An information-theoretic approach to language complexity: variation in naturalistic corpora

Inaugural-Dissertation

zur

Erlangung der Doktorwürde

der Philologischen Fakultät

der Albert-Ludwigs-Universität

Freiburg i.Br.

vorgelegt von

Katharina Luisa Ehret

aus Freiburg i.Br.

WS 2015/2016

Erstgutachter/in: Prof. Dr. Benedikt Szmrecsanyi
Zweitgutachter/in: Prof. Dr. Dr. h.c. Bernd Kortmann

Vorsitzende/r des Promotionsausschusses
der Gemeinsamen Kommission der
Philologischen, Philosophischen und Wirtschafts-
und Verhaltenswissenschaftlichen Fakultät: Prof. Dr. Joachim Grage

Datum der Disputation: 07.10.2016

To my parents

Contents

List of Figures **iii**

List of Tables **v**

Preface and Acknowledgements **vii**

1. Introduction **1**

2. Literature review and theoretical background **11**
2.1. Linguistic Complexity 11
2.1.1. Defining linguistic complexity 11
2.1.2. Explaining linguistic complexity 22
2.2. Information theory 31
2.2.1. Information theory, Shannon entropy and Kolmogorov complexity 31
2.2.2. Information-theoretic complexity 36

3. Experimenting with the compression technique **43**
3.1. Parallel texts 43
3.1.1. Method and data 43
3.1.2. The Gospel of Mark 50
3.2. Parallel, semi-parallel and non-parallel texts 56
3.2.1. Method and data 56
3.2.2. Alice's Adventures in Wonderland 66
3.2.3. Newspaper texts 74
3.3. Summary 81

4. Excursion: Targeted file manipulation **83**
4.1. Method and data 83
4.2. Analysing morphological markers 92
4.3. Analysing constructions 97
4.4. Summary 104

5. Exploring compression algorithms **105**
5.1. Comparing and interpreting algorithmic complexity 105
5.1.1. Interpreting compressed strings 105
5.1.2. Comparing compressed strings 116
5.2. Summary 128

6. Case studies **131**

6.1. Assessing complexity variation in British English registers 131
6.1.1. Method and data 131
6.1.2. Register variability 137
6.2. Measuring complexity in learner English 141
6.2.1. Method and data 141
6.2.2. Complexity variation in learner essays 147
6.2.3. Does mother tongue matter? 151
6.3. Summary 159

7. Summary and Conclusion **161**
7.1. Results 162
7.2. Discussion 168

Bibliography **175**

Appendix A. Tables **187**
A.1. Targeted file manipulation: Tukey's HSD tables 187
A.2. Case studies: Standard deviation in the ICLE national subsets . . 194

Appendix B. R scripts **197**
B.1. Basic web scraper 197
B.2. Multiple distortion and compression script 199
B.3. Random sampling function 207

Appendix C. Python script **209**

Appendix D. Shell scripts **211**
D.1. Fix fullstops 211
D.2. Remove corpus markup 212
D.3. Remove punctuation 213
D.4. Remove UTF-8 characters 213

Appendix E. Zusammenfassung **215**

List of Figures

2.1. Extra-linguistic factors impacting on complexity variation 22
2.2. Diagram of a communication system 32
2.3. Entropy of a two-choice situation 34

3.1. Overall complexity hierarchy in Mark 51
3.2. Morphological by syntactic complexity in Mark 54
3.3. Real time drifts in English 55
3.4. Overall complexity hierarchy of the parallel Alice corpus 67
3.5. Overall complexity hierarchy of the semi-parallel Alice corpus . . . 68
3.6. Morphological by syntactic complexity in the parallel Alice corpus 69
3.7. Morphological by syntactic complexity in the semi-parallel Alice corpus 71
3.8. Overall complexity hierarchy in the Euro-Congo-Tunisia news corpus 75
3.9. Overall complexity hierarchy in the Euro-Congo news corpus . . . 76
3.10. Morphological by syntactic complexity in the Euro-Congo-Tunisia news corpus 77
3.11. Morphological by syntactic complexity in the Euro-Congo news corpus 78

4.1. Morphological by syntactic complexity of morphological markers in Alice 94
4.2. Morphological by syntactic complexity of morphological markers in Mark 96
4.3. Morphological by syntactic complexity of morphological markers in the Euro-Congo news corpus 97
4.4. Morphological by syntactic complexity of constructions in Alice . . 99
4.5. Morphological by syntactic complexity of constructions in Mark . . 102
4.6. Morphological by syntactic complexity of constructions in the Euro-Congo news corpus 103

5.1. Unique strings in the original Alice lexicon 110
5.2. Compressed strings by linguistic category in the three Alice lexica 125

6.1. Overall complexity hierarchy of written BNC registers 138
6.2. Morphological by syntactic complexity of written BNC registers . . 140
6.3. Overall complexity hierarchy of ICLE learner groups 148
6.4. Morphological by syntactic complexity of ICLE learner groups . . 151
6.5. Overall complexity hierarchy of ICLE learner groups in national varieties 155

6.6. Overall complexity hierarchy of German learner groups in ICLE . . 156
6.7. Morphological by syntactic complexity of ICLE learner groups by national variety 157
6.8. Morphological by syntactic complexity of German learner groups in ICLE 158

7.1. Overall complexity ranking of newspapers 167
7.2. Analytical pipeline 171

E.1. Globale Komplexitätshierarchie der Zeitungsgenres 222

List of Tables

2.1. Overview of relevant empirical studies on complexity 26

3.1. Number of words and sentences in the Gospel of Mark 45
3.2. Spelling variance in the Gospel of Mark 46
3.3. Segmentable inflected word tokens in Mark 52
3.4. Word order patterns in Mark 52
3.5. Number of words and sentences in the Alice database 57
3.6. Number of sentences in the newspaper corpora 58
3.7. Standard deviations of complexity scores in the parallel Alice corpus 61
3.8. Standard deviations of complexity scores in the Euro-Congo corpus 62
3.9. Standard deviations of complexity scores in the Euro-Congo-Tunisia corpus 63
3.10. Standard deviations of complexity scores in the semi-parallel Alice corpus 64
3.11. Standard deviations of file sizes in the semi-parallel Alice corpus . 65
3.12. Syntactic complexity rankings in parallel and semi-parallel Alice corpora 73
3.13. Syntactic complexity rankings in parallel and semi-parallel Alice copora 80

