results in errors. These can influence the observed differences between original and lemmatized texts. The frequencies and the types of errors depend on the lemmatization tool and the level of difficulty. Both taggers employ statistical models which are trained on samples of manually lemmatized texts and provide high accuracy on words already seen in

[6] file: Rcode/Chapter5/entropyMorphEnglish.R

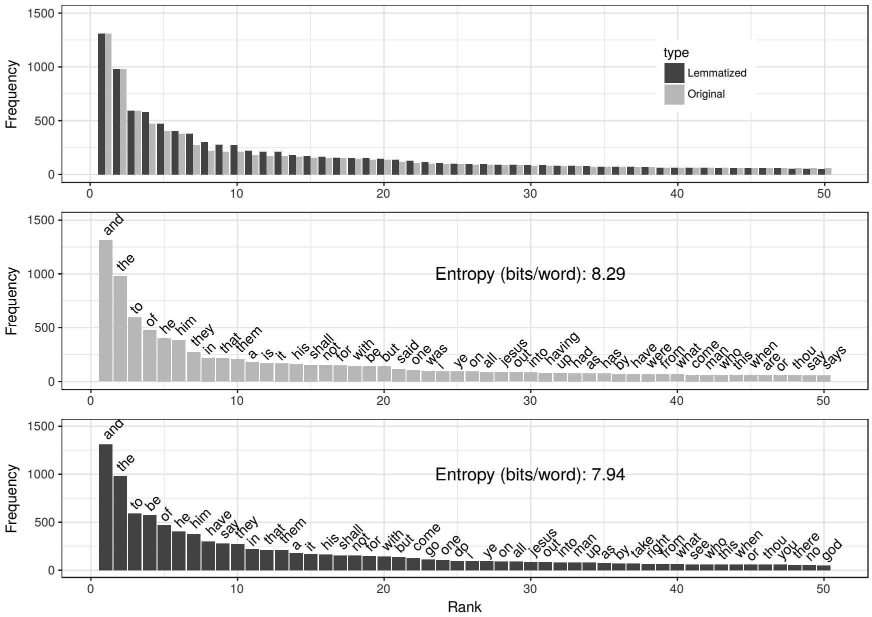

Figure 5.4: Word frequency distributions for an original and lemmatized English corpus. The example is constraint to the first 50 ranks. Unigram entropies are unscaled shrinkage values for the whole distribution.The panels give the original (light grey) and lemmatized (dark grey) distributions side by side (top), as well as the original (middle) and lemmatized (bottom) separately with respective word types above bars.

the training set (close to 100%). The words not seen in the training set are harder to lemmatize, and hence are expected to result in more erroneous lemmas. Table 5.2 shows the percentage of unknown word types by corpus and tagger.

Despite some differences in the percentages of unknown tokens, the overall effect of errors on the entropy estimation is expected to be similar across languages. Both taggers will transform fewer word types to lemmas than they actually should. In consequence, there is less difference between original and lemmatized frequency distributions than there should be. This, in turn, results in an underestimation of the actual unigram entropy difference. Also, there are generally more unknown words in languages with many inflectional categories. For example, for both taggers the percentage of unknown tokens is higher for Polish than for English. It is thus expected that our estimations are less reliable for languages with abundant inflections.

Table 5.2: Information on lemmatizers, including languages, ISO codes, taggers, number of tokens per parallel corpus, number and percentage of unknown tokens. Adopted from Bentz et al. (2017b).

Language	ISO	Tagger	No. Tokens	unknown	%
Bulgarian	bul	TreeTagger	13993	497	3.6
Czech	ces	BTagger	12020	3068	25
Dutch	nld	TreeTagger	16732	1089	6.5
English	eng	BTagger	16781	2140	13
English	eng	TreeTagger	16781	486	2.9
Estonian	est	BTagger	12807	3116	24
Estonian	est	TreeTagger	12807	1621	12.7
Finnish	fin	TreeTagger	11841	1130	9.5
French	fra	TreeTagger	17602	983	5.6
German	deu	TreeTagger	15732	911	5.8
Hungarian	hun	BTagger	12491	3694	30
Italian	ita	TreeTagger	15314	888	5.8
Latin	lat	TreeTagger	11427	266	2.3
Macedonian	mkd	BTagger	15033	3370	22
Polish	pol	BTagger	13188	4026	30
Polish	pol	TreeTagger	13188	1670	12.7
Romanian	ron	BTagger	16278	3766	23
Russian	rus	TreeTagger	12152	957	7.9
Slovak	slk	TreeTagger	11700	304	2.6
Slovene	slv	BTagger	13075	2847	22
Spanish	spa	TreeTagger	15581	907	5.8
Swahili	swh	TreeTagger	12281	638	5.2

With all these caveats in mind, the neutralization of inflections allows us to estimate the differences in entropies between the original and the lemmatized (inflections neutralized) corpus versions, i.e.

$$\Delta \hat{H}_{\text{orig}\rightarrow\text{lem}} = \hat{H}_{\text{orig}} - \hat{H}_{\text{lem}}. \tag{5.4}$$

The entropy values for the original compared to the lemmatized corpora are given in Figure 5.5.[7] Panel a) of Figure 5.5 illustrates the range of entropy values for the original corpora (from a minimum of circa 8.25 bits/word in English to a maximum of circa 10.25 bits/word in Finnish) and the corresponding entropy values after lemmatization.

Panel b) sorts languages according to the entropy difference $\Delta \hat{H}_{\text{orig}\rightarrow\text{lem}}$. English has the lowest entropy difference, followed by Bulgarian and Dutch. In these

[7] file: Rcode/Chapter5/entropyMorphLemma.R

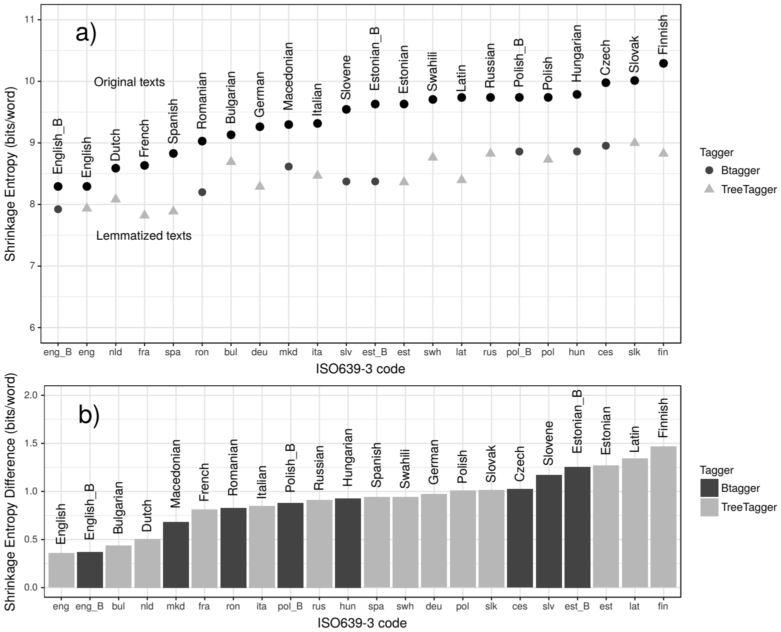

Figure 5.5: Inflectional marking and unigram entropy values. a) Unigram entropy values for original corpora (black dots), corpora lemmatized with the BTagger (dark grey dots), and the TreeTagger (light grey triangles) across 19 languages. Languages are ordered by there original shrinkage entropy from lowest (English) to highest (Finnish). b) Unigram entropy differences, again ordered from lowest to highest.

languages, inflectional morphology plays a minor role for encoding information. The middle range is populated by such languages as French, Russian, Swahili, and German. The biggest entropy differences are found in Estonian, Classical Latin, and Finnish, which heavily rely on inflectional markers.

Normalizing the results in panel b), we can further estimate the entropy share of inflectional marking per language, i.e. $\hat{S}_{\text{orig}\rightarrow\text{lem}}$. The percentages are visualized in Figure 5.6. The unigram entropy share of inflection is highest for Finnish (14.2%), Classical Latin (13.8%), and Estonian (13.1 to 13.2%). It is lowest for English (4.3 to 4.5%), Bulgarian (4.8%), and Dutch (5.9%). This again reflects the differing importance of inflectional marking across these languages. A caveat to keep in mind is that the scale from lowest to highest inflectional entropy share will be influenced by the performance of lemmatization tools. This can distort the ranking of languages, particularly in the middle range. For example, Polish (10.4%) is ranked lower than Spanish (10.7%). However, the percentage of tokens unknown to the TreeTagger is more than twice as high for Polish (12.7%) than for Spanish

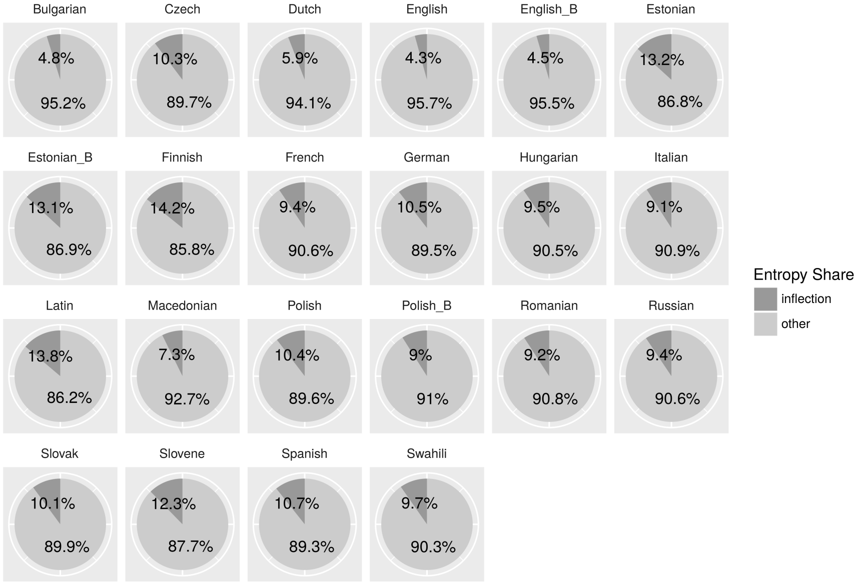

Figure 5.6: Entropy share of inflectional marking across 19 languages. Some languages are represented twice, with the extension _B indicating that the result is based on the BTagger.

(5.8%), and this brings down the estimated entropy share for Polish. Improvement of automated lemmatization tools will increase the precision of quantitative and corpus-based measures such as the entropy share.

Finally, we can estimate how much of the unigram entropy variance between languages is due to differences in inflectional marking, yielding $\widehat{EV}_{\text{orig}\rightarrow\text{lem}}$. Namely, the original variance in entropies is $\text{Var}(\hat{H}_{\text{orig}}) = 0.31$ and the variance that remains after lemmatization is $\text{Var}(\hat{H}_{\text{lem}}) = 0.14$. The variance explained by inflectional differences is then $\frac{0.31-0.14}{0.31} = 0.55$. In other words, neutralization of inflections across these 19 languages reduces the variance in entropy values by 55%. Thus, we can conclude that differences in inflectional marking are the most important factor driving differences in unigram entropies – at least for the 19 languages analysed here – since they explain more than half of the variance between languages.

5.2.5 Derivation, compounds, contractions, and clitics

We can also manipulate corpora by neutralizing derivations, compounds and contractions/clitics, to then get unigram entropy differences, entropy shares, and explained entropy variances for these word-formation processes. However, automatically neutralizing this range of processes is not as straightforward as neutralizing inflectional morphology only. At this point, there are no cross-linguistic computational tools equivalent to the BTagger and TreeTagger for such a task. As an alternative, Bentz et al. (2017b) use manual neutralization to calculate differences in word frequency distributions. This data is here reanalysed with a focus on unigram entropies.

In Bentz et al. (2017b), the parallel corpora for English and German are compiled using parts of the *Open Subtitles Corpus* (OSC),[8] the EPC, UDHR and the Book of Genesis.[9] On the upside, this text sample is balanced between spoken and written language as well as different registers (colloquial, political, legal, religious). On the downside, the sample has to be kept small (9211 tokens in English, 8304 in German), in order to enable maximally informed, manual neutralization of word-formation patterns.

Derivational morphology

Derivational morphology, just like inflectional morphology, has an impact on the range of different word types a language uses. For instance, there is a range of Germanic and Latin prefixes and suffixes that are used to derive new word types from roots in English (e.g. *in-alien-able* → *alien*, *hope-ful-ly* → *hope*, *childhood* → *child*). For German the respective decisions are in some cases more difficult since several derivational affixes can be attached to the same root (e.g. *Anerkennung* → *kennen*, *Errungenschaften* → *ringen*) and can be mixed with compounding (e.g. *Dringlichkeitsdebatte* → *Dringensdebatte*) or inflectional morphology (e.g. *abgeändert* → *ändert*).

Clitics and contractions

The parallel text sample includes the *Open Subtitles Corpus*, consisting of (scripted) spoken language. As a result, there is a range of contractions and clitics (e.g. *you've* → *you have*, *you're* → *you are* , *I'll* → *I will*, *won't* → *will not*, *parliament's* → *parliament*). The *'s* genitive is included both under inflection and under clitics. From a theoretical perspective, it is often categorized as a phrasal clitic, namely, it does not attach exclusively to nouns, but rather to noun phrases. However, in language production, it is in most cases directly following nouns and is thus likely perceived as an element very similar to noun inflection by learners and speakers.

[8] 2013, http://opus.lingfil.uu.se/OpenSubtitles2013.php

[9] This is not the PBC but another Bible corpus at http://homepages.inf.ed.ac.uk/ s0787820/bible/

We find similar clitics and contractions in the German sample (*geht's* → *geht es*, *rührt's* → *rührt es*, *dir's* → *dir es*, *beim* → *bei dem*, *ins* → *in das*).

Compounds

The last word-formation pattern considered here is compounding. In both English and German different parts of speech can be compounded to yield complex word types (e.g. noun-noun, adjective-noun, preposition-noun, among others). To neutralize them, these compounds are split into separate word types (e.g. *daytime* → *day time*, *downstairs* → *down stairs*, *gentlemen* → *gentle men*). An exception are proper names such as *Hellfish*, which are not "de-compounded". Similar principles apply to German (e.g. *Arbeitsschutzregelungen* → *Arbeit schutz regelungen*, *kräuterstinkender* → *kräuter stinkender*).

Results: entropy shares

Both the English and the German corpus are neutralized for derivations, compounds, contractions and clitics – in separate steps and according to the principles outlined above. These are discussed in more detail in Bentz et al. (2017b). For comparison purposes, inflections are also neutralized manually. The results in terms of entropy shares are summarized in Figure 5.7.[10]

Of all word-formation processes, inflections have the highest entropy share for both German ($\hat{S}^{\text{deu}}_{\text{orig}\to\text{lem}} = 7.9\%$), and English ($\hat{S}^{\text{eng}}_{\text{orig}\to\text{lem}} = 5.4\%$). Compare this to the results for the automatically lemmatized corpora in Section 5.2.4, where we found 10.5% and 4.3 to 4.5% respectively. There is a discrepancy of 2.6% for German and 0.9 to 1.1% in English. Such discrepancies can be due to either differences in the corpus samples or differences in the lemmatization principles. The second highest entropy share is found for derivation in German (2.1%) and for clitics/contractions in English (1.1%). This seems to reflect a propensity to use more derivational morphology in German and more contractions and clitics in English. However, to make general statements about English and German information encoding, such patterns will have to be tested on bigger, more representative corpora.

Moreover, in English, compounds have a very small, but positive, entropy share (0.3%), while in German the entropy share is actually negative (-0.6%). This is because, in the English corpus, the unigram entropy decreases when compounds are split into separate word tokens, whereas in the German corpus the same modification actually leads to a unigram entropy *increase*. From an

[10] file: Rcode/Chapter5/entropyWordFormation.R

information-theoretic perspective, there is a fundamental difference between inflectional and derivational processes, on one hand, and compounding, on the other.

To understand this better, notice that neutralization of inflections and derivations either leads to an increase in token frequencies and a decrease in the number of types (e.g. neutralizing *go*, *goes*, *went*, *gone* to *go*), or to the replacement of one word type for another (e.g. replacing the word type *goes* by its lemma *go* if there is no other inflected form of that lemma in the text). In the first scenario, the unigram entropy decreases. In the second scenario, it stays the same. However, the first scenario is much more likely – especially for neutralization of many different word types – and lemmatization will thus quite generally decrease the unigram entropy of the text.

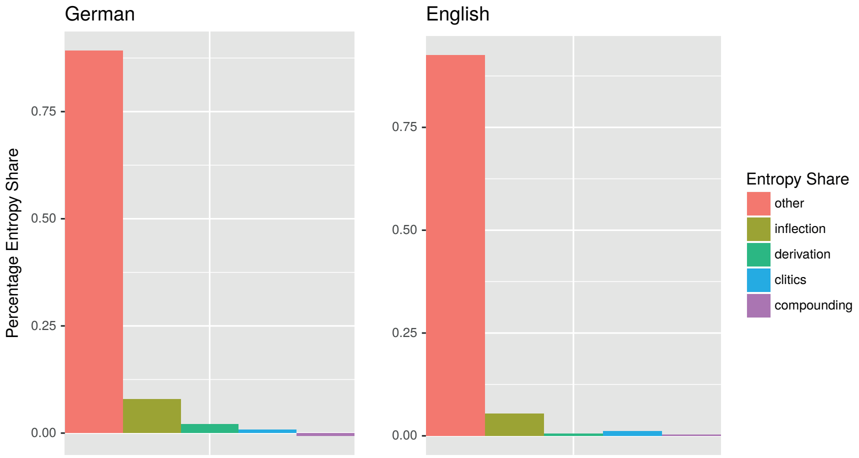

Figure 5.7: Entropy shares of inflections, derivations, clitics/contractions and compounds in German and English. This is illustrated in a bar plot rather than a pie chart, since compounding takes a negative entropy share in German (-0.6%).

In contrast, assume the compound *daytime* occurs only once in an English corpus. If it is split into *day* and *time*, then there are two possible scenarios: a) if the word types *day* and *time* already occurred in the corpus independent of the compound, then the additional tokens of *day* and *time* will just be added to the already existing token frequencies. This corresponds to a net increase in tokens by two and to a net decrease of types by one, and hence reduces the entropy (everything else being equal); b) if the word types *day* and *time*, on the other hand, are not represented in the original corpus, then splitting *daytime* will remove one word type, but also create two new ones. This leads to a net increase of word types by one, and hence to an increase in entropy. In the analyses above, scenario a) seems to be prevalent in the English corpus, and scenario b) in the German corpus.

Thus, inflection and derivation add information to a text by default, while compounding can result in either adding or removal of information. As a consequence, when we reverse inflectional and derivational processes by neutralization, we inevitably remove information from the corpus. Reversing compounding, on the other hand, can lead to *either* removal *or* adding of information. Again, these processes have to be tested with bigger, and more representative corpora to derive general conclusions about English and German information encoding.

To conclude, all the word-formation processes en bloc amount to an overall entropy share of 10.8% in German (disregarding the -0.6% of compounding) and 7.4% in English. The vast “residues” of entropy shares (89.2% and 92.6%) are then covered by the base vocabulary. Thus, from an information-theoretic perspective, we still keep around 90% of the information in English and German corpora even if all word-formation processes are neutralized, illustrating that the basic lexicon is by far the most important dimension of information encoding.

Results: explained entropy variance

Apart from analysing entropy shares per language, we can also ask how much the original difference between the German and English corpus is driven by the differences in information encoding strategies. The original unigram entropy value for the English corpus is $\hat{H}^{\text{eng}}_{\text{orig}} = 8.83$, and for German $\hat{H}^{\text{deu}}_{\text{orig}} = 9.39$. This amounts to a unigram entropy difference of $\Delta\hat{H} = 0.56$.

Manual neutralization of inflections in the English sample corpus yields an entropy of 8.36 bits/word, and for German 8.65 bits/word. Hence, the after-neutralization entropy difference is 0.29 bits/word. The proportion of entropy difference explained by inflectional morphology $\hat{EV}_{\text{orig}\rightarrow\text{lem}}$ is thus $\frac{0.56-0.29}{0.56} = 0.48$. This means 48% of the unigram entropy difference between English and German is explained by differences in the productivity of inflectional markers. This is similar to the result across 19 languages above, where we had 55% variance explained by inflection.

In comparison, the English entropy value after neutralization of derivations is 8.78 bits/word and the German one is 9.19 bits/word. This amounts to an after-neutralization entropy difference of 0.41 bits/word. The variance explained by derivational morphology $\hat{EV}_{\text{orig}\rightarrow\text{root}}$ is thus $\frac{0.56-0.41}{0.56} = 0.27$. This means another 27% of entropy difference is explained by derivational morphology. Furthermore, the entropy of the English distribution after neutralization of clitics and contractions is 8.74 bits/word, and for German 9.31 bits/word, which yields an after-neutralization entropy difference of 0.57 bits/word. Note that this difference is actually slightly *bigger* than the original difference of 0.56 bits/word. Neutralization of clitics and contractions thus yields a negative explained variance value, namely $\frac{0.56-0.57}{0.56} = -0.02$, i.e. –2%. This means the differences in unigram entropies increase rather than decrease by means of neutralizing clitics and contractions in English and German.

Looking at compounding, this effect is even more pronounced. Namely, the entropy difference before and after decompounding amounts to 0.64 bits/word. As for clitics/contractions, this is actually bigger than the original 0.56 difference and yields a variance explained value of $\frac{0.56-0.64}{0.56} = -0.14$. Thus, decompounding increases the unigram entropy difference by 14%. Again, this is due to systematic differences in how English and German use compounding to encode information.

5.2.6 Tone

Another dimension of information encoding is tone, that is, the systematic harnessing of pitch accents to distinguish words that would otherwise sound the same. This is a widespread strategy across languages of the world, particularly prevalent in languages of Mesoamerica, Sub-Saharan Africa, and Southeast Asia. Probably the most prominent examples are Mandarin (cmn) and Cantonese Chinese (yue), alongside other Sino-Tibetan languages. Consider the following example given in Yip (2002, p. 2):

[yau] in Cantonese

high level	'worry'
high rising	'paint (noun)'
mid level	'thin'
low level	'again'
very low level	'oil'
low rising	'have'

Here, the word *yau* can have six different meanings, depending on the pitch accent applied. In written language, tones are indicated according to marking systems which vary across areas of the world. In the following paragraph, the longer discussion of Yip (2002, p. 19-21) is briefly summarized.

In African linguistics, tones are typically indicated by diacritics. For example, a high tone on the vowel [a] is written as *á*, mid tone as *ā*, low tone as *à*, falling tone from high to low as *â*, and the inverse denoted as *ǎ*. In some cases, only high tones are marked, while the mid tone and low tone are left unmarked.

In the linguistic tradition of Asia, tones are indicated with diacritics in some Latin transliterations such as Pinyin. Importantly, this usage diverges considerably from the African system. Namely, high tone is indicated as *ā*, high rising as *á*, low falling as *à*, and low falling-rising as *ǎ*. In general, however, numerical tone marking is more widespread in this area of the world. Numerical tone markers are incorporated by so-called “Chao tone letters”, which are actually numbers, not letters. They normally consist of two or three digits indicating the start and end pitch of a syllable and a pitch change in the case of complex tone contours. The digits run from 1 (lowest tone) to 5 (highest tone). Thus, a constant high tone is indicated as a^{55} and low tone as a^{11}. Pitch contour examples include high rising a^{35}, and low falling a^{31}. Complex contours include low falling-rising a^{214}, and low rising-falling a^{231}.

Numerical tone marking is also the default for written languages of Mesoamerica, but again there are important differences. Firstly, tone numbers are reversed, with 1 indicating the highest tone, and 5 indicating the lowest tone. Secondly, steady pitch is commonly indicated with a single digit. Thirdly, while tone contours are also indicated with two digits, these are sometimes delimited by an additional hyphen as in high rising a^{3-2}. All three systems are summarized giving some of the most prominent tone examples below.

Tone	Africa	Asia	Mesoamerica
high	*á*	a^{55} (*ā*)	a^{1}
low	*à*	a^{11}	a^{5}
rising	*ǎ*	a^{35} (*á*)	a^{3-2} (a^{32})
falling	*â*	a^{31} (*à*)	a^{2-3} (a^{23})

The corpus samples used in this book include texts of all three macroareas and there is a considerable number of texts that represent tone languages. An important caveat is that tone is not always indicated in writing. Ultimately, this can mean that we underestimate the unigram entropies of tone languages, since we lack a dimension of information encoding in the written representation which is present in spoken language.

An exhaustive assessment of the information encoding potential of tone goes beyond the analyses presented in this book. However, to get an impression of the scale of tone marking strategies and their impact on unigram entropies, three languages are chosen to represent the relevant macroareas: Lingala (lin), a Bantu language of Sub-Saharan Africa; Hakka Chinese (hak), one of the Sinitic languages transliterated to Latin with tone indication; and Usila Chinantec (cuc), an Otomanguean language of Southern Mexico. For Lingala, the UDHR is the text basis, while for Hakka Chinese and Usila Chinantec this is the PBC. Example sentences are given below.

(25) Lingala (lin, UDHR 01)
Bato nyɔ́nsɔ na mbótama bazalí nsɔ́mí mpé bakókání na limɛmya mpé makokí .

(26) Hakka Chinese (hak, PBC 40001001)
Yâ-sû Kî-tuk he Thai-ví ke heu-thoi , Thai-ví he Â-pak-lâ-hón ke heu-thoi . Yâ-sû ke kâ-phú he án-ngiòng :

(27) Usila Chinantec (cuc, PBC 40001001)
I^{4}la^3 ti^{2}ton^3 la^{4}jang^{34} sa^{1}jeun^3 quian^1 Jesucristo a^{3}lang^{43} jon^{43}tyie^1 A^{3}brang^{23} jian^3 Da^{3}vei^{23} .

In the Lingala example, the only tone distinction is between stress (high tone indicated by acute accent) and no stress (no tone) (Divuilu, 2005, p. 40), while in the Pinyin transliteration of Hakka Chinese there are several diacritics for differing tones. In Usila Chinantec, a range of tone numerals are used instead.

With regards to unigram entropy, the crucial question is how many different word types are merged to a single word type when tonal marking is removed in these texts. For example, depending on different tone contours, the word type *chie* in Usila Chinantec can mean *chie*^{23} ‘sth. goes’, *chie*^{34} ‘sth. will go’, or *chie*^{32} ‘illness’, and chie^{43} ‘straight’ (Skinner and Skinner, 2000, p. xvi). While in the first two cases falling tone contours reflect different grammatical tenses of the verb ‘to go’ , in the latter two cases rising tone contours distinguish lexical meanings of a noun and an adjective. If tone marking is removed from a Usila Chinantecan text, these distinctions are lost, and we are left with *chie*, ambiguously representing any of these meanings.

To assess the impact that tone marking has on unigram entropies, all diacritics in Lingala and Hakka Chinese, as well as digits marking tone in Usila Chinantec are removed. We can then calculate the entropy share $\hat{S}_{\text{orig}\rightarrow\text{no tone}}$ and explained variance $\widehat{EV}_{\text{orig}\rightarrow\text{no tone}}$ of tone markings.[11] The results are given in Figure 5.8.

Lingala has the lowest entropy share of tones with a mere 0.4%, while in Hakka Chinese this amounts to 1.2%, and in Usila Chinantec to 4.7%. The latter is comparable to the entropy share of inflections in English (4.3 to 4.5% and 5.4% respectively). That is, from a unigram perspective, tone markings encode about as much information in Usila Chinantec as inflections encode in English. For the

[11] file: Rcode/Chapter5/entropyTones.R

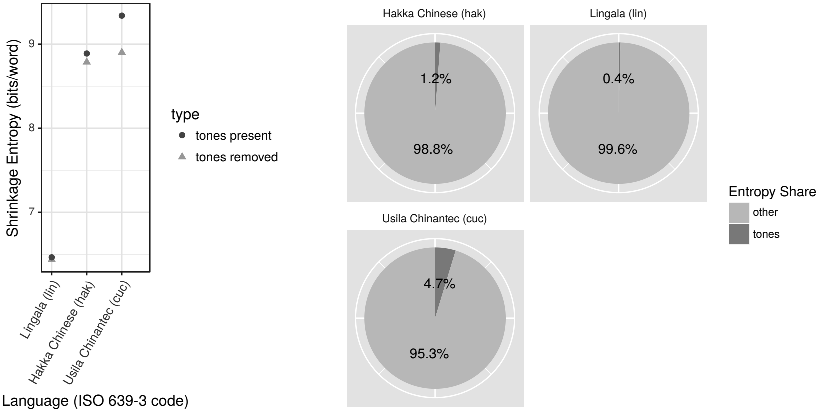

Figure 5.8: Unigram entropy differences and shares of tone markings. The left panel gives shrinkage entropies before and after removal of tone markings in Lingala (lin), Hakka Chinese (hak), and Usila Chinantec (cuc). The panels on the right illustrate the corresponding entropy shares.

other two languages, the information encoding potential of tones is hence comparatively small.

Furthermore, the variance in original unigram entropy values across these three languages is $\mathrm{Var}(\hat{H}_{\mathrm{orig}}) = 2.39$ bits/word. After removal of tone marking it is $\mathrm{Var}(\hat{H}_{\mathrm{orig}}) = 1.94$. The variance explained by tone differences is thus $\widehat{EV}_{\mathrm{orig}\rightarrow\mathrm{no\ tone}} = \frac{2.39-1.94}{2.39} = 0.19$. Hence, 19% of the unigram entropy difference in Lingala, Hakka Chinese, and Usila Chinantec is explained by varying strategies of how productively tones are employed.

Summary: word-formation and tone entropies

In the previous sections, different word-formation patterns and tone marking strategies were tested for their impact on unigram entropies in individual languages, as well as their potential to explain variance in unigram entropies across different languages. Three sets of parallel texts were used: a) 19 languages of the PBC and UDHR, b) English and German texts of the OSC, UDHR, EPC, and the Book of Genesis, and finally c) the UDHR in Lingala as well as the PBC in Hakka Chinese and Usila Chinantec. In the first experiment, texts were automatically neutralized for inflections only (i.e. lemmatized). In the second, this was done manually for inflections, derivations, clitics/contractions and compounds. In the third, texts were neutralized for tone marking.

The experiment involving lemmatization of 19 languages illustrated that entropy shares of inflections range from around 4% (English) to around 15% (Finish). Moreover, variance in unigram entropy values is explainable to 55% by differences in inflectional strategies. It emerges that inflectional marking is the most important dimension for explaining unigram entropy differences in parallel texts across languages. Of course, a sample of 19 languages of three different families (Indo-European, Atlantic-Congo, and Uralic) and with a strong bias towards overrepresenting European languages, cannot be seen as a balanced representation of languages across the world. Only the development of cross-linguistic corpora and computational tools will enable more representative typological analyses across more variegated samples.

In the experiment with English and German corpora, the entropy share of inflections is highest (5.4% and 7.9% respectively), while derivation, clitics/contractions and compounds play less of a role. The exact ordering of entropy shares of word-formation patterns varies even between these two closely related languages. This might reflect subtle differences in the productivities of word-formation patterns more generally. Again, this will have to be tested with bigger and more representative corpora to be conclusive. Manual neutralization yields an explained unigram entropy variance of 48% for inflections (similar to the 55% across 19 languages) and 27% for derivations. Moreover, neutralization of clitics and contractions *increases* the entropy difference by 2%, and for compounds by 14%, and thus actually yields negative explained variances. However, with regards to the question what drives differences in entropy values across different languages, the sign of the percentage-wise change is only secondary, with the magnitude of the effect on unigram entropy variance being more important. Based on the analyses in this section, we can conclude that inflections have the strongest impact on entropy variation (55% variance explained across 19 languages and 48% of difference explained for English and German), followed by derivations (27% difference explained for German and English), compounds (14%), and clitics and contractions (2%).

Another productive strategy to encode information at the world level – instead or in conjunction with changes in morphological material – is to apply differing pitch accents to syllables and words, i.e. tone. The potential of tones to encode information at the unigram level was measured using texts from Lingala, Hakka Chinese and Usila Chinantec. While tone marking has a relatively small entropy share for Lingala and Hakka Chinese (around 1%), in Usila Chinantec it reaches roughly the level of entropy share covered by inflectional marking in English, i.e. around 5%. This is an important result to keep in mind for later analyses, especially since not all texts indicate tonal differences by either diacritics or tone numbers. The unigram entropy variance explained by tone marking across the three languages analysed here is 19%. This is in the same ballpark of the variance explained by derivational marking between English and German (27%).

The question remains of how to explain the “residue” of entropy share for individual languages and the cross-linguistic variation in unigram entropies. For example, if we apply neutralization of both inflections and derivations to German and English texts, we will reduce the entropy difference by around 75%. What about the missing 25% of unexplained difference? As outlined above, neutralization of compounds and clitics will not help to further reduce the difference. Lexical distinctions in the basic vocabulary are a much more likely candidate. For example, the deity in the PBC is variously called *god*, *lord* or *father* in English. This might contrast with other languages which consistently use a single expression.

Finally, it should not be forgotten that the overarching assumption of constant content in parallel texts only holds approximately. There are still differences in *what* exactly is encoded in any given translation, and *what not*. This is in some cases related to rules of discourse conventionalized in a given language. For instance, in Example (23) the Iñupiatun affix *-ni* is a marker of evidentiality, i.e. indicating that somebody is reporting something about Jesus, rather than it being first hand knowledge. The fact that the stories in the Bible are told by a narrator is explicitly communicated here. In many other languages, this will be rather implicit information to be inferred by the reader. In other cases, mentioning or excluding certain information can be a stylistic choice of the translator. In yet other cases, certain verses, and parts of such, might not be available in every given translation to be compared.

Unfortunately, differences relating to the basic vocabulary, to explicit and implicit encoding, and to individual stylistic choices are harder to systematically quantify and evaluate. This is why using parallel texts is advisable in the first place.

5.3 Register and style

The last factor to be discussed as potential language “internal” effect influencing unigram entropies is variation connected to register and style – at the level of whole corpora. Texts of different registers and styles are known to exhibit systematic variation in the range of vocabulary, morphological marking and syntactic structures used. For example, Baayen (1994, 2008) demonstrates that derivational suffixes such as the Romance suffix *-ity*, and the Germanic suffix *-ness* reveal different degrees of productivity according to stylistic factors and different text types. Along similar lines, historical letters (Säily and Suomela, 2009) and Present Day

English texts (Säily, 2011) are shown to exhibit significant differences in the productivity of *-ity* according to whether these were written by women or men.

Given such findings, it is conceivable that different registers and styles are directly or indirectly linked to unigram entropy differences, especially via differences in the vocabulary and the productivity of inflectional and/or derivational morphology. The main set of corpora used for unigram entropy estimation throughout this book represents three different registers and styles: The UDHR is a legal document, the PBC religious writing, and the EPC consists of written speeches and discussions. Since unigram entropies are centred and scaled per corpus (Section 4.4), the mean values are the same across the three corpora, namely zero (see Figure 5.9).[12]

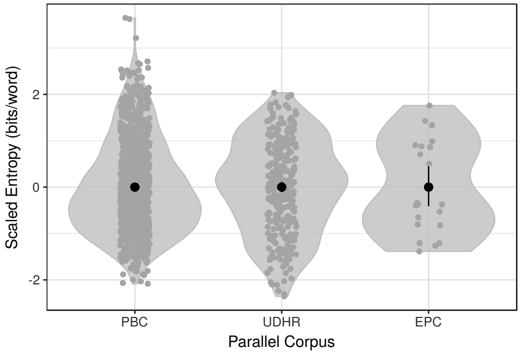

Figure 5.9: Violin plots of scaled unigram entropies per corpus. Scaled unigram entropies (y-axis) are categorized by corpus, i.e. PBC (Parallel Bible Corpus), UDHR (Universal Declaration of Human Rights), and EPC (European Parallel Corpus). Black dots indicate mean values with confidence intervals. Light grey violins outline symmetric density distributions of entropic values. Individual data points are plotted in grey, with jitter added for better visibility.

However, the density distributions still differ somewhat, with the PBC displaying the highest density below zero, i.e. a skew towards lower unigram entropy, while the UDHR displays a slight skew towards higher unigram entropy, and the EPC displays a bimodal density distribution with the lowest density around the mean value. Such differences in densities can derive from variation in the sample of families and geographic areas that a corpus represents. We will get back to these in the next chapter.

[12] file: Rcode/Chapter5/entropyCorpora.R

It is vital to remember that the overall aim is to compare languages – represented by texts – according to their lexical diversities, i.e. unigram entropy values. We want to be able to say that language A has a higher/lower unigram entropy value than language B. In the optimal case, such a ranking is independent of the register and style used in a specific corpus. However, the skews in unigram entropy densities per corpus suggest that, even for centred and scaled corpora, there might still be a slight bias.

Given this state of affairs, it is important to clarify how much our choice of corpus changes the relative ranking of languages according to unigram entropies. For the same set of languages the EPC might, for instance, give us a different ranking of values than the PBC or the UDHR. One way of getting an impression of the consistency of ranking is to correlate unigram entropies for languages represented in any two of the corpora. Namely, a Spearman rank correlation of $r = 1$ would indicate that the ranking is fully consistent, and hence that register difference – and in fact any other systematic difference arising from specific properties of the corpora – does not have an impact on the ranking of values. A Spearman rank correlation close to $r = 0$, on the other hand, would indicate that the two corpora give us completely different rankings, and hence register difference has a strong impact. Figure 5.10 visualizes the correlations for every possible pair of parallel corpora.

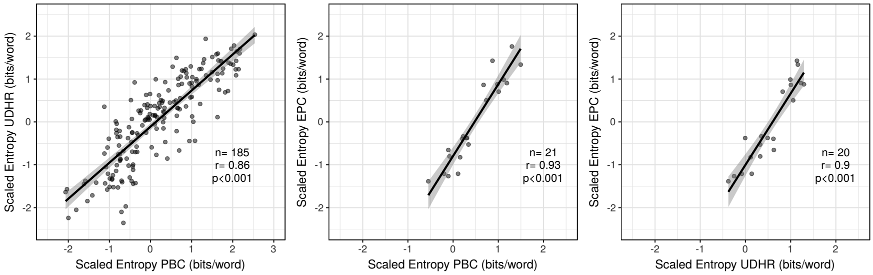

Figure 5.10: Correlations of unigram entropy values for corpus pairs. The panels compare (from left to right) unigram entropies for 185 languages that are available in both the PBC and the UDHR, 21 languages overlapping in the PBC and EPC, and 20 languages overlapping in the UDHR and EPC. Linear models are given as black lines with 95% confidence intervals (grey).

All three correlations are strong, reflected in Spearman coefficients of $r = 0.86$, $r = 0.93$, and $r = 0.9$ respectively. This goes to show that there is generally strong “agreement” between different corpora on which languages have high or low unigram entropies. Though the unigram entropy value of a given language might differ depending on the corpus we use to represent it, the ranking of this language in comparison to other languages is likely to be consistent as long as the register is the same or similar across the languages compared. Take English and German as examples again. The highest scaled unigram entropy for English is found in the UDHR with −0.38 (compared to the EPC with –1.39 and the PBC with −0.55). This value is even slightly higher than the German EPC value of −0.39. However, the value for the German UDHR is 0.62, and thus considerably higher than for English.

In sum, as long as we use texts of same parallel corpus to rank languages according to unigram entropies, the ranking is likely to be consistent with what we would find for other parallel corpora.

5.4 Summary

This chapter dealt with descriptive – language "internal" – factors shaping word frequency distributions and hence unigram entropies of different texts and languages. Namely, variance due to writing systems and scripts, word-formation patterns, tone marking, as well as registers and styles was investigated and the exact impact on unigram entropies quantified and discussed.

As a first observation, it turns out that writing systems, and the scripts derived from them, can have an impact on unigram entropy values. However, for the nine different scripts tested here, this effect is restricted to a maximum percentage-wise change of 3.2%, as exemplified by Korean written in Latin versus Hankul. For all other scripts this effect is exceedingly small (below 1%), and in some cases nonexistent.

Word-formation patterns, on the other hand, emerge as strong descriptive predictors of lexical diversities in parallel texts. Namely, neutralization of inflectional markers in a sample of 19 languages is shown to reduce the entropy variance by 55%. In a similar vein, manual neutralization of inflections, derivations, compounds and clitics/contractions in English and German amount to changes in the unigram entropy variance by 48%, 27%, 14% and 2% respectively. This leaves around 10-25% of unexplained variance for other factors such as the lexicon. In a typologically very different set of languages, the effect of tone marking on unigram entropy values was evaluated. Removal of tone markers reduces the variance in unigram entropies by 19%. Thus, mainly inflection and derivation, and to a smaller extent also compounding, cliticization/contraction and tone emerge as descriptive factors that account for the biggest part of differences in unigram entropies across parallel texts and the languages represented by them.

Moreover, the consistency of unigram entropy rankings was tested by Spearman rank correlations between pairs of corpora. These analyses demonstrate that though different registers can be associated with higher or lower entropy values, they do not strongly interact with the cross-linguistic ranking of languages. Conceptually, registers and styles are not independent of the other descriptive predictors. Registers and styles will have an impact on lexical diversities only *via* the specific choice of lexicon, productivity of word-formation patterns, or tonal distinctions associated with them. For example, there is very likely a unigram entropy difference between spoken and written language due to different ranges of vocabulary, and the complexity of morphological structures used. It is difficult to imagine how register and style by themselves might have an impact on lexical diversity. In the domain of descriptive factors, scripts, word-formation, and tone have a more basic status than register and style. Namely, the former mediate the effect of the latter.

As a general restriction, all these effects were tested based on small subsamples of the original 1833 texts and 1217 languages. Most of these languages were standard European languages. Clearly, the importance of particular descriptive factors might change for other subsamples of languages. Therefore, building typologically balanced cross-linguistic corpora, and developing automated tools to process them, are two points high up on the agenda towards developing a corpus-based, quantitative and reproducible language typology.

In conclusion, using information on scripts, word-formation, and tone marking, we can “explain” a good part of the variation in unigram entropies across languages. However, at the face of it, explanations with reference to the “internal” structural properties of languages just constitute further descriptions of diversity from a different angle. As such, they help to disentangle the pathways and causes of change, but are arguably not *causal* explanations by themselves. We might ask: why do parallel corpora in Iñupiatun have high unigram entropy and in Hawaiian low unigram entropy? An answer referring to the fact that Iñupiatun is a polysynthetic language and Hawaiian is an analytic language is not fully getting to the core of the matter. It merely translates an information-theoretic observation into a structural observation using linguistic terminology. That is, it further elaborates *what* is different. To be clear, such links between quantitative measures and linguistic analyses are valuable by themselves. They constitute the foundational work for a corpus- and usage-based language typology, and hence a viable object of study for a coherent research project. Nevertheless, it is argued throughout this book that descriptive factors are not explaining *why* there is a difference. A satisfying causal explanation has to go beyond descriptive analyses of the set of linguistic interactions $\mathcal{L}(t)$, and link them with factors pertaining to the population of speakers $\mathcal{S}(t)$ and their learning and usage preferences.

6 Explanatory Factors: Language “External” Effects

The previous chapter shed some light on *descriptive*, language “internal” factors, and how they relate to lexical diversities in different languages. This is an important first step towards understanding the mechanisms at play when lexical diversities change. For example, it was demonstrated that productive inflectional morphology systematically increases the number of word types, and spreads token frequencies more uniformly across them, which yields higher unigram entropies. However, as pointed out in Section 3.3, this is not enough information by itself to constitute a causal theory of language change and evolution. Instead, *explanatory* factors need to be established, namely, links between lexical diversities and the structure of speaker and learner populations. Eventually, this will elicit the pressures of language learning and usage that drive the evolution of particular encoding strategies.

6.1 Population size

A series of qualitative and quantitative studies in the past 20 years have embarked on illustrating a link between the structural characteristics of languages and the size of the populations they are spoken by. Answers are sought at different levels of linguistic structure, from phoneme inventories to morphosyntax. Some of the relevant studies are discussed in the following.

Phoneme inventories

Starting with Hay and Bauer (2007) it is argued that population sizes are positively correlated with phoneme inventory sizes, meaning that languages spoken by *bigger* populations are predicted to have *more diverse* phoneme inventories. Notably, the correlation is shown to hold for both consonants and vowels, and across different language families. Hay and Bauer (2007) base their claim on a sample of 216 languages of 42 language families, of which a majority is of the widespread and well-documented type represented, for instance, by English, Hindi and Mandarin.

Expanding the language sample to 504 languages of 50 families, Atkinson (2011) replicates the overall result of a link between bigger populations and more diverse phoneme inventories. Besides population size, his statistical model also includes the “distance from origin”, i.e. the distance from Africa, as a predictor.

DOI 10.1515/9783110560107-006

This also turns out to be significant. The explanation given for this result is that language populations spread from Africa and left a trace in form of a so-called “serial founder” or “bottleneck” effect, such that higher distance from the origin predicts lower levels of diversity. In this perspective, a series of population bottlenecks leads to a gradual reduction of diversity as sub-populations split from respective founder populations. Thus, reduction of phoneme diversity is seen in parallel to the well-known reduction of genetic diversity from Africa towards Asia, and into the Americas.

The statistical claims of Hay and Bauer (2007) and Atkinson (2011) are (largely) confirmed by Jaeger et al. (2011), though with a grain of salt regarding the exact statistical models to be fitted, and their reliability given sparse data. The population size effect on phoneme inventory size remains positive in all their models. However, while it is significant in a model that takes grouping at the level of family, subfamily, and genus into account, it ceases to be significant in models that also take countries and continents as grouping factors into account – as a proxy for “language contact”. Along similar lines, another replication study based on 969 languages (Moran et al., 2012) calls the validity of Hay and Bauer (2007)’s findings into question. It argues that there is considerable between-family and between-genus variation that is not taken into account in the original model. Considering these grouping factors, the overall effect of population size on phoneme inventory size is argued to be relatively small, though still statistically significant.

Finally, in the most extensive study to date – comprising more than 3153 languages of 109 families – Wichmann et al. (2011) again replicate the link between population size and phoneme inventory size, including controls for grouping at the level of language families. In addition, they take into account another potential confound: the link between phoneme inventory size and average word length. Languages with shorter words might evolve bigger phoneme inventories to counter-balance the loss of information encoding potential – a hypothesis brought forward by Nettle (1998, 1995). Confirming this link, Wichmann et al. (2011, p. 21) argue that the correlation between population size and phoneme inventory size could be mediated by the effect that larger populations (probably through pressures of language learning) reduce word stems to regular and shorter canonical forms. Shorter stems, in turn, would require more phonemes to disambiguate the otherwise increasing number of homonyms.

Whatever the exact explanation for the positive correlation between population and phoneme inventory size, the statistical analyses so far indicate that the overall association holds. There is, however, considerable variation within and between different families and geographic areas, which further complicates the search for a coherent explanation.

Morphological complexity

An idea akin to the population size and phoneme diversity hypothesis is that population size negatively correlates with the complexity of inflectional morphology. Lupyan and Dale (2010) quantitatively demonstrate this link. See also Nettle (2012) for a recent overview on the topic. In a nutshell, languages spoken by bigger populations tend to be those with reduced morphological marking. Overall, Lupyan and Dale (2010)'s sample covers more than 2000 languages for which linguistic information is available in 28 WALS chapters. These cover diverse topics such as "inflectional morphology", "number of cases", "person marking on verbs", "coding of evidentiality", "coding of negation", and "distance distinctions in demonstratives". For 23 of the 28 categories, population size (logarithmically transformed) turns out to be a significant predictor of lower complexity, i.e. less differentiated marking strategies. This result is interpreted as reflecting learning pressures associated with bigger populations. The linking hypothesis is that bigger populations are by trend those that also have "recruited" more adult L2 speakers in the past (Lupyan and Dale, 2010; Dale and Lupyan, 2012; Lupyan and Dale, 2015). This is further discussed below.

Rates of change

Before large scale empirical data and statistical models were employed to link population size and linguistic structure, computational modelling studies laid the theoretical groundwork. For instance, Nettle (1999) proposed population size as a predictor for the rate of language change. Simply speaking, smaller groups of speakers should be more transparent for change and hence perpetuate innovations more quickly. This line of reasoning predicts that small languages are expected a) to exhibit higher rates of lexical change, b) to have more borrowed items in their lexicon,[1] but also c) to be more likely to preserve marked grammatical structures (e.g. unusual word orders) (Nettle, 1999, p. 134). It might be argued that the last prediction was confirmed – though rather indirectly – by Lupyan and Dale (2010). Namely, if complex and opaque morphology is considered a "marked" structure, then it holds true that it tends to be preserved in smaller populations. The argument in a), on the other hand, could not be verified in more direct assessments based on large-scale data (Wichmann and Holman, 2009; Wichmann et al., 2008).

[1] Though borrowing of loanwords has to be linked with language contact, which in turn can increase the overall population size.

In fact, a recent study by Bromham et al. (2015) explicitly falsifies the assumption under a). By means of phylogenetic modelling of changes in cognate lists for 20 Polynesian languages, they show that bigger populations tend to gain new words and smaller populations are prone to lose words. They conclude that smaller populations in their sample do not have higher rates in terms of uptake of innovation. Interestingly, Bromham et al. (2015) also find no evidence that bigger populations have higher rates of word gain due to borrowing of words from other populations. Rather, there seems to be a genuinely higher potential of word creation in the bigger populations analysed.

Population size and lexical diversity

The aforementioned computational and statistical analyses suggest that population size is a potential predictor of lexical diversity. To test this, the scaled unigram entropy estimations for 1217 languages are merged with population size data from one of the last openly available versions of the Ethnologue (Lewis et al., 2013). This yields a sample of 1395 texts and 1193 languages for which both unigram entropies and population sizes per language are available.

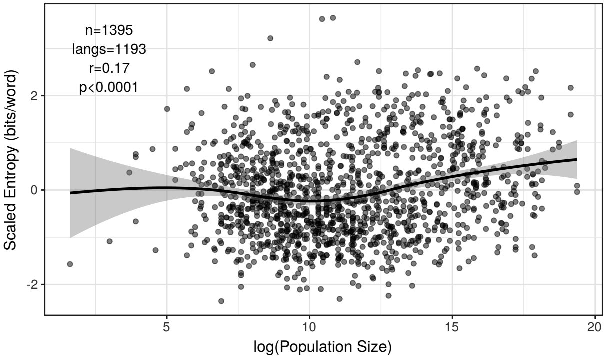

Figure 6.1: Population size and scaled unigram entropy. Logarithmically transformed population sizes (L1 speakers) are given on the x-axis. Scaled unigram entropies are given on the y-axis. The number of texts, languages, as well as the Pearson correlation coefficient and p-value are given in the panel. A local regression smoother with confidence intervals (grey) is overlaid.

Note that the 17th version of the Ethnologue generally gives population sizes as the number of native (L1) speakers. If a language is spoken in one country only, the number of speakers is given for that country as “population”. If the language is spoken in several countries, the total number of L1 speakers is given as "population total all countries".[2] If available, the population sizes used here are the total population sizes across countries. Otherwise, the population sizes of single countries are given. Only population sizes bigger than zero are taken into account, meaning that extinct languages are not included. Figure 6.1 plots these population sizes versus scaled unigram entropy values.[3]

There is a slightly positive trend, i.e. bigger populations are associated with higher lexical diversities. The Pearson coefficient for the correlation is small ($r = 0.17$), but significant ($p < 0.0001$). This simple correlation analysis suggests that, indeed, population size is a predictor of lexical diversity. However, this needs to be confirmed in more advanced statistical models, involving competing predictors as well as controls for genealogical and geographic proximity (Chapter 8).

6.2 Adult learner percentages

The idea that adult language learning shapes linguistic structure has been around in sociolinguistics for several decades. One of the core arguments in this context is that languages in contact are prone to lose abundant and fine-grained morphological distinctions (Thomason and Kaufman, 1988; Wray and Grace, 2007; McWhorter, 2007, 2011; Trudgill, 2011). Language "contact" always involves learning and usage of (at least) two languages, be it by infants and small children (i.e. bilingualism), or adults ("non-native" learning).

However, the difference between these "two types of contact" is argued to play a crucial role for the linguistic outcome (Trudgill, 2011, p. 15). Namely, morphological simplification is associated particularly with adult learning, while early bilingualism can even lead to borrowing of additional morphological material (see Chapter 10 for further discussion). Lupyan and Dale (2010) associated their analyses with this line of reasoning via the linking hypothesis that population sizes are a reflection language contact involving adults. Bentz and Winter (2012) and Bentz and Winter (2013) directly used L2 percentages in 69 languages to show a negative relationship between language contact and case marking. This is, languages with higher L2 percentages tend to have fewer or no nominal case markers. This effect is likely related to the difficulty adult learners have with case distinctions, irrespective of whether morphological case is used in their native language.

[2] http://www.ethnologue.com/17/about/language-info/#Population

[3] file: Rcode/Chapter6/entropyPopSize.R

Moreover, Bentz et al. (2015) extended this argument to L2 percentages and lexical diversity. The dataset of L2 speaker information in Bentz et al. (2015) is used here as well. It contains languages for which numbers of L1 and L2 speakers in the linguistic community were available at the time of collection. This information is found for 226 languages using the *SIL Ethnologue* (Lewis et al., 2013), the *Rosetta project website*,[4] the *UCLA Language Materials Project*,[5] and the *Encarta*.[6]

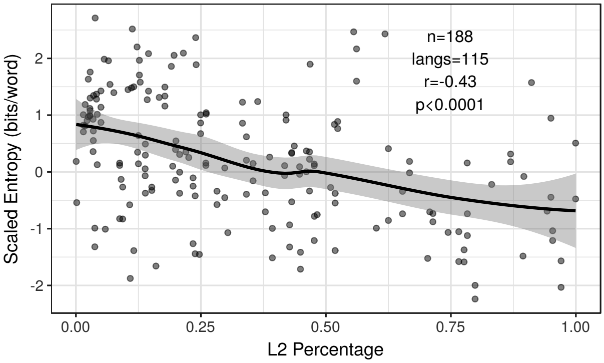

Figure 6.2: L2 speaker percentages and scaled unigram entropies. L2 speaker percentages are given on the x-axis. Scaled entropies are given on the y-axis. The number of texts, languages, as well as the Pearson correlation coefficient and p-value are given in the panel. A local regression smoother with confidence intervals (grey) is overlaid.

Whenever L1 and L2 speaker numbers differ in the sources, the average is calculated as an estimate. This smooths some of the incommensurable values that are certainly to be found in sources like Ethnologue. Note that Sanskrit and Esperanto are excluded from the sample. Sanskrit is an extreme outlier in the Indo-European family. In the database of Bentz et al. (2015), it is listed with a very high ratio of L2 to L1 speakers. This is due to the fact that it is learned and used almost exclusively as the language of liturgy in Hinduism. In this sense, there are very few native speakers of Sanskrit, while many students learn it in schools as L2. Arguably, this is not the kind of L2 learning and usage scenario that is supposed to reduce morphological complexity. Esperanto, on the other hand, is an artificial language with a high ratio of L2 speakers. However, since it is a constructed language, there is no point to be made about potential shaping of its linguistic structure due to

[4] www.rosettaproject.org

[5] www.lmp.ucla.edu

[6] http://en.wikipedia.org/wiki/Encarta

natural processes of language change (though such processes might have been at play in its very recent history).

Based on the remaining averaged speaker numbers, we can calculate the L2 speaker percentages for each of the 226 languages. Merging this L2 speaker dataset with unigram entropies yields a sample of 188 texts and 115 languages. Figure 6.2 illustrates the relationship between L2 speaker percentages and entropies.[7]

There is a clear negative trend, meaning that bigger L2 percentages are associated with lower unigram entropies. This is reflected in a medium Pearson correlation coefficient ($r =$ -0.43) of high significance ($p <$ 0.0001). Note that this sample of languages is much smaller than the sample with population size information *by a factor of 10*. The correlation between population size and unigram entropies for this smaller sample is a minor $r =$ 0.02, and not significant ($p =$ 0.78). This illustrates that the correlation with L2 percentages is much stronger than the correlation with pure population size – at least for this sample of 115 languages. Again, though, these results have to be corroborated in more elaborate models.

6.3 Language status

Besides information on population sizes and L2 speakers, the Ethnologue also categorizes languages according to their "status", defined as the level of "intergenerational transmission" of a language. The Ethnologue adheres to the so-called *Expanded Graded Intergenerational Disruption Scale* (EGIDS), which is a discrete scale of 13 levels from "international language" to "extinct language".[8] The individual levels are given in Table 6.1.

These levels provide an approximation of the probability that intergenerational transmission of a language is successful. As such, they also indirectly reflect the social prestige associated with a language. For example, a national language is more likely to be associated with high intergenerational transmission and high social prestige, than a developing language or a language threatened by extinction. Using the Ethnologue, we arrive at a sample of 1397 texts and 1197 languages for which both unigram entropies and language status information is available. Figure 6.3 plots the language status levels versus entropies per language.[9]
Since language status is an ordinal variable that can be ordered numerically from highest to lowest (in this case 1 to 12), it is possible to use it as a continuous

[7] file: Rcode/Chapter6/entropyL2.R

[8] https://www.ethnologue.com/about/language-status

[9] file: Rcode/Chapter6/entropyLangStatus.R

Table 6.1: Information on language status from the Ethnologue (17th version).

Level	Label	Description
0	International	The language is widely used between nations in trade, knowledge exchange, and international policy.
1	National	The language is used in education, work, mass media, and government at the national level.
2	Provincial	The language is used in education, work, mass media, and government within major administrative subdivisions of a nation.
3	Wider Communication	The language is used in work and mass media without official status to transcend language differences across a region.
4	Educational	The language is in vigorous use, with standardization and literature being sustained through a widespread system of institutionally supported education.
5	Developing	The language is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable.
6a	Vigorous	The language is used for face-to-face communication by all generations and the situation is sustainable.
6b	Threatened	The language is used for face-to-face communication within all generations, but it is losing users.
7	Shifting	The child-bearing generation can use the language among themselves, but it is not being transmitted to children.
8a	Moribund	The only remaining active users of the language are members of the grandparent generation and older.
8b	Nearly Extinct	The only remaining users of the language are members of the grandparent generation or older who have little opportunity to use the language.
9	Dormant	The language serves as a reminder of heritage identity for an ethnic community, but no one has more than symbolic proficiency.
10	Extinct	The language is no longer used and no one retains a sense of ethnic identity associated with the language.

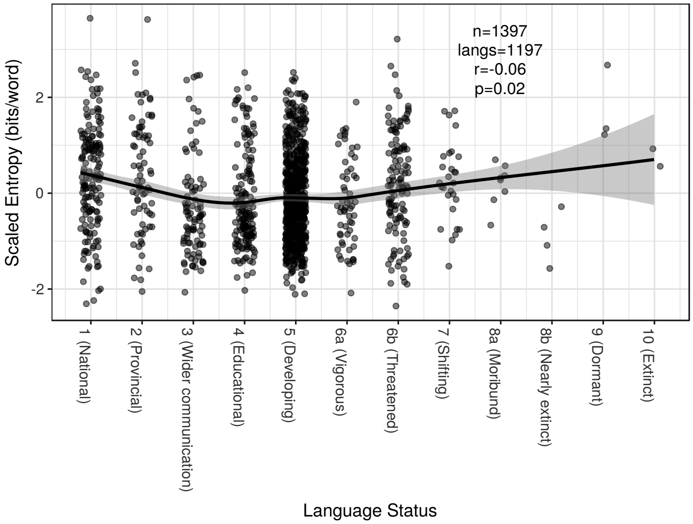

Figure 6.3: Language status and unigram entropies. The discrete status of a language (x-axis) plotted against the unigram entropy per language (y-axis). Language status is assessed on a discrete scale from “1 (National)” to “12 (Extinct)”. Note that though the EGIDS names a level “0 (International)”, this level is not assigned to any language in the sample. For better visibility, points are jittered around the discrete categories on the x-axis. The number of texts, languages, as well as the Pearson correlation coefficient and p-value are given in the panel. A local regression smoother with confidence intervals (grey) is overlaid.

variable in a Pearson correlation. There is an overall slightly negative correlation ($r = -0.09$, $p = 0.02$), meaning that languages of higher status have slightly higher entropies than languages of lower status (note that higher status languages have lower numbers).

6.4 Summary

Three explanatory, language “external” factors were discussed here and preliminarily tested for their association with lexical diversity. This includes population size, L2 speaker percentages, and language status. All three are significant predictors of lexical diversities by themselves, i.e. using simple Pearson correlations. However, it should be kept in mind that the effect sizes for population size and language status are small, and statistical significance is likely due to the usage of relatively extensive samples.

Having said that, population size and language status display a positive association with scaled unigram entropies (remember that scale for language status is inverted), meaning that bigger populations, and languages of higher status are – by trend – associated with higher lexical diversities. For L2 speaker percentages, the association runs the other way around: higher relative numbers of L2 speakers are associated with lower lexical diversities.

However, before we can further interpret these results, we need to combine the individual predictors in a single statistical model to account for potential non-independence, i.e. multicollinearity. Also, a further source of non-independence that needs to be taken into account is the grouping at the family and area level.

7 Grouping Factors: Language Families and Areas

In the previous two chapters, descriptive and explanatory factors of variation in lexical diversity were discussed. Remember from Section 3.3 that a third kind of factor is also relevant: genealogical and geographic *grouping*. Due to phenomena such as population drift and contact, languages spoken by different communities can cluster together in groups at different levels. At the phylogenetic or genealogical level, we talk about language *families* (or *stocks*) (e.g. Indo-European) and language *genera* (e.g. Romance). At the level of geography, we talk about language *areas* or language *regions* (e.g. Europe, Greater Mesopotamia, Mesoamerica, etc.).

Crucially, due to these different levels of grouping, most languages and the measurements taken from them, are not *independent* data points in a statistical sense. It has been pointed out in several quantitative typological studies (Dryer, 1989; Bickel, 2013; Jaeger et al., 2011; Cysouw, 2010; Moran et al., 2012) that systematic variation at the level of language families and language areas has to be taken into account in statistical models. Only this will allow us to extrapolate our findings beyond the subsample of languages we are currently looking at. Systematic variation in lexical diversity per family and area is discussed in turn.

7.1 Unigram entropy by family

Language families and genera are the outcome of proto-languages splitting into separate branches and drifting apart at different rates over hundreds and thousands of years. Despite the manifold pathways of divergence, a common root can sometimes be reconstructed by means of analysing regular sound changes and/or comparing phonological, lexical, morphological, and syntactic similarities. The existence of deep-rooted ancestral relationships, as in the case of, for instance, the Indo-European, Sino-Tibetan, Austronesian, and Afroasiatic language families are relatively uncontroversial.

An important repercussion is that languages belonging to a given family are not "blank slates" with regards to their structural properties (Bickel, 2013). Rather, they adhere (to some degree) to the principle of "identity by descent". For instance, Indo-European languages are known to have emanated from a proto-language that very likely had morphological markers for up to eight or nine different nominal cases. Hence, any of today's Indo-European languages had a "head start" in terms of developing and maintaining morphological case marking, as compared to, for example, Sino-Tibetan languages. Similarly, it is conceivable that lexical diversities are preserved over time and that languages of

DOI 10.1515/9783110560107-007

certain families are therefore higher up on the scale than others, due to descent from a proto-language with high lexical diversity. To test this, we can harness the *AUTOTYP* database (Nichols et al., 2013), which provides information on language "stocks", i.e. families for which there is linguistic evidence of common descent, as well as Glottolog (Hammarström et al., 2016), which gives "top-level" family information. Merging unigram entropies with AUTOTYP information yields a sample of 1049 texts and 731 languages grouped into 145 stocks. For Glottolog this yields 1398 texts and 1195 languages grouped into 105 families.[1] Figure 7.1 plots unigram entropies grouped by stocks and families.[2] Only those represented by more than five data points are included.[3]

It is apparent that language stocks and families differ widely with regards to average entropy values. In the AUTOTYP classification, these range from a mean value of $\mu = -1.28$ ($\sigma = 0.57$) for Adamawa-Ubangi (part of Atlantic-Congo in Glottolog) to $\mu = 2.16$ ($\sigma = 0.39$) for Dravidian languages. According to a Wilcoxon rank sum test, this difference in mean values is significant ($p < 0.001$).[4] In the Glottolog, the range is from a mean value of $\mu = -0.81$ ($\sigma = 0.69$) for Austroasiatic to $\mu = 1.71$ ($\sigma = 0.43$) for Quechuan languages, again with a significant difference in the means ($p < 0.001$). By means of phylogenetic signal analyses, Bentz et al. (2015) illustrate that such grouping structure at the stock and family level is likely due to the relative stability of lexical diversities over time. For Austronesian, Indo-European and Bantu languages it is shown that lexical diversities of the UDHR and PBC have a generally high phylogenetic signal. This means that lexical diversities cluster on family trees (built by using cognate data) as we expect assuming that they have evolved along the branches. This further confirms that phylogenetic grouping plays a role for understanding variation in lexical diversities.

Another pattern emerging from these plots is that language families which are generally associated with high morphological complexity, e.g. Dravidian, Quechua, Turkic, Uralic, etc. have high average unigram entropy values, while those generally associated with low morphological complexity, e.g. Adamawa-Ubangi, Austroasiatic, and Sino-Tibetan tend to rank among the lower unigram

[1] Note that this is 106 if "NA" is considered a group.

[2] file: Rcode/Chapter7/entropyFamilies.R

[3] Merging the unigram entropy file with the AUTOTYP and Glottolog databases by using ISO codes can lead to an increase in the number of data points. For example, for Basque the AUTOTYP lists different varieties under the same ISO code (eus), which will then result in multiple rows with the same unigram entropy and ISO code.

[4] Statistical significance is here only tested for the minimum and maximum mean values. Conceptually, it only matters whether *any* of the means differ, not whether *all* the means differ.

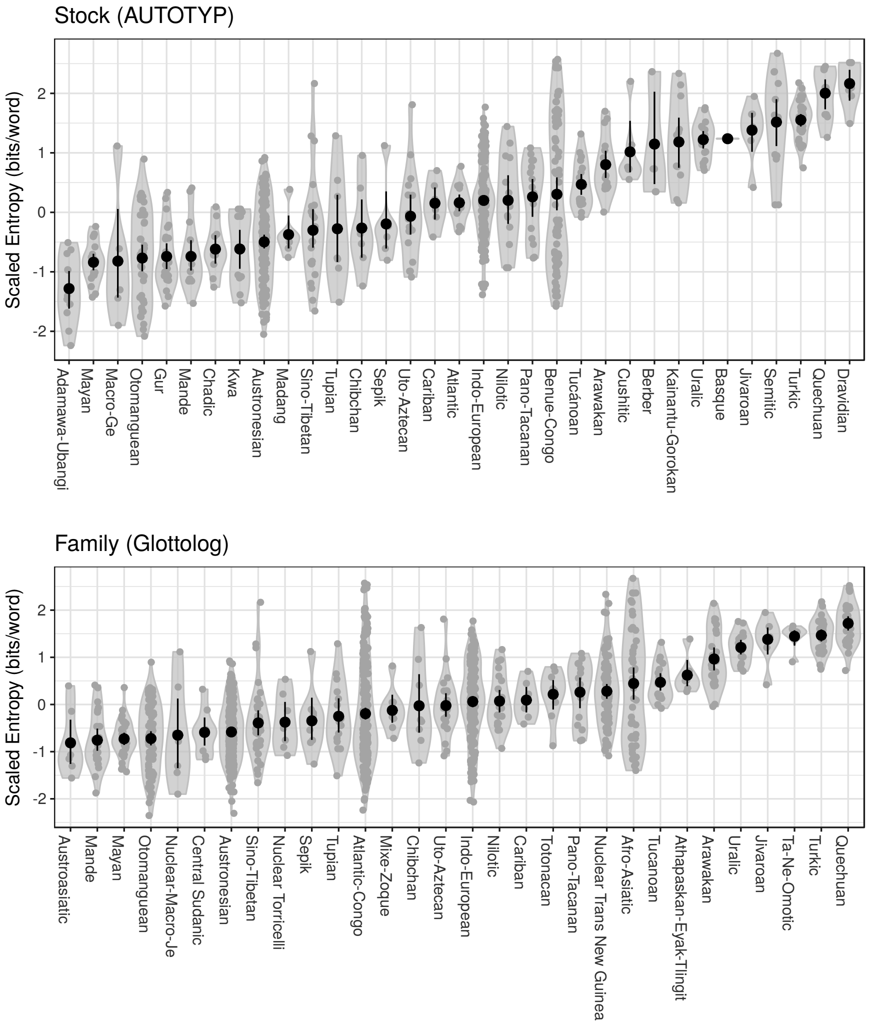

Figure 7.1: Unigram entropies grouped by language stocks and families. The x-axis gives AUTOTYP stock names (top) and Glottolog top-level family names (bottom). The y-axis gives scaled unigram entropy values. Mean values per stock and family are given as black dots with confidence intervals. Only families and stocks with more than five data points are plotted. Light grey violins outline symmetric density distributions of entropic values. Individual data points are plotted in grey, with jitter added for better visibility.

entropy families. This seems to confirm the analyses in Chapter 5, namely, that differences in inflectional productivity are strong drivers of unigram entropy variance. In fact, unigram entropies of parallel texts (among other corpus-based measures) are strongly correlated with a morphological complexity measure derived from the WALS (Bentz et al., 2016). An interesting caveat is that some of the lower unigram entropy families are also associated with tone marking, e.g. Mayan, Otomanguean, Austroasiatic, Sino-Tibetan, etc. From an information-theoretic point of view, this raises the possibility that tone marking is an alternative strategy to encode information, particularly prone to appear when inflectional marking is lost – or did not exist in the first place. Since tone marking is not necessarily represented in written language, this can mean that unigram entropies of tone languages are underestimated. However, remember that the entropy share of tone marking for the three languages analysed in Chapter 5 was only between 1% to 4%. So even if tone marking was always included in the texts, this would most likely not drastically increase family averages of unigram entropies.