4.1. Text frequency of morphorphological markers and constructions per text type 84
4.2. Number of sentences and words per text genre 85
4.3. Standard deviations of complexity scores by text and construction 90
4.4. Standard deviations of complexity scores by text and construction 91
4.5. Syntactic correlation of morphological markers 92
4.6. Morphological correlation of morphological markers 93
4.7. Morphological ranking of morphological markers 95
4.8. Syntactic ranking of morphological markers 95
4.9. Syntactic correlation of constructions 98
4.10. Morphological correlation of constructions 98
4.11. Morphological ranking of constructions 100
4.12. Syntactic ranking of constructions 101

5.1. Distribution of unique strings in the original Alice lexicon 109
5.2. String length in the original Alice lexicon 111
5.3. Compressed strings by linguistic category in the original Alice lexicon 114
5.4. Distorted passages from Alice 117

5.5. Distribution of unique strings in the morphologically distorted Alice lexicon . . . 119
5.6. String length in the morphologically distorted Alice lexicon . . . 120
5.7. Distribution of unique strings in the syntactically distorted Alice lexicon . . . 122
5.8. String length in the syntactically distorted Alice lexicon . . . 123
5.9. Compressed strings by linguistic category in the three Alice lexica 127

6.1. Overview of the written macro-registers and the newspaper micro-registers in the BNC . . . 132
6.2. Standard deviations of file sizes in the BNC . . . 135
6.3. Standard deviations of complexity scores in the BNC . . . 136
6.4. Number of argumentative essays according to years of studying English at school and university . . . 142
6.5. Learner groups by years of instruction in English at school and university . . . 143
6.6. Standard deviations of file sizes in ICLE . . . 145
6.7. Standard deviations of complexity scores in ICLE . . . 146
6.8. SLA measures versus Kolmogorov measures . . . 150
6.9. ICLE learner groups by national background . . . 153

A.1. Tukey's HSD for morphs in Alice . . . 188
A.2. Tukey's HSD for morphs in Mark . . . 189
A.3. Tukey's HSD for morphs in the Euro-Congo corpus . . . 190
A.4. Tukey's HSD for constructions in Alice . . . 191
A.5. Tukey's HSD for constructions in Mark . . . 192
A.6. Tukey's HSD for constructions in the Euro-Congo corpus . . . 193
A.7. Standard deviations of complexity scores in the ICLE national subset 195
A.8. Standard deviations of complexity scores in the German ICLE subset 196

Preface and Acknowledgements

Partial summaries of the research discussed in this study have appeared as Ehret & Szmrecsanyi (2016a), Ehret & Szmrecsanyi (2016b), Ehret (2014). I am grateful for the generous funding provided by the Cusanuswerk (Bonn) without which none of this work would have been possible. I would also like to acknowledge and thank the following individuals:

– Benedikt Szmrecsanyi, for helpful discussion, valuable feedback, advice, support, and for being a really good supervisor.

– Christoph Wolk, for discussion, extensive advice on statitistics and, for help with advanced `R` magic.

– Jens Stimpfle, for practical support with `Python` and shell programming as well as the retrieval of the `gzip` lexicon.

– Annemarie Verkerk and the Max Planck Institute for Psycholinguistics (Nijmegen) for providing me with some of the *Alice's Adventures in Wonderland* texts.

– Bernd Kortmann and Christian Mair, for helpful advice and feedback.

– Alex Housen, Lourdes Ortega, Amir Zeldes, and Kimmo Kettunen, for helpful comments and feedback.

– The audiences of the following conferences and workshops, for their critical questions, feedback and comments:

  - IClaVE 8, Leipzig, May 2015.
  - Colloquium on "Cross-linguistic aspects of complexity in SLA", VUB Brussels, December 2014.
  - ISLE 3, Zurich, August 2014.
  - ICAME 35, Nottingham, May 2014.
  - Workshop on "Theoretical and Computational Morphology: New Trends and Synergies", ICL 19, Geneva, July 2013.

– Martin J. Smith, for proofreading, language discussions, and his friendship.

– Lorna Syme, for proofreading.

– Liane Neubert and Martin Böke, for their support, especially during the final stages of my PhD.

– My parents, family and friends, in Germany and Cameroon, for their unfailing support, their love, and for believing in me.

The usual disclaimers apply. All errors are solely mine.

1. Introduction

Marrying quantitative corpus linguistics to information theory, this work contributes to the ongoing linguistic complexity debate by exploring a hitherto underresearched methodology which uses compression algorithms to assess linguistic complexity in corpora. The methodology has the potential of being a radically objective and powerful tool in linguistic complexity research that can serve both as a complementary diagnostic in traditional research as well as a full-blown, independent analysis tool. The central aims are primarily to advance the development and applicability of the method, and secondly to gain understanding of the underlying compression algorithm, and hence, information-theoretic complexity. Therefore, this work is primarily methodological in nature.

The point of departure is the current typological complexity debate and quest for complexity metrics which was set in motion by the provocative claim that some languages are simpler than others (McWhorter 2001b). This claim challenged the assumption that, on the whole, all languages are equally complex (e.g. Crystal 1987; Hockett 1958). Numerous volumes on the topic have since been published (e.g. Dahl 2004; Kortmann & Szmrecsanyi 2012; Miestamo et al. 2008; Sampson et al. 2009) and linguistic complexity continues to be one of the most hotly debated notions among the contemporary linguistic community. Complexity research pivots around three questions:

(i) How can linguistic complexity be defined?

(ii) How can linguistic complexity be measured?

(iii) How can linguistic complexity variation be explained?

Despite the extensive study of linguistic complexity from various angles, no unanimous answer to these questions has been found. Rather, an abundance of definitions has been put forward, each of which is equally valid within its context of research but fails to be universally accepted or applicable. Generally, however, a distinction between *absolute complexity* and *relative complexity* is made (Miestamo 2006; Miestamo et al. 2008). Absolute complexity is a theory-oriented notion and is interested in the complexity inherent in a linguistic system, while relative complexity notions define complexity in relation to a language user. The latter, therefore, tend to be more

applied and usage-oriented than the former. In terms of complexity measures, the *status quo* is similar. Various measures have been proposed but they draw either on empirically expensive evidence or are highly selective and subjective in nature.