Another interesting observation is that some families and stocks have a wide range (i.e. standard deviation) of unigram entropies, while for others this is rather narrow. For example, Atlantic-Congo languages cover almost the entire world wide spectrum from circa –2 to circa 2 scaled bits/word ($\sigma = 1.04$), while Uralic ($\sigma = 0.33$) and Turkic ($\sigma = 0.35$) are confined to a relatively narrow spectrum, and are hence more homogeneous with regards to unigram entropies. Thus, there seem to have been pressures at play in the history of families like Atlantic-Congo which drove individual languages to maximally diverge, while in the case of Turkic and Uralic the languages rather converged to a similar unigram entropy.

7.2 Unigram entropy by area

In parallel to language families, language areas are also associated with the presence or absence of linguistic features, and areal patterning has been invoked as a key to understanding linguistic diversity (Bickel, 2017; Nichols, 1992; Dryer, 1989). Looking at the feature of morphological case marking again, we find that it is not evenly distributed across areas of the world, but prevalent in some (e.g. Eurasia) and only marginal in others (e.g. Africa) (Bickel and Nichols, 2009). Bickel (2017) attributes such patterning, at least in part, to potentially “far-reaching” spreads of language populations in linguistic prehistory. As a consequence, the connectedness of even vast geographic regions can be reflected in the presence or absence of linguistic features.

To test whether such areal clusters affect lexical diversities, the same datasets as for language families and stocks are used. Besides information on stocks, AUTOTYP also includes information on 23 areas which are relevant for diffusion of linguistic features. Glottolog, on the other hand, categorizes languages into only six macroareas. Figure 7.2 plots unigram entropies by areas and macroareas – again only the ones represented by more than five data points.

Just as for families and stocks, there is systematic variation in mean unigram entropies by areas and macroareas as well. According to the AUTOTYP grouping, Oceania has the lowest average entropy ($\mu = -0.67$, $\sigma = 0.72$) and Eastern North America has the highest ($\mu = 1.73$, $\sigma = 1.7$). The Wilcoxon rank sum test for minimum and maximum means indicates a significant difference ($p < 0.0001$). In contrast, for Glottolog macroareas we find that North America has the lowest ($\mu = -0.38$, $\sigma = 0.86$) and South America the highest average ($\mu = 0.57$, $\sigma = 1.02$), with the difference being significant again ($p < 0.0001$). The apparent discrepancy between the rankings of North America derives from the fact that the rather coarse-grained macroareas of Glottolog collapse Mesoamerica and North America into a single category "North America", whereas the more fine-grained categorization of AUTOTYP keeps these separate. Note that Mesoamerica, with the Otomanguean and Mayan languages most prominently represented, has generally low unigram word entropies. This illustrates the impact of granularity when geographically grouping together individual languages and whole language families.

7.3 Global patterns of entropy

7.3.1 Latitude: the Low-Entropy-Belt

Interestingly, there seem to be geographical patterns of variation in unigram entropies even beyond the level of areas and macroareas, namely, on a global scale. Figure 7.3 plots 1398 texts of 1195 languages onto a world map by using latitudes and longitudes per language as provided in Glottolog 2.7.[5] It appears that languages around or just above the equator display systematically lower entropies than languages further up north or further down south. This pattern emerges since most areas around the equator, i.e. Oceania, African Savannah, Mesoamerica and Southeast Asia (and language families associated with them) have low average unigram entropies. This global pattern is here called the "Low-Entropy-Belt". It reflects the fact that unigram entropy is increasing away from the equator, i.e. towards higher and lower latitudes. Some preliminary statistical analyses are reported in Bentz (2016).

[5] file: Rcode/Chapter7/entropyWorldMap.R

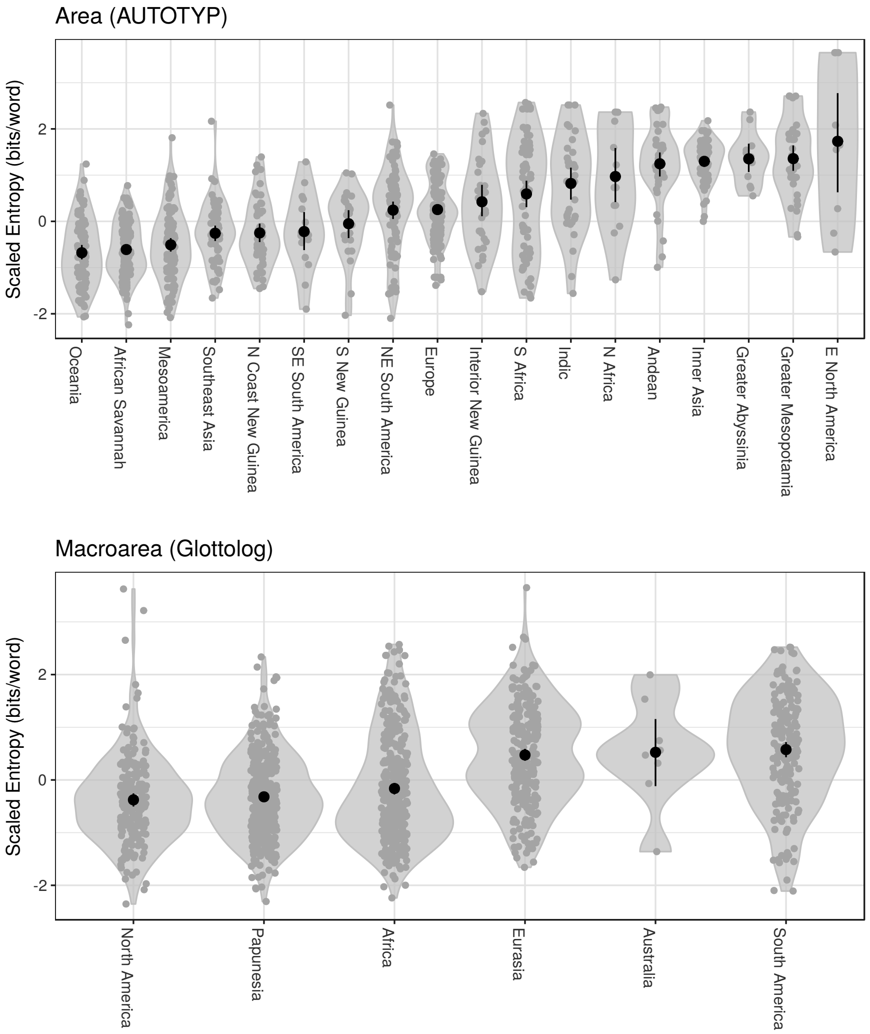

Figure 7.2: Unigram entropies grouped by language areas and macroareas. The x-axis gives AUTOTYP area names (top) and Glottolog macroarea names (bottom). Scaled unigram entropies are represented on the y-axis. Mean values per area and macroarea are given as black dots with confidence intervals. Light grey violins outline symmetric density distributions of entropic values. Individual data points are plotted in grey, with jitter added for better visibility.

A scenario explaining this global pattern is one in which – as a general trend – language populations around the equator have faced particular ecological conditions and, as a consequence, evolved subsistence strategies which are more prone to result in population and language contact. In fact, climatic factors such as mean temperature and precipitation emerge as some of the strongest predictors for lexical diversity and morphosyntactic complexity in a recent meta-study (Lewis and Frank, 2016). However, it is difficult to imagine direct causal links between climatic factors and lexical diversity. Rather, there might be a link between subsistence strategy and prolonged periods of language contact involving adult learning. Areas such as Mesoamerica, African Savannah, and Oceania seem predestined for such large-scale language contact scenarios, potentially even reaching back into human prehistory.

At this point, the Low-Entropy-Belt is a statistically significant pattern at a global scale, which, however, is driven by particular language families and areas having high or low average unigram entropy. In other words, the pattern holds *between* rather than *within* families and areas (Bentz, 2016). This makes sense given that very few families actually span latitudes from the equator to the latitudinal extremes. Further statistical analyses are necessary to disentangle the exact geographic dimensions – in conjunction with family and areal grouping – which give rise to this global pattern.

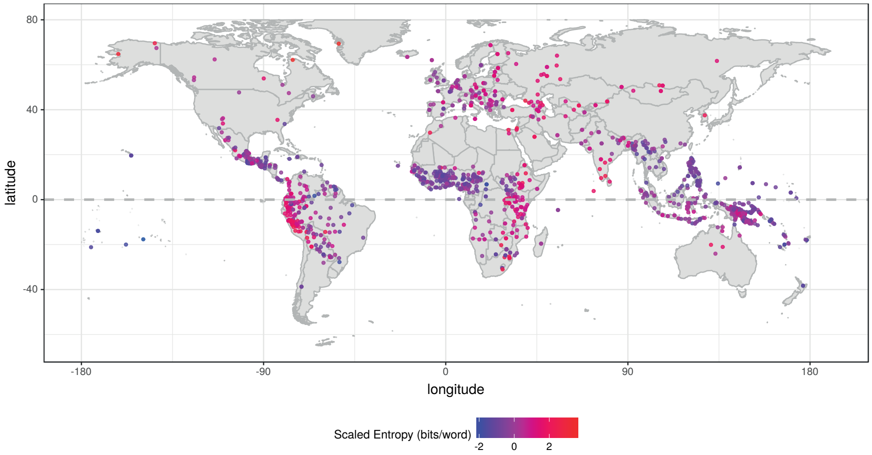

Figure 7.3: World map with unigram entropies. The equator is indicated as dashed grey line. Scaled unigram entropies are denoted by colour from low (blue) to red (high).

7.3.2 Altitude: “high” entropy

Another geographic dimension that plays a role for local and global areal effects is altitude. As argued in Nichols (2013, 2016), the “vertical archipelago” fuels the creation of so-called linguistic enclaves. Languages spoken in high, rugged mountain ranges are less likely to undergo intensive contact than languages in lowland areas – everything else being equal. In the case of Nakh-Daghestanian languages, for instance, communities living in the highlands are likely to learn additional languages in order to establish trade relations on lowland markets. This creates an asymmetrical contact scenario with influx of highland L2 speakers into lowland languages, but not the other way around. Nichols (2013) demonstrates that this contact scenario is associated with reduction of morphological opacity in lowland languages, while highland languages retain the original levels.

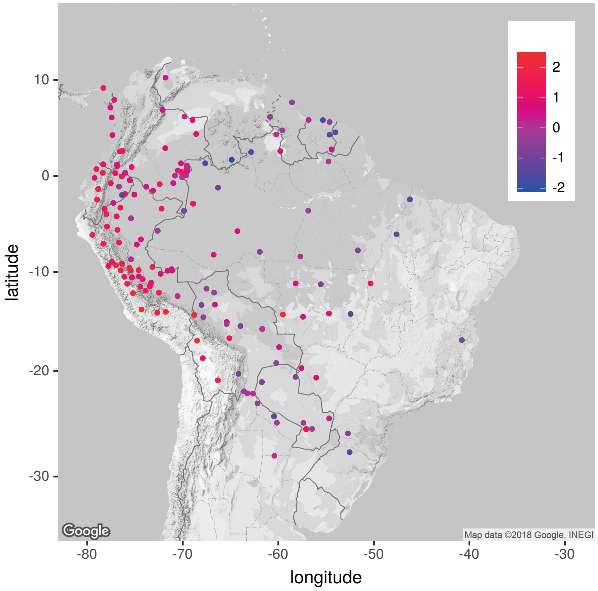

Figure 7.4: South American languages and unigram entropies. Subsample of 191 texts and 161 languages categorized as South American in Glottolog 2.7. Scaled unigram entropies are denoted by colour from low (blue) to red (high).

Though the exact sociolinguistic scenarios deriving from geographic isolation in high altitudes are likely to differ across languages of the world, the overall effect on the morphological complexity of languages seems to hold. Apart from the Caucasus area, this is also illustrated for South America, with the Andean region accommodating some of the highest morphological complexity languages on the continent (Nichols and Bentz, 2018). These findings are based on three measures relating to morphological opacity, inventory size of morphological markers, and unigram entropies of parallel texts. The same pattern also emerges in the unigram entropy sample used here. Figure 7.4 zooms into the world map and focuses on languages categorized as "South American" in Glottolog. In line with Nichols and Bentz (2018), languages located in, or close to, the Andean area tend to display particularly high unigram entropies, while languages located further away from the Andes display lower values. Note, however, that most of the Andean languages in Bolivia and Peru belong to either the Quechuan or Aymaran families. Again, further statistical analyses are needed to test whether the altitude effect also holds *within* these particular families as well as families and areas of the world more generally.

7.4 Summary

Both language families and areas are grouping factors strongly associated with variance in lexical diversities. Given the overall range of scaled entropy values from around –2.5 to 2.5, it is noteworthy that average values per family widely differ from around –1 (Adamawa-Ubangi, Austroasiatic) to around 2 (Dravidian, Quechuan). In parallel, average values per area differ from around −0.7 (Oceania) to 1.7 (Eastern North America). These minimum and maximum mean values differ significantly from a statistical point of view. This requires taking families and areas as grouping factors into account when statistically modelling predictors of lexical diversity.

Apart from families and areas established in databases like AUTOTYP and Glottolog, geographic factors more generally start to be considered as drivers of linguistic complexity and diversity. Baechler and Seiler (2016), for instance, is a collection of articles investigating language variation in the light of isolation. In future research, there might be more large-scale – and even global – geographic predictors of linguistic diversity emerging. Latitude and altitude were here briefly discussed as such.

8 Predicting Lexical Diversity: Statistical Models

The previous two chapters discussed *explanatory* and *grouping* factors relevant for the modelling and prediction of lexical diversity across languages of the world. In short, explanatory factors relate to population characteristics. Three of these are considered in the current account: population size, L2 speaker percentages, as well as the status of a language. Grouping factors, on the other hand, relate to geographic and genealogical clustering, most prominently at the level of language families and geographic areas. Another potential confound, which was discussed as language "internal" effect in Section 5.3, is variance between corpora of the same language, e.g. due to different registers and styles.

In the following, two statistical models are built to establish robust predictors of lexical diversity: 1) a *multiple regression model*, taking into account all three explanatory factors in a single model, and 2) a linear *mixed-effects model*, expanding the model structure to take into account potential variation due to grouping by families and areas, as well as corpus type. The first model helps to understand which explanatory factor is most important when overlap in variance explained with the other factors is given. The second model further establishes whether explanatory factors are still significant predictors after systematic differences between the grouping levels have been adjusted.

8.1 Multiple regression: combining explanatory factors

Linear regression is a method for predicting values of a given *dependent* variable by using values of a second variable – called *independent* or *predictor* variable – assuming a linear relationship between them. In our case, we want to predict the estimated scaled unigram entropy $\hat{H}^{\text{scaled}}$ of a parallel text (representing a language) given the value of the predictor variable. The linear regression model is specified as

$$\begin{aligned}\hat{H}_i^{\text{scaled}} &= \beta_0 + \beta_x x_i + \epsilon_i,\\ \epsilon_i &\sim N(0, \sigma_\epsilon^2),\end{aligned} \tag{8.1}$$

where each i^{th} unigram entropy value is predicted by the value of predictor variable x assuming a linear relationship with a slope of β_x (henceforth called coefficient), and the y-axis intercept of β_0. The prediction error ϵ (also called residual error) is assumed to be normally distributed with mean 0 and variance σ_ϵ^2. Multiple regression follows the same rationale, though integrating multiple predictor

DOI 10.1515/9783110560107-008

variables in a single model. Baayen (2008, p. 165-236) and Baayen (2014) further explains these with linguistic examples. Having several predictors in the same model gives an indication of whether all the predictors considered are actually necessary, in the sense of contributing independently to explaining the variance in a given dependent variable. Here, logarithmically transformed population size ($\log(x_i)$), adult speaker proportions (y_i), and language status (z_i) are the independent predictors. The model is specified as follows.

$$\begin{aligned}\hat{H}_i^{\text{scaled}} = \beta_0 + \beta_x \log(x_i) + \beta_y y_i + \beta_z z_i + \epsilon_i,\\ \epsilon_i \sim N(0, \sigma_\epsilon^2).\end{aligned} \tag{8.2}$$

Thus, the unigram entropy is predicted by the intercept β_0, plus the coefficients β_x, β_y, and β_z, multiplied by the respective values of the predictor variables. Again, the model assumes normality of the residuals ϵ_i. The crucial statistical question is whether a given coefficient is significantly different from zero, essentially meaning that the respective predictor significantly contributes to predicting the value of the dependent variable.

The sample used for multiple regression analysis contains 178 texts of 110 languages, stemming from 30 families (stocks) and 18 areas. It is constrained by the number of texts for which unigram entropies, information on all three predictors, as well as stock and area information from AUTOTYP is available. AUTOTYP is chosen here over Glottolog 2.7 as it has finer-grained information on linguistic areas. In this sample, the range of values of the predictor variables is reduced compared to the original sample. For example, only seven out of the originally 12 categories of language status are represented.

The model in Equation 8.2[1] is fitted to the empirical data using the package *lme4* (Bates et al., 2012) in *R* (R Core Team, 2013).[2] Multiple regression models are based on several preconditions: linearity, normality of the residuals, homoscedasticity, and absence of multicollinearity. Checks that these preconditions are met can be found in Appendix 13. The result of the multiple regression analysis is given in Table 8.1.

The estimated coefficient of L2 percentage is negative and highly significant ($p < 0.001$). This agrees with the negative Pearson correlation coefficient ($r = -0.43$, $p < 0.0001$) found in Chapter 5. This means that an increase of the predictor variable (L2%) by exactly one unit corresponds to an estimated decrease in the dependent variable by 1.7 units. In other words, if a language went from

[1] As well as further simpler models involving only subsets of the predictors.

[2] file: Rcode/Chapter8/entropyMultiReg.R

Table 8.1: Results of the multiple regression model.

Predictor	Coefficient	Estimate	SE	t-value	p-value
Intercept	β_0	-0.691692	0.680491	-1.016	0.6195
L2%	β_y	-1.698911	0.273932	-6.202	<0.001 ***
Status	β_z	0.009587	0.053025	0.181	0.8567
log(PopTotal)	β_x	0.094451	0.038878	2.429	0.0161 *
					R^2=0.2054

*** $p < 0.001$; ** $p < 0.01$; *$p < 0.05$

0% to 100% adult L2 speakers then we would predict a scaled unigram entropy decrease of 1.7. This corresponds to roughly 34% of the global unigram entropy scale from ca. –2.5 to 2.5. To get an impression of the effect size, consider the unigram entropy of the PBC in Turkish (tur: 2.08). This is predicted to change into a value close to Italian (ita: 0.34) if only L2 speakers were learning the language. Similarly, for English (eng:–0.55) unigram entropy is predicted to change into a value close to Tok Pisin (tpi: –2.08) given only L2 learners in the population.

The estimated coefficient of language status, on the other hand, is close to zero and not significant ($p = 0.86$). This is in disagreement with the significant negative Pearson correlation coefficient found earlier ($r = -0.06, p = 0.02$), and will be further discussed below.

Logged population size has a positive coefficient significantly different from zero ($p < 0.05$). This is again in agreement with the Pearson correlation coefficient found earlier ($r = 0.17$, $p < 0.0001$). The coefficient value of the multiple regression means that an increase in (logged) population size by one unit corresponds to an estimated increase in scaled unigram entropy by 0.09 units, i.e. 1.8% of the global entropy scale. As an example, assume a language population would go from 10,000 to 100,000, i.e. increase by a factor of ten, then the logged increase would be roughly 2 units, such that we expect an increase in scaled unigram entropy by 3.6% of the global scale. Arguably, this is a very small effect – though still statistically significant.

Finally, the R^2 value of the overall multiple regression model is 0.21, meaning that all the predictors together explain 21% of the variance in scaled unigram entropies. Further analyses of the data reveal that the non-significance of the coefficient for language status is probably due to the fact that language status is correlated with logged population size ($r = -0.54$, $p < 0.0001$). This is to be expected, as languages of higher status, e.g. national languages, have more speakers. As a consequence, population size explains variation in unigram entropies that overlaps with variation explained by language status and the latter becomes non-significant.

Interestingly, there is no correlation between population size and L2 percentage for the currently used sample of languages ($r = -0.07$, $p = 0.38$). This is somewhat surprising, as we might assume that bigger populations also have more L2 speakers. Given this lack of correlation between population size and L2 percentage, the results suggest that L2 percentage is a much stronger predictor for scaled unigram entropies than population sizes. However, it is important to keep in mind that the original sample of 1833 texts and 1217 languages was here reduced to 178 texts and 110 languages due to data sparsity. This means, in turn, that the results of the multiple regression are less generalizable than the simple Pearson correlation analyses.

8.2 Mixed-effects regression: controlling for non-independence

Due to the strong genealogical and geographical effects on lexical diversity, it is necessary to modify the multiple linear regression model by including information on language families and areas. It has been pointed out in several typological studies (Dryer, 1989; Jaeger et al., 2011; Bickel, 2013; Cysouw, 2010; Moran et al., 2012) that grouping at the phylogenetic and geographical level undermines the independence assumption of individual data points.

Moreover, scaled unigram entropy values are grouped according to corpora. This is because each data point corresponds to a specific text. Texts, in turn, can come from three different corpora (PBC, UDHR, EPC) and hence potentially exhibit systematic variation in lexical diversity according to register and style.

To take into account non-independence of data points, mixed-effects models – combining so-called fixed and random effects – have been suggested as a viable method (Bates et al., 2012; Baayen, 2008; Baayen et al., 2008; Barr et al., 2013; Bates et al., 2015; Baayen, 2014). A mixed-effects model is here applied to the unigram entropy data using the *lme4* package in *R* again. The sample is the same as for the multiple regression model, consisting of 178 texts, 110 languages, 30 stocks, and 18 areas. Since language status was not significant in the multiple regression model, only L2% and logged population size are considered as fixed effects. Additionally, both random slopes and random intercepts are considered for grouping factors. Models are built stepwise, starting with the random effects, adding in fixed effects when the optimal random effects structure is found.[3] The decision of including specific random intercepts/slopes and fixed effects is based

[3] file: Rcode/Chapter8/entropyMixedEffects.R

Table 8.2: Results of stepwise linear mixed-effects regression. Models that improve in terms of AIC and likelihood ratio tests are marked in bold face.

Fixed	Random		AIC	β_y	SE	t-value	p-value	R^2	
	intercept	slope						f[‡]	f+r
–	s[†]	–	468	–	0.23	0.6	–	0	0.7
–	**s,a**	**–**	**452**	**–**	**0.27**	**0.3**	**<0.001 *****	**0**	**0.7**
–	s,a,c	–	451	–	0.33	-0.27	0.06	0	0.73
–	s,a	$s_{L2\%}$	451	–	0.25	1.43	0.07	0	0.68
–	s,a	$a_{L2\%}$	453	–	0.26	-0.47	0.16	0	0.71
–	s,a	s_{logPop}	455	–	0.28	-0.17	0.16	0	0.72
–	s,a	a_{logPop}	453	–	0.24	1.18	0.24	0	0.68
logPop	s,a	–	453	0.03	0.03	1.13	0.27	0.004	0.68
L2%	**s,a**	**–**	**443**	**-0.92**	**0.25**	**-3.74**	**<0.001 *****	**0.04**	**0.67**

*** p<0.001; ** p<0.01; *p<0.05

[†] s: stock, a: area, c: corpus

[‡] f: fixed effect only, f+r: fixed and random effects

on the AIC as a criterion for model improvement (Baayen, 2014). If the AIC decreases by two points or more, then the new model with respective random and fixed effects is kept. P-values are calculated with likelihood ratio tests which assess whether the new model is significantly better than the preceding model in terms of goodness of fit. Stepwise model building is necessary, since the "maximal" model (Barr et al., 2013) with all possible fixed effects included as well as random intercepts and slopes does not converge. Stepwise model building and results are illustrated in Table 8.2.

The best model arrived at in terms of AIC (443, see last row of Table 8.2) includes only L2 percentage as fixed effect and random intercepts by stock and area. Random intercepts by corpus as well as random slopes by stock, area, and corpus (for both L2 percentage and population size) do not significantly decrease the AIC, reflected in p-values of likelihood ratio tests being bigger than 0.05. Inclusion of logged population size as a fixed effect does not improve the model once the best random effects structure is found. The specification of the optimal model is

$$\begin{aligned}
\hat{H}_i^{\text{scaled}} = \beta_0 + \beta_{0s} + \beta_{0a} + \beta_y y_i + \epsilon_i,\\
\epsilon_i \sim N(0, \sigma_\epsilon^2),\\
\beta_{0s} \sim N(0, \sigma_{\beta_{0s}}^2),\\
\beta_{0a} \sim N(0, \sigma_{\beta_{0a}}^2),\\
\epsilon_i \perp \beta_{0s}, \beta_{0a},
\end{aligned} \tag{8.3}$$

where β_{0s}, and β_{0a} are the random intercepts per stock and area. The fixed effect coefficient we are mainly interested in is β_y, i.e. the linear coefficient of L2 percentage predicting unigram entropy. Again, this coefficient should be significantly different from zero. The coefficients of the best model are estimated based on the *Restricted Maximum Likelihood* (REML) method. In parallel to the residual errors ϵ_i, the values of random intercepts are assumed to be distributed normally. Additionally, the condition in the last line states that the residual errors have to be orthogonal to the random effects, i.e. uncorrelated. The assumptions for this mixed-effects model are checked in Appendix 14.

The estimated coefficient associated with L2 percentage ($\hat{\beta}_y$) is negative in the optimal model (last row of Table 8.2). This is in line with the earlier finding of a negative Pearson correlation coefficient and a negative multiple regression coefficient. Thus, the overall trend holds: texts written in languages with higher L2 speaker percentages tend to have lower unigram entropies. The estimated coefficient is $\hat{\beta}_y = -0.92$. One unit increase in L2 percentage is associated with a 0.92 decrease in scaled unigram entropy. Going from 0% to 100% L2 speakers we expect a decrease in scaled entropy by around 20% – even *after* controlling for the fact that different families, areas, and corpora have higher or lower mean unigram entropy values.

Table 8.2 also gives R^2 values per model, i.e. the percentage of variance explained. This can be assessed for both the fixed effects only (f) and the fixed combined with random effects (f+r) using package *MuMin* in R (Bartoń, 2015). Generally, the variance explained is expected to increase when adding random intercepts and slopes to account for grouping structure in the data. Remember that the R^2 of the multiple regression model was 0.21, meaning that the fixed effects explained 21% of the variance in lexical diversities across texts. In the mixed-effects model, including just random intercepts by stock (simplest model in the first row of Table 8.2) already increases the variance explained to 70%. This confirms that stocks carry a lot of information about lexical diversity. In the final model, the variance explained by the fixed effect (L2 percentage) is 4%, and the overall variance explained – including random effects – is 67%. Hence, the best model arrived at here explains around two thirds of the variance in lexical diversities of 178 texts and 110 languages across the world. This leaves one third of variance to be explained by other predictors and/or random noise.

In the following sections, the results of the optimal mixed-effects model are further discussed and interpreted based on diagnostic plots by stocks, areas, and corpora.

8.2.1 Visualization by family

The main effect emerging from the statistical models is that L2 speaker percentages predict scaled unigram entropies of texts, even if systematic variation between stocks and areas is accounted for. A visual way of checking whether this negative association holds within groups is to facet a scatterplot by the respective grouping factor. For language stocks (henceforth referred to as families) this can be seen in Figure 8.1. In the sample used for mixed-effects modelling, 30 language families are found. However, only the ones represented by more than five texts are plotted here. This includes a category “Other” which summarizes all the remaining families.[4]

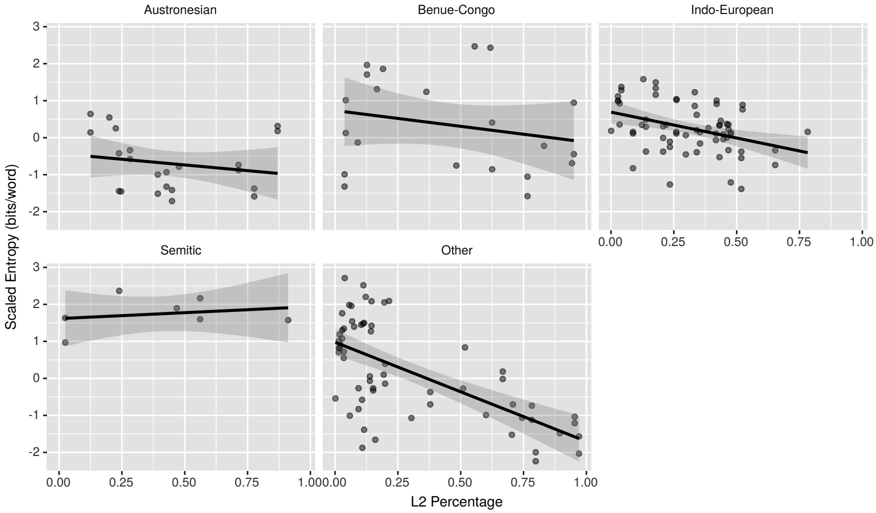

Figure 8.1: Unigram entropy and L2 speaker percentage by language family. Scatterplots of the relationship between L2 speaker percentages and scaled unigram entropies faceted by language family. Family names are given above the boxes. Linear regression models are fitted per family (black lines) with 95% confidence intervals (transparent grey).

These families differ in terms of intercepts. For example, the intercept with the y-axis is lower for Austronesian than for Benue-Congo and Semitic. This is the reason for random intercepts by family improving the mixed-effects models sig-

[4] Note that this is a convention for plotting. In the mixed-effects models languages are strictly grouped by family.

nificantly. Notably, the negative slope holds for four of the five family groups (including “Other”). A counter-example are Semitic languages, for which the slope of a linear model is rather positive. However, Semitic has only six members, and just about makes the criterion of having more than five members. For such underrepresented groups random variation will play a bigger role, and hence variation in slopes is less meaningful (see also Jaeger et al., 2011). Semitic languages might be subgrouped under the top-level family or “quasi-stock” of Afroasiatic, rather than being construed as a separate stock or family (Nichols, 1997).

In fact, only three families are well represented in the sample: Austronesian, Benue-Congo and Indo-European. All three of these display the negative association between L2% and unigram entropy This gives us some confidence that the effect does not only hold *between* but also *within* language families. This is a visual confirmation of the mixed-effects result that random slopes by family do not play a role for model improvement. However, this way of viewing the data also illustrates a limitation, namely, that only few families are well sampled. To make stronger claims about a universal trend, we need more data for more families.

8.2.2 Visualization by area

A similar scatterplot for language areas is given in Figure 8.2. There are 10 areas (including “Other”) with more than five members. Here, the texts are distributed somewhat more evenly across the groups, with African Savannah, Europe, Indic, Oceania, and South Africa having reasonable sample sizes (bigger than 15). Again, the negative trend holds in the majority of areas, namely, for all except Greater Abyssinia (East Africa) and Southeast Asia. This is the reason why random slopes per area do not improve the model. Random intercepts, on the other hand, improve the model, since they systematically differ between areas.

8.2.3 Visualization by corpus

There are only three corpora and they are hence all well represented in terms of numbers of texts, though the EPC is much smaller than the other two by a factor of more than ten. The negative trend clearly holds across all three, while the slope is steeper for the EPC than for the PBC and the UDHR. Note that the negative relationship between L2 percentage and unigram entropy is particularly strong for “Standard European” languages, as is evidenced by the relatively steep negative slopes for the Indo-European family and the area “Europe”. Since the EPC consists of texts written in these languages, it is not surprising that it exhibits a stronger

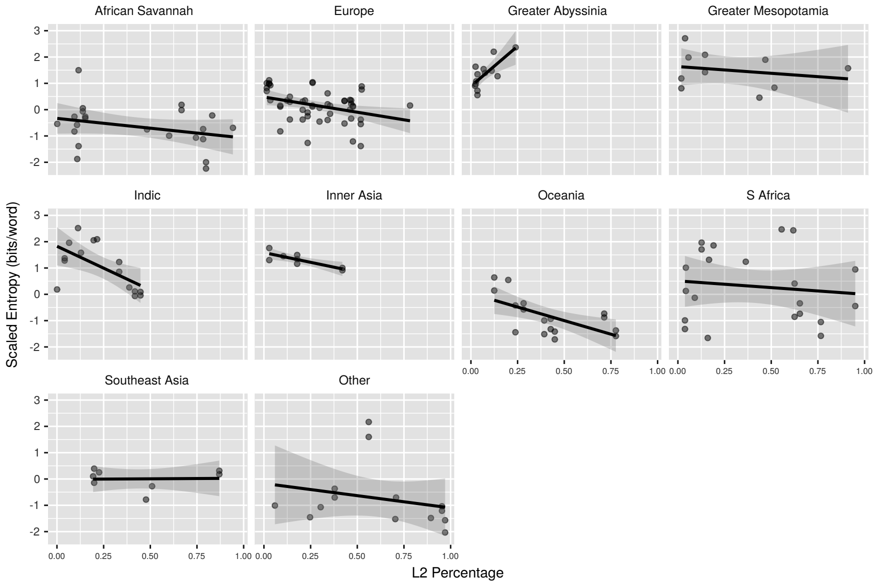

Figure 8.2: Unigram entropy and L2 speaker percentage by language area. Area names are given above the boxes. Linear regression models are fitted per family (black lines) with 95% confidence intervals (transparent grey).

negative slope than the other two corpora, which sample from a much wider range of languages. The model improvement by adding random y-intercepts by corpus is minor ($\Delta\text{AIC} < 2$) compared to the improvement when random intercepts by area ($\Delta\text{AIC} = 16$) are introduced.