In addressing these issues, I explore and extend an unsupervised, algorithmic measure—in the following also referred to as *compression technique*—that has its roots in information theory and was first proposed by the Finnish mathematician Juola (1998). Essentially, this measure boils down to the notion of *Kolmogorov complexity*, which defines the complexity of a given text sample as the length of the shortest possible description of this text sample. Imagine a sort of *I spy* game in which two different objects have to be described. The goal of this particular game is to use as few words as possible to fully describe the two objects. On the basis of these descriptions, the complexity of the objects can be determined. The more words are needed to describe the object—while being as concise as possible—the more complex this object is. In plain English, the longer the shortest description of an object is, the more complex is this object.

In this spirit, the linguistic complexity in text samples is measured by approximating their information content, or Kolmogorov complexity, with compression algorithms. The basic idea is that text samples which can be compressed comparatively better, i.e. more efficiently, are linguistically comparatively less complex. Kolmogorov complexity is an absolute notion of complexity and based on the form of structures, not on their function or meaning. In other words, Kolmogorov complexity is agnostic about deep linguistic form-function pairings. In the realm of linguistics then, Kolmogorov-based information-theoretic complexity is a measure of *structural surface redundancy*. This is another way of saying, in very simplified terms, that it measures the recurrence and repetition of orthographic character sequences (structures) in a text. Kolmogorov complexity conflates, to some extent, the following notions of complexity:

- ✔ *Quantitative complexity:* the number of grammatical contrasts, markers or rules in a linguistic system. More rules are equated with more complexity (Dahl 2004; McWhorter 2001b; Shosted 2006).
- ✔ *Irregularity-based complexity:* the number of irregular grammatical markers in a linguistic system. Irregular markers are regarded as more complex than regular markers (Kusters 2003; McWhorter 2001b; Trudgill 2004).

Importantly, it does not encompass:

- ✘ *Redundancy-based complexity:* linguistic markers, forms or categories without grammatical and communicative function (McWhorter 2001b; Seuren & Wekker 1986; Trudgill 1999).

- ✘ *L2 acquisition difficulty:* linguistic features which are difficult to acquire for adult language learners are complex (Kusters 2003; Szmrecsanyi & Kortmann 2009; Trudgill 2001).

On the methodological plane, the central characteristics and advantages of this measure can be summarised as follows.

Objective.

One of the inherent features and major assets of the compression technique is its unparallelled objectivity. The compression technique does neither require the *a priori* categorisation of linguistic features into complex or simple, nor the subjective selection of some features over others common to most traditional complexity metrics. In fact, the compression technique is agnostic about form-meaning relationships and possesses no linguistic knowledge of the texts it is applied to, i.e. its measurements are unsupervised and radically objective.

Economical.

The compression technique is an economical means of measuring linguistic complexity because it is easily implemented, and can in principle be applied to any orthographically transcribed text database. Thus, the measure does not rely on empirical evidence which is labour-intensive to obtain and expensive to reproduce.

Usage-based.

The compression technique is a usage-based methodology and is based on naturalistic, authentic language samples rather than, for instance, on paradigmatic analyses.

Usage-based approaches to linguistics have become increasingly popular in recent years and are, essentially, concerned with the interaction between grammar and the mind. More specifically, they are interested in the effect of language use on the cognitive representation of language (Bybee 2006: 712). The underlying principle of all of these theories is that grammar emerges directly from, and is influenced by, actual language usage (e.g. Bybee 2006, 2010; Ellis 1998; Langacker 1987, 1988; Tomasello 2003). Language is seen as a symbolic dimension consisting of form-function units which are connected to the meaning they transmit, and the communicative situation they are used in. As such, grammar emerges through the use of symbolic units and their grammaticalisation (Behrens 2009: 384–385; Tomasello 2003). Bybee (2006) for instance, defines grammar as the cognitive representation of a speaker's experience with language (Bybee 2006: 711), i.e. language usage determines, shapes and changes a speaker's linguistic system. Usage-based theories are particularly prominent and have produced

invaluable insights in the field of language acquisition research. In a usage-based framework, language is acquired solely through exposure and the help of general human cognitive processes and capabilities such as categorisation, generalisation and analogy, which permit the human mind to construct a grammar of the input language (Bybee 2006: 711; see also Tomasello 2003: 3–4). The usage-based approach can be summarised with the following quote: "language structure emerges from language use" (Tomasello 2003: 5, 327).

Language phenomena, their patterning and usage can be studied in naturalistic text corpora which sample written or spoken language produced by language users. This means that compression algorithms—which work on orthographically transcribed texts and return measurements that are directly based on naturalistic language—constitute an inherently usage-based means for studying linguistic complexity.

Holistic.

As mentioned above, the compression technique does not require manually selected features as input but works directly on the data. Since algorithmically measured complexity is not restricted to specific linguistic features, the compression technique constitutes a holistic means of measuring linguistic complexity.

It goes without saying that the compression technique is not flawless (for a more detailed discussion on drawbacks and advantages refer to Section 7.2).

Agnostic.

Compression algorithms are completely agnostic about deep linguistic structures such as form-function pairings and possess no knowledge of the compressed texts.

Text-dependent.

The methodology is strictly text-based and its measurements are to some extent text-dependent, i.e. they depend on orthographic transcription conventions and, in the case of non-parallel corpora, on the propositional content of the texts. Orthography is important because variant spellings of the same word (e.g. *neighbourhood* and *neighborhood*) affect the compressibility of texts and increase their complexity. Content control is a crucial factor that needs to be considered when working with non-parallel samples in order to ensure that the measurements are comparable and reliable. Thus, the compression technique cannot be used with randomly chosen texts (see also Chapter 3, Section 3.2).

Input-dependent.

Furthermore, the performance of the compression technique relies on the quality and quantity of the input data. Generally, the compression technique produces more reliable measurements with larger datasets, and text samples that are compared should be of the same size. In terms of data quality, the input texts need to be carefully prepared, i.e. any non-textual information should be removed and orthography has to be normalised.

Morphology-sensitive.
In this work, Kolmogorov-based information-theoretic complexity is defined as a measure of structural surface redundancy and refers to the recurrence of orthographically transcribed character sequences in a text. As structural redundancy and morphological complexity are somewhat correlated, the compression technique has a slight tendency of favouring morphological complexity. Therefore, large amounts of structural redundancy in a text can affect the measurements of overall complexity.

Going beyond the mere application of the compression technique, the current work provides a first in-depth analysis of Kolmogorov complexity in linguistic terms and assesses the metric from a linguistically responsible perspective by exploring the workings of compression algorithms. It furthermore aims at the development and advancement of the method. Specifically, the analysis is guided by the following research questions and objectives, which will be discussed in turn below.

(i) Can compression algorithms be applied to data other than parallel copora?

(ii) Can compression algorithms measure the complexity of specific linguistic features?

(iii) What do compression algorithms, linguistically speaking, actually measure?

(iv) How well do compression algorithms capture intra-linguistic, i.e. within language, complexity variation in naturalistic corpora?

The first question regards the applicability of the compression technique to different data types. Previous research using compression algorithms for measuring linguistic complexity restricted the analysis to parallel text databases—basically translational equivalents of one text in different languages—in order to rule out differences in the propositional content of the texts. While such parallel text databases are ideal for the study of cross-linguistic complexity variation, the restriction of the method to parallel corpora poses a severe limitation to algorithmic complexity research of other

areas (e.g. intra-linguistic complexity research). It is therefore imperative to test the applicability of the compression technique to data other than parallel corpora. Thus, I assess to which extent the propositional content of the data analysed influences the results and apply the compression technique to different data types, i.e. parallel, semi-parallel and non-parallel databases, as well as naturalistic large-scale corpora of English. While compression algorithms work well on various data types, content control and data sparsity are issues to be considered when working with non-parallel and naturalistic corpora. This is another way of saying that the compression technique cannot reliably measure samples of randomly chosen texts, and generally returns more robust results when applied to larger datasets. The Lord's prayer counting 52 words for instance, is too small, but the Gospel of Mark counting about 15,000 words in the English Standard Version is sufficiently large to return reliable measurements.

Second, can compression algorithms be used to measure the complexity of specific linguistic features? Hitherto algorithmic complexity research focused on measuring complexity from the bird's eye perspective, i.e. the complexity of morphology and syntax were measured in their entirety as sub domains of a language. The present work introduces a new, modified version of the classic compression technique which allows the measurement of morphological and syntactic complexity from the jeweller's eye perspective (see also Ehret 2014). Put differently, the compression technique can be used to measure specific morphosyntactic features in English.

The third question is concerned with what I dub the *black box conundrum*. Although it is generally good news that the compression technique returns results which are linguistically meaningful and interpretable, the responsible linguist should dig deeper and try to understand *why* the method works and *what* the method actually does. It is therefore one of the major objectives of this work to determine what exactly compression algorithms like `gzip` measure and, figuratively speaking, to take a look into the black by analysing the workings of the algorithm. On the basis of `gzip` 's lexicon output—a collection of compressed text sequences—I provide a detailed analysis of algorithmically recognised strings and define information-theoretic, Kolmogorov-based complexity in linguistic terms.

Finally, the fourth question, while being of a methodological nature is at the same time the most "linguistic" research question in this set, because it relates to the measurement of intra-linguistic complexity variation. This question has so far not been addressed in the literature—which focuses on the cross-linguistic measurement of Kolmogorov complexity. The present work, pertaining to the Freiburg school of complexity research (c.f. Kortmann & Szmrecsanyi 2009, 2012; Szmrecsanyi & Kortmann 2009), seeks to fill this gap and shed light on information-theoretic, intra-linguistic, as opposed to cross-linguistic, complexity variation in English. In two case studies (see Chapter 6) I show that complexity variability in British Eng-

lish registers as well as learner Englishes can be successfully measured with the compression technique.

In essence, this work demonstrates first, how the compression technique can be applied to non-parallel corpora, and second, how it can be applied to different text types and varieties of English.

This work is structured as follows.

Chapter 2 gives a detailed overview of the literature and theoretical background on linguistic complexity research as well as information theory. Firstly, the origin of the current linguistic complexity debate will be traced from the middle ages to the present day and major concepts and notions of linguistic complexity will be discussed. Furthermore, different metrics of complexity and factors held responsible for complexity variation will be presented. It is worth noting that I am motivated by and focus on typological and sociolinguistic complexity research. Second language acquisition (e.g. Larsen-Freeman 2006; Ortega 2003) or generative approaches (e.g. Newmeyer & Preston 2014) will therefore not be reviewed in detail. Secondly, concepts of information theory which are indispensible for the understanding of the methodology will be introduced. Starting with Shannon's mathematical theory of communication, this section will describe how information can be quantified and measured. This is followed by a review of Kolmogorov-based information-theoretic metrics so far explored in the literature, and a concise definition of information-theoretic complexity in the context of this work.

Chapter 3 is devoted to the validation and extension of the compression technique in three steps. Up to now, the application of the compression technique was restricted to parallel text corpora, i.e. translational equivalents of the same text in different languages. This chapter demonstrates that the compression technique yields linguistically meaningful results and can also be used with semi-parallel and non-parallel texts.

In the first step, the compression technique as proposed by Juola (2008) is applied to a parallel text database comprising the Gospel of Mark[1] in a handful of historical varieties of English and six other languages (Esperanto, Finnish, French, German, Hungarian, and Latin). It is demonstrated that compression algorithms can be utilised to assess the linguistic complexity of parallel texts on an overall, morphological and syntactic level. For instance, according to my measurements, Hungarian and Finnish are overall rather complex languages while all the English Bible versions, with the exception of the West Saxon version, are overall comparatively simple. The comparison of morpholo-

[1]Throughout this work the Gospel of Mark will also be referred to as "Mark".

gical and syntactic complexity of the English Bible versions furthermore depicts the development of English from a morphologically complex language, to a syntactically complex language over time. In fact, the algorithm captures both cross-linguistic as well as intra-linguistic complexity variation and provides results which dovetail with findings of previous research (Bakker 1998; Nichols 1992).

In the second step, an extended, statistically more robust version of the compression technique is introduced and applied to a parallel and—after permutation—semi-parallel database of *Alice's Adventures in Wonderland* by Lewis Carroll[2] which spans nine European languages (Dutch, English, Finnish, French, German, Hungarian, Italian, Romanian, and Spanish). The overall, morphological and syntactic complexity of the nine languages in both datasets is measured, and complexity rankings are established on the basis of these measurements. Subsequently, the rankings of the parallel and semi-parallel datasets are compared. The results show that the compression technique yields linguistically meaningful results with both parallel and semi-parallel data.