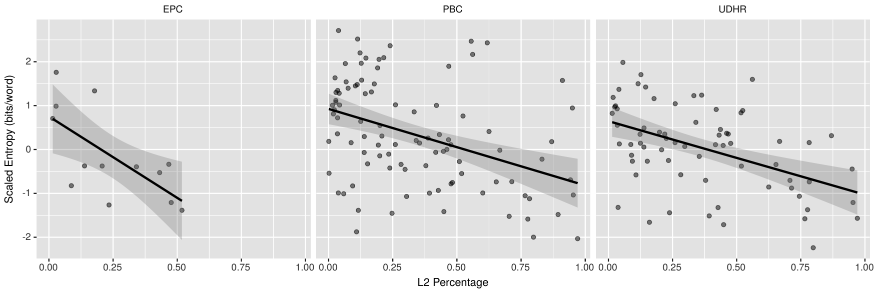

Figure 8.3: Unigram entropies and L2 speaker percentage by corpus. Corpus abbreviations are given above the panels. Linear regression models are fitted per family (black lines) with 95% confidence intervals (transparent grey).

8.3 Summary

To sum up the results of the multiple regression analyses: we can state that – taken explanatory factors on their own – both L2 percentage and population size emerge as significant predictors of lexical diversity. As an approximation, increasing L2 speaker percentages from 0% to 100% yields an estimated *decrease* of unigram entropy by 34% on the globally covered scale. This is considerable effect, roughly corresponding to the scaled unigram decrease between Turkish and Italian, or English and Tok Pisin. For logged population size, on the other hand, there is an estimated *increase* in unigram entropy by around 3.6%, which is statistically significant though small in comparison. Language status is not significant when combined with the other two predictors in a multiple regression. It is correlated with population size, thus "losing out" in explaining variance.

In mixed-effects regression analyses, when grouping factors are introduced by means of random intercepts – and thus adjusting for idiosyncrasies of families and areas – population size also drops out as a significant predictor. In the best model arrived at by stepwise model building, L2 percentage still has a significant negative coefficient predicting unigram entropy. The effect size is similar to the multiple regression model. We predict a 20% decrease in unigram entropy as L2% goes from 0% to 100%. Furthermore, L2 percentage explains 4% of variance in unigram entropies, and the overall model 67%.

The variance explained by even the simplest model, with only random intercepts by stock, is already at 70%. Note that this is expected just by virtue of families (and areas) having vastly differing mean unigram entropies, as elaborated in Chapter 7. In the end, does this mean that population size and L2 percentage are uninteresting with regards to explaining lexical diversity, since families and areas cover most of the variance already? To answer this question, we need to take a step back and consider the conceptual underpinnings of the languages as complex adaptive systems framework again.

9 Explaining Diversity: Multiple Factors Interacting

The conceptual framework laid out in Chapter 3 views languages as the accumulation of linguistic interactions at a specific point in time, i.e. $\mathcal{L}(t)$. These interactions, in turn, are performed by a speaker population $\mathcal{S}(t)$ according to their network competences $\mathcal{NC}(t)$. Linguistic inquiry often focuses either on language in this usage-based and "externalized" sense or language in a universal competence sense. In contrast, the CAS model does not necessarily focus on either of these. Rather, the *co-evolution* of linguistic interactions and learning preferences is the subject of study.

From this perspective, it was pointed out in Section 3.3 that there are three conceptually different levels of explanation: *descriptive* factors, *explanatory* factors, and *grouping* factors. We might ask: why do Finnish and Hungarian have higher lexical diversity than English and Dutch? There are three ways of addressing this question.

Firstly, a *descriptive* answer highlights the link between information-theoretic and linguistic concepts. In Section 5.2.4, a tight link between unigram entropy and productivity of inflectional marking was established. Languages using more inflectional marking, such as Hungarian or Finnish, have higher unigram entropy than languages using less inflectional marking, such as English or Dutch – ceteris paribus. This is not so much an answer as a new perspective on the original problem. It translates the information-theoretic concept of unigram entropy into the linguistic concept of inflectional marking and constitutes a first step towards understanding the linguistic phenomena in the light of standard information theory.

Secondly, an answer with reference to *grouping factors* could be: we know that Uralic languages have higher average unigram entropy than Indo-European languages. Hence, Hungarian and Finnish are expected to have higher unigram entropies than English and Dutch. This circumvents the original question by opening the synchronic problem to a diachronic, phylogenetic dimension. There are two interesting problems arising from this. First, we can further ask *why* Uralic languages have systematically higher entropy than Indo-European languages. Of course, this can go on *ad infinitum*, just pushing the answer to deeper levels of grouping. Second, there is no a priori reason to assume that lexical diversity has been stable between Proto-Uralic, Proto-Indo-European and the languages descending from them. In fact, in the case of Romance languages descending from Latin lexical diversity has systematically decreased across the board (Bentz and Berdicevskis, 2016).

DOI 10.1515/9783110560107-009

Phylogenetic relatedness by itself is only a sufficient answer under the assumption of so-called neutral drift, that is, purely random variance in frequencies of forms leading to the structural divergence of languages. As discussed in Chapter 3, only recently have the methods to model such processes for natural language become available (Blythe, 2012; Yanovich, 2016; Newberry et al., 2017; Kauhanen, 2017). More research is needed to understand exactly how much of linguistic diversity can be explained by neutral drift and how much of it is related to selection and adaptation to the properties of speaker/signer populations.

Thirdly, *explanatory factors*, as conceptualized here, relate to the population of language users and, ultimately, to their network competences, which reflect particular scenarios of learning and usage. For example, if we could conclusively show that adult language learning has played a more prominent role in the history of English and Dutch than in the history of Finnish and Hungarian, then learning constraints might emerge as part of the reason for lower unigram entropy in English and Dutch. Arguably, this is also "just" a translation of an information-theoretic problem into a sociolinguistic and psycholinguistic problem. Crucially, however, this translation promises to break the circularity of invoking an explanation at the same level (i.e. descriptive to descriptive, or grouping to grouping). Diachronically speaking, if we can establish a link between changes in $\mathscr{L}(t)$ to changes in $\mathscr{S}(t)$, then we are getting closer to a causal explanation.

This is not to say that studies on descriptive factors and grouping factors are per se less interesting, or less important than studies on explanatory factors. They are just conceptually different. Ultimately, a coherent theory of language change and evolution will elicit how these three levels interact.

9.1 Explanation and grouping

Against this backdrop we can reinterpret the statistical results of the previous chapter. It was established that both population size and L2 speaker percentages are significant predictors of lexical diversity when taken on their own. However, when idiosyncrasies of specific families, areas, and corpora are accounted for, population size ceases to be significant, while L2 percentage stays significant. The best mixed-effects model in terms of model fit explained 67% of variance in lexical diversity, mainly due to strong effects of random intercepts by language family and area. In comparison, the explanatory factor L2 percentage explained only 4% of the variance. Does this mean that, after all, the explanatory factors are negligible?

The answer – in the context of the CAS model supported here – should be *no*. Dismissing the effect of L2 percentage and even population size, on the basis of family and area variation would be misleading. Namely, this would favour grouping factors as an explanation over sociolinguistic and psycholinguistic factors. However, the grouping of lexical diversity on a phylogenetic and geographical level is in need of explanation itself. Is it the outcome of shallow and deep historic phenomena relating to population drift and isolation? This could explain *where* and *when* we expect to find certain linguistic features, but not *why* (Bickel, 2015). The question *why* has to be answered with reference to either neutral drift models or explanatory factors reflecting causal theories of learning and usage – or both.

Instead of replacing explanatory factors with grouping factors in our interpretation of statistical models, we should consider the interaction between them as informative. Explanatory factors might also give us a window into geographic and genealogical grouping. Pressures by adult language learning, for instance, might not only constitute a factor *alongside* variation between families and areas, but actually drive the geographic and phylogenetic grouping in the first place. In this view, differences *between* and *within* groups are not just unwanted variation that we need control for in order to interpret the main effect, but are potentially informative as to how the main effect caused the grouping in the first place. This line of reasoning is intimately related to the phenomenon called *Simpson's paradox*, also known as *Galton's problem* in evolutionary biology.

9.1.1 Simpson's paradox in language typology

Simpson's paradox translated to typology states that a specific effect can hold *between* different groups (e.g. families or areas), while *within* groups the effect might not play a role, or even show an inverted pattern (Jaeger et al., 2011; Moran et al., 2012). In the following, this is illustrated with generated data. Three separate scenarios are considered: a) there is a between-group effect, but no within-group effect, b) there is a within-group effect, but no between-group effect, and c) there is both a within-group and a between-group effect.[1]

Scenario A: between-group effect only

For five independent groups x (predictor) and y (dependent) variables are generated by drawing 100 numbers randomly from a Gaussian normal distribution with differing means. In this scenario, there is no correlation between x and y values within the five groups. However, due to systematic differences in mean x

[1] Rcode/Chapter9/simpsonsParadox.R

and y values per group there is a strong negative correlation between groups (see Figure 9.1 left panel).

Fitting a linear regression model through all the data points (black dashed line) yields a negative coefficient (slope) that is strongly significant ($\hat{\beta}_{\text{lm}} = -0.89$, $p < 0.001$). However, if we fit a linear mixed-effects model with random intercepts per group to the data, then the coefficient is not significant anymore ($\hat{\beta}_{\text{lme}} = -0.01$, $p = 0.82$). Due to the generated structure of the data, there is a strong between-group effect, but no within-group effect at all. This is reflected in the ceasing significance of the slope coefficient when random intercepts are introduced.

In a nutshell, the mixed-effects model takes into account non-independence of measurements by virtue of adjusting the respective intercept of each group. This can be seen by looking at the so-called Best Linear Unbiased Predictors (BLUPs) of the mixed-effects model. In the case of group A this is 3.96, meaning that the intercept of this group is adjusted downwards by circa 4 points on the y-axis. In comparison, the BLUP for group E is –3.96, meaning that this group is adjusted upwards by circa 4 points. The same is done for the other groups. This way all groups end up at intercepts of around 6 and the negative coefficient across all data points ceases to be significant. For mathematical details and further examples see also Baayen et al. (2008).

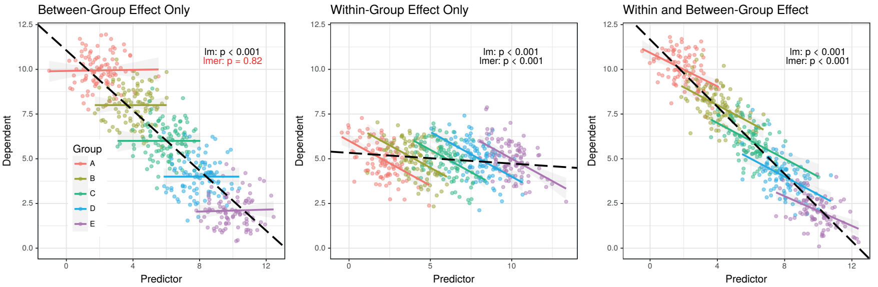

Figure 9.1: Illustration of Simpson's paradox. Simulated data for a predictor (x-axis) and a dependent variable (y-axis) in three different scenarios. Groups are indicated by different colours. Linear model lines are plotted across all data points (black dashed) and for each group separately (coloured lines).

Scenario B: within-group effect only

The simulation procedure is here the same as for Scenario A, except that there is now a correlation of -0.5 seeded into the x and y values within each group, while the mean y-values for all groups are the same (i.e. $\mu = 5$, see Figure 9.1 middle panel). Fitting a linear regression model through all data points yields a small but significant negative coefficient ($\hat{\beta}_{\text{lm}} = -0.06$, $p < 0.001$). In this case, the coefficient is still significant after random intercepts are introduced in a mixed-effects model ($\hat{\beta}_{\text{lme}} = -0.49$, $p < 0.001$). In fact, the estimated coefficient is now very close to the correlation coefficient originally seeded into the groups.

Scenario C: within-group and between-group effect

In the last scenario, there is a correlation of -0.5 seeded into the x and y values within each group *and* the mean y-values for all groups are different (see Figure 9.1 right panel). As a result, the linear regression model yields a significant negative coefficient ($\hat{\beta}_{\text{lm}} = -0.94$, $p < 0.001$) and the coefficient is again significant and close to the seeded correlation coefficient in the mixed-effects model ($\hat{\beta}_{\text{lme}} = -0.51, p < 0.001$).

Imagine Scenario A underlies our actual data with groups being families. If we went with an interpretation solely based on significance of the fixed effect in the mixed-effects model, then we would conclude that there is no interesting pattern to be observed, since the negative association does not hold within families. However, clearly, the fact that there is systematic between-family variation, i.e. that family means of the dependent variable are negatively correlated with group means of the predictor variable, is still in need of explanation. As a consequence, changes in the significance of the fixed effect due to adjustments by random intercepts and slopes, at least when applied to typological data, do *not* imply that the main effect is not interesting anymore, or not "in need of explanation". In fact, as long as there is no conclusive alternative hypothesis as to *why* the means by group differ, it is valid to consider the fixed effect examined in the model as a reason for the grouping in the first place (see also Jaeger et al., 2011, p. 296).

To go back to the empirical data: Indo-European languages have a higher mean L2 percentage ($\mu_{\text{IE}} = 0.31$) than Uralic languages ($\mu_{\text{Uralic}} = 0.02$)[2] and lower unigram entropy ($\mu_{\text{IE}} = 0.26$ versus $\mu_{\text{Uralic}} = 1.12$). So there is systematic *between*-group variation. If we had only these two families in our sample and we adjusted for differences in mean unigram entropy, then the overall effect might cease to be significant. Still, it is viable to conjecture that the systematic between-group differences are related to Indo-European languages on average experiencing more pressure from adult language learning, while Uralic languages might have a history of native language transmission. Whether or not these effects also show up within a given family is driven by a multitude of further factors.

[2] Though this sample only contains Hungarian and Finnish.

For example, language change phenomena often have an S-shaped, non-linear character (Blythe and Croft, 2012). This seems to be reflected also in change allegedly caused by adult learning pressure (Bentz and Winter, 2013, p. 12). It is conceivable that changes in L2 percentage are not effective until a certain threshold is exceeded. For example, Bentz and Winter (2013) illustrate that there is a sudden drop in the likelihood of having a morphological case-marking system once languages pass the threshold of 50% L2 speakers. Note that Uralic languages in our sample are far below this threshold, with 2% L2 speakers on average, whereas Indo-European languages (31%) are closer to it. Some Indo-European languages like Afrikaans (65%) and English (52%) exceed the threshold, while others like Polish (8%) and Romanian (13%) clearly stay below. As a consequence, there is also more deviation in L2 percentage within Indo-European languages ($\sigma = 0.18$) than between the two Uralic languages Hungarian and Finnish ($\sigma = 0.01$). Given these differences, it makes sense that the likelihood to observe a within-family effect is higher in Indo-European than in Uralic.

Overall, this suggests that a lack of within-group variation does not diminish the necessity to explain between-group effects. The same mechanism of change might have been at play within groups, but to different degrees and potentially also at different times in the history of a language family or area. This directly links to another important caveat: the problem of *time depth*.

9.2 Lexical diversity through time

The parallel texts and the information on language populations used here represent *synchronic* data. Parallel texts such as the UDHR, PBC and EPC reflect languages as they are *now* and resources such as the Ethnologue only give recent numbers of L1 and L2 speaker populations. Hence, the above results of multiple and mixed-effects regressions are a cross-section of *diachronic* processes. A viable question is *if* and *how* these synchronic results extrapolate back in time.

Modern versions of glottochronology are a quantitative way of estimating time depths of splits in language family trees. There are different flavours depending on the data (e.g. Swadesh lists, cognate judgements) and the methods used (e.g. Bayesian, Maximum likelihood, etc.). All of these accounts are controversial within the traditional comparative community, mainly because some of the simplifying assumptions underlying automated methods, for example, constancy of change rates, are at odds with observations from actual language history. Also, for the Indo-European language family it has been shown that different sets of cognate data and different tree priors in Bayesian analyses give diverging results for the earliest splits. These range from circa 8000-9500 BP (Gray and Atkinson, 2003; Bouckaert et al., 2012) to circa 6500-5500 BP (Chang et al., 2015; Rama, 2016). Thus, time depths based on the currently available automated methods have to be taken as crude approximations.

With this caveat in mind, some glottochronological studies are briefly reviewed here, particularly the ones relevant to the question how stable language populations are over time. The comparative data needed for such estimations is available via databases such as the *Automated Similarity Judgement Program* (ASJP, Wichmann et al. 2013). This project has collected Swadesh lists of 40 to 100 lexical items per language for more than 4000 languages. Given lists of base vocabulary, *edit distances* between languages can be calculated. The edit distance (per word) is the minimum number of changes necessary to turn a word in one language into the conceptually equivalent word in another language. This metric is also referred to as *Levenshtein distance*. Further modifications to account for different word lengths and synonyms yield the *normalized and divided Levenshtein distance* (LDND) (Holman et al., 2008).

The idea of modernized glottochronology (Holman et al., 2011; Wichmann et al., 2008; Serva and Petroni, 2008) is that LDNDs between word lists for any two languages give an indication of the time of divergence since their last common ancestor. For example, averaging across Indo-European language pairs Serva and Petroni (2008) argue that the time depth of the Indo-European family is proportional to the logarithmically transformed inverse of the LDND and estimate it to circa 5500 years BP, thus being roughly in line with some of the Bayesian accounts (Chang et al., 2015; Rama, 2016). Holman et al. (2011) apply the LDND method across language families of the world and estimate the time depths of the earliest splits. These are shown to approximate dates given based on other sources, such as archaeological, epigraphic (i.e. based on inscriptions), or historical data.

Crucially, Wichmann and Holman (2009) model the average ratios of population sizes (for pairs of languages within a family) as a function of the average LDNDs. This relationship can be used to estimate how far back in time population ratios might extrapolate. They show that mean population ratios between languages of the same areas (Africa, Eurasia, Australia, New Guinea, and the Americas) extrapolate into the past (with diminishing accuracy) *by several thousand years* (Wichmann and Holman, 2009, p. 267). Of course, population sizes – and specifically L2 speaker percentages – of particular languages can drastically fluctuate over time due to migration, expansion and trade. For example, the immense speaker populations of English and Spanish today are largely due to expansions of the British and Spanish empires in the last circa 500 years. These are extreme cases of growth in a relatively short amount of time. However, across a sample of dozens of Indo-European languages, such fluctuations average out and pairwise ratios of speaker populations are stable over hundreds of years, or even millennia.

What about the diachronic stability of lexical diversity as measured on the basis of parallel texts? One way of getting an impression of diachronic stability is to estimate so-called *phylogenetic signals*. Assume we have a phylogenetic tree for a given language family (built on cognate data or other lexical material) and a trait value per language represented on the tips of the tree (lexical diversity values in our case). Phylogenetic signal reflects how well these trait values fit the phylogeny under a given evolutionary process (e.g. Brownian motion). There is a range of phylogenetic signal metrics reviewed in Münkemüller et al. (2012). One of them, Pagel's λ (Freckleton et al., 2002; Pagel, 1999), was used in Bentz et al. (2015) to estimate phylogenetic signals of lexical diversities. λ-values can range from zero to one. $\lambda = 0$ indicates a general mismatch between the phylogeny and the empirical trait values, while $\lambda = 1$ indicates that lexical diversities of languages follow the expectations given the phylogenetic tree. In terms of time depth, $\lambda = 1$ might indicate that lexical diversities extrapolate back to early splits on the tree, while in the case of $\lambda = 0$ there is no evidence that they extrapolate back at all.

Phylogenetic signal analyses are performed in Bentz et al. (2015) for lexical diversities within the Indo-European, Austronesian, and Bantu (Atlantic-Congo) families. It is shown that unigram entropies have relatively strong phylogenetic signals, with Austronesian languages having the strongest signal ($\lambda = 1$), followed by Bantu ($\lambda = 0.85$) and Indo-European ($\lambda = 0.64$). This suggests that – as a general trend – lexical diversities of extant languages still reflect splits of considerable phylogenetic depth, i.e. hundreds or thousands of years ago.

A disadvantage of the phylogenetic signal method is that the analyses are again based on synchronic data and only indirectly infer diachronic pathways of change based on simplifying assumptions, such as constant rates of change in a Brownian motion model. Deviations from strict Brownian motion are going to be reflected in phylogenetic signals. However, inferring the exact underlying evolutionary model from phylogenetic signals is currently difficult to impossible (Revell et al., 2008). As a consequence, we need further independent evidence to corroborate how stable lexical diversities are over time.

Another, more direct way of measuring the diachronic stability of lexical diversities is to harness historical corpora. In Bentz et al. (2014), lexical diversity is analysed by applying type accumulation curves and Zipf-Mandelbrot parameters to translations of the Book of Genesis into Old English and Modern English. This study measures a decrease in lexical diversity of 23% between the two periods. This is further shown to be related to the loss of nominal case marking and verbal inflections. For example, while the lemma *land* 'land/country' occurs with dative (*land-e*) and genitive (*land-es*) inflections in the Old English text, in Modern English it occurs bare of these markers. The loss of inflected word types thus significantly reduces the lexical diversity of Modern English compared to its ancestor of around 1000 years ago. Note that other Germanic languages are more conservative in this regard. For instance, Modern German preserves the genitive in *des Landes* and the dative (though increasingly sounding archaic) at least in written language, e.g. *auf dem Lande*. Trudgill (2011) gives an extensive discussion of morphological changes in further Germanic languages and argues that different patterns of preservation and loss are related to pressures from adult learning and usage.

In a more recent study, word unigram entropies are compared for another branch of the Indo-European family: Romance languages (Bentz and Berdicevskis, 2016). In particular, parallel texts in Modern Romance varieties are compared to a Classical Latin translation. Of course, Classical Latin as handed down to us by famous Roman scholars and poets is a conservative written variety that only indirectly reflects the spoken Vulgar Latin that evolved into Modern Romance languages (Herman, 2000). Nonetheless, it is informative to assess how lexical diversity has changed from written Latin to written Modern Romance. An example for the loss of morphological elaboration similar to the Old English to Modern English scenario is given in Figure 9.2.

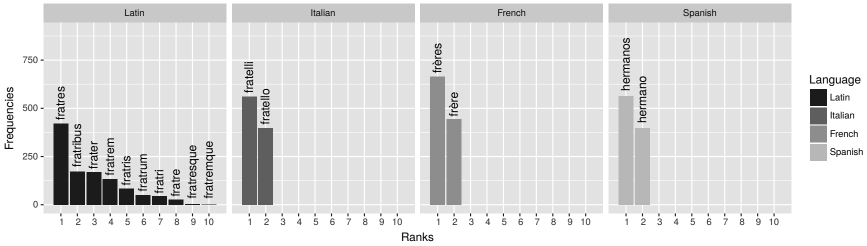

Figure 9.2: Word types of the lemma for 'brother' in Classical Latin as well as Modern Romance translations of the Bible. Types (x-axis) are ordered by their token frequency (y-axis).

In the Classical Latin translation, the lemma *frater* 'brother' is inflected for case and number,[3] thus creating a wide range of word types with relatively low token frequencies. In comparison, in Italian, French and Spanish there is only a distinction between singular and plural,[4] and the two respective word types occur more

[3] The addition of -*que* in the two lowest frequency forms is an enclitic indicating the coordinating conjunction 'and', rather than a case or number inflection. However, it qualifies as being part of a word type by using the criterion of white spaces in writing.

[4] In French, the distinction is only maintained in writing, the pronunciation is the same for both forms.

frequently. As a general trend, we get longer-tailed distributions of token frequencies over word types for Classical Latin than for Modern Romance languages. This is also reflected in unigram word entropies (Figure 9.3). The Latin text has an entropy of around 10.5 bits/word compared to around 9.5 bits/word for the Romance varieties analysed in Bentz and Berdicevskis (2016). Moreover, using lemmatization tools as discussed in Chapter 5, it is shown that the reduction of lexical diversity between Latin and the Romance languages is largely due to differences in inflectional marking. Namely, around one bit of information originally encoded in Latin inflections is lost – or replaced by other coding strategies such as word order – in Italian, French and Spanish.

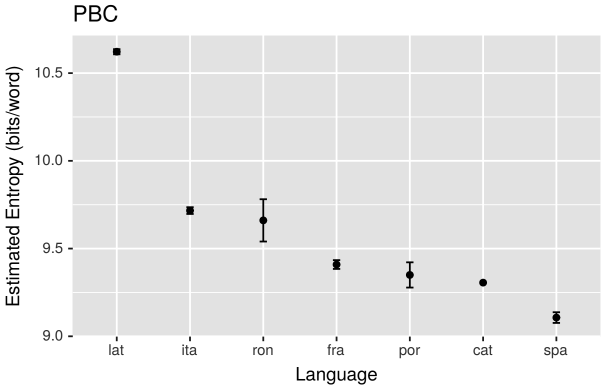

Figure 9.3: Estimated unigram word entropies for Latin and Modern Romance languages based on the Parallel Bible Corpus. The x-axis gives ISO codes for Latin (lat), Italian (ita), Romanian (ron), French (fra), Portuguese (por), Catalan (cat), and Spanish (spa). In some cases, several translations are available for the same ISO code. Mean values are then given with 95% confidence intervals.

Thus, in the roughly 2000 years since the expansion of the Roman empire, the Vulgar Latin varieties which evolved into Modern Romance languages lost around one bit of information formerly stored in inflectional markers and hence 10 to 15% of their lexical diversity. Again, it has been conjectured that this is related to the “recruitment” of substratum adult learners in the various outposts of the Roman empire (Herman, 2000; Bentz and Christiansen, 2010).

Overall, the population ratio analyses of Wichmann and Holman (2009) and the phylogenetic signal analyses of Bentz et al. (2015) suggest that both population structure and lexical diversity are preserved over hundreds or even thousands of years – across the board. Given some well-documented cases, e.g. the development from Old English towards Modern English or from Latin towards Romance languages, we can observe change in lexical diversity in "real time". These examples illustrate that lexical diversity can drop considerably within centuries or millennia. Such rapid changes, however, are likely related to extreme examples of social upheaval and change.

Thus, the results of statistical analyses in Chapter 8, which associate synchronic population data with synchronic linguistic data are not confined to a shallow interpretation in the here and now, but reflect past language change. This has particularly interesting implications for the interpretation of the mixed-effects results. Namely, it suggests that *within*-family effects are only detectable given certain preconditions: a) enough time-depth for the effects of adult learning to become reflected in language structure, but also b) not too much time-depth in order for the phylogenetic signals of lexical diversity and L2 percentage to still be reflected in synchronic data.

For example, let us assume the Indo-European family indeed has a time depth of circa 6000 years BP as argued by some of the studies mentioned above. For this particular family we find the negative association between L2 percentage and unigram entropy (see Figure 8.1). This is due to big Indo-European languages such as English, Spanish, French and German having relatively many adult learners and relatively low entropies, compared to smaller languages such as Icelandic, Lithuanian and Latvian, which have low numbers of adult learners and high entropies. As pointed out above, the "recruitment" of L2 speakers in the big languages has largely happened in expansions over the last couple of centuries or millennia at most. Within this time-frame, the lexical diversity of some Germanic and Romance languages has changed considerably. Thus, in the evolution of Indo-European languages, different favourable factors coincide and allow us to detect within-family variation in synchronic data.

In sum, this suggests that synchronic data on language and population structure are not necessarily "shallow". Instead, they can extrapolate back in time for hundreds and thousands of years. The question for mixed-effects analyses is then whether within-group variation is preserved at time-depths where effects are still detectable.

9.3 Multiple Factor Model

Explaining lexical diversities of languages around the world requires an understanding of the interplay between explanatory factors and grouping factors. Only this enables us to give causal answers to the question *why* specific languages score higher or lower on this measure. Ultimately, the question boils down to asking how learning and usage patterns came to be reflected in language populations at large and how they perpetuated through time to shape patterns of groupings at different geographical and genealogical levels.

In the actual statistical modelling we have generally not included descriptive factors as outlined in Chapter 5. An exception is the consideration of variation between registers and styles by means of having random intercepts by corpus. In theory, we could have considered further descriptive factors, such as scripts and word-formation patterns, as predictors in the model. Given the results of Section 5.1 and Section 5.2 it is clear that different scripts and different word-formation patterns are linked with lexical diversity. Namely, it was shown that scripts can change unigram entropies by around 10-15% (in extreme cases) and inflectional marking by 55% across 19 languages. Hence, it is to be expected that the degree of inflectional morphology will be a strong predictor of lexical diversity across languages. Cross-linguistic information on morphological marking strategies could be taken from databases like the WALS (Dryer and Haspelmath, 2013). However, conceptually it is problematic to put such language "internal" predictors into a statistical model alongside explanatory predictors, since they compete for variance explained, albeit at different levels of explanation.

For example, it is to be expected from studies such as Lupyan and Dale (2010) and Bentz and Winter (2013) that L2 speaker percentages are correlated with quantitative measures of morphological "complexity", such as the number of nominal case markers or inflectional marking of tense, mood and aspect on the verb. Hence, variance explained by L2 percentage will overlap with variance explained by indicators of morphological complexity. This means they are not independent predictors. Instead, higher numbers of adult learners are most likely part of the *reason* for lower morphological complexity (Lupyan and Dale, 2010; Bentz and Winter, 2013, 2012). Again, this is not to say that systematic links between descriptive factors and lexical diversity are not worth exploring. On the contrary, such analyses can further elicit the exact mechanisms that link explanatory factors with lexical diversity.

Overall, the research agenda emerging from the CAS model is to establish links between explanatory factors at the population level – and ultimately at the level of individual learning and usage patterns – and quantitative measures of linguistic structure. Descriptive factors can further help to disentangle the mechanisms underlying the co-evolution of population structure and language structure. Grouping factors then reflect the pathways of this co-evolution in the shallow and deep past. A schematic depiction of this research agenda, with the links established in previous studies is found in Figure 9.4.

The figure lists the explanatory and descriptive factors considered to date. Links are drawn with different types of arrows depending on whether there is strong evidence for a link (unbroken arrow) or an alleged link with some support (dashed).

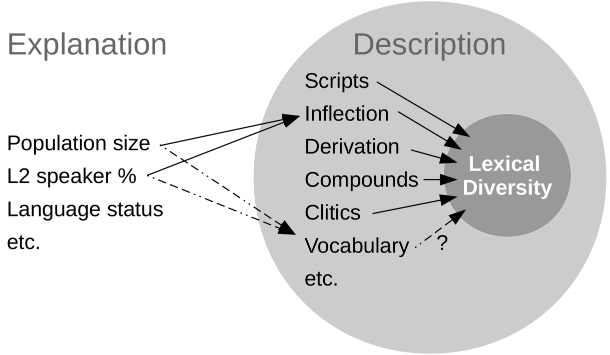

Figure 9.4: Interaction of descriptive and explanatory factors. Descriptive factors are here represented language "internally", i.e. inside the grey circle. Explanatory factors are represented "externally", i.e. outside the circle. The links between explanatory and descriptive factors, as well as the links between descriptive factors and lexical diversity are indicated by arrows according to the strength of connection (medium: dashed, strong: unbroken).

With regards to *population size*, some limited evidence for a link with lexical diversity has been uncovered in the current study. There is a significant positive Pearson correlation (Section 6.1) and a small but significant positive coefficient in the multiple regression (Section 8.1), which ceases to be significant when random intercepts and slopes by family and area are introduced. We are thus dealing with an effect that differs considerably between families and areas. Furthermore, population size is correlated with language status. Hence, population size and language status are not two separate predictors, but rather share the same link with lexical diversity. In the literature, Sinnemäki (2009) establishes a link between population size and the inflectional marking of core arguments and Lupyan and Dale (2010) argue for a link between population size and morphological complexity more generally. This is indicated by an unbroken arrow between population size and inflection in Figure 9.4.

Interestingly, the results by Lupyan and Dale (2010) actually predict a *negative* correlation with lexical diversity, not a *positive* one as found here. Namely, in their study, bigger populations are correlated with less inflectional marking, and less inflectional marking is expected to result in lower lexical diversity (Section 5.2.4).

In contrast, the relationship between population size and lexical diversity found here is *positive* (Section 6.1), i.e. bigger populations are associated with higher lexical diversity. These partly contradictory results might be due to differences in the language samples used, or due to genuine differences in the mechanisms linking population size with lexical diversity.

For example, Bromham et al. (2015) illustrate for 20 Polynesian languages that population size is positively correlated with the rate by which languages gain new words. Interestingly, this effect is argued to be independent of loanword borrowing. If this trend extrapolates across languages, it could drive a positive association between population size and the size of basic vocabularies: bigger populations would be expected to have bigger basic vocabularies and hence higher lexical diversity (everything else being equal). This might counterbalance the negative effect of reduced morphology on lexical diversity.

Hence, there might be multiple, competing effects linking population size with lexical diversity, ranging from extension of the base vocabulary to inflectional marking strategies. Another, rather stylistic effect, can also not be conclusively excluded: for corpora like the UDHR and PBC it seems likely that a considerable proportion of the religious and legal vocabulary is not directly translatable into smaller languages whose speakers do not deal with such concepts. In these cases, translators are likely to repeatedly use periphrastic constructions built on the lexical material available in the language. Such circumscription of abstract concepts might, in turn, decrease the lexical diversity of the translations.

Language status "loses out" on population size in the multiple regression model, while the simple Pearson correlation is negative and just about significant. Given the coding of the data, this means that higher status is linked with higher lexical diversity. However, there is, to my knowledge, currently no theoretical explanation for a link between language status and lexical diversity in the literature. Since population size and language status share explained variance, the alleged links given for population size seem valid for language status too. Namely, it is conceivable that language status is reflected in lexical diversity via the vocabulary used in the parallel texts. Languages of "high" status (remember that status is here defined on an endangerment scale and not necessarily equivalent with higher social status), such as national languages, probably have a wider vocabulary at their disposal to encode specialized information about legal (UDHR) and religious matters (PBC), as compared to local languages used for genuinely different purposes.

Finally, *adult language learning*, as reflected in L2 percentages, emerges as a strong predictor of lexical diversity, both in the multiple regression (Section 8.1) and in the mixed-effects regression (Section 8.2) models. Though in the latter the significance is reduced. This means that the effect largely – but not entirely

– holds *between* and *within* different levels of grouping. Thus, languages with higher numbers of adult learners tend to be those with lower lexical diversity. The size of this effect is moderate to strong considering that going from 0% L2 to 100% L2 speakers predicts a decrease in unigram entropy by 37%.