In the third step, the compression technique is utilised with two non-parallel corpora of newspaper texts which sample the same nine languages as the Alice database. After measuring the overall, morphological and syntactic complexity of the nine languages, the rankings of the newspaper measurements are compared to the ranking of the parallel Alice database, which serves as reference. I find that the rankings of morphological and syntactic complexity in the newspaper corpora are largely congruent with the Alice database. Yet, the overall complexity measures correspond only moderately well. All in all, the algorithmic measurement of complexity in non-parallel texts is possible but requires a certain amount of content control. This is another way of saying that the propositional content of the components of a non-parallel corpus should be similar, as random texts cannot be reliably measured with the compression technique.

Chapter 4 is an excursion into previously unexplored territory and presents a new flavour of the compression technique; targeted file manipulation. Essentially, targeted file manipulation combines the systematic removal of target structures from a text with the compression technique, i.e. distortion and subsequent compression. Expanding on previous work by the author (Ehret 2014), the contribution of a handful of morphological markers and functional constructions to the complexity in three different English texts is analysed. On an interpretational level, the textual complexity of these individual features

[2]Throughout this work *Alice's Adventures in Wonderland* will also be referred to as "Alice".

on the morphological and syntactic level is derived from their complexity contribution to the text. The focus is thereby put on assessing the extent of intertextual variation in regard to the features' textual complexity.

In general terms, I show that the presence of more morphological marker types leads to an increase in morphological complexity of the texts while, at the same time, it facilitates the algorithmic prediction of (morpho)syntactic patterns. Furthermore, the results imply that invariant grammatical markers such as the future marker *will* increase simplicity. These findings hold across different texts, i.e. intertextual variation is negligible.

In methodological terms, targeted file manipulation is shown to be an effective method for measuring complexity trends of specific linguistic features in English texts. Expressly, algorithms can be utilised for the detailed analysis of morphological and syntactic complexity. Thus, I fill a hitherto unaddressed gap in information-theoretic complexity research which surveys algorithmic complexity from a bird's eye perspective (e.g. Ehret & Szmrecsanyi 2016b; Juola 2008; Sadeniemi et al. 2008).

Chapter 5 explores the actual algorithm of the open source compression program `gzip` , which is used as compressor in the studies presented in this work, and aims at defining information-theoretic complexity in linguistic terms. To this end, `gzip` 's lexicon—a line-by-line output of compressed text sequences—is retrieved for *Alice's Adventures in Wonderland*. The first section of this chapter gives a detailed description of the distribution of compressed strings in the lexicon of the original Alice text. Furthermore, every string is manually analysed and annotated for linguistic (e.g. lexical words or phrasal constructions such as *to see*, *looked anxiously␣*) and non-linguistic category (e.g. random strings such as *gree␣*). Finally, the lexica of a syntactically and a morphologically distorted version of Alice are described, annotated and compared to the original Alice lexicon.

The results reveal that compression algorithms such as `gzip` do capture recurring linguistic (surface) structures such as suffixes, verbs or whole phrases. Needless to say, the algorithm does not systematically select or prefer linguistically meaningful units over random strings. In the light of these findings, Kolmogorov-based information-theoretic complexity as measured with lexicon-based algorithms is defined as a measure of structural surface redundancy. This chapter furthermore shows that the process of distortion affects the compressibility of texts as intended.

Chapter 6 presents two case studies in which the compression technique is

used to assess intra-linguistic complexity variation in the *British National Corpus* (BNC) and the *International Corpus of Learner English* (ICLE). On a methodological plane, this chapter is concerned with the applicability of the compression technique and the extent to which algorithmic measurements can approximate complexity in large-scale naturalistic corpora.

In the first case study, the complexity variation on the overall, morphological and syntactic tier in twenty different written registers of British English as sampled in the BNC is measured. I find that the registers analysed vary in their complexity such that less formal registers like email or letters are generally less complex than more formal registes such as newspapers. This ties in with the register variation along the involved-abstract dimension reported in Biber (1988) and, to some extent, in Szmrecsanyi (2009).

In the second case study, I assess the overall, morphological and syntactic complexity of essays written by students with different levels of instructional exposure in English and evaluate the influence of the learners' national background / mother tongue on the complexity of their text production. All other things being equal, the amount of instruction received in English is taken as a proxy for the proficiency of the learners. The results indicate that higher amounts of instructional exposure leads to increased overall and morphological complexity. The production of less advanced learners, in contrast, is marked by increased syntactic complexity. Furthermore, Kolmogorov measures of learner language systematically correlate with SLA measures of complexity. Some evidence is found which suggests that the complexity of learner essays in ICLE is influenced by the learners' mother tongue background but that, all in all, the relationship between learner essay complexity and instructional exposure is rather stable across different backgrounds. Further research is needed to clarify the relationship between national background and the complexity of learner language.

On the whole, this chapter provides empirical evidence for the applicability of the compression technique to naturalistic corpus resources since the results are in line with what more traditional research reports.

Chapter 7 provides a short summary and discussion of the results considering the research questions introduced above. Specifically, I will evaluate the applicability of the compression technique and its significance for linguistic complexity research. I will conclude by pointing out advantages and drawbacks of the method.

2. Literature review and theoretical background

2.1. Linguistic Complexity

2.1.1. Defining linguistic complexity

Linguistic complexity is one of the most hotly debated notions in present-day linguistics. Some ingredients of the present debate can already be found in philosophical approaches to language which emerged among eighteenth and nineteenth century philosophers and scholars, including prominent figures like Herder and Humboldt. Even though a sense of differing values of languages was already present during the middle ages—some languages, in particular Latin and Greek, were considered more appropriate than others—the concept of superiority of some languages over others arose with the emergence of European nation states and the concomitant growing nationalism (Leavitt 2011: 16–19). By the eighteenth century, philosophers in France, Britain and Germany roughly distinguished languages along a scale from savage and crude to refined and sophisticated: "[...] one important tendency in the second half of the century was to define poles of language types. At one extreme were languages that were wilder, closer to savagery and nature [...]. At the other extreme were languages that were highly, even too refined [...]" (Leavitt 2011: 70). While these classifications of languages were rather a by-product of nationalist ideas and philosophies, they foreshadowed the evaluative judgements inherent in later classifications of languages. A first scientific categorisation of languages was put forward by the brothers Friedrich and August von Schlegel. In his treatise 'Über die Weisheit und Sprache der Indier' (1808), Friedrich von Schlegel proposes a two-fold classification dividing languages into flective and affixive types (Schlegel 1808). Inherent in this classification is the evaluative judgement that inflectional (Indo-European) languages such as Greek are superior and are therefore to be preferred whereas languages such as Chinese are inferior (Schlegel 1808: 44–59). In this spirit, von Schlegel writes about the Chinese language: "Die Sprache dieser sonst so verfeinerten Nation stuende also grade auf der untersten Stufe" (1808: 49). August von Schlegel later added a third category, "languages without grammatical structure" (Schlegel 1818: 14), i.e. isolating. Like his brother, he judges inflectional languages as superior and claims that isolating languages due to their lack

of grammar must place a great obstacle to the intellectual development of peoples and their cultures.

> Les langues [...] si divisent en trois classes: les langues sans aucune structure grammaticale, les langues qui emploient des affixes, et les langues a inflexions 6. Les langues de la première classe n'ont qu'une seule espèce de mots, incapables de recevoir aucun développement ni aucune modification. [...] Il n'y a dans ces langues ni déclinaisons, ni conjugaisons, ni mots dérivés, ni mots composés autrement que par simple juxta-position, et toute la syntaxe consiste à placer les élémens inflexibles du langage les uns à côté des autres. De telles langues doivent présenter de grands obstacles au développement des facultés intellectuelles [...] Je pense, cependant, qu'il faut assigner le premier rang aux langues à inflexions.
>
> (Schlegel 1818: 14–15)

In a similar vein, Wilhelm von Humboldt (1994, 1836) categorises languages on a scale from primitive to perfected, and claimed that the language of a nation and their thought or 'ideas' are invariably connected such that "[...] alles durch Rede Gewirkte aber immer ein zusammengesetztes Erzeugniss des Geistes und der Sprache ist. Jede Sprache muss in dem Sinne aufgefasst werden, in dem sie durch die Nation gebildet ist [...]" (von Humboldt 1994: 55). He thus ascribed differences between languages to the differing mental capacities of the nations who speak it.

> Ueberall ist in den Sprachen das Wirken der Zeit mit dem Wirken der Nationaleigenthümlichkeit gepaart [...]. Auf diese Weise nun ist eine fortschreitende Entwicklung des Sprachvermögens, und zwar an sichren Zeichen, erkennbar, und in diesem Sinn kann man mit Fug und Recht von stufenartiger Verschiedenheit unter Sprachen reden.
>
> (von Humboldt 1994: 52–53)

Against this backdrop of (d)evaluative judgements about the complexity of languages the hypothesis that all languages are of equal complexity sprang up and was commonly agreed on throughout the twentieth century (Akmajian et al. 1997; Bickerton 1995; Crystal 1987; Edwards 1994; Fortson 2004; Hockett 1958; O'Grady et al. 1997; Wells 1954). One of the first chapters in the *Cambridge Encyclopedia of Language* (Crystal 1987) is dedicated to the topic of language complexity stating that all languages are overall considered to be of equal complexity. According to Crystal there is no such thing as a 'primitive' language; all natural languages have equally complex grammars which have no limitations as regards their expressiveness.

> It comes near to stating the obvious that all languages have developed to express the needs of their users, and that in a sense all languages

> are equal. [...] All languages have a complex grammar: there may be relative simplicity in one respect (e.g. no word-endings), but there seems always to be relative complexity in another (e.g. word-position).
>
> (Crystal 1987: 6)

This statement further hints at another assumption going hand in hand with the equal-complexity hypothesis, namely that even though some languages appear to be simpler in one linguistic domain, this simplicity is inevitably compensated for in another domain. This trade-off relationship between sub-domains of a language, specifically morphology and syntax, is more explicitly formulated in the oft-times quoted passage by structuralist linguist Charles Hockett (1958).

> Objective measurement is difficult, but impressionistically it would seem that the total grammatical complexity of any language, counting both morphology and syntax, is about the same as that of any other. This is not surprising, since all languages have about equally complex jobs to do, and what is not done morphologically has to be done syntactically. Fox, with a more complex morphology than English, thus ought to have somewhat simpler syntax; and this is the case. Thus one scale for the comparison of the grammatical systems of different languages is that of average degree of morphological complexity— carrying with it an inverse implication as to degree of syntactical complexity.
>
> (Hockett 1958: 180–181)

While lacking empirical back-up, the long-standing assumption that all natural languages are of equal linguistic complexity—in current publications referred to as equi-complexity dogma (Kusters 2003: 5), principle of invariance of language complexity (Sampson 2009: 1) or very short ALEC (All Languages are Equally Complex) (Deutscher 2009: 234)—had remained unchallenged for much of the twentieth century. Whatever the reasons for the tenacity of this assertion may have been, the alleged truism has very recently been questioned and scrutinised (Kusters 2003; McWhorter 2001b; Shosted 2006). Most notably it was criticised by McWhorter in a somewhat provocative article suggesting that creoles have the world's simplest grammars (McWhorter 2001b), thereby kicking off a heated debate between the defenders of the equal-complexity hypothesis and its challengers. The major objections against the finding that, after all, some languages are simpler than others, seem to be motivated by devaluative judgements about language complexity originally postulated by nineteenth century scholars. These scholars adhered, as Szmrecsanyi & Kortmann (2012) put it succinctly, to the idea that "complex is beautiful, simple is retarded" (Szmrecsanyi & Kortmann 2012: 8). However, these are antiquated attitudes and present-day linguistics is not interested in proving the value of one language over another. Kusters (2003) writes about his definition of complexity that

"no evaluative judgement is assumed: neither complexity nor simplicity are taken to be of a higher value" (Kusters 2003: 8). Thus both simple and complex languages "[...] fulfill all of the functional needs of human language" (McWhorter 2002: 36). Particularly since McWhorter's (2001b) article and the ensuing debate in the special issue of *Linguistic Typology* 5 2 / 3, the notion of linguistic complexity has been investigated with renewed vigour, both in the typological and sociolinguistics camp (e.g. Dahl 2004; Kortmann & Szmrecsanyi 2012; Miestamo et al. 2008; Sampson et al. 2009) and in second language acquisition research (e.g. Larsen-Freeman 1978, 2006; Ortega 2003, 2012; Pallotti 2015). Most recently, linguistic complexity has also been investigated from a generative and psycholinguistic perspective (e.g. Culicover 2013; Newmeyer & Preston 2014; Järvikivi et al. 2014).