A link between reduced inflectional marking and imperfect language learning by adults is established in both qualitative (Kusters, 2003; McWhorter, 2007; Wray and Grace, 2007; Bentz and Christiansen, 2010; Trudgill, 2011) and quantitative studies (Lupyan and Dale, 2010; Dale and Lupyan, 2012; Bentz and Winter, 2012, 2013, 2014). Moreover, it was shown here that inflectional morphology has a strong impact on lexical diversity (Section 5.2.4). Thus, the link between L2 percentage and lexical diversity via inflectional marking is well supported. It is conceivable that there are further links involving derivational morphology, compounds and clitics, since all of these can change lexical diversity. Besides simplification of morphology due to imperfect learning by adults, the borrowing of loanwords is also a well-known phenomenon in contact linguistics (Thomason and Kaufman, 1988). Borrowing of vocabulary could have a positive effect on lexical diversity and counterbalance the reduction caused by the loss of inflectional morphology in certain contact scenarios.

Note that there has been no link established so far between any of the explanatory predictors and scripts. There are certainly ways in which population size, language contact and language status interact with the rise and fall of script types. However, these links were not a topic of the current study.

The other way around, there is some evidence that population size and language contact have an impact on base vocabulary, but no direct link between basic vocabulary and lexical diversity has been established. This is because it is notoriously hard to measure the difference that an expansion of the basic vocabulary would have on lexical diversity, abstracting away from all word-formation patterns. In order to measure this effect we would have to neutralize all inflections, derivations, compounds, clitics, and other structural features beyond these (e.g. tone) in a parallel corpus. There is currently, to my knowledge, no computational tool which would consistently achieve this across a range of languages.

Of course, there are many more descriptive, explanatory, and grouping factors that have not been considered in the current account. For example, another interesting descriptive property of languages is the systematic relationship between the length of words and their frequencies (Zipf, 1949; Piantadosi et al., 2011; Ferrer-i-Cancho et al., 2015; Bentz and Ferrer-i-Cancho, 2016) or the shapes of spectrograms of languages indicating the rate of repetition of information (Moscoso del Prado, 2011). There are also further explanatory factors relating to language learning, usage, and attitude, which coould be explored in future studies. For example, whether a language is taught in school or not and whether a language is written, spoken, or both. Finally, geographical grouping can go beyond language areas, namely extending to whole latitudinal and longitudinal bands, or climatic zones. For further reference, a meta-study by Ladd et al. (2015) reviews the available correlational studies looking at language "internal" and "external" factors.

9.4 Summary

To summarize, when interpreting multiple regression and mixed-effects regression results, it is important to keep track of the conceptual underpinnings reflected in the predictor variables. On one hand, descriptive factors are helpful to understand *what* exactly is different between languages, both in information-theoretic and linguistic terms. Explanatory factors, on the other hand, are a first step to establish causal links between the linguistic structures in focus and the learning and usage patterns shaping them. Grouping factors are then a synchronic reflection of diachronic changes in linguistic structure. Different levels of grouping are indicative of the co-evolutionary pathways that explanatory and descriptive factors have taken. Hence, instead of viewing all three types of factors as equivalent and competing for variance explained, they should be teased apart. Disentangling the mutual dependencies between descriptive, explanatory and grouping factors sheds new light on how the diversity of language structures found across the world evolved.

10 Further Problems and Caveats

The core part of this book was to establish statistical associations between information-theoretic properties of languages, on one hand, and characteristics of the populations using them, on the other. Lexical diversity was chosen to be modelled and explained. The predictors of main interest were population size, second language learner percentages, and language status. The statistical links between these predictor variables and lexical diversity can be seen as synchronic reflections of the co-evolutionary pathways that language structure and population structure have taken. However, to get the full picture of how and why languages evolve in certain directions, population characteristics have to be translated into biases of language learning and usage. We need to understand why certain historical scenarios have certain effects on lexical diversity, while others have opposite effects, or no effects at all. Hence, an important psycholinguistic question is how lexical diversity relates to language learning in children and adults. While an exhaustive overview of all the literature that has been written on the topic goes beyond the scope of this book, Section 10.1 sketches some of the core findings relevant.

Furthermore, the concept of lexical diversity is based on the assumption that word tokens and types exist as meaningful units of information encoding. Clearly, there are limitations to this account. Firstly, the structure and distribution of words can change over time within the same language. For example, when processes of grammaticalization blur the boundaries between free and bound lexical material, such that formerly independent tokens merge to built new word types. A brief note on grammaticalization is given in Section 10.2. Secondly, information encoding in natural languages certainly happens “beyond words”. This can be interpreted in two senses: firstly, it refers to *other structural levels*, e.g. phonemes, morphemes, phrases. More generally, it also hints at the fact that language is embedded in a rich array of *multi-modal perceptions* and our general knowledge of the world. Information-theoretic accounts have sometimes been accused of trying to exclude this aspect of language by solely focusing on the code and ignoring its meaningful interpretation. However, as has been pointed out in Chapter 4, this is not necessarily the case. Issues relating to information encoding “beyond words” are further discussed in Section 10.3.

DOI 10.1515/9783110560107-010

10.1 Language learning and lexical diversity

Lexical diversity is defined as the distribution of word token frequencies over word types. Given constant content, a language with few word types and high token frequencies has low lexical diversity and a language with many types and generally low token frequencies has high lexical diversity. The unigram entropy based on Shannon's definition is used throughout this book to measure these differences in word frequency distributions. To get an intuition of the potential repercussions on language learning consider the following thought experiment.

In a treatise on *Word and Object*, Willard van Orman Quine has famously described the problem of translating from one language into another with a parabola about a fictitious language – in the following referred to as *Quinean*. Imagine a linguist joins a speaker of Quinean for a hunt in the woods. Suddenly a rabbit runs past and the Quinean speaker utters "gavagai" (Quine et al., 2013, p. 25). According to Quine's rationale, it is impossible to determine exactly what this utterance means and translate it perfectly into English – or any other language for that matter. It could mean 'rabbit', or 'look over there', or 'there is a light brown rabbit running towards the east', or any other of an infinite number of potential meanings. The linguist can narrow down the meaning by using a battery of tests usually employed by fieldworkers, but any further inquiry ultimately faces the same issues of interpretability which hamper a one-shot-interpretation in the first place.

10.1.1 Learnability vs. expressivity

Examine this example from an information-theoretic point of view. First, assume that speakers of Quinean just utter *gavagai* in regular time intervals, completely independently of what is happening around or inside of them. In a corpus of this language, we would find that they use this single word type over and over again. Its token frequency would be maximal, i.e. equivalent to the overall number of tokens uttered. In this type of Quinean, there is no choice for the speaker and hence no uncertainty for the hearer. As a result, there is no information encoding potential. *Gavagai* could mean anything and hence means nothing. This version of Quinean has zero unigram entropy.[1] Let us call this version of the language *Minimum Entropy Quinean* (Quinean$^{\text{Min}}$).

In contrast, imagine that Quinean is an extremely expressive language. Namely, there is a specific word type for any conceivable concept. *Gavagai* actu-

[1] Though if the speaker can choose between uttering *gavagai* and not uttering it, then it carries a maximum of one bit of information.

ally means ‘there is a *light* brown rabbit running towards the east’, while *govagai* means ‘there is a *dark* brown rabbit running towards the east’, and *govatai* means ‘there is a dark brown rabbit running towards the *west*’, and so on. In a corpus of this version of Quinean, there is likely a high number of different word types encoding even minor shades of meaning. In the theoretical maximum case, the number of word types is equal to the overall number of tokens, i.e. each word type would only occur once as a hapax legomenon. This language would have maximum information encoding potential at the level of words.[2] Let us call this *Maximum Entropy Quinean* (Quinean$^{\text{Max}}$).

Now, to learn Quinean$^{\text{Min}}$ we merely have to remember the word type *gavagai* and utter it regularly. Purely in terms of memory storage and retrieval, it is maximally learnable. In contrast, to learn Quinean$^{\text{Max}}$ we have to remember a panoply of word types – a potentially infinite number – and how to match them onto the corresponding concepts. This makes it minimally learnable, or, in fact, genuinely unlearnable. In theory, there is an inverse relationship between entropy and learnability – at least in terms of memory storage. Minimum entropy corresponds to maximum learnability and maximum entropy corresponds to minimum learnability.[3]

Another term sometimes used in this context is *compressibility*. A corpus of Quinean$^{\text{Min}}$ is highly compressible, since the word type *gavagai* only has to be stored once and every token can be replaced by a pointer to the storage location of the type, rather then being stored separately. A corpus of Quinean$^{\text{Max}}$, on the other hand, is *incompressible* at the level of words, since every word type, e.g. *gavagai*, *govagai*, *govatai*, etc. is unique and has to be stored separately.[4] Thus, Quinean$^{\text{Min}}$ is highly learnable, compressible, but not expressive, while Quinean$^{\text{Max}}$ is hard to learn, incompressible, but highly expressive.

The evolutionary trade-off between these competing pressures has been modelled computationally and tested experimentally (Kirby et al., 2008, 2015; Tamariz and Kirby, 2015). Tamariz and Kirby (2015), for instance, argue that the compressibility of drawings is the outcome of learning pressure over several generations of cultural transmission. In Kirby et al. (2008), it is shown experimentally that human participants are inclined to reduce the number of different word types

[2] The exact entropy value will depend on the number of types/tokens.

[3] Though see also Takahira et al. (2016) for another information-theoretic conceptualization of “learnability”.

[4] Standard compression algorithms are based on character encodings, rather than whole word encodings. Hence, they would further compress the Quinean$^{\text{Max}}$ corpus in this example. For instance, by harnessing the fact that all these word types end in -*ai*, which is redundant information.

in an artificial language – if only pressure for learning is given. Starting with a scenario of random character strings matched onto moving objects of different colours and shapes, participants have to learn the string/scene mappings. For example, a blue bouncing circle is associated with the random string *manehowu*, while a red bouncing circle is associated with *wuneho* (see supplementary material in Kirby et al., 2008). Note that this is close to Quinean$^{\mathrm{Max}}$ in terms of unigram entropy: in most cases, there will be one unique random letter string matching each scene, i.e. meaning. The output of one generation of learners, when tested on the mappings, is given to the next generation as input to learn. It turns out that over several generations, the learners reduce this high-entropy language to a low-entropy language, which is, however, essentially uninformative as to the meaning encoded. Namely, the originally 27 different character strings are reduced to 2-5 highly ambiguous strings[5] over 10 generations.

In the second experiment of Kirby et al. (2008), pressure to maintain expressivity is introduced. If a participant overgeneralizes strings to different meanings, all but one of these meanings is removed when the data is given to the next participant. In this case, the languages do not collapse into low-entropy states. Rather, of the 27 strings 12-23 are preserved over 10 generations. The core finding is that systematic structure within the strings, i.e. word internal structure, starts to evolve to encode colour, shape and movement. For example, the "prefixes" *n*-, *l*-, and *r*- emerge to indicate black, blue, and red colour respectively, while the "suffixes" -*ki*, -*plo*, and -*pilu* emerge to encode particular movements (Kirby et al., 2008, Fig. 5).

A similar trade-off is investigated in an iterated learning experiment by Berdicevskis (2012) and computationally modelled in Kirby et al. (2015). Here, the terms learnability, compressibility and expressivity are explicitly used. The computational model in the latter paper again illustrates that if languages are selected exclusively for learnability, then they move towards low-entropy states. If only expressivity matters, then "holistic" languages, i.e. high-entropy languages with one-string-one-meaning mappings are maintained. Again, the central finding is that only if both pressures are present at the same time, then languages develop compositionality to overcome the fundamental information-theoretic trade-off between expressivity and compressibility. This is further illustrated with another iterated learning experiment in Kirby et al. (2015). The major difference to the earlier experiments in Kirby et al. (2008) is that now participants are not on their own, but have to use the learned string/meaning mappings in a communicative

[5] There are 4 chains of transmission in the experiment, i.e. 4 separate runs of cultural transmission.

task involving another learner. This is a more natural and realistic way to introduce pressure for communicative success (Berdicevskis, 2012). In this setting, again the languages become compositional.

Overall, these experiments and computational models suggest that natural languages should fall in the middle range between extremely compressible and extremely expressive languages. In other words, we neither expect to find Quinean$^{\text{Min}}$ nor Quinean$^{\text{Max}}$ in the real world, but rather languages falling in between. The analyses in Chapter 4, specifically Figure 4.8, illustrate that this is the case indeed. Namely, the unigram entropies of natural languages display a unimodal distribution along the spectrum from 6 to 13 bits/word (unscaled) and from around –3 to 3 (scaled).

While this spectrum is relatively narrow, there are also clear outliers. Some languages are considerably more compressible than others based on the text samples used here. Are those languages also less "expressive"? Purely based on the unigram entropy of words – yes. This seems to contradict the fact that the content of the parallel texts is assumed to be (relatively) constant, i.e. the texts supposedly express the same overall meaning. However, information can also be encoded in how words combine to multiword expressions, phrases, and sentences. Clearly, information encoding does not only happen at the word level. In other words, lacking expressivity at the level of lexical diversity might be counterbalanced by expressivity at other levels of encoding, i.e. in the syntactic or pragmatic dimension. Possible information encoding patterns beyond the word and their relation to expressivity will be further discussed in Section 10.3.

Having said this, with reference to lexical diversity, there are clear differences between languages. For example, the Bible in Hawaiian has a relatively low scaled unigram entropy (–1.8) and is hence more compressible at the level of unigrams, than the Bible in Iñupiatun, which has one of the highest entropies (3.2). Figure 10.1[6] visualizes the difference between Iñupiatun and Hawaiian. Note that in this figure, only the first 100 ranks (i.e. highest frequent word types) are shown. Iñupiatun has overall 20268 word types in this Bible text of 50000 tokens, while Hawaiian has only 1521. On the other hand, Hawaiian has much higher token frequencies than Iñupiatun on average ($\mu_{\text{haw}} = 32.9$, $\mu_{\text{esk}} = 2.5$).

Coming back to the problem of language learning: a learner of Hawaiian will come across the same word types much more often than a learner of Iñupiatun. Of course, translations of the Bible and the legalese of the UDHR are not the type of input a child or adult learner typically comes across anyways. It is conceivable that in other text corpora – more realistically reflecting the Hawaiian language as

[6] Rcode/Chapter10/freqDistsHawEsk.R

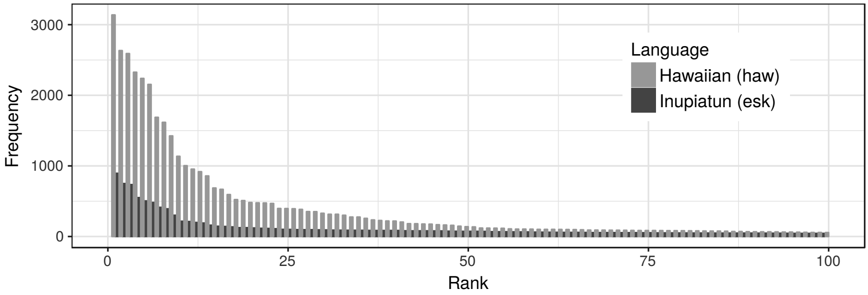

Figure 10.1: Word frequency distributions for the Northwest Alaska Iñupiatun (esk) and Hawaiian (haw) Bible. This visually illustrates the difference between distributions of word tokens (y-axis) over word types (x-axis) for Iñupiatun (dark grey) and Hawaiian (light grey). Note that only the first 100 ranks are shown. Iñupiatun has 20268 types (i.e. ranks) overall, whereas Hawaiian has only 1521.

learned by children and adults – the unigram entropy might have a higher value and the differences between the two languages might not be as stark. For instance, Nettle and Romaine (2000, p. 56) point out that traditional Hawaiian fishermen have an extensive vocabulary for hundreds of different species of fish, which are not even adequately described by marine biologists yet. If the texts used for analyses are about fishing, then the Hawaiian unigram entropy is likely to increase compared to a language like English.

However, in the massively parallel text collections currently available, we find a good extrapolation quality of entropy rankings per language across different corpora (Section 5.3). Also, a spoken corpus was included in the study comparing English and German word frequency distributions (Section 5.2.4) and this did not change the basic fact that German has longer-tailed distributions and higher unigram entropy. It is likely that overall unigram entropies are comparable to what we find in the limited corpora currently available, even if more variegated and spoken material will certainly help to enrich the picture.

Leaving aside discussions about unigram entropy values for particular languages, we can state more generally that being faced with low versus high unigram entropy distributions will make a difference for the learner – everything else being equal. This directly follows from the vast literature on the impact that frequencies have on first language (Ambridge et al., 2015; Lieven, 2010; Tomasello and Tomasello, 2003; Roy et al., 2009; Diessel, 2007; Stoll et al., 2017) and second language acquisition (Ellis, 2002; Ellis and Collins, 2009; Goldschneider and DeKeyser, 2001; Larsen-Freeman, 1975, 1976). This is not to say that frequencies in the input are the only – or even the dominant – factor involved in learning. A recent large-scale longitudinal study on word learning indicates that factors relating to spatial, temporal, and linguistic distinctiveness of the input are even more important for "the birth of a word" than pure frequencies (Roy et al., 2015). However, frequency still emerges as one of the most robust predictors of learning success in the data accrued over the past decades.

Moreover, frequency effects are not limited to the learning of basic vocabulary, but are also reflected in learning of morphologically complex forms (Lieven, 2010; Ambridge et al., 2015; Ellis, 2002). This is important, since it was shown in Section 5.2.4 and based on Bentz et al. (2017b), that around 50% of the cross-linguistic variance in word frequency distributions can be attributed to morphological marking.

When looking at the link between morphological marking and lexical diversity, we can add a further complication to the picture. Arguably, there is a difference between word forms built on the basis of regular morphology and ones that are irregular, i.e. have to be stored whole. For example, learning the word forms *bake-d*, *accept-ed* and *claim-ed* might be easier once the "rule" of attaching *-ed/d* to the root for past tense formation is discovered. Whereas remembering the irregular forms *went*, *caught* and *was* just requires "brute force" rote learning. This is the rationale of dual-route accounts (Clahsen et al., 1997; Clahsen, 1999; Pinker and Ullman, 2002; Marslen-Wilson and Tyler, 2007; Taft, 2004; Marcus et al., 1995). However, even if this distinction is valid (see Behrens and Tomasello, 1999, for some reasons why it might not be), there are still many irregular forms that need to be remembered, especially in languages with complex morphology. Take the up to 37 inflectional noun classes in German (Steiner and Prün, 2007), which will make it harder for a learner to find systematic patterns of noun morphology. Even in English, which is sometimes given as an example of a morphologically fairly regular language, the highest frequency forms are irregular (Lieberman et al., 2007). While it has been shown convincingly that even irregular forms can display subtle patterns that learners can pick up on (Cuskley et al., 2015), the burden is still on the learner to extract these patterns from a high entropy word distribution in a language like Iñupiatun, whereas in the low-entropy case of Hawaiian, probabilities are high that the same word form will occur in different contexts.

Apparent differences between low- and high-entropy word frequency distributions have led Gries (2012, p. 500) to propose that simple token frequency counts are not sufficient for predicting learning outcomes. Instead, the *skewness* of type-/token distributions, as reflected by unigram entropies, has to be taken into account. Goldberg et al. (2004) illustrate this point experimentally. In their experiment, participants have to learn the meanings of novel verbs and the constructions they occur in, from visual scenes. It turns out that constructional meanings are learned and remembered more easily when the five different novel verb types are presented in a lower-entropy (8-2-2-2-2), rather than a higher-entropy (4-4-4-2-2) distribution.

Overall, evidence from computational modelling and artificial language learning research seems to suggest that the learnability of word frequency distributions should negatively correlate with their entropy. However, note that in the results of our statistical models (Chapter 8), the presence of adult second language learners in relation to first language learners emerged as a predictor of lexical diversity. This raises a further important question: is there a *qualitative* difference between so-called "native" (L1) and "non-native" (L2) learning, specifically with regards to word frequencies? There is a multitude of studies attesting to differences and similarities in what is traditionally called "native" and "non-native" language learning. However, there are fewer studies more specifically related to the question of how frequencies affect word learning in children and adults. These are of central importance here.

10.1.2 Learning in children and adults

Vocabulary learning

The investigations by Roy et al. (2009, 2015) are a case in point. In an ultra-dense corpus of video and audio recordings of a single child's language learning (between an age of 9-24 months), both studies show an overall effect of frequency of occurrence in the caregivers speech on the *Age of First Production* (AoFP). Words that are more frequent in the caregivers' speech tend to be produced earlier by the child. However, as pointed out above, the newer of the two studies, Roy et al. (2015), also takes into account the *distinctiveness* of input words. Distinctiveness here refers to words being uttered in a confined spatial, temporal or linguistic context. For example, the word *breakfast* is mostly uttered by caregivers in a confined space (the kitchen) at a certain time of day (in the morning). This makes it easier for the child to map the string *breakfast* onto the activity of eating in the morning. Such distinctiveness features emerge as more robust predictors of AoFP than pure frequencies.

The importance of spatial, temporal, and linguistic contexts in word learning might be one of the reasons for why child language and adult language learning could differ. Children are, in most societies, typically surrounded by caregivers that engage in permanent interactions and shared tasks, thus creating a rich, multi-modal context for word learning. In some L2 learning scenarios, e.g. in a classroom, there are no native speakers permanently around for creating this multi-modal learning environment. If this rich context is missing, pure frequencies in the input might come to the foreground again. For a recent discussion on this topic see also Qureshi (2016).

Learning of morphology

With regards to frequency effects in learning word-formation patterns, the literature on regularization of inconsistent marking strategies in both children and adults is most insightful (Hudson Kam and Newport, 2005; Reali and Griffiths, 2009; Hudson Kam and Chang, 2009; Kam and Newport, 2009; Cuskley et al., 2015).

To start with, it is important to point out that regularization is only relevant for changes in lexical diversity if it leads to an overall *reduction* of the number of word types. For example, if a learner consistently uses the word form *go* instead of *went*, then this reduces the overall number of word types in their production.[7] However, if *goed* is consistently used instead of *went*, then this does not reduce the number of word types.[8] Instead, the overregularized form *goed* just *replaces* the irregular *went*. In other words, it is important to tease apart *replacement* and *reduction* in what is often ambiguously referred to as "regularization". Only reduction plays a role for lexical diversity, for example, when originally inflected forms are reduced to a word type that already existed independent of the marking. This is a non-trivial and important distinction that needs to be drawn to not confuse different notions of "regularization". Henceforth, the *went* → *go* type of change is called *reduction* and the *went* → *goed* type of change *replacement*.

Keeping this distinction in mind, a study using an artificial language learning paradigm (Hudson Kam and Newport, 2005) illustrates that children tend to regularize determiners more than adults. "Regularization" here refers to reduction, that is, the consistent usage of a single determiner in conjunction with different nouns when the input for learning is inconsistent with regards to determiner usage. Namely, in Experiment 2 of Hudson Kam and Newport (2005), five to seven year old children as well as adults have to learn a simple artificial language from spoken input in which nouns either occur accompanied by a determiner 100% of the time (consistent condition), or 60% of the time (inconsistent condition). For example, in the consistent condition, the noun *mɛlnag* 'car' consistently occurs with a determiner as in *po mɛlnag* 'a car', while in the inconsistent condition both *mɛlnag* on its own and *po mɛlnag* are encountered. While adult learners produce determiners consistently when presented with consistent input and inconsistently when presented with inconsistent input (thus probability-matching

[7] If *go* is used also independent of the past tense context, which is very likely.

[8] Unless *goed* is already used in another context, which is unlikely.

their production to the input), children tend to overgeneralize determiner usage in the inconsistent condition (Hudson Kam and Newport, 2005, p. 181).

The linguistic behaviour of children in this experiment can be seen as the simplest case of regularization by reduction. Namely, there is a binary choice between using and not using the determiner *po*. In the inconsistent input language, the probability of *po* is 0.6 and the probability of not using it accordingly 0.4. We thus have an entropy of 0.97 bits in the determiner system of the input language. Children increase the usage frequency of the determiner in relation to the input frequency (0.71 versus 0.29) and thus decrease the entropy to 0.87 bits. Adults, in contrast, probability-match the input frequency and thus rather preserve the entropy of the original system. The authors interpret this as evidence that children might be more strongly involved in the regularization of languages than adults.

However, two follow-up studies (Hudson Kam and Chang, 2009; Kam and Newport, 2009) found that adults also regularize determiners. This time, the experiments involved different conditions with multiple determiners differing according to their frequencies in the input. In this more complex setup, adults also started to regularize determiners. Importantly, regularization here again refers to reduction of different determiner forms to a single base form, which in these experiments is attested in adult production as well. Nevertheless, children are more consistent in their reduction of several forms to a smaller set of repeated forms, even in less complex conditions. Hudson Kam and Chang (2009) and Kam and Newport (2009) interpret this as the outcome of memory constraints. Children are known to have lower capacities in terms of memory storage and retrieval and they are hence more prone to overgeneralize a highly frequent default form. Adults have higher memory capacity, meaning that probability matching is possible beyond what is attested for children and overgeneralization with replacement of low frequency forms kicks in only when more complex input is faced.

Overall, these studies together strongly suggest that both adults and children reduce the entropy of determiner distributions, though the effect is more pronounced for children. The effect of entropy reduction by learning pressure is not confined to determiners. It extends to the distribution of verb forms via so-called *optional infinitives* (Freudenthal et al., 2015; Wexler, 1994) in children. For instance, children regularly use *go* in third person contexts, e.g. *that go there* instead of *that goes there*. The same pattern is also attested for noun forms. Children learning German sometimes use the nominative form instead of forms inflected for accusative case. For example, in the sentence *das kind kriegt den elefant* 'the child gets the elephant', uttered by a three year old child immersed into German from birth (Eisenbeiss et al., 2006, p. 16). In Standard German, *elefant* would have to be marked by the accusative suffix *elefant-en*.[9]

Both reduction of verbal morphology and noun morphology are observed in adult production too. Second language learners of German frequently omit nominal bound morphology (Parodi et al., 2004), irrespective of their first language (Turkish, Korean or Romance). Likewise, Greek L2 learners of Turkish produce both omission and substitution errors in verbs and nouns (Papadopoulou et al., 2011), and English learners of Turkish omit verb (Haznedar, 2003) and noun morphology (Gürel, 2000), even when fairly advanced (Haznedar, 2006). Hence, both child L1 and adult L2 learners struggle to learn the panoply of word forms in morphologically rich languages. Such learning pressure can lead to a reduction of entropy in the distribution of word forms.

This raises an interesting paradox: if both children and adults are prone to reduce the entropy of word frequency distributions across languages, then how can high-entropy languages like Iñupiatun exist in the first place?

10.1.3 Esoteric and exoteric linguistic niches

The hypothesis that imperfect second language learning simplifies languages (especially morphology) is getting more and more prominent in sociolinguistic frameworks (Thomason and Kaufman, 1988; Trudgill, 2002; McWhorter, 2002, 2007; Wray and Grace, 2007; Lupyan and Dale, 2010; Trudgill, 2011; Dale and Lupyan, 2012; Bentz and Winter, 2013; Bentz et al., 2015). In an overview paper, Wray and Grace (2007) have outlined the potential "consequences of talking to strangers". They distinguish between *esoteric* (intra-group) languages and *exoteric* (inter-group) languages. Esoteric languages are learned and used in relatively closed groups of intimates. Exoteric languages, on the other hand, are learned and used to communicate across different groups. Pidgin languages are an extreme case of the latter type where different populations meet with little or no linguistic common ground.

Given the mutual familiarity and the rich, shared cultural background of speakers in the esoteric niche, information might not have to be neatly packaged and easily encodable and decodable in such small-group languages. Wray and Grace (2007) speculate that an easily understandable language might not even be desirable in such a setting, since a strict delimitation of *who is an insider* and *who is an outsider* might play an important role. In fact, Evans (2011, p. 13) reports

[9] Though many German speakers would by now probably accept the unmarked form as grammatical.

how a fieldworker witnessed the speakers of the language Selepet in Papua New Guinea deciding at a meeting that the word for ‘no’, *bia*, should be replaced by *bunge* in order to distinguish themselves from other villages speaking similar varieties. Such sociolinguistic phenomena might not be restricted to the replacement of particular words, but could extend to further linguistic structures, as further outlined by Evans:

> William Thurston studied “esoterogeny” – the engendering of difference in linguistic obscurity – with Anem speakers of New Britain, off the New Guinea mainland. He found that “esoterogenic” languages tend to streamline pronunciation in ways that make the overall structure harder to see, comparable to saying *dja* for *didja* from *did you* in English. They replace clear regular relationships with “suppletive” (totally irregular) ones, revelling in alternations like *good:better* at the expence of the more transparent *big:bigger* style.
> (Evans, 2011, p. 13)

Interestingly, from an information-theoretic point of view, we might speculate that such “streamlined pronunciation” compresses information and increases the number of word types in a language. This, in turn, might be reflected in higher unigram entropy (see also the notes on grammaticalization of frequently co-occurring words below). Changes of the “suppletive” type, on the other hand, might lead to less regularity and hence higher entropy at the within-word level. Thus, sociolinguistic scenarios of tight-knit esoteric communities might be associated with increasing compression and higher entropy, at different levels of language structure. The exact mechanisms of such processes and their information-theoretic outcomes are an interesting avenue for further research.

In contrast, exoteric languages are by their very nature outward-oriented and inclusive. They are used for building trade relations, cross-cultural exchange, or the integration of migrants and enable communication between peoples and nations. Ease of learning and usage might come to the foreground in such an open setting. Against the backdrop of the esoteric/exoteric distinction, Trudgill (2011, p. 185) names the five major social factors involved in complexification: small population size, dense social networks, large amounts of shared information and low contact (i.e. adult L2 learning). The counterparts of these factors might then drive simplification: big population size, loose social networks, limited shared information, and high contact.

10.1.4 Small-scale multilingualism

A note of caution is in place here. As Trudgill (2011, p. 32) is well aware, “language contact” is often used as a cover term for language learning and usage scenarios that might in fact differ considerably in terms of their impact on language structure. While many sociolinguists agree that contact is associated with simplification, typologists (especially when working on non-European languages) sometimes come to rather opposing conclusions, namely, that complex features can be accrued in certain areas *because of* extensive contact. In Thomason and Kaufman (1988, p. 74-75), a systematic overview on language contact phenomena and their outcomes is given. According to their "borrowing scale", strong cultural pressure can lead to borrowing of "inflectional affixes and categories (e.g. new cases)" and hence add morphological complexity to the target language. As a concrete example, Aikhenvald (2007, p. 245-247) discusses how East Tucanoan languages in the Vaupés linguistic area of north-western Brazil have extensive structural impact on Tariana, a neighboring Arawakan language.

The apparent disagreement between the sociolinguistic and typological perspective is discussed in more detail by Lüpke (2016) and termed *Trudgill's conundrum*. This study takes the perspective of so-called *small-scale multilingualism* in diverse areas of West Africa, Amazonia, Northern Australia and Melanesia. Arguably, language communities in these areas are likely to be better models of language contact scenarios throughout human prehistory than the large-scale language expansions associated with the spread of agriculture and other technological innovations in recent history. Lüpke (2016, p. 63) argues that small-scale settings of contact share a set of characteristics, including a geographically confined basis, shared cultural traits, complex dynamics of exchange via dialectic relationships and, most importantly, "extensive multilingualism instead of or alongside a *lingua franca*". Due to especially the last point, it is conceivable that languages involved in such multilingual networks might score high on quantitative measures of contact such as percentage of L2 speakers, while not necessarily displaying patterns of simplification.

For instance, adults living in the Warruwi community in north-west Arnem land (Australia) are reported to typically speak three to eight indigenous languages on a daily basis (Singer and Harris, 2016). This suggests that languages in this area will score high in terms of percentages of L2 speakers. At the same time, these languages are also considered morphologically complex, since they feature elaborate morphological marking on the verb and systems of four to five genders (Singer and Harris, 2016, p. 26). They thus potentially run counter the purported relationship between higher contact in terms of adult learning and morphological simplification.

In fact, the convenience samples in Bentz and Winter (2013), Bentz et al. (2015), and in the analyses of this book mainly include languages of large families such as Indo-European, Austronesian and Atlantic-Congo. While the results hold for these families, it is possible that including more languages of small-scale multilingual areas such as the Warruwi community and the Vaupés basin might add new facets to the statistical picture.

10.1.5 Are children better learners than adults?

Besides the more general discussion about the outcomes of different language contact scenarios, the question about the exact cognitive mechanisms underlying morphological simplification and complexification is also still open. With regards to morphological simplification, Lupyan and Dale (2010, p. e8559) put forward the hypothesis that while adult learners have particular problems with learning the often redundant information encoded in morphological markers (e.g. person agreement on the verb), children might profit from such redundancy. Along similar lines, Trudgill (2011) posits that systematic, qualitative differences in learning by children and adults must account for the observed discrepancy between simplification and complexification in historical language contact scenarios. He explains this with reference to

> [...] the well-known fact, obvious to anyone who has been alive and present in normal human societies for a couple of decades or so, that while small children learn languages perfectly, the vast majority of adults do not, especially in untutored situations.
> Trudgill (2011, p. 35-36)

To underline this claim, he cites a classic experimental study by Johnson and Newport (1989), in which the English proficiency of 46 native speakers of Korean and Chinese is tested. The participants in this study arrived in the US at different ages. The proficiency test includes grammaticality judgements on morphological marking of past tense, plural, third person singular, and present progressive. The study reports that age of arrival in the US is the prime predictor of proficiency. In particular, learners that arrived in the US in between the age of three to seven perform almost identical to the "native" speaker control group, while all the other age groups (8-10, 11-15, 17-39) perform significantly worse. Importantly, Johnson and Newport (1989, p. 82) argue that this effect is *independent* of overall time of exposure to English, since age of first arrival and overall exposure time are carefully controlled for and do not correlate. However, it is generally hard to estimate and compare the exact amount of exposure to English for different learners in a real-world setting, since this will depend on a host of further factors relating to education, attitude and cultural assimilation.