The current work contributes to the ongoing typological complexity debate and will therefore only review the relevant typological and sociolinguistic literature. This line of research is primarily motivated by the question "Are all languages equally complex?", and is interested in linguistic complexity *per se*. A detailed review of second language acquisition research on complexity as well as generative and purely psycholinguistic studies is therefore outside the scope of this work. Suffice it to say that second language acquisition research is interested in linguistic complexity as a means for assessing aspects of second language acquisition, i.e. it aims to describe and assess second language performance, production and proficiency (Ortega 2012: 128). Consequently, the measures of complexity commonly used in second language acquisition studies differ considerably from the metrics of typologically motivated complexity research. The former mainly focus on syntactic measures, which are often considered indicators for overall language proficiency (Ortega 2003: 492). Commonly used metrics are, for example, the length of T-units, clauses and sentences, or the degree of clausal embedding and coordination, whereby a higher degree of any of these measures usually indicates a higher degree of (inter)language complexity (Ortega 2012: 127, 139). While recent formal and psycholinguistic approaches also aim at measuring linguistic complexity as such—rather than, say, benchmark proficiency—, they are predominantly concerned with cognitive aspects of linguistic complexity, i.e. they favour experimental set-ups and focus on the mental representation of, or capacity, to process language (Newmeyer & Preston 2014; Järvikivi et al. 2014). [1]

The central issues in the typological complexity debate are (i) finding a generally applicable definition of what exactly linguistic complexity is, (ii) measuring this complexity, and (iii) explaining variation of linguistic complexity.

Before tackling the issue of defining linguistic complexity, a few lines will be dedicated to the distinction between *global* and *local complexity*

[1] The structure of this section is based on Kortmann & Szmrecsanyi (2012) and provides an extension and critique of their review.

(Miestamo 2008: 29–32). As a heritage of the equal-complexity hypothesis some studies (Juola 2008; McWhorter 2001b, 2008) aim at establishing metrics in order to measure global / overall, linguistic complexity of languages. Addressing overall complexity, however, is a somewhat complex task and studies set to work on it are faced with two related problems: firstly the matter of representativity, and secondly the issue of comparability (Miestamo 2008). In order to measure overall linguistic complexity, a metric encompassing all linguistic levels of a language would need to be defined, and within each of these levels all grammatical aspects would need to be quantified. The scope of such a study would most likely surpass the workload doable by any team of linguists. Therefore, assessing the overall complexity of a language let alone a set of several languages is a sheer impossible task. If we assume, for the sake of the argument, that a study of such an extent was manageable, we are still left with the problem of comparability across different languages. Comparability requires this overall metric to implement a categorisation of all aspects of grammar cross-linguistically and then subsume each aspect under one measure. However, even if all categories could be attested for in all languages in the sample, the question remains how the individual grammatical aspects should be weighted. As Deutscher puts it:

> Since it is not possible to collapse the list of complexity measures [...] into one overall figure, no non-arbitrary single measure of overall complexity can be defined. The overall complexity of a language A can only be viewed as a vector (one-dimensional matrix) of separate values ($A_1 ... A_n$), each representing the measure for one of the n subdomains. In set-theoretic terms, the n different subdomains will give n distinct total orders on the set of languages, and as these orders do not necessarily coincide, the result will only be a partial order on the set of languages. This means that it will not be possible to compare any two given languages in the set for overall complexity.
>
> (Deutscher 2009: 294–250)

All in all, measuring overall complexity is at best a can of worms or a "chase after a non-existent wild goose" (Deutscher 2009: 251). Therefore, most research focuses on measuring complexity in different sub-domains of language. So far the following levels of language have been subject to complexity analyses:

(i) *Phonology.* The size of phoneme inventories, the number of phonetic contrasts and vowel / consonant distinctions as well as the presence / absence of a tonal system are indicators for phonological complexity (e.g. Nichols 2009; Shosted 2006; Trudgill 2004).

(ii) *Syntax.* Syntactic complexity is often measured by counting the number of word order rules, or the degree of clausal embedding / subordination (e.g. Karlsson 2009; Sinnemäki 2008).

(iii) *Morphology.* Inflectional morphology, i.e. the number of inflectional markers / bound morphemes but also the presence of homonymy, fusion and allomorphy are taken as indicators for complexity on the morphological level (e.g. Gil 2008; Kusters 2003, 2008; Szmrecsanyi & Kortmann 2009).

(iv) *Semantics and lexicon.* The number of syllables or monosyllabic words, but also the number of verbal derivations, root distinctions and complex lexical patterns such as compounding, reduplication or verb serializations are measures of semantic and lexical complexity (e.g. Fenk-Oczlon & Fenk 2008; Nichols 2009; Riddle 2008).

(v) *Pragmatics.* Pragmatic complexity is relatively rarely explored. Bisang (2009) also calls it 'hidden complexity' as it refers to, for instance, the lack of explicit marking or multi-functional structures which are considered as complex due to the fact that one structure expresses a wide range of different meanings (Bisang 2009).

Despite the differing terminology and definitions of complexity across individual studies, two major concepts of complexity can generally be distinguished: absolute complexity and relative complexity (Miestamo 2006, 2008, 2009). Absolute complexity is a theory-oriented notion of complexity and is understood as independent and unrelated to a language user. It refers to the amount of features / parts in a linguistic system or, in information-theoretic terms, to the length of the description of a given linguistic phenomenon (Dahl 2004). As such, absolute complexity can be considered 'objective' in the sense of its being concerned with the complexity inherent in a linguistic system of a given language and therefore autonomous of any agent. Relative complexity on the other hand is a 'subjective', agent-related kind of complexity. This approach defines complexity in terms of cost, processing or acquisition difficulty as experienced by and relative to a language user (i.e. speaker, hearer or learner) (Miestamo 2006, 2008). Mostly, this language user is a second language learner and, in fact, relative complexity is often equated to *second language acquisition difficulty* (Kusters 2003, 2008; Szmrecsanyi & Kortmann 2009; Trudgill 2001). In the following paragraphs some popular notions of absolute and relative complexity will be presented. Although complexity metrics are as numerous as the publications on this topic, most metrics fall into one of the categories introduced below. I will further review some of the previous research on complexity which, essentially, established these metrics. A tabular survey of relevant typological and sociolinguistic complexity research as well as some second language acquisition approaches to complexity published since 2000 is given in Table 2.1 at the end of this chapter.[2]

[2]Note that, for reasons of relevance, formal approaches to complexity will not be reviewed.