A multitude of further studies since the early 90s have aimed to replicate Johnson and Newport's findings, with varying success. A recent meta-study by Qureshi (2016) reviews the arguments raised for and against systematic differences between younger (in this case below an age of 15) and older learners (above 15). In this context, there is an important distinction drawn between second language learning (SL) and foreign language learning (FL). While the former refers to a naturalistic situation of being faced with another language spoken in the immediate social environment, the latter rather refers to the instructed and formal context of learning a language in the classroom. Based on a meta-analysis of data from overall 26 studies Qureshi concludes:

> A medium to large effect size is observed for age and second language proficiency in SL contexts [...] The fact that early learners outperform late starters in SL contexts might be attributed to many factors, including the nature of immersion and the amount of input. For example, younger learners in SL settings have a greater opportunity to be formally and informally immersed in the target language [...] They also receive significant exposure to input from native speakers [...], something that is not present in FL contexts. [...] In FL contexts, on the other hand, an 'early advantage' for early starters is missing.
> (Qureshi, 2016, p. 157)

This outcome can be seen as roughly confirming Johnson and Newport (1989)'s findings. However, the meta-study also finds significant differences in learning outcomes depending on the exact conditions in which participants were tested, and this tells us a cautionary tale.

Moreover, the cognitive reasons for this purported difference are also still unclear. Interestingly, under the controlled conditions of artificial language learning experiments discussed above (Hudson Kam and Newport, 2005; Hudson Kam and Chang, 2009; Kam and Newport, 2009), which are arguably closer to the FL than SL setting, adults in fact *outperform* children in terms of converging onto a target language given the same input and exposure. In the second language learning scenario, on the other hand, there seems to be an early learner advantage emerging with regards to ultimate attainment.

In conclusion, whether children can be said to generally learn languages with "ease" and "perfectly", while adults do not, is still an open question. The only clear outcome emerging from the review of the language learning literature is that a simple distinction between "natives" and "non-natives" overly simplifies the complex picture of how age and exposure effects interact to yield different learning outcomes. For further discussion see also the recent overview article by Kempe and Brooks (2018).

10.1.6 Learner age and exposure

Rather than drawing a categorical distinction between "natives" and "non-natives" it might be more helpful to think about language learning along (at least) two continuous dimensions: *age* and *exposure*. For example, while children learning a first language overregularize and reduce morphological markers, L1 adults largely converge to the common usage of word forms in the population of speakers (though there is certainly still variance between speakers). Hence, the ratio of L1 adults to L1 children might be responsible for variation in the usage of forms (Briscoe, 2000b, p. 248), not only the ratio of L2 to L1 speakers. Specifically, this can mean that more children in a population might be linked with a higher probability of substitution and omission errors in the overall "corpus" of the population. On the other hand, it is attested that extensive bilingualism involving early learners might cause net increases in the usage of inflections (see Aikhenvald, 2003, p. 3, Nichols, 1992 and Trudgill, 2011, p. 40), whereas adult L2 learners seem generally more likely to omit or substitute morphological markers and therefore push morphological simplification. Thus, we can construe a matrix of morphological marking strategies emerging from different combinations of *age* and *exposure*. Table 10.1 condenses the results and implications of the language learning studies discussed above.

Table 10.1: Morphological outcomes according to exposure and age.

	high exposure (L1)	**low exposure (L2)**
children	simplification (complexification)	simplification
adults	complexification (?)	simplification

Based on the literature about overregularization in early language learning, we might speculate that there would be a high pressure for morphological simplification in a population if it mainly consisted of children learning the language, though there can be traces of complexification if the respective children are bilinguals borrowing morphological material into the language. Moreover, simplification is likely the outcome of children learning a low-exposure second language. L1 adults, having had high exposure throughout their life time, will put the least pressure on a language to simplify and might, in fact, further complexify the language due to compressing word forms via "streamlined pronunciation". On the other hand, adults learning a low-exposure second language are most likely to push this language in the direction of morphological simplification. Overall, populations with high rates of low-exposure learners (be it children or adults) are predicted to be most susceptible to morphological regularization and loss. In contrast, languages learned and used by high-exposure children, growing up to become high-exposure adults, are likely to introduce the lowest pressure for simplification and might even complexify their first language due to processes which are not well understood yet.

At the face of this, the distinction between children and adults might turn out to be secondary and the amount of exposure might emerge as the most important predictor for simplifying and complexifying processes in language change. Along those lines, in a preliminary analysis by Bentz and Berdicevskis (2016), the difference between L1 and L2 learners is conceptualized as the amount of exposure that participants receive to learn an artificial language. This is based on an experiment designed to further elicit the exact pathways of inflectional loss via imperfect learning (Berdicevskis and Semenuks, to appear). It is shown that having low-exposure learners in a transmission chain can reduce the unigram entropy of the artificial language due to reductions in the morphological marking system. The interaction between low- and high-exposure learners across several generations thus emerges as an important parameter determining the outcome of language change. Yet another linking mechanism between learning outcomes and language change is proposed in Atkinson (2016). Here, it is argued that the accommodation of high-exposure "native" speakers to the low-exposure "non-native" speakers drives the simplification of the language system, rather than the low-exposure learners by themselves.

Finally, the sociolinguistic setting of accommodation to non-standard varieties is also observed in research on so-called multiethnolects emerging in cities across Europe due to recent migration patterns (Wiese, 2006; Wiese and Rehbein, 2016; Wiese and Pohle, 2016). With regards to the question of whether children or adults drive language change, this line of inquiry suggests that the answer quite literally lies in between. Namely, it is adolescents accommodating to and adopting patterns of change that have come about by language contact. Another interesting finding is that even speakers exposed to the standard language from early childhood onwards will adopt and carry on the local multiethnic variety, thus further blurring the distinction between "native" and "non-native" usage patterns.

10.2 A note on grammaticalization

The last section discussed empirical studies on language learning and concluded that an important parameter for predicting outcomes of contact is the level of exposure that learners have when immersed in a population of language users. With regards to lexical diversity, we might speculate that low-exposure learners will be biased in their "sampling" of the range of word forms used by the population and might hence reduce their own usage to a set of salient base forms. High exposure, on the other hand, might increase the likelihood of "sampling" a wider range of word forms and hence help to preserve lexical diversity. In the latter case, lexical diversity might even systematically increase over time due to phenomena such as *grammaticalization*.

Looking at the diachronic evolution and change of inflectional markers, it has been observed that they are often derived from formerly free morphemes that are merged to verbs, nouns and adjectives (Heine and Kuteva, 2007; Hopper and Traugott, 2003; Lehmann, 1985). A typical example of grammaticalization is the Old English noun *līc* 'body', which, when co-occurring with adjectives, was phonetically reduced and became the productive derivational suffix *-ly* used to build new adverbs. Likewise, the inflectional future in Romance languages such as Italian *canterò* 'I will sing' derives from Latin *cantare habeo* 'I have to sing'. Examples for the evolution of noun morphology are the Hungarian inflectional elative and inessive case markers, which derive from a noun originally meaning 'interior' (Heine and Kuteva, 2007, p. 66).[10]

Such grammaticalization processes can drive languages to display a panoply of different word forms, as exemplified by conjugation classes. This is the case for the Latin to Italian example above, where grammaticalization has led to the evolution of an inflectional future tense with different forms according to person and number, e.g. *canterò*, *canterai*, *canterà*, *canteremo*, *canterete*, *canteranno*. Crucially, grammaticalization happens due to cognitive entrenchment of lexical items that frequently co-occur (Bybee, 2006, 2003). Over several generations of language learning and usage this might gradually increase the range of word forms in a language and as a consequence its lexical diversity. As a proof of concept, Bentz and Buttery (2014) use a simple "grammaticalization" model to simulate how languages might evolve from displaying short-tailed (low-entropy) towards long-tailed (high-entropy) word frequency distributions. This is achieved simply by merging highly co-occurring word tokens. For example, the highly frequent English bigram *of the* is merged to *ofthe*. Interestingly, this exact contraction is attested in German, where *von dem* is often rendered as *vom*.

Furthermore, in particular the grammaticalization sub-process of cliticization is typically associated with the loss of phonetic material, also called *erosion*. As argued by Schiering (2006, 2010), this is not necessarily a universal feat of cliti-

[10] Though it might be argued that these constitute "fused postpositions" rather than inflectional cases (Spencer, 2008).

cization, but rather dependent on the prosodic system, i.e. regular stress patterns, of the respective language. With this caveat in mind, we might speculate that if stress patterns do allow erosion, then it is more likely to occur between intimates, where a rich, shared background is given and communication is more robust to the loss of phonetic material. In contrast, the lingua franca situation will require information to be spelled out more explicitly. For example, the Standard German phrase *hast du es ihm gesagt?*, literally 'have you it him told?', can be reduced in colloquial German and German dialects to something akin to *hast's'm g'sagt?* Thus eroding (some) vowels and compressing the information into a shorter phonetic string. This strategy leads to faster transmission of the message, though to the expense of making it harder for the hearer to parse and decode it. Presumably, this strategy only works if learners have enough exposure to the communication system to even understand highly compressed messages. This would make it favourable in the esoteric niche and disfavoured in the exoteric niche.

To sum up the previous sections, the striking difference in lexical diversity of languages like Hawaiian and Iñupiatun might, at least partly, derive from the historical and social contexts in which these languages evolved. While learning of long-tailed word form distributions will always pose difficulties to learners, be it children or adults, early learners might have the advantage of longer and more intensive exposure. They eventually converge onto the usage of forms common in the society they are surrounded by. After convergence, as high-exposure adults, they might even further drive lexical diversity by merging free forms that are frequently co-occurring. Late learners, on the other hand, are at a disadvantage just by virtue of being less exposed to the long tail of word forms. Crucially, this means that we do not necessarily have to posit a qualitative difference between L1 and L2 learning, or "native" and "non-native" learners for that matter. It might be sufficient to further elicit which populations are prone to feature high- and low-exposure learners. This has the potential to explain differential pathways of how lexical diversity can change and evolve.

10.3 Beyond words

As pointed out earlier in this book, using words as basic information encoding units is problematic from a typologically informed point of view. Both theoretical considerations and computational modelling suggest that, after all, words might not play an indispensable role for natural language processing and human comprehension alike. However, note that differences in word entropies are likely to have reflections in language data quite generally, independent of the exact definition of information encoding units.

For example, instead of looking at single word tokens, i.e. unigrams, we could follow Grefenstette (2010)'s rationale and investigate combinations of unigrams, e.g. bigrams, instead. In the English PBC, the bigrams with the highest frequencies are: *and he* (93), *and they* (79), *to him* (76). In German these are: *und sagte* 'and said' (24), *und die* 'and the' (19), *zu ihm* 'to him' (19). These examples already suggest that English has systematically higher bigram frequencies, i.e. lower bigram entropy, than German – in parallel to unigram entropies. In fact, Figure 10.2[11] illustrates that there is generally a strong positive (though non-linear) relationsip between the scaled entropy of unigrams and bigrams for languages of both the UDHR and PBC.

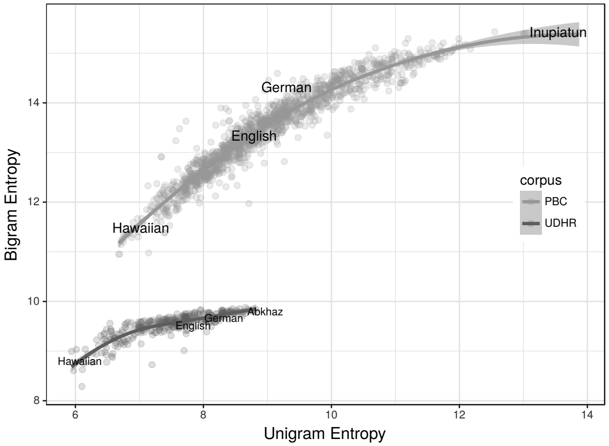

Figure 10.2: Entropies (H_shrink) for unigram distributions (x-axis) versus bigram distributions (y-axis). Values are calculated for both the languages of the UDHR (dark grey) and the PBC (light grey). Non-linear loess smoothers are overlaid with 95% confidence intervals. The language names for Hawaiian, English, German, and Iñupiatun of the PBC, and for Hawaiian, English, German, and Abkhaz of the UDHR are plotted in rough agreement with the corresponding points.

Hawaiian, English, German and Iñupiatun[12] are here pointed out as examples. Note that bigram entropies apparently hit a ceiling towards an entropy of ca. 16 bits for the PBC and at around 10 bits for the UDHR. However, the crucial point of

[11] Rcode/Chapter10/entropyBigrams.R

[12] Abkhaz is given for the UDHR instead.

Figure 10.2 is that entropies at one level of information encoding, i.e. unigrams, are systematically linked to entropies at another level, i.e. bigrams. It is an interesting avenue for further research to look at entropies beyond unigrams. This avenue of research is explored in a range of recent studies (Montemurro and Zanette, 2011, 2016; Koplenig et al., 2017; Bentz et al., 2017a).

From the perspective of a learner or an automated comprehension system it can make sense to perceive highly entrenched expressions such as *he says* as a single unit, thus rendering the white space as an idiosyncrasy of orthography. However, faced with different distributions of characters, unigrams, bigrams, or n-grams more generally, there is still an overarching observation: in some languages the rate of repetition of information encoding units is higher than in others. For future research, it will be interesting to see how entropy changes with an increasing window size of information encoding units, i.e. from characters and words, to phrases and sentences. It seems reasonable to assume that information encoding potentials at different levels of linguistic structure will be interdependent and potentially balance out – to a certain extent.

10.3.1 Complexity trade-offs and equi-complexity

This idea has deep roots in the history of linguistic reasoning. Especially in the discussions surrounding *language complexity*, the hypothesis that complexities at different levels of language structure balance out is often either wholeheartedly supported or fiercely rejected. In connection with the recent rise of interest in language complexity (Sampson et al., 2009; Dahl, 2004; Newmeyer and Preston, 2014b; Miestamo et al., 2008) came the questioning of the equi-complexity hypothesis. Sampson (2009) and Newmeyer and Preston (2014a) give an overview of the historical literature on the topic.

Generally speaking, there seems to have been a strong consensus among linguists at least since the 1950s that a) complexities at different levels of linguistic structure, such as phonemes, syllables, morphemes, words and phrases, correlate with each other; and that b) all languages are ultimately equally complex. The former is henceforth referred to as the *trade-off hypothesis* and the latter as the *equi-complexity hypothesis*. These hypotheses are often seen as strongly interlinked. However, as pointed out by Fenk-Oczlon and Fenk (2011, 2014), they are, in fact, logically independent of each other. Different levels of complexity can be correlated *within* languages, while still summing up to a difference in overall complexity *between* languages.

With regards to the *trade-off hypothesis* Fenk-Oczlon and Fenk (2008) illustrate across different samples of languages that the following relationships hold (among others): *fewer phonemes* per syllable correspond to *more syllables* per word and *fewer syllables* per word correspond to *more words* per clause. These correlations suggest clear trade-offs between information encoding at different structural levels. If there are few phonemes to encode information at the syllable level, then this is counterbalanced by having more syllables per word. Likewise, if there are fewer syllables per word, then this is counterbalanced by having more words per clause. Clearly, there will be lower and upper limits to the range of phonemes per syllable, syllables per word, and words per clause, due to languages generally being shaped to fit information encoding along a communication channel, which has also been conceptualized as a *bottleneck* (Christiansen and Chater, 2016). However, within that range, languages are able to adapt to pressures of learning. In fact, even within the same language, speakers might harness the flexibility of encoding strategies to keep the information flow relatively constant. This proposal has been put forward by Fenk and Fenk (1980) and elaborated in Fenk-Oczlon (2001). It has become known as the *Uniform Information Density* hypothesis (Frank and Jaeger, 2008; Jaeger, 2010). For critical reviews of this proposal see Ferrer-i-Cancho et al. (2013) and Ferrer-i-Cancho (2017c).

A further well-known trade-off that has been suggested in the literature refers to core argument marking: languages tend to either mark core arguments by case inflections (i.e. morphologically), or by rigid word order (i.e. syntactically). Sinnemäki (2008, 2014) has provided quantitative evidence for this observation. In a stratified sample of 50 languages, the presence of case marking strongly predicted the absence of rigid word order and the presence of rigid word order strongly predicted the absence of case marking (Sinnemäki, 2014, p. 192).

10.3.2 Entropy at different levels of language structure

More specifically with regards to information theory, Juola (1998, 2008) has developed a measure of morphological complexity and a measure of linguistic complexity more generally by applying off-the-shelf file compression algorithms to parallel texts. This follows the rationale outlined earlier: entropy is the upper bound on compressibility. A string of characters, or vocabulary of character strings (i.e. word types), that has zero entropy is maximally compressible. A string of characters of equal probability, or a vocabulary of character strings of equal probability, has maximum entropy and is minimally compressible. Entropy reflects the information encoding potential of a code, i.e. a distribution of information encoding units and their probabilities. Lossless compression is only possible to the point where the code is optimally used and this optimal usage is represented by the entropy of the code.

Against this theoretical backdrop, Juola (2008) applies a standard compression algorithm to 24 parallel translations of the Bible in 14 different languages to assess how much redundant, i.e. compressible, information there is in each. The first main finding is that after compression the size of the texts (in bytes) bears 94% less variance than before. This is interpreted as evidence for the equi-complexity hypothesis. All texts, and by extension the languages they represent, are roughly equivalent in terms of compressed file size, i.e. after a near-optimal code per text has been found. Furthermore, Juola argues that it is possible to tease apart the influence of morphological, syntactic and pragmatic information by separately and randomly deleting 10% of characters, 10% of word tokens, and 10% of verses and then again assessing file size differences before and after compression.

To see this, remember that Iñupiatun is a high-entropy language with many word types, whereas Hawaiian is a low-entropy language with fewer word types. Randomly deleting characters from word tokens in a given text will create new word types and get the text closer to the maximum entropy state. In the case of Iñupiatun this will have less of an effect than for Hawaiian, since Iñupiatun is already closer to the maximum entropy state than Hawaiian. In other words, considerably worse compression ratios after random deletion of characters indicate less complex “morphology”. Note, however, that this a crude definition of morphology. Remember from Section 5.2.4 that inflectional marking accounts for around 50% in the variance of word frequency distributions (across 19 languages) and inflectional and derivational morphology together explain around 75% of the variance in word frequency distributions for English and German. This means there is around 25% to 50% variance in word frequency distributions that is not due to morphological marking, but rather to the basic vocabulary, as well as other factors not accounted for. These will also play a role for the compressibility of texts. A more precise term for Juola’s measure of morphological complexity is then complexity of *within-word information*.

Similarly, if a language uses certain multiword expressions frequently, e.g. the periphrastic future tense construction *is going to* in English, then random deletion of words will cause more of a loss in compressibility than in a language with fewer multiword expressions. Juola (2008) interprets this as a reflexion of “syntax”, though it is unclear how the usage of multiword expressions is related to more traditional notions of syntax, or typological notions of word order such as the order of subject, verb, and object. In the multiword sense, English has relatively complex “syntax”, due to a relatively high number of fixed multiword expressions and constructions. This will here be referred to as complexity of *between-word information*.

The trade-off between these within-word and between-word complexities can be seen in Figure 10.3. This figure is based on the numbers in Table 9 of Juola

(2008). The x-axis reflects the ratio of compressed byte size (CBS) after characters are randomly replaced (CBS_c) to the compressed byte size of the original text (CBS_o), i.e. $\frac{\mathrm{CBS}_c}{\mathrm{CBS}_o}$. For example, in the original table by Juola, the Bible in Basic English (BBE) has a ratio of 1.18, i.e. the compressed byte size after character replacement is 1.18 times bigger than in the original text. This is most likely explained by the fact that randomly replacing characters leads to the creation of more word types. The effect is strongest for languages that have relatively few word types to start with, e.g. the Bible in Basic English (BBE) and Bahasa Indonesian (bis). Importantly, note that higher values of the ratio given by Juola correspond to *less* within-word information. In order to make higher values on the x-axis correspond to *more* within-word information, the inverse of these values is plotted in Figure 10.3.[13] Thus the Bible in Basic English gets a value of $\frac{1}{1.18} = 0.85$.

On the other hand, random deletion of word tokens decreases the information encoding potential of the texts. The third column of Table 9 in Juola (2008) gives the compressed byte sizes after 10% of word tokens have been randomly deleted, divided by the compressed byte size of the original texts, i.e. $\frac{\mathrm{CBS}_w}{\mathrm{CBS}_o}$. These ratios range from 0.94 to 0.98, meaning that texts are generally smaller in compressed byte size after random word token deletion. This is likely related to the fact that the deletion of word tokens generally decreases the information encoded in the text. However, for texts in which multiword expressions are more prominent, word token deletion has less of an effect, i.e. decreases the compression size less, and the values stay closer to one. Thus, higher values on the y-axis correspond to *more* between-word information.

Within-word information correlates negatively with between-word information across the 24 parallel texts representing 14 different languages. The Pearson correlation is strong, despite few data points ($r = -0.75, p < 0.0001$). Of course, considering that 8 of the texts are actually English versions and 19 are written in Indo-European languages, independence of data points is not given.

A recent study by Ehret and Szmrecsanyi (2016a) uses a complexity metric inspired by Juola (2008)'s and confirms these negative correlations for 16 translations of the Bible, 9 translations of Alice in Wonderland, and 9 non-parallel newspaper texts. In contrast to Juola (2008), however, Ehret and Szmrecsanyi (2016a) do not argue that overall complexity scores are essentially the same for all languages. Rather, they rank languages from highest to lowest adjusted overall complexity scores and hence provide evidence against the equi-complexity hypothesis.

[13] Rcode/Chapter10/JuolaAnalyses.R

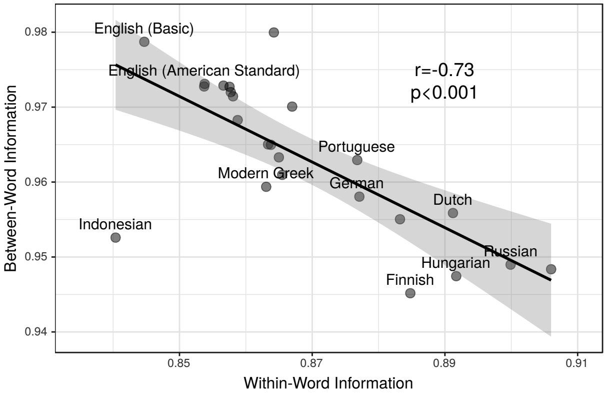

Figure 10.3: Trade-off between within-word and between-word information encoding for Bible texts based on Juola (2008). The x-axis represents the within-word information index and the y-axis the between-word information index as derived on the basis of compression ratios. A linear model with 95% confidence intervals is overlaid. Some data points are labelled by language name.

Overall, Juola (1998, 2008) takes the credit for being the first to encounter the problem of measuring language complexity within an information-theoretic, quantitative, and hence reproducible account. The methods and overall results of his and also Ehret and Szmrecsanyi (2016a)'s analyses are intuitively plausible. Still, more work needs to be done to yield interpretations of these results from a more linguistically informed angle. Ziv-Lempel (Ziv and Lempel, 1977) and related compression algorithms applied to text files purely aim to compress a string of character encodings (unicode or otherwise) to a string of minimum possible length, without loss of information. It is, as yet, unclear how this exactly relates to linguistically meaningful concepts, such as grapheme diversity, inflectional marking, or word order (beyond multiword constructions). The analyses in Ehret (2016) as well as Koplenig et al. (2017) are a further important steps in this direction.

Finally, Montemurro and Zanette (2011) can be seen, to date, as one of the most extensive accounts of cross-linguistic estimation of overall entropy. They estimate the overall entropy of 8 languages from more than 5000 texts. Following a similar information-theoretic rationale as Juola (1998, 2008), their approximations are based on an entropy rate estimator also going back to the work by Ziv and Lempel (Kontoyiannis et al., 1998). They tease apart the amount of information encoding potential carried by unigrams and information encoding potential carried by regularities beyond the unigram level. Interestingly, it turns out that both the overall entropies (denoted H) and unigram entropies (denoted H_s) differ between languages, thus again rejecting the equi-complexity hypothesis. Also, the difference between these two types of entropy, i.e. $H - H_s$, is shown to be remarkably constant across the different languages. This finding is affirmed by Montemurro and Zanette (2016) with Bible translations into 75 languages and by Bentz et al. (2017a) with a sample of the Parallel Bible Corpus featuring 1259 languages.

10.4 Summary

Several problems and caveats relating to the methods and assumptions of this book have been discussed in this chapter. First of all, in order for an information-theoretic account of linguistic diversity to be complete solid links need to be established between entropic measures and language learning. A preliminarily overview of the relevant literature on language learning by children and adults was given here. More specifically, it is important to understand the role unigram entropies play for word and word-form learning. In this context, the experimental literature suggests that the learning pressures that children and adults impose on a language might, after all, turn out to be similar. A high-entropy, long-tailed distribution of information encoding units is going to be harder to learn than a low-entropy, short-tailed distribution – everything else being equal.

This raises an apparent paradox: if both children and adults find it hard to learn high-entropy languages, then how can these come into existence in the first place? And how come they are maintained over generations and generations of learning? The solution offered is that learning pressures can differ not only along the dimension of child versus adult learning, but also along other dimensions, such as exposure. Children might be prone to overgeneralize particular morphological markers, but they will eventually converge onto the usage patterns of the adults they are surrounded by. Later in life they might become high-exposure adults who, via processes such as grammaticalization, even increase lexical diversity. Hence, in societies with high-exposure learners, speakers, and signers, i.e. close-knit, small groups of intimates, lexical diversity might get closer to the maximum value sustainable with regards to memory (and other) constraints. On the other hand, in big, open societies with more low-exposure learners (children and adults) the pressure onto the language to move towards low lexical diversity is bound to be higher. This can derive simply from the fact that, in such open societies, the average amount of exposure to the long-tailed distribution of word forms is shorter.

A further problem discussed in this chapter is related to the so-called "indeterminacy of words". That is, the issue of coherently defining what a "word" really is, while taking into account the psycholinguistic, descriptive and technical perspectives alike. It seems impossible to define the concept of a "word" while accommodating for all of these perspectives at the same time. At first sight, this seems discouraging with regards to developing objective measures of cross-linguistic variance in lexical diversities. Maybe it is consoling that the problem of overarching and clear-cut definitions is not confined to the language sciences, but permeates science in general. There is no coherent, universally agreed-upon definition of the term “species”, but this does not prevent biologists from determining the diversity of species in habitats, in fact, by using the exact same entropy measures that were applied here too.

From a less theoretical and more pragmatic point of view, the question is not so much *what* a word exactly is, or if an overarching definition is even possible, but *how* our decisions when defining the unit to be measured affect the outcome of the measurement. In this sense, there is an overall picture emerging: some languages have a higher rate of repetition of word types than others.

Clearly, it is expected that the rate of repetition can vary not only between languages, but also within a language at different levels of information encoding. Some preliminary evidence for such trade-offs was discussed here. One of the most interesting avenues for further research is to measure such information encoding trade-offs across large language samples.

11 Conclusions: Universality and Diversity

In the introduction to this book, some examples were given for the astonishing diversity of linguistic encoding strategies we find across languages of the world. The age-old and ongoing debate about words for snow in different cultures reflects one facet of this diversity at the level of vocabulary. Likewise, there are languages with excessive morphological marking, an extreme example being the alleged 1.5 million verb forms in the Caucasian language Archi. Such claims raise interest among editors of the Guinness Book of Records and eyebrows among typologists. Further information encoding strategies, such as compounding, can likewise increase or decrease lexical diveristy. To the bemusement of the general public, the German government introduced the *Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz* in the 1990s, and hence took the compounding capacity of the language to its extremes. The representativeness of such extreme examples aside, there is no denying that languages, that is, their speakers and signers, "choose" widely differing strategies to encode information.

Going beyond cherry-picked examples of diversity, this book aimed to establish methods to measure the cross-linguistic difference in linguistic encoding at the level of words. Lexical diversity can be pinned down precisely given a defined set of texts and a workable definition of the basic information encoding unit: the word. Shannon entropy is a measure reflecting the information encoding capacity of any code, human languages included. A minimum word entropy language would repeat the same word over and over again, thus being maximally learnable but minimally expressive. A maximum entropy language would have a separate word for each concept, thus being minimally learnable but maximally expressive. Of course, as pointed out in the very first sentence of this book: *words have two sides*. One side is their information encoding capacity reflected by repeating patterns realized in their physical appearance, that is, the sounds and characters they are shaped of. The other side is the co-activation with other visual, auditory, and olfactory input they cause in human brains, that is, their meaning. As recent studies have suggested, the systematic entrenchment of such co-activation in word learning relies on a multi-modal and rich environment. There are natural constraints to the range of word-based encoding strategies languages can adopt. The limit of human memory is just one of the most obvious.

As a consequence, linguistic encoding strategies are the outcome of multiple competing factors. They have to be expressive, but also learnable and usable, often under time pressure. Considering such competing pressures, it is not surprising that languages do not seem to exploit the full range of possible entropies. In terms of word entropies, the 1217 languages included in this book only cover

DOI 10.1515/9783110560107-011

around 40% of the theoretically possible spectrum. In order for a language to move towards the extreme ends of this spectrum, it would have to undergo lopsided pressure from a single factor. This can be simulated in artificial learning experiments and computational models. For instance, if only pressure for learning is given, artificial languages tend to collapse into low-entropy states. In natural languages, however, this is prevented by a drive for expressivity, counterbalancing the learnability pressure.

The interaction of multiple pressures is probably also the reason for why word entropies across many languages display a unimodal density distribution. In most cases, learning pressures and expressivity pressures (among others) keep each other in check, and the majority of languages fall close to the mean entropy around 9 bits/word. Only if a particular pressure is prevalent throughout the history of a language, will that language be pushed away from the mean. Hawaiian and Iñupiatun were given as interesting outliers at the low- and high-entropy ends.

Overall, the factors that shape lexical diversity were subsumed under three categories: *descriptive*, *explanatory* and *grouping* factors. In the first category, we find such linguistic properties as script, word-formation, as well as register and style. Explanatory factors, on the other hand, are related to population structure, and, ultimately, to learning and usage preferences of subpopulations. Moreover, grouping factors are geographic and genealogical patterns that arise from the drift and dispersal of populations throughout time and space. If we want to answer *why* Hawaiian and Iñupiatun have vastly divergent lexical diversities, we could refer to the relative lack of morphological marking in Hawaiian, keeping the number of word forms small, while in Iñupiatun morphological marking is more productive, creating a vast array of word forms. This constitutes a *descriptive* answer. An answer with reference to population structure could be that Hawaiian had a higher percentage of low-exposure learners in its history compared to Iñupiatun. Low-exposure learning scenarios are likely to introduce learning pressure to the encoding system. This type of argumentation was called the *explanatory* perspective here. Finally, rather than having to do with historically shallow processes of language change, the reason for the discrepancy might reach further back in time to Proto-Austronesian and Proto-Eskimo-Aleut. Maybe the social settings in which these proto-languages evolved were vastly different in terms of esoteric and exoteric learning and usage scenarios. Any languages evolving from these ancestor populations, not just Hawaiian and Iñupiatun, would then potentially still reflect such biases. This requires us to also take genealogical grouping into account.

In this account, we have not included any measure of entropy beyond the word level. It seems very likely that languages will balance out some of the differences at other levels of encoding. This is strongly supported by quantitative studies that have ventured to measure the complexity of languages and their subsystems. However, these studies also suggest that the trade-off is unlikely to be perfect, in the sense of leading to overall equi-complexity.

Maintaining that a language like Hawaiian has lower entropy and is hence less expressive than a language like Iñupiatun, will leave a bad taste in the mouth of many linguists. It feels too much like 19th century linguistic and anthropological theories of Western supremacy. However, the overarching framework of languages as complex adaptive systems does not support an absolute statement of the type: complex languages are better than simple languages – or the other way around. Rather, languages adapt to the sociolinguistic niches they are spoken, written, and signed in. From a CAS perspective, low-entropy languages have adapted to be learnable and high-entropy languages have adapted to be expressive. Arguing that one language is better than another is like arguing that birds are better than fish – pointless.

This also sheds new light on the "universalist" versus "variationist" opposition. There seems to have been a clear preference for the former in the last decades of linguistic inquiry. One of the most ubiquitous statements in introductions to language evolution papers and textbooks is that "language is what makes us human". Hence, surely there must be something universal across all human languages, unique to us, something that we can put our finger on when we claim that human languages are categorically different from any other communication system. From this perspective, we are biased to think that variation in languages is just an epiphenomenon, a veil over the underlying unity.

However, maybe this veil is the most interesting aspect of human languages after all. Maybe the diversity of human languages is exceptional among communication systems. Compared to the vast majority of animals, humans have adapted to live in almost any environment on the face of this planet. Their most versatile and powerful tool, language, has adapted to fit the changing requirements of the human niche. In this case, a scientific approach to languages does not require forcing ourselves to believe that they are ultimately all the same, but to embrace, protect, and explain their diversity.

12 Appendix A: Advanced Entropy Estimators

The following discussion is largely based on Hausser and Strimmer (2009), and reprinted with slight modifications from Bentz et al. (2017a) Appendix B.

The Miller-Madow estimator

The *Miller-Madow* (MM) estimator (Hausser and Strimmer, 2009, p. 1471) aims to reduce the estimation bias by adding a correction to the ML estimated entropy such that

$$\hat{H}^{MM} = \hat{H}^{ML} + \frac{M_{>0} - 1}{2N}, \tag{12.1}$$

where $M_{>0}$ refers to the number of types with token frequencies > 0, i.e. V in our definition. Note that the corrective $\frac{V-1}{2N}$ is relatively big for the $N < V$ scenario, and relatively small for the $N > V$ scenario. Hence, it counterbalances the underestimation bias in the $N < V$ scenario of small text sizes.

Bayesian estimators

Another set of estimators derives from estimating p_i within a Bayesian framework using the Dirichlet distribution with $a_1, a_2, \ldots, a_V$ as priors such that

$$\hat{p}(w_i)^{Bayes} = \frac{f_i + a_i}{N + A}, \tag{12.2}$$

where a_i values essentially "flatten out" the distribution of frequency counts to overcome the bias towards short tailed distributions (with small V). In parallel to N we have $A = \sum_{i=1}^{V} a_i$ added to the denominator (Agresti and Hitchcock, 2005, p. 302-303).

Now, depending on which priors exactly we choose, we end up with different estimated entropies (see also Table 1 in Hausser and Strimmer (2009, p. 1471)). A uniform *Jeffreys prior* of $a_i = 1/2$ gives us $\hat{H}^{Jeff}$, a uniform *Laplace prior* of $a_i = 1$ gives us $\hat{H}^{Lap}$, a uniform *Perks prior* of $a_i = 1/V$ gives us $\hat{H}^{SG}$, after Schürmann and Grassberger (1996), who proposed to use this prior. Finally, the so-called minimax prior of $a_i = \sqrt{N/V}$ yields $\hat{H}^{minmax}$.

Furthermore, the most recent – and arguably least biased – entropy estimator based on a Bayesian framework is the *Nemenman-Shafee-Bialek* (NSB) estimator.

DOI 10.1515/9783110560107-012

Nemenman et al. (2002, p. 5) illustrate that the entropies estimated with the other priors proposed above will be strongly influenced by the prior distributions and only recover after a relatively big number of tokens has been sampled. Instead of directly using any specific Dirichlet prior, they form priors as weighted sums of the different Dirichlet priors, which they call *infinite Dirichlet mixture priors*. The resulting entropy estimates $\hat{H}^{NSB}$ turn out to be robust across the whole range of sample sizes.

The Chao-Shen estimator

Chao and Shen (2003, p. 432) propose to overcome the problem of overestimating the probability of each type (in their case species instead of word types) by first estimating the so-called sample coverage as

$$\hat{C} = 1 - \frac{m_1}{N}, \tag{12.3}$$

where m_1 is the number of types with frequency 1 in the sample (i.e. *hapax legomena*). The idea is that the number of types not represented by tokens is roughly the same as the number of types with frequency 1. In this case, the sample coverage reflects the conditional probability of getting a new type if a token is added to the sample N. This probability is then multiplied with the simple ML estimate $\hat{p}(w_i)^{ML}$ to get the so-called *Good-Turing* estimated probability of a type

$$\hat{p}(w_i)^{GT} = \left(1 - \frac{m_1}{N}\right) \hat{p}(w_i)^{ML}. \tag{12.4}$$

Furthermore, Chao and Shen (2003, p. 431) suggest to use the *Horvitz-Thompson estimator* to modify the estimated entropy $\hat{H}^{ML}$. This estimator is based on the rationale that if N tokens have been sampled with replacement, then the probability of the i^{th} type not being represented by a specific token is $1 - \hat{p}(w_i)^{GT}$. Thus, the probability of the i^{th} type not being represented by any token is $(1 - \hat{p}(w_i)^{GT})^N$, and, inversely, the probability of it being included in a sample of N tokens is $1 - (1 - \hat{p}(w_i)^{GT})^N$. The full specification of the Horvitz-Thompson estimator, with Good-Turing probability estimates, is then

$$\hat{H}^{CS} = -K \sum_{i=1}^{r} \frac{\hat{p}(w_i)^{GT} log_2(\hat{p}(w_i)^{GT})}{1 - (1 - \hat{p}(w_i)^{GT})}. \tag{12.5}$$

The James-Stein shrinkage estimator

Finally, Hausser and Strimmer (2009, p. 1472) put forward an entropy estimator based on the so-called *James-Stein shrinkage*. According to this approach the estimated probability per type is

$$\hat{p}(w_i)^{\mathit{shrink}} = \lambda\hat{p}(w_i)^{\mathit{target}} + (1-\lambda)\hat{p}(w_i)^{\mathit{ML}}, \tag{12.6}$$

where $\lambda \in [0,1]$ is the shrinkage intensity and $\hat{p}(w_i)^{\mathit{target}}$ is the so-called "shrinkage target". Hausser and Strimmer (2009, p. 1473) suggest to use the maximum entropy distribution as a target, i.e. $\hat{p}(w_i)^{\mathit{target}} = \frac{1}{V}$. This yields

$$\hat{p}(w_i)^{\mathit{shrink}} = \frac{\lambda}{V} + (1-\lambda)\hat{p}(w_i)^{\mathit{ML}}. \tag{12.7}$$

The idea here is that the estimated probability $\hat{p}(w_i)^{\mathit{shrink}}$ consists of two additive components, $\frac{\lambda}{V}$ and $(1-\lambda)\hat{p}(w_i)^{\mathit{ML}}$ respectively. In the full shrinkage case ($\lambda = 1$) Equation 12.7 yields

$$\hat{p}(w_i)^{\mathit{shrink}} = \frac{1}{V}, \tag{12.8}$$

i.e. the maximum entropy. In the lowest shrinkage case ($\lambda = 0$) Equation 12.7 yields

$$\hat{p}(w_i)^{\mathit{shrink}} = \hat{p}(w_i)^{\mathit{ML}}, \tag{12.9}$$

i.e. the ML estimation that is biased towards low entropy. Given empirical data, the true probability is very likely to lie somewhere in between these two cases and hence $0 < \lambda < 1$. In fact, Hausser and Strimmer (2009, p. 1481) show that the optimal shrinkage $\lambda^{\star}$ can be calculated analytically and without knowing the true probabilities p_i. Given the optimal shrinkage, the probability p_i^{shrink} can then be plugged into the original entropy equation to yield

$$\hat{H}^{\mathit{shrink}} = -K\sum_{i=1}^{r}\hat{p}(w_i)^{\mathit{shrink}} log_2(\hat{p}(w_i)^{\mathit{shrink}}). \tag{12.10}$$

13 Appendix B: Multiple Regression Assumptions

Linearity

The multiple regression run here assumes that a linear relationship holds between the dependent variable ($\hat{H}^{\text{scaled}}$) and the predictor variables (population size, L2 percentage, language status). That this is the case can be seen in Figure 13.1.[1] Here, a local regression smoother (loess) is overlaid onto the linear regression line. The confidence intervals of these should always overlap. The only (almost) significant divergence from linearity is to be found for the relationship between scaled entropy and language status. The local regression smoother is relatively sensible to variation in the data. Accounting for these idiosyncratic patterns by introducing non-linear relationships would overfit the data.

Normality

Another model assumption is that the errors (residuals) are approximately normally distributed. This assumption can be checked with reference to Figure 13.2. There is a slight left-skew away from the mean value. Arguably, this is a minor deviation.

Homoscedasticity

It is further assumed that the variation of residuals is relatively uniform across fitted values, i.e. the deviation from fitted values should not exhibit any clear trends. This is also called *homoscedasticity* assumption. To check this, fitted values of the multiple regression model are plotted versus their residuals (Figure 13.3). The confidence intervals of a local regression model through these data points should always include the zero line. This is the case here, meaning that the residuals do not display any so-called heteroscedasticity.

Multicollinearity

Since there are three predictors in the multiple regression model, it needs to be checked whether correlations between them (e.g. the one between population size and language status) might cause an undue inflation of the variance. The variance inflation factor (VIF) for the model – calculated with *R* package *fmsb* (Nakazawa, 2015) – is relatively low (1.28). A value of 2 is normally given as threshold where we need to start worrying about multicollinearity.

[1] file: Rcode/Chapter8/entropyMultiReg.R

DOI 10.1515/9783110560107-013

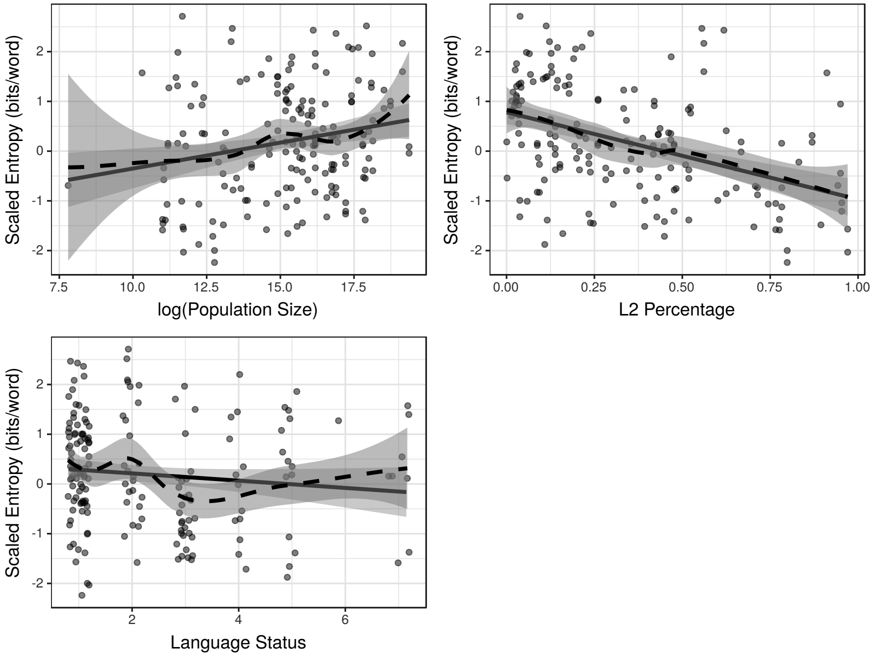

Figure 13.1: Linearity assumption. Relationship between scaled unigram entropy and population size (upper left panel), L2% (upper right panel), and language status (lower left panel). The linear models (full lines) and their 95% confidence intervals (transparent grey) generally overlap with the local regression smoother "loess" (dashed line) which is sensitive to nonlinearities in the data.

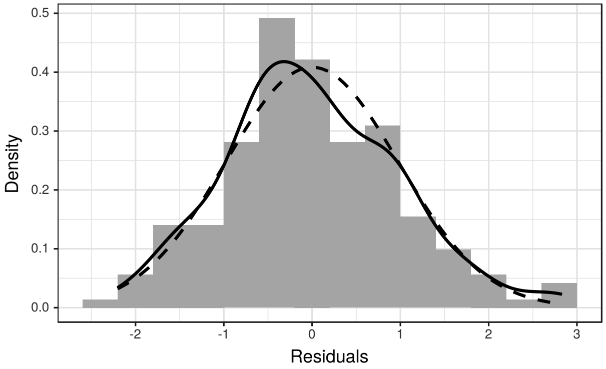

Figure 13.2: Normality assumption. Distribution of residuals for the multiple regression model. The approximated empirical density curves (full line) follow closely the theoretical density curves (dashed).

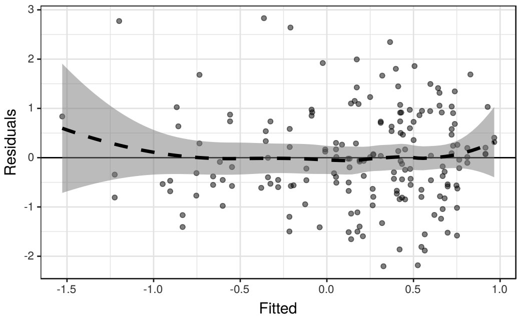

Figure 13.3: Homoscedasticity assumption. Plot of fitted values versus residuals for the multiple regression model. The dashed line indicates a local regression model with 95% confidence intervals (transparent grey).

14 Appendix C: Mixed-effects Regression Assumptions

Just as for multiple regression, the requirements of linearity, homoscedasticity, and normality of the residuals have to be met in the case of a mixed-effects regression (Winter, 2013; Baayen et al., 2008; Jaeger et al., 2011). Additionally, the so-called *Best Linear Unbiased Predictors* (BLUPs), i.e. the adjustments of the intercepts and slopes of the random effects, should be normally distributed (Jaeger et al., 2011, p. 294).

Linearity

The linearity of the relationship between L2 percentage and scaled entropies is already illustrated in the Appendix13. The linearity assessed for the multiple regression also applies here, since we are dealing with the same data.

Normality

Normality of residuals can be checked with reference to Figure 14.1.

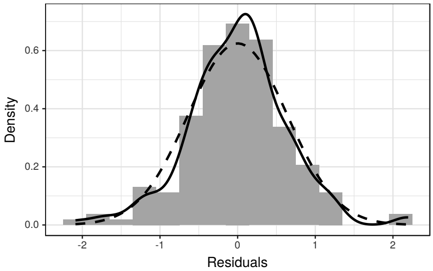

Figure 14.1: Normality assumption. Distribution of residuals for the mixed- effects regression model. The approximated empirical density curve (black line) generally follows the theoretical density curve (dashed).

Homoscedasticity

Homoscedasticity means that the variation of residuals is relatively uniform across fitted values. To check this, fitted values of the multiple regression model are plotted versus their residuals (Figure 14.2). The confidence intervals of a local regression model should include the zero line, which is the case here.

DOI 10.1515/9783110560107-014

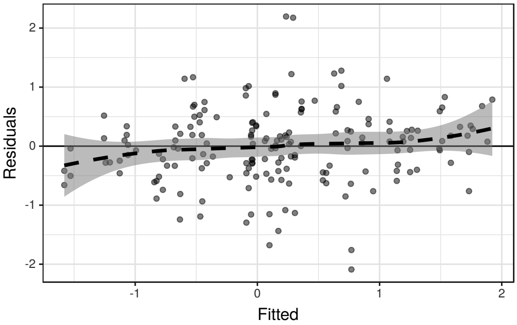

Figure 14.2: Homoscedasticity assumption. Plot of fitted values versus residuals for the multiple regression model. The black dashed line indicates a local regression model with 95% confidence intervals (transparent grey).

Normality of BLUPs

Besides assumptions that have to be met in linear regression models in general, mixed-effects models also require that the adjustments by random intercepts and slopes (BLUPs) are normally distributed. This can be checked by means of quantile-quantile plots (Figure 14.3). The points representing BLUPs should roughly fall on a line in between the standard normal quantiles from –2 to 2 (on the y-axis) (Jaeger et al., 2011, p. 294). More precisely, it should be possible to fit a line through all the confidence intervals of BLUPs. This is possible here.

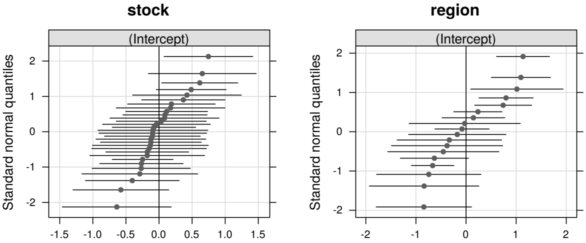

Figure 14.3: Distribution of BLUPs. Plot of the adjustments by random intercepts for stocks and regions (areas). The black lines associated with each dot indicate 95% confidence intervals.

Bibliography

Ackerman, F. and Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language*, 89(3):429–464.

Agresti, A. and Hitchcock, D. B. (2005). Bayesian inference for categorical data analysis. *Statistical Methods and Applications*, 14(3):297–330.

Aikhenvald, A. Y. (2003). *A grammar of Tariana, from Northwest Amazonia*. Cambridge University Press, Cambridge.

Aikhenvald, A. Y. (2007). Typological distinctions in word-formation. *Language typology and syntactic description*, 3:1–65.

Altmann, E. G., Dias, L., and Gerlach, M. (2017). Generalized entropies and the similarity of texts. *Journal of Statistical Mechanics: Theory and Experiment*, page 014002.

Altmann, E. G. and Gerlach, M. (2016). Statistical laws in linguistics. In *Creativity and Universality in Language*, pages 7–26. Springer.

Ambridge, B., Kidd, E., Rowland, C. F., and Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of child language*, 42(02):239–273.

Atkinson, M. D. (2016). *Sociocultural determination of linguistic complexity*. PhD thesis, The University of Edinburgh.

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, 332:346–349.

Baayen, H. R. (2001). *Word frequency distributions*. Kluwer, Dordrecht, Boston & London.

Baayen, H. R. (2008). *Analyzing linguistic data: A practical introduction using R*. Cambridge University Press, Cambridge.

Baayen, H. R., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Baayen, R. H. (1994). Derivational productivity and text typology. *Journal of Quantitative Linguistics*, 1(1):16–34.

Baayen, R. H. (2014). Multivariate statistics. In Podesva, R. J. and Sharma, D., editors, *Research methods in linguistics*, chapter 16, pages 337–373. Cambridge: Cambridge University Press.

Baayen, R. H., Shaoul, C., Willits, J., and Ramscar, M. (2016). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience*, 31(1):106–128.

Baechler, R. and Seiler, G., editors (2016). *Complexity, Isolation, and Variation*, volume 57 of *linguae & litterae*. Walter de Gruyter, Berlin/Boston.

Baixeries, J., Elvevåg, B., and Ferrer-i Cancho, R. (2013). The evolution of the exponent of Zipf's law in language ontogeny. *PloS ONE*, 8(3):e53227.

Bane, M. (2008). Quantifying and measuring morphological complexity. In *Proceedings of the 26th west coast conference on formal linguistics*, pages 69–76. Somerville, MA, USA: Cascadilla Proceedings Project.

Baroni, M. (2009). Distributions in text. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics. An international handbook*, chapter 39, pages 803–821. Mouton de Gruyter, Berlin, New York.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

DOI 10.1515/9783110560107-014

Bartoń, K. (2015). *MuMIn: Multi-Model Inference*. R package version 1.15.1.

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2015). Parsimonious mixed models. *arxiv pre-print*, page 1506.04967.

Bates, D., Maechler, M., and Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes*. R package.

Bauer, T. (1996). Arabic writing. In Daniels, P. T. and Bright, W., editors, *The world's writing systems*, chapter 50, pages 559–568. Oxford University Press, Oxford.

Baxter, G. J., Blythe, R. A., Croft, W., and McKane, A. J. (2006). Utterance selection model of language change. *Physical Review E*, 73(4):046118.

Baxter, G. J., Blythe, R. A., Croft, W., and McKane, A. J. (2009). Modeling language change: An evaluation of Trudgill's theory of the emergence of New Zealand English. *Language Variation and Change*, 21(02):257–296.

Beckner, C., Ellis, N. C., Blythe, R., Holland, J., Bybee, J., Christiansen, M. H., Larsen-Freeman, D., Croft, W., and Schoenemann, T. (2009). Language is a complex adaptive system. *Language Learning*, 59(December):1–26.

Behr, F. H., Fossum, V., Mitzenmacher, M., and Xiao, D. (2003). Estimating and comparing entropies across written natural languages using ppm compression. In *Data Compression Conference, 2003. Proceedings. DCC 2003*, page 416. IEEE.

Behrens, H. and Tomasello, M. (1999). And what about the Chinese? *Behavioral and Brain Sciences*, 22(06):1014–1014.

Bentz, C. (2016). The low-complexity-belt: evidence for large-scale language contact in human prehistory? In Roberts, S. G., Cuskley, C., McCrohon, L., Barceló-Coblijn, L., Feher, O., and Verhoef, T., editors, *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*.

Bentz, C., Alikaniotis, D., Cysouw, M., and Ferrer-i Cancho, R. (2017a). The entropy of words – learnability and expressivity across more than 1000 languages. *Entropy*, 19(6):275.

Bentz, C., Alikaniotis, D., Samardžić, T., and Buttery, P. (2017b). Variation in word frequency distributions: Definitions, measures and implications for a corpus-based language typology. *Journal of Quantitative Linguistics*, 24(2-3):128–162.

Bentz, C. and Berdicevskis, A. (2016). Learning pressures reduce morphological complexity: linking corpus, computational and experimental evidence. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), 26th International Conference on Computational Linguistics*.

Bentz, C. and Buttery, P. (2014). Towards a computational model of grammaticalization and lexical diversity. In *Proceedings of the 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL) at EACL*, pages 38–42.

Bentz, C. and Christiansen, M. H. (2010). Linguistic adaptation at work? The change of word order and case system from Latin to the Romance languages. In Scott-Phillips, T. C., Tamariz, M., Cartmill, E. A., and Hurford, J. R., editors, *The evolution of language. Proceedings of the 8th international conference (EVOLANG8)*, pages 26–33, Singapore. World Scientific.

Bentz, C. and Ferrer-i-Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In Bentz, C., Jäger, G., and Yanovich, I., editors, *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. University of Tübingen.

Bentz, C., Kiela, D., Hill, F., and Buttery, P. (2014). Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. *Corpus Linguistics and Linguistic Theory*, 10(2):175–211.

Bentz, C., Ruzsics, T., Koplenig, A., and Samaržić, T. (2016). A comparison between morphological complexity measures: typological data vs. language corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), 26th International Conference on Computational Linguistics*.

Bentz, C., Verkerk, A., Kiela, D., Hill, F., and Buttery, P. (2015). Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PLoS ONE*, 10(6):e0128254.

Bentz, C. and Winter, B. (2012). The impact of L2 speakers on the evolution of case marking. In Scott-Phillips, T. C., Tamariz, M., Cartmill, E. A., and Hurford, J. R., editors, *The Evolution of Language. Proceedings of the 9th International Conference (EVOLANG9)*, pages 58–64, Singapore. World Scientific.

Bentz, C. and Winter, B. (2013). Languages with more second language speakers tend to lose nominal case. *Language Dynamics and Change*, 3:1–27.

Bentz, C. and Winter, B. (2014). Languages with more second language learners tend to lose nominal case. In Wichmann, S. and Good, J., editors, *Quantifying Language Dynamics: On the Cutting Edge of Areal and Phylogenetic Linguistics*, pages 96–124. Brill.

Berdicevskis, A. (2012). Introducing pressure for expressivity into language evolution experiments. In Scott-Phillips, T. C., Tamariz, M., Cartmill, E. A., and Hurford, J. R., editors, *The Evolution of Language. Proceedings of the 9th International Conference (EVOLANG9)*, Singapore. World Scientific.

Berdicevskis, A. and Semenuks, A. Different trajectories of morphological overspecification and irregularity under imperfect language learning. In Arkadiev, P. and Gardani, F., editors, *Morphological complexity*. Oxford University Press, Oxford.

Berwick, R. C., Friederici, A. D., Chomsky, N., and Bolhuis, J. J. (2013). Evolution, brain, and the nature of language. *Trends in cognitive sciences*, 17(2):89–98.

Biberauer, T., Holmberg, A., Roberts, I., and Sheehan, M., editors (2010). *Parametric variation: Null subjects in minimalist theory*. Cambridge University Press, Cambridge.

Bickel, B. (2013). Distributional biases in language families. In Bickel, B., Grenoble, L. A., Peterson, D. A., and Timberlake, A., editors, *Language Typology and Historical Contingency*. John Benjamins Publishing Co.

Bickel, B. (2015). Distributional typology: statistical inquiries into the dynamics of linguistic diversity. In Heine, B. and Narrog, H., editors, *Oxford Handbook of Linguistic Analysis*. Oxford University Press, Oxford, 2nd edition.

Bickel, B. (2017). Areas and universals. In Hickey, R., editor, *The Cambridge Handbook of Areal Linguistics*, Cambridge Handbooks in Language and Linguistics, pages 40–54. Cambridge University Press, Cambridge.

Bickel, B. and Nichols, J. (2007). Inflectional morphology. In Shopen, T., editor, *Language typology and syntactic description*, pages 169–240. Cambridge University Press, Cambridge.

Bickel, B. and Nichols, J. (2009). The geography of case. In Malchukov, A. and Spencer, A., editors, *The Oxford Handbook of Case*, pages 479–493. Oxford: Oxford University Press.

Bickel, B. and Zúñiga, F. (2017). The 'word' in polysynthetic languages: phonological and syntactic challenges. In Fortescue, M., Mithun, M., and Evans, N., editors, *Oxford Handbook of Polysynthesis*, pages 158–186. Oxford University Press.

Bickerton, D. (1981). *Roots of language*. Karoma, Ann Arbor.

Bickerton, D. (1990). *Language and species*. University of Chicago Press, Chicago/London.

Blevins, J. P. (2016). *Word and paradigm morphology*. Oxford University Press, Oxford.

Blythe, R. A. (2012). Neutral evolution: a null model for language dynamics. *Advances in Complex Systems*, 15(03n04):1150015.

Blythe, R. A. and Croft, W. (2012). S-curves and the mechanisms of propagation in language change. *Language*, 88(2):269–304.

Blythe, R. A. and Croft, W. A. (2009). The speech community in evolutionary language dynamics. *Language Learning*, 59(s1):47–63.

Boas, F. (1911). *Handbook of American Indian languages. Part 1.* Government Print Office, Washington.

Bochkarev, V., Solovyev, V., and Wichmann, S. (2014). Universals versus historical contingencies in lexical evolution. *Journal of The Royal Society Interface*, 11(101):20140841.

Boeckx, C. and Piattelli-Palmarini, M. (2005). Language as a natural object–linguistics as a natural science. *The linguistic review*, 22(2-4):447–466.

Bolhuis, J. J., Tattersall, I., Chomsky, N., and Berwick, R. C. (2014). How could language have evolved? *PLoS biology*, 12(8):e1001934.

Borgwaldt, S. R., Hellwig, F. M., and de Groot, A. M. (2004). Word-initial entropy in five languages: Letter to sound, and sound to letter. *Written Language & Literacy*, 7(2):165–184.

Borgwaldt, S. R., Hellwig, F. M., and De Groot, A. M. (2005). Onset entropy matters–letter-to-phoneme mappings in seven languages. *Reading and Writing*, 18(3):211–229.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., and Atkinson, Q. D. (2012). Mapping the origins and expansion of the Indo-European language family. *Science*, 337:957–960.

Bright, W. (1996). The Devanagari script. In Daniels, P. T. and Bright, W., editors, *The world's writing systems*, chapter 31, pages 384–390. Oxford University Press.

Briscoe, E. J. (1998). Language as a complex adaptive system: co-evolution of language and of the language acquisition device. In van Halteren, H., editor, *8th Meeting of Comp. Linguistics in the Netherlands*, pages 3–40.

Briscoe, T. (2000a). Evolutionary perspectives on diachronic syntax. In Pintzuk, S., Tsoulas, G., and Warner, A., editors, *Diachronic syntax: Models and Mechanisms*, pages 75–108. Oxford University Press, Oxford.

Briscoe, T. (2000b). Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device. *Language*, 76(2):245–296.

Briscoe, T. (2003). Grammatical assimilation. In Christiansen, M. H. and Kirby, S., editors, *Language evolution*, pages 295–316. Oxford University Press, Oxford.

Briscoe, T. (2005). Coevolution of the language faculty and language(s) with decorrelated encodings. In Tallerman, M., editor, *Language Origins: Perspectives on Evolution*, pages 310–333. Oxford University Press.

Briscoe, T. (2009). What can formal or computational models tell us about how (much) language shaped the brain. In Bickerton, D., Kirby, S., and Szathmary, E., editors, *Biological Foundations and Origins of Syntax: Proceedings of the Ernst Strungmann Forum Workshop*, pages 369–382. MIT Press.

Bromham, L., Hua, X., Fitzpatrick, T. G., and Greenhill, S. J. (2015). Rate of language evolution is affected by population size. *Proceedings of the National Academy of Sciences*, 112(7):2097–2102.

Brown, P. F., Pietra, V. J. D., Mercer, R. L., Pietra, S. A. D., and Lai, J. C. (1992). An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1):31–40.

Bybee, J. (2003). Mechanisms of change in grammaticization: The role of frequency. In Joseph, B. D. and Janda, J., editors, *The Handbook of Historical Linguistics*, pages 602–623. Oxford: Blackwell.

Bybee, J. (2007). *Frequency of use and the organization of language*. Oxford University Press, Oxford.

Bybee, J. L. (2006). From usage to grammar: The mind's response to repetition. *Language*, 82(4):711–733.

Chand, V., Kapper, D., Mondal, S., Sur, S., and Parshad, R. D. (2017). Indian english evolution and focusing visible through power laws. *Languages*, 2(4).

Chang, W., Cathcart, C., Hall, D., and Garrett, A. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*, 91(1):194–244.

Chao, A. and Shen, T.-J. (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and ecological statistics*, 10(4):429–443.

Chater, N. and Christiansen, M. H. (2012). A solution to the logical problem of language evolution: language as an adaptation to the human brain. In Tallerman, M. and Gibson, K., editors, *Oxford Handbook of Language Evolution*, pages 626–639. Oxford: Oxford University Press.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124.

Chomsky, N. (1957). *Syntactic structures*. Mouton, The Hague.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.

Chomsky, N. (2005). Three factors in language design. *Linguistic inquiry*, 36(1):1–22.

Chomsky, N. (2010). Some simple evo devo theses: How true might they be for language. In Larson, R. K., Déprez, V., and Yamakido, H., editors, *The evolution of human language: biolinguistic perspectives*, pages 45–62. Cambridge: University Press Cambridge.

Chomsky, N. (2011). Language and other cognitive systems. What is special about language? *Language Learning and Development*, 7(4):263–278.

Christiansen, M. H. and Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(5):489–509.

Christiansen, M. H. and Chater, N. (2015). The language faculty that wasn't: a usage-based account of natural language recursion. *Frontiers in Psychology*, 6:1182.

Christiansen, M. H. and Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39:1–72.

Christiansen, M. H. and Kirby, S. (2003). Language evolution: consensus and controversies. *TRENDS in Cognitive Science*, 7(7):300–305.

Chumakina, M. (2011). Morphological complexity of Archi verbs. In Authier, G. and Maisak, T., editors, *Tense, aspect, modality and finiteness in East Caucasian languages*, pages 1–25. Universitätsverlag Dr. N. Brockmeyer, Bochum.

Clahsen, H. (1999). Lexical entries and rules of language: a multidisciplinary study of German inflection. *Behavioral and Brain Sciences*, 22(6):991–1060.

Clahsen, H., Eisenbeiss, S., and Sonnenstuhl-Henning, I. (1997). Morphological structure and the processing of inflected words. *Theoretical Linguistics*, 23(3):201–250.

Clark, R. and Roberts, I. (1993). A computational model of language learnability and language change. *Linguistic Inquiry*, 24(2):299–345.

Cornish, H., Tamariz, M., and Kirby, S. (2009). Complex adaptive systems and the origins of adaptive structure: What experiments can tell us. *Language Learning*, 59(s1):187–205.

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory*. John Wiley & Sons, New Jersey.

Croft, W. (2000). *Explaining language change: An evolutionary approach*. Pearson Education Limited, Edinburgh.

Cubberley, P. (1996). The Slavic alphabets. In Daniels, P. T. and Bright, W., editors, *The world's writing systems*, chapter 27, pages 346–355. Oxford University Press, Oxford.

Cuskley, C., Colaiori, F., Castellano, C., Loreto, V., Pugliese, M., and Tria, F. (2015). The adoption of linguistic rules in native and non-native speakers: Evidence from a wug task. *Journal of Memory and Language*, 84:205–223.

Cysouw, M. (2010). Dealing with diversity: Towards an explanation of NP-internal word order frequencies. *Linguistic Typology*, 14:253–286.

Cysouw, M. and Wälchli, B. (2007). Parallel texts. Using translational equivalents in linguistic typology. *Sprachtypologie & Universalienforschung STUF*, 60.2.

Dahl, Ö. (2004). *The growth and maintenance of linguistic complexity*. John Benjamins Publishing, Amsterdam/Philadelphia.

Dahl, Ö. (2007). From questionnaires to parallel corpora in typology. *STUF-Sprachtypologie und Universalienforschung*, 60(2):172–181.

Dale, R. and Lupyan, G. (2012). Understanding the origins of morphological diversity: The Linguistic Niche Hypothesis. *Advances in Complex Systems*, 15(3):1150017/1–1150017/16.

Daniels, P. T. and Bright, W., editors (1996). *The world's writing systems*. Oxford University Press, New York/Oxford.

de Boer, B. and Thompson, B. (2018). Biology-culture co-evolution in finite populations. *Scientific Reports*, 8(1):1209.

Deacon, T. (1997). *The symbolic species*. New York: Norton.

Deacon, T. W. (2010). What is missing from theories of information? In Davies, P. and Gregersen, N. H., editors, *Information and the nature of reality. From physics to metaphysics*, pages 146–170. Cambridge University Press: Cambridge, UK.

Dediu, D., Janssen, R., and Moisik, S. R. (2017). Language is not isolated from its wider environment: Vocal tract influences on the evolution of speech and language. *Language & Communication*, 54:9–20.

Dediu, D. and Ladd, D. R. (2007). Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, aspm and microcephalin. *Proceedings of the National Academy of Sciences*, 104(26):10944–10949.

Derungs, C. and Samardžić, T. (2017). Are prominent mountains frequently mentioned in text? Exploring the spatial expressiveness of text frequency. *International Journal of Geographical Information Science*, 32(5):856–873.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, 25(2):108–127.

Divuilu, F. (2005). *Lingala-English, English-Lingala Dictionary*. Congolese Voluntary Organisation.

Dixon, R. M. W. (1994). *Ergativity*. Cambridge University Press.

Dębowski, Ł. (2017). Is natural language strongly nonergodic? A stronger theorem about facts and words. *ArXiv e-prints*.

Dębowski, Ł. (2016). Consistency of the plug-in estimator of the entropy rate for ergodic processes. In *Proceedings of the 2016 IEEE International Symposium on Information Theory. (ISIT)*, pages 1651–1655.

Dryer, M. S. (1989). Large linguistic areas and language sampling. *Studies in language*, 13(2):257–292.

Dryer, M. S. and Haspelmath, M., editors (2013). *World Atlas of Language Structures online*. Max Planck Digital Library, Munich.

Dye, M., Johns, B., and Jones, M. N. (2016). The structure of names in memory: Deviations from uniform entropy impair memory for linguistic sequences. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society, Philadelphia, PA*.

Dye, M., Milin, P., Futrell, R., and Ramscar, M. (2017a). Cute little puppies and nice cold beers: An information theoretic analysis of prenominal adjectives. In *CogSci 2017: Computational Foundations of Cognition. 39th Annual Meeting of the Cognitive Science Society, London, UK*. Cognitive Science Society.

Dye, M., Milin, P., Futrell, R., and Ramscar, M. (2017b). A functional theory of gender paradigms. In Kiefer, F., Blevins, J. P., and Bartos, H., editors, *Perspectives on Morphological Organization: Data and Analyses*, pages 212–239. Brill.

Dình Hoà, N. (1996). Vietnamese. In Daniels, P. T. and Bright, W., editors, *The world's writing systems*, chapter 59, pages 691–695. Oxford University Press.

Ehret, K. (2016). *An information-theoretic approach to language complexity: variation in naturalistic corpora*. PhD thesis, Albert-Ludwigs-Universität Freiburg.

Ehret, K. and Szmrecsanyi, B. (2016a). An information-theoretic approach to assess linguistic complexity. In Baechler, R. and Seiler, G., editors, *Complexity, isolation and variation*. de Gruyter, Berlin.

Ehret, K. and Szmrecsanyi, B. (2016b). Compressing learner language: An information-theoretic measure of complexity in SLA production data. *Second Language Research*, pages 1–23.

Eisenbeiss, S., Bartke, S., and Clahsen, H. (2006). Structural and lexical case in child German: Evidence from language-impaired and typically developing children. *Language Acquisition*, 13(1):3–32.

Ellis, N. and Collins, L. (2009). Input and second language acquisition: The roles of frequency, form, and function. *The Modern Language Journal*, 93(3):329–335.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in second language acquisition*, 24(02):143–188.

Ellis, N. C. (2013). Emergentism. In Chapelle, C. A., editor, *The Encyclopedia of Applied Linguistics*, pages 1–10. Wiley Online Library.

Evans, N. (2011). *Dying words: Endangered languages and what they have to tell us*. The Language Library. John Wiley & Sons.

Farrar, K. and Jones, M. C. (2002). Introduction. In Jones, M. C. and Esch, E., editors, *Language change: The interplay of internal, external and extra-linguistic factors*, pages 1–19. Mouton de Gruyter, Berlin/New York.

Feldman, L. B. and Barac-Cikoja, D. (1996). Serbo-Croatian: A biscriptal language. In Daniels, P. T. and Bright, W., editors, *The world's writing systems*, chapter 64, pages 769–772. Oxford University Press.

Fenk, A. and Fenk, G. (1980). Konstanz im Kurzzeitgedächtnis - Konstanz im sprachlichen Informationsfluß. *Zietschrift für experimentelle and angewandte Pshychologie*, XXVII(3):400–414.

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In Bybee, J. L. and Hopper, P. J., editors, *Frequency and the Emergence of Linguistic Structure*, pages 431–448. John Benjamins, Amsterdam.

Fenk-Oczlon, G. and Fenk, A. (2008). Complexity trade-offs between the subsystems of language. In Miestamo, M., Sinnemäki, K., and Karlsson, F., editors, *Language complexity: typology, contact, change*, pages 43–65. John Benjamins, Amsterdam.

Fenk-Oczlon, G. and Fenk, A. (2011). Complexity trade-offs in language do not imply an equal overall complexity. In Solovyev, V. and Polyakov, V., editors, *Text Processing and Cognitive Technologies*, number 20 in XIII International Conference Cognitive Modeling in Linguistics, pages 164–167.

Fenk-Oczlon, G. and Fenk, A. (2014). Complexity trade-offs do not prove the equal complexity hypothesis. *Poznan Studies in Contemporary Linguistics*, 50(2):145–155.

Ferrer-i-Cancho, R. (2005). Zipf's law from a communicative phase transition. *European Physical Journal B*, 47:449–457.

Ferrer-i-Cancho, R. (2017a). The optimality of attaching unlinked labels to unlinked meanings. *Glottometrics*, 36:1–16.

Ferrer-i-Cancho, R. (2017b). Optimization models of natural communication. *Journal of Quantitative Linguistics*, pages 1–31.

Ferrer-i-Cancho, R. (2017c). The placement of the head that maximizes predictability. an information theoretic approach. *arXiv preprint arXiv:1705.09932*.

Ferrer-i-Cancho, R., Bentz, C., and Seguin, C. (2015). Compression and the origins of Zipf's law of abbreviation. *arXiv preprint*, page 1504.04884.

Ferrer-i-Cancho, R., Dębowski, Ł., and del Prado Martín, F. M. (2013). Constant conditional entropy and related hypotheses. *Journal of Statistical Mechanics: Theory and Experiment*, L07001.

Ferrer-i-Cancho, R. and Díaz-Guilera, A. (2007). The global minima of the communicative energy of natural communication systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06):P06009.

Ferrer-i-Cancho, R. and Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences USA*, 100:788–791.

Fitch, W., Hauser, M. D., and Chomsky, N. (2005). The evolution of the language faculty: clarifications and implications. *Cognition*, 97(2):179–210.

Fitch, W. T. (2010a). *The evolution of language*. Cambridge University Press, Cambridge.

Fitch, W. T. (2010b). Three meanings of 'recursion': key distinctions for biolinguistics. In Larson, R. K., Déprez, V., and Yamakido, H., editors, *The evolution of human language: biolinguistic perspectives*, pages 73–90. Cambridge University Press, Cambridge.

Fitch, W. T. (2018). The biology and evolution of speech: A comparative analysis. *Annual Review of Linguistics*, 4(1).

Frank, A. and Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the 30th annual meeting of the cognitive science society*, pages 933–938. Cognitive Science Society Washington, DC.

Freckleton, R. P., Harvey, P. H., and Pagel, M. (2002). Phylogenetic analysis and comparative data: A test and review of evidence. *The American Naturalist*, 160(2):712–726.

Freudenthal, D., Pine, J. M., Jones, G., and Gobet, F. (2015). Simulating the cross-linguistic pattern of Optional Infinitive errors in children's declaratives and Wh-questions. *Cognition*, 143:61–76.

Futrell, R., Mahowald, K., and Gibson, E. (2015). Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 91–100.

Gao, Y., Kontoyiannis, I., and Bienenstock, E. (2008). Estimating the entropy of binary time series: Methodology, some theory and a simulation study. *Entropy*, 10(2):71–99.

Geertzen, J., Blevins, J. P., and Milin, P. (2016). Informativeness of linguistic unit boundaries. *Italian journal of linguistics*, 28(1):25–47.

Gell-Mann, M. (1992). Complexity and complex adaptive systems. In John A. Hawkins, M. G.-M., editor, *The Evolution of Human Languages, SFI Studies in the Sciences of Complexity*, pages 3–18.

Gell-Mann, M. (1994). *The Quark and the Jaguar*. W. H. Freeman and Company, New York.

Gesmundo, A. and Samardžić, T. (2012). Lemmatisation as a tagging task. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 368–372. Association for Computational Linguistics.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., Gibson, M., Piantadosi, S. T., and Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, 114(40):10785–10790.

Gil, D. (2009). How much grammar does it take to sail a boat? In Sampson, G., Gil, D., and Trudgill, P., editors, *Language complexity as an evolving variable*, pages 19–34. Oxford University Press, Oxford.

Göksel, A. and Kerslake, C. (2005). *Turkish: A comprehensive grammar*. Routledge, London and New York.

Goldberg, A. E., Casenhiser, D. M., and Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive linguistics*, 15(3):289–316.

Goldschneider, J. M. and DeKeyser, R. M. (2001). Explaining the natural order of L2 morpheme acquisition in English: A meta-analysis of multiple determinants. *Language learning*, 51(1):1–50.

Gray, R. D. and Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439.

Greenberg, J. H. (1960). A quantitative approach to the morphological typology of language. *International journal of American linguistics*, 26(3):178–194.

Grefenstette, G. (2010). Estimating the number of concepts. In de Schrywer, G.-M., editor, *A way with words: Recent advances in lexical theory and analysis*, pages 143–156. Kampala.

Grefenstette, G. and Tapanainen, P. (1994). What is a word, what is a sentence?: problems of tokenisation.

Gries, S. T. (2012). Frequencies, probabilities, and association measures in usage-/ exemplar-based linguistics: Some necessary clarifications. *Studies in Language*, 11(3):477–510.

Gürel, A. (2000). Missing case inflection: Implications for second language acquisition. In Howell, C., Fish, S. A., and Keith-Lucas, T., editors, *Proceedings of the 24th Annual Boston University Conference on Language Development*, pages 379–390, Somerville, MA. Cascadilla Press.

Ha, L. Q., Stewart, D. W., Hanna, P., and Smith, F. J. (2006). Zipf and type-token rules for the English, Spanish, Irish and Latin languages. *Web Journal of Formal, Computational and Cognitive Linguistics*, 8.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational*

Linguistics on Language technologies, pages 1–8. Association for Computational Linguistics.

Hale, J. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S., editors (2016). *Glottolog 2.7*. Jena: Max Planck Institute for the Science of Human History.

Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 45(1):31–80.

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579.

Hausser, J. and Strimmer, K. (2009). Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *The Journal of Machine Learning Research*, 10:1469–1484.

Hausser, J. and Strimmer, K. (2014). *entropy: Estimation of Entropy, Mutual Information and Related Quantities*. R package version 1.2.1.

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press, Oxford.

Hawkins, J. A. (2014). *Cross-linguistic Variation and Efficiency*. Oxford University Press, Oxford.

Hay, J. and Bauer, L. (2007). Phoneme inventory size and population size. *Language*, 83(2):388–400.

Haznedar, B. (2003). Missing surface inflection in adult and child L2 acquisition. In Liceras, J. M., editor, *Proceedings of the 6th Generative Approaches to Second Language Acquisition Conference (GASLA)*, pages 140–149, Somerville, MA. Cascadilla Proceedings Project.

Haznedar, B. (2006). Persistent problems with case morphology in L2 acquisition. *Interfaces in multilingualism: Acquisition and representation*, pages 179–206.

Heine, B. and Kuteva, T. (2007). *The genesis of grammar: A reconstruction*. Oxford: Oxford University Press.

Herman, J. (2000). *Vulgar Latin*. The Pennsylvania State University Press, University Park, PA.

Holland, J. H. (2005). Language acquisition as a complex adaptive system. In Minett, J. W. and Wang, W. S. Y., editors, *Language acquisition, change and emergence*, pages 411–435. Hong Kong: City University of Hong Kong Press.

Holland, J. H. (2006). Studying complex adaptive systems. *Journal of Systems Science and Complexity*, 19(1):1–8.

Holland, J. H. (2012). *Signals and boundaries: Building blocks for complex adaptive systems*. Mit Press.

Holman, E. W., Brown, C. H., Wichmann, S., Müller, A., Velupillai, V., Hammarström, H., Sauppe, S., Jung, H., Bakker, D., Brown, P., et al. (2011). Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, 52(6):841–875.

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., and Bakker, D. (2008). Advances in automated language classification. *Quantitative Investigations in Theoretical Linguistics*, pages 40–43.

Hopper, P. J. and Traugott, E. C. (2003). *Grammaticalization*. Cambridge University Press.

Hudson Kam, C. L. and Chang, A. (2009). Investigating the cause of language regularization in adults: Memory constraints or learning effects? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3):815.

Hudson Kam, C. L. and Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1(2):151–195.

Jackendoff, R. and Pinker, S. (2005). The nature of the language faculty and its implications for evolution of language (Reply to Fitch, Hauser, and Chomsky). *Cognition*, 97(2):211–225.

Jaeger, T. and Levy, R. P. (2006). Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems*, pages 849–856.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62.

Jaeger, T. F., Graff, P., Croft, W., and Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology*, 15:281–320.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 19(1):57–84.

Johnson, J. S. and Newport, E. L. (1989). Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. *Cognitive psychology*, 21(1):60–99.

Jones, M. C. and Esch, E. (2002). *Language change: The interplay of internal, external and extra-linguistic factors*. Mouton de Gruyter, Berlin, New York.

Jones, M. C. and Singh, I. (2005). *Exploring language change*. Routledge, New York.

Jost, L. (2006). Entropy and diversity. *OIKOS*, 113(2).

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, 5(3):206–213.

Juola, P. (2008). Assessing linguistic complexity. In Miestamo, M., Sinnemäki, K., and Karlsson, F., editors, *Language complexity: typology, contact, change*, pages 89–108. Amsterdam: John Benjamins.

Kam, C. L. H. and Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive psychology*, 59(1):30–66.

Kanwal, J., Smith, K., Culbertson, J., and Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, 165:45–52.

Kauhanen, H. (2017). Neutral change. *Journal of Linguistics*, 53(2):327–358.

Kempe, V. and Brooks, P. J. (2018). Linking adult second language learning and diachronic change: A cautionary note. *Frontiers in Psychology*, 9:480.

Kibrik, A. E. (1998). Archi. In Spencer, A. and Zwicky, A. M., editors, *The handbook of morphology*, pages 455–476. Blackwell, Oxford.

King, R. (1996). Korean writing. In Daniels, P. T. and Bright, W., editors, *The world's writing systems*, chapter 17, pages 219–227. Oxford University Press.

Kirby, S., Cornish, H., and Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686.

Kirby, S., Dowman, M., and Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences of the United States of America*, 104(12):5241–5245.

Kirby, S. and Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In Cangelosi, A. and Parisi, D., editors, *Simulating the evolution of language*, pages 121–147. Springer.

Kirby, S., Tamariz, M., Cornish, H., and Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Kontoyiannis, I., Algoet, P. H., Suhov, Y. M., and Wyner, A. J. (1998). Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Transactions on Information Theory*, 44(3):1319–1327.

Koplenig, A. (2015). Using the parameters of the Zipf-Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes: a large-scale corpus analysis. *Corpus Linguistics and Linguistic Theory*.

Koplenig, A., Meyer, P., Wolfer, S., and Müller-Spitzer, C. (2017). The statistical trade-off between word order and word structure–large-scale evidence for the principle of least effort. *PloS ONE*, 12(3):e0173614.

Kretzschmar, W. A. (2015). *Language and Complex Systems*. Cambridge University Press.

Krupnik, I. and Müller-Wille, L. (2010). Franz Boas and Inuktitut terminology for ice and snow: from the emergence of the field to the "Great Eskimo Vocabulary Hoax". In *SIKU: Knowing our ice*, pages 377–400. Springer.

Kusters, W. (2003). *Linguistic complexity. The influence of social change on verbal inflection*. PhD thesis, Universiteit Leiden.

Ladd, D. R., Roberts, S. G., and Dediu, D. (2015). Correlational studies in typological and historical linguistics. *Annual Review of Linguistics*, 1(1):221–241.

Lanz, L. A. (2010). *A Grammar of Inupiaq Morphosyntax*. ERIC.

Larsen-Freeman, D. E. (1975). The acquisition of grammatical morphemes by adult ESL students. *TESOL quarterly*, 9(4):409–419.

Larsen-Freeman, D. E. (1976). An explanation for the morpheme acquisition order of second language learners. *Language Learning*, 26(1):125–134.

Larsen-Freeman, D. E. (1997). Chaos/complexity science and second language acquisition. *Applied linguistics*, 18(2):141–165.

Lehmann, C. (1985). Grammaticalization: Synchronic variation and diachronic change. *Lingua e Stile*, 20:303–318.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Lewis, M. L. and Frank, M. C. (2016). Linguistic niches emerge from pressures at multiple timescales. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Lewis, M. P., Simons, G. F., and Fenning, C. D., editors (2013). *Ethnologue: Languages of the world*. SIL International, Dallas, Texas, 17th edition.

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., and Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, 449(11):713–716.

Lieven, E. (2010). Input and first language acquisition: Evaluating the role of frequency. *Lingua*, 120(11):2546–2556.

Lightfoot, D. W. (1979). *Principles of diachronic syntax*. Cambridge University Press, Cambridge.

Lüpke, F. (2016). Uncovering small-scale multilingualism. *Critical Multilingualism Studies*, 4(2):35–74.

Lupyan, G. and Dale, R. (2010). Language structure is partly determined by social structure. *PloS ONE*, 5(1):e8559.

Lupyan, G. and Dale, R. (2015). The role of adaptation in understanding linguistic diversity. In De Busser, R. and LaPolla, R. J., editors, *Language Structure and Environment*. John Benjamins Publishing Company, Amsterdam.

MacLean, E. A. (2012). *North Slope Iñupiaq to English Dictionary*. Alaska Native Languages Archives, University of Alaska Fairbanks.

Magga, O. H. (2006). Diversity in Saami terminology for reindeer, snow, and ice. *International Social Science Journal*, 58(187):25–34.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., and Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.

Mair, V. H. (1996). Modern Chinese writing. In Daniels, P. T. and Bright, W., editors, *The world's writing systems*, chapter 15, pages 201–208. Oxford University Press.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. In Jackson, W., editor, *Communication Theory*, pages 468–502. Butterworths Scientific Publications, London.

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., and Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology*, 29:189–256.

Marslen-Wilson, W. D. and Tyler, L. K. (2007). Morphology, language and the brain: The decompositional substrate for language comprehension. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 362:823–836.

Martin, L. (1986). "eskimo words for snow": A case study in the genesis and decay of an anthropological example. *American anthropologist*, 88(2):418–423.

Mayer, T. and Cysouw, M. (2014). Creating a massively parallel bible corpus. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014*, pages 3158–3163. European Language Resources Association (ELRA).

McCarthy, P. M. and Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4):459–488.

McCarthy, P. M. and Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

McWhorter, J. H. (2002). What happened to English? *Diachronica*, 19(2):217–272.

McWhorter, J. H. (2007). *Language interrupted: Signs of non-native acquisition in standard language grammars*. Oxford University Press, New York.

McWhorter, J. H. (2011). *Linguistic simplicity and complexitiy: Why do languages undress?* Mouton de Gruyter, Boston.

Michalke, M. (2014). *koRpus: An R Package for Text Analysis*. (Version 0.05-5).

Miestamo, M., Sinnemäki, K., and Karlsson, F. (2008). *Language complexity: Typology, contact, change*. John Benjamins Publishing.

Milin, P., Kuperman, V., Kostic, A., and Baayen, R. H. (2009). Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. In Blevins, J. P. and Blevins, J., editors, *Analogy in grammar: Form and acquisition*, pages 214–252. Oxford University Press, Oxford.

Mitchell, D. (2015). Type-token models: a comparative study. *Journal of Quantitative Linguistics*, 22(1):1–21.

Moberg, J., Gooskens, C., Nerbonne, J., and Vaillette, N. (2006). Conditional entropy measures intelligibility among related languages. In *Proceedings of Computational Linguistics in the Netherlands*, pages 51–66. Rodopi Amsterdam.

Moisik, S. R. and Dediu, D. (2017). Anatomical biasing and clicks: Evidence from biomechanical modeling. *Journal of Language Evolution*.

Montemurro, M. A. and Zanette, D. H. (2011). Universal entropy of word ordering across linguistic families. *PLoS ONE*, 6(5):e19875.

Montemurro, M. A. and Zanette, D. H. (2016). Complexity and universality in the long-range order of words. In *Creativity and Universality in Language*, pages 27–41. Springer, Heidelberg, Germany.

Moran, S., McCloy, D., and Wright, R. (2012). Revisiting population size vs. phoneme inventory size. *Language*, 88(4):877–893.

Moscoso del Prado, F. (2011). The mirage of morphological complexity. In *Proc. of the 33rd Annual Conference of the Cognitive Science Society*, pages 3524–3529.

Moscoso del Prado Martín, F., Kostić, A., and Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, 94(1):1–18.

Münkemüller, T., Lavergne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K., and Thuiller, W. (2012). How to measure and test phylogenetic signal. *Methods in Ecology and Evolution*, 3(4):743–756.

Murphy, L. (2013). *R package 'likelihood': Metods for maximum likelihood estimation*.

Nakazawa, M. (2015). *fmsb: Functions for Medical Statistics Book with some Demographic Data*. R package version 0.5.2.

Nemenman, I., Shafee, F., and Bialek, W. (2002). Entropy and inference, revisited. *Advances in neural information processing systems*, 1:471–478.

Nettle, D. (1995). Segmental inventory size, word length, and communicative efficiency. *Linguistics*, 33(2):359–367.

Nettle, D. (1998). Coevolution of phonology and the lexicon in twelve languages of West Africa. *Journal of Quantitative Linguistics*, 5(3):240–245.

Nettle, D. (1999). Is the rate of linguistic change constant? *Lingua*, 108:119–136.

Nettle, D. (2012). Social scale and structural complexity in human languages. *Phil. Trans. R. Soc. B*, 367(1597):1829–1836.

Nettle, D. and Romaine, S. (2000). *Vanishing voices. The extinction of the world's languages*. Oxford University Press.

Newberry, M. G., Ahern, C. A., Clark, R., and Plotkin, J. B. (2017). Detecting evolutionary forces in language change. *Nature*, 551(7679):223.

Newmeyer, F. J. and Preston, L. B. (2014a). Introduction. In Newmeyer, F. J. and Preston, L. B., editors, *Measuring grammatical complexity*, pages 1–14. Oxford University Press.

Newmeyer, F. J. and Preston, L. B. (2014b). *Measuring Grammatical Complexity*. Oxford University Press.

Nichols, J. (1992). *Linguistic diversity in space and time*. University of Chicago Press, Chicago.

Nichols, J. (1997). Modeling ancient population structures and movement in linguistics. *Annual review of anthropology*, 26(1):359–384.

Nichols, J. (2013). The vertical archipelago: Adding the third dimension to linguistic geography. In Stukenbrock, A., Szmrecsanyi, B., Auer, P., and Hilpert, M., editors, *Space in language and linguistics: Geographical, interactional, and cognitive perspectives*, pages 38–60. De Gruyter.

Nichols, J. (2016). Complex edges, transparent frontiers: grammatical complexity and language spreads. In Baechler, R. and Seiler, G., editors, *Complexity, isolation, and variation*, pages 117–139. De Gruyter.

Nichols, J. and Bentz, C. (2018). Morphological complexity of languages reflects the settlement history of the americas. In Harvati, K., Jäger, G., and Reyes-Centeno, H., editors, *New perspectives on the peopling of the Americas*. Kerns Verlag, Tübingen.

Nichols, J., Witzlack-Makarevich, A., and Bickel, B. (2013). The AUTOTYP genealogy and geography database: 2013 release.

Niyogi, P. and Berwick, R. C. (1997). Evolutionary consequences of language learning. *Linguistics and Philosophy*, 20:697–719.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, 401:877–884.

Papadopoulou, D., Varlokosta, S., Spyropoulos, V., Kaili, H., Prokou, S., and Revithiadou, a. (2011). Case morphology and word order in second language Turkish: Evidence from Greek learners. *Second Language Research*, 27(2):173–204.

Parodi, T., Schwartz, B. D., and Clahsen, H. (2004). On the L2 acquisition of the morphosyntax of German nominals. *Linguistics*, 42(3):669–705.

Pereira, F. (2000). Formal grammar and information theory: together again? *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 358(1769):1239–1253.

Pereltsvaig, A. (2012). *Languages of the world: an introduction*. Cambridge University Press, Cambridge.

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

Piattelli-Palmarini, M. and Uriagereka, J. (2004). The immune syntax: the evolution of the language virus. In Jenkins, L., editor, *Variation and universals in biolinguistics*, pages 341–377. Oxford: Elsevier.

Pinker, S. (2003). Language as an adaptation to the cognitive niche. In Kirby, Simon; Christiansen, M. H., editor, *Language evolution*, pages 16–37. Oxford University Press.

Pinker, S. and Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, 13(04):707–727.

Pinker, S. and Jackendoff, R. (2005). The faculty of language: what's special about it? *Cognition*, 95(2):201–236.

Pinker, S. and Ullman, M. T. (2002). The past and the future of the past tense. *TRENDS in Cognitive Science*, 6(11):456–463.

Pintzuk, S., Tsoulas, G., and Warner, A. (2000). *Diachronic syntax: Models and mechanisms*. Oxford University Press, Oxford.

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B. D., Köhler, R., Krupa, V., Macutek, J., Pustet, R., Uhlirova, L., and Vidya, M. N. (2009). *Word frequency studies*. Mouton de Gruyter, Berlin & New York.

Popescu, I.-I., Altmann, G., and Köhler, R. (2010). Zipf's law - another view. *Quality & Quantity*, 44(4):713–731.

Posey, D. A. (2002). *Kayapó ethnoecology and culture*, volume 6 of *Studies in environmental anthropology*. Routledge, London and New York.

Pukui, M. K. and Elbert, S. H. (1975). *New pocket Hawaiian dictionary: with a concise grammar and given names in Hawaiian*. University of Hawaii Press.

Pullum, G. K. (1989). The great Eskimo vocabulary hoax. *Natural Language & Linguistic Theory*, pages 275–281.

Quine, W. V. O., Churchland, P. S., and Føllesdal, D. (2013). *Word and object*. MIT press.

Qureshi, M. A. (2016). A meta-analysis: Age and second language grammar acquisition. *System*, 60:147–160.

R Core Team (2013). R: A language and environment for statistical computing.

Rama, T. (2016). Ancestry sampling for Indo-European phylogeny and dates. In Bentz, C., Jäger, G., and Yanovich, I., editors, *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. Universitätsbibliothek Tübingen.

Rao, R. P., Yadav, N., Vahia, M. N., Joglekar, H., Adhikari, R., and Mahadevan, I. (2010). Entropy, the indus script, and language: A reply to r. sproat. *Computational Linguistics*, 36(4):795–805.

Rao, R. P. N. (2010). Probabilistic analysis of an acient undeciphered script. *Computer*, pages 76–80.

Rao, R. P. N., Yadav, N., Vahia, M. N., Joglekar, H., Adhikari, R., and Mahadevan, I. (2009). Entropic evidence for linguistic structure in the Indus script. *Science*, 324(29):1165.

Reali, F. and Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, 111(3):317–328.

Reali, F. and Griffiths, T. L. (2010). Words as alleles: connecting language evolution with bayesian learners to models of genetic drift. *Proceedings of the Royal Society of London B: Biological Sciences*, 277(1680):429–436.

Regier, T., Carstensen, A., and Kemp, C. (2016). Languages support efficient communication about the environment: Words for snow revisited. *PLoS ONE*, 11(4):e0151138.

Regier, T., Kemp, C., and Kay, P. (2015). Word meanings across languages support efficient communication. *The handbook of language emergence*, pages 237–263.

Revell, L. J., Harmon, L. J., and Collar, D. C. (2008). Phylogenetic signal, evolutionary process, and rate. *Systematic Biology*, 57(4):591–601.

Ritt, N. (2004). *Selfish Sounds and Linguistic Evolution: A Darwinian Approach to Language Change*. Cambridge University Press.

Roberts, I. (2007). *Diachronic syntax*. Oxford University Press, Oxford.

Roberts, I. and Holmberg, A. (2010). Introduction: Parameters in minimalist theory. In Biberauer, T., Holmberg, A., Roberts, I., and Sheehan, M., editors, *Parametric variation: Null subjects in minimalist theory*, pages 1–58. Cambridge University Press, Cambridge.

Roberts, I. and Roussou, A. (2003). *Syntactic change: A minimalist approach to grammaticalization*. Cambridge University Press, Cambridge.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., and Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, 112(41):12663–12668.

Roy, B. C., Frank, M. C., and Roy, D. (2009). Exploring word learning in a high-density longitudinal corpus. In *Proceedings of the 31st Meeting of the Cognitive Science Society. Amsterdam, The Netherlands*. Cognitive Science Society, Inc.

Ruegsegger, M. and Ruegsegger, J. (1955). *Vocabulario de Zapoteco del dialecto Miahuatlan del estado del Oaxaca*. Instituto Lingüístico de Verano, A.C.

Säily, T. (2011). Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory*, 7(1):119–141.

Säily, T. and Suomela, J. (2009). Comparing type counts: The case of women, men and -ity in early English letters. *Language and Computers*, 69(1):87–109.

Salas, A. (2006). *El mapuche o araucano. Fonología, gramática y antología de cuentos*. Santiago: Centro de Estudios Públicos.

Sampson, G. (2009). A linguistic axiom challenged. In Sampson, G., Gil, D., and Trudgill, P., editors, *Language complexity as an evolving variable*, chapter 1, pages 1–18. Oxford University Press.

Sampson, G., Gil, D., and Trudgill, P. (2009). *Language complexity as an evolving variable*. Oxford University Press.

Sapir, E. (1921). *Language: An Introduction to the Study of Speech*. Hartcour, Brace & World, New York.

Schiering, R. (2006). *Cliticization and the evolution of morphology: A cross-linguistic study on phonology in grammaticalization*. PhD thesis, Universität Konstanz.

Schiering, R. (2010). Reconsidering erosion in grammaticalization. evidence from cliticization. In Stathi, K., Gehweiler, E., and König, E., editors, *Grammaticalization: current views and issues*, volume 119, pages 73–101. John Benjamins Publishing.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, volume 12, pages 44–49.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

Schürmann, T. and Grassberger, P. (1996). Entropy estimation of symbol sequences. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 6(3):414–427.

Scott-Phillips, T. (2015). *Speaking Our Minds: Why human communication is different, and how language evolved to make it special*. Palgrave MacMillan.

Scott-Phillips, T. C. and Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, 14(9):411–417.

Serva, M. and Petroni, F. (2008). Indo-European languages tree by Levenshtein distance. *EPL (Europhysics Letters)*, 81(6):68005.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell Systems Technical Journal*, 27:379–423.

Shannon, C. E. (1951). Prediction and entropy of printed English. *The Bell System Technical Journal*, 30(1):50–65.

Shannon, C. E. and Weaver, W. (1949). *The mathematical theory of communication*. The University of Illinois Press, Urbana.

Singer, R. and Harris, S. (2016). What practices and ideologies support small-scale multilingualism? A case study of Warruwi Community, northern Australia. *International Journal of the Sociology of Language*, 2016(241):163–208.

Sinnemäki, K. (2008). Complexity trade-offs in core argument marking. In Miestamo, M., Sinnemäki, K., and Karlsson, F., editors, *Language complexity: Typology, contact, change*, pages 67–88. John Benjamins Publishing.

Sinnemäki, K. (2009). Complexity in core argument marking and population size. In Sampson, G., Gil, D., and Trudgill, P., editors, *Language complexity as an evolving variable*, pages 126–141. Oxford University Press.

Sinnemäki, K. (2014). Complexity trade-offs: a case study. In Newmeyer, F. J. and Preston, L. B., editors, *Measuring linguistic complexity*, chapter 9, pages 179–201. Oxford University Press.

Skinner, L. E. and Skinner, M. B. (2000). *Diccionario Chinanteco de San Felipe Usila, Oaxaca*. Instituto Lingüístico de Verano, A.C., Coyoacán, México.

Smith, K., Brighton, H., and Kirby, S. (2003). Complex systems in language evolution: the cultural emergence of compositional structure. *Advances in Complex Systems*, 6(04):537–558.

Smith, K. and Kirby, S. (2008). Cultural evolution: implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1509):3591–3603.

Spencer, A. (2008). Does Hungarian have a case system? In Corbett, G. G. and Noonan, M., editors, *Case and grammatical relations. Studies in honor of Bernard Comrie*, pages 35–56. Benjamins Amsterdam.

Sproat, R. (2014). A statistical comparison of written language and nonlinguistic symbol systems. *Language*, 90(2):457–481.

Steels, L. (2000). Language as a complex adaptive system. In *Parallel Problem Solving from Nature PPSN VI*, pages 17–26. Springer.

Steiner, P. and Prün, C. (2007). The effects of diversification and unification on the inflectional paradigms of German nouns. In Grzybek, P. and Köhler, R., editors, *Exact methods in the study of language and text*, pages 623–632. Mouton de Gruyter, Berlin & New York.

Stockhammer, R. (2014). *Grammatik. Wissen und Macht in der Geschichte einer sprachlichen Instution*. Suhrkamp.

Stoll, S., Mazara, J., and Bickel, B. (2017). The acquisition of polysynthetic verb forms in Chintang. In Fortescue, M., Mithun, M., and Evans, N., editors, *Oxford Handbook of Polysynthesis*. Oxford University Press.

Szmrecsanyi, B. (2009). Typological parameters of intralingual variability: Grammatical analyticity versus syntheticity in varieties of English. *Language Variation and Change*, 21(03):319–353.

Szmrecsanyi, B. (2012). Analyticity and syntheticity in the history of English. In Nevalainen, T. and Traugott, E. C., editors, *The Oxford Handbook of the History of English*, chapter 52, pages 655–665. Oxford University Press.

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology*, 57A(4):745–765.

Takahira, R., Tanaka-Ishii, K., and Dębowski, Ł. (2016). Entropy rate estimates for natural language – a new extrapolation of compressed large-scale corpora. *Entropy*, 18(10):364.

Tamariz, M. and Kirby, S. (2015). Culture: copying, compression, and conventionality. *Cognitive Science*, 39:171–183.

Thomason, S. G. and Kaufman, T. (1988). *Language contact, creolization, and genetic linguistics*. University of California Press, Berkeley, Los Angeles, Oxford.

Threatte, L. (1996). The Greek alphabet. In Daniels, P. T. and Bright, W., editors, *The world's writing systems*, chapter 22, pages 271–280. Oxford University Press.

Tomasello, M. and Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Trudgill, P. (2002). *Sociolinguistic variation and change*. Georgetown University Press, Washington, DC.

Trudgill, P. (2011). *Sociolinguistic typology: social determinants of linguistic complexity*. Oxford University Press, Oxford.

Tweedie, F. J. and Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32:323–352.

Wexler, K. (1994). Optional infinitives, head movement and the economy of derivations1. In Lightfoot, D. and Hornstein, N., editors, *Verb movement*, chapter 14. Cambridge University Press.

Wichmann, S., Müller, A., Wett, A., Velupillai, V., Bischoffberger, J., Brown, C. H., Holman, E. W., Sauppe, S., Molochieva, Z., Brown, P., Hammarström, H., Belyaev, O., List, J.-M., Bakker, D., Egorov, D., Urban, M., Mailhammer, R., Carrizo, A., Dryer, M. S., Korovina, E., Beck, D., Geyer, H., Epps, P., Grant, A., and Valenzuela, P. (2013). The ASJP database (version 16).

Wichmann, S., Rama, T., and Holman, E. W. (2011). Phonological diversity, word length, and population sizes across languages: The asjp evidence. *Linguistic Typology*, 15(2):177–197.

Wichmann, S. r. and Holman, E. W. (2009). Population size and rates of language change. *Human Biology*, 81(2-3):259–274.

Wichmann, S. r., Stauffer, D., Schulze, C., and Holman, E. W. (2008). Do language change rates depend on population size? *Advances in Complex Systems*, 11(3):1–20.

Wiese, H. (2006). “Ich mach dich Messer”: Grammatische Produktivität in Kiez-Sprache (“Kanak Sprak”). *Linguistische Berichte*, 2006(207):245–273.

Wiese, H. and Pohle, M. (2016). “Ich geh Kino” oder “... ins Kino”? Gebrauchsrestriktionen nichtkanonischer Lokalangaben. *Zeitschrift für Sprachwissenschaft*, 35(2):171–216.

Wiese, H. and Rehbein, I. (2016). Coherence in new urban dialects: A case study. *Lingua*, 172:45–61.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *arXiv preprint*, page 1308.5499.

Winter, B. (2014). Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. *BioEssays*, 36(10):960–967.

Wray, A. (2014). Why are we so sure we know what a word is? In Taylor, J., editor, *The Oxford Handbook of the Word*, chapter 42. Oxford University Press, Oxford.

Wray, A. and Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, 117:543–578.

Yang, C. D. (2000). Internal and external forces in language change. *Language Variation and Change*, 12:231–250.

Yanovich, I. (2016). Genetic drift explains Sapir’s “drift” in semantic change. In *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*, pages 321–329.

Yip, M. (2002). *Tone*. Cambridge University Press, Cambridge.

Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language*. Harvard University Press, Cambridge (Massachusetts).

Zipf, G. K. (1935). *The psycho-biology of language*. The M.I.T Press, Cambridge (Massachusetts).

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley, Cambridge (Massachusetts).

Ziv, J. and Lempel, A. (1977). A universal algorithm for sequential data compression. *IEEE Transactions on information theory*, 23(3):337–343.

Index

DOI 10.1515/97831104957-015