Quantitative complexity. The motto of this metric is "more is more complex" (Arends 2001: 180). In other words, quantitative complexity is interested in the amount of grammatical contrasts, markers or rules in a linguistic system. The more grammatical contrasts, markers or rules a given language possesses, the more complex is this language: "The guiding intuition is that an area of grammar is more complex than the same area in another language to the extent that it encompasses more overt distinctions and / or rules than another grammar [...]" (McWhorter 2001b: 135). Quantitative complexity is an absolute notion of complexity. Studies using quantitative complexity metrics are, for instance, McWhorter (2001b, 2012) where more marked elements in the phonemic inventory, more syntactic rules and semantic distinctions are taken as indicators of more complexity. Specifically, McWhorter (2001b) established this metric to demonstrate that creoles which are relatively 'young' languages are grammatically less complex than older languages. In his typological approach to complexity, Shosted (2006) counts structural units of two linguistic domains, i.e. morphology and phonology, and quantifies their combinatory possibilities in order to measure the correlation between the two categories. While these studies refer to and quantify the actual number of features in a linguistic system, Dahl (2004) suggests to measure complexity in information-theoretic terms. The complexity of a given linguistic (sub)system or phenomenon would then be "the length of the shortest possible specification or description of it" (Dahl 2004: 21).

Information-theoretic complexity. This notion of complexity, which is by its very nature absolute, is a relative newcomer to the linguistic complexity debate. Although Dahl (2004) urged to ground and define linguistic complexity in information-theoretic terms, this metric has so far received little attention by linguists and literature is not as vastly available as for the more traditional complexity metrics. Information-theoretic complexity, i.e. Kolmogorov complexity, was first implemented by the Finnish mathematician Patrick Juola (1998) and further explored by computer scientists and a handful of linguists (Bane 2008; Ehret & Szmrecsanyi 2016b; Juola 2008; Sadeniemi et al. 2008). It is an unsupervised, algorithmic measure of complexity which is largely based on (ir)regularity and redundancy of linguistic surface structures. An in-depth discussion of this metric will be given in section 2.2.[3]

Irregularity-based complexity. This metric could be subsumed under the motto 'more irregular is more complex' and considers, for instance, irregularities in a language's phoneme inventory or the presence of more

[3]N.b. there are other information-theoretic notions of complexity, which are, however, not relevant to the current work as they are not Kolmogorov-based.

irregular markers such as morphological and / or derivational inflections as more complex:

> Grammars differ in the degree to which they exhibit irregularity and suppletion. English's small set of irregular plurals like children and people is exceeded by the vast amount of irregularity in German plural marking: a masculine noun may take -e for the plural, but almost equally will also take an umlaut [...]. Then grammars differ in the degree to which they exhibit suppletion. Suppletion is moderate in English, especially evident in the verb 'to be' which distributes various Old English roots across person, number, and tense: am, are, is, was, were, been, be. But the Caucasian language Lezgian has no fewer than sixteen verbs that occur in suppletive forms.
>
> (McWhorter 2012: 246)

Sometimes, the notion of (non-)transparency (Nichols 2013; Trudgill 2004) or opaqueness (Mühlhäusler 1974) is mentioned in the same breath as irregularity-based complexity. Nichols (2013), for example, assessing the impact of geographical isolation on language complexity refers to non-transparency between surface structures and underlying forms, i.e. broadly speaking irregularities. Similarly, Trudgill (2004) analyses the relation between society type and linguistic structures. Specifically, he investigates how contact or isolation as well as community size and structure impact on phoneme inventories in Polynesian languages (Trudgill 2004). He finds that simplicity in a language is increased if irregularities are regularised, i.e. are reduced (Trudgill 2004: 307) and argues that regularisation leads to an increase in transparency. It is beyond argument that more irregularities increase the complexity of a linguistic system and, thus might ultimately increase its non-transparency. However, the terms (non-)transparency and opaqueness in this context are problematic. While irregularities of a linguistic system are quantified, the literature implies that these irregularities are non-transparent / opaque structures and therefore difficult to process or learn. This means that irregularities are often evaluated in regard to a language user: "Imperfect learning, that is, leads to the removal of irregular and non-transparent forms which naturally cause problems of memory load for adult learners" (Trudgill 2004: 307). On the basis of this observed 'confusion' in the literature, Szmrecsanyi & Kortmann (2012) argue that—partly being an absolute measure of complexity—irregularity-based complexity is a 'hybrid notion' due to the fact that irregularities as defined by theory are ultimately posing difficulty to a language user (Szmrecsanyi & Kortmann 2012: 12). In spite of this fact, I recommend to consider irregularity-based complexity as being primarily absolute as the irregularities observed are inherent in and a property of a linguistic system

and thus independent of a language user. Needless to say, one should, of course, investigate whether and to which extent this notion as well as all the other absolute notions of complexity are related to language users (Dahl 2004: 40). Furthermore, I argue that transparency-based complexity should be regarded as a separate notion of complexity (see below).

Redundancy-based complexity. Redundancy-based or ornamental complexity is concerned with linguistic markers, forms or categories that have no apparent grammatical function and are dispensable to communication and, therefore, redundant (McWhorter 2001b, 2012; Trudgill 1999): "There are many features commonly found in grammars which are a product of a gradual evolution of a sort which proceeded quite independently of communicative necessity, and must be adjudged happenstance accretion. [...] Crucially, this added complexity emerges via chance, not necessity" (McWhorter 2001b: 129). Trudgill (1999), in a study on the functionality of grammatical gender, writes:

> We know that languages drag along with them a certain amount of, as it were, unnecessary historical baggage. [...] at least in some languages, there is much more of this afunctional historical baggage than has sometimes been thought. For example, the presence of different declensions for nominal forms and different conjugations for verbal forms in inflecting languages would appear to provide good evidence that languages can demonstrate large amounts of complex and non-functional differentiation which provide afunctionally large amounts of redundancy [...].
>
> (Trudgill 1999: 148)

Apart from different conjugations or declensions, redundant features may also include, for instance, "evidential marking, ergativity, inalienable possessive marking, and inherent reflexive marking" (McWhorter 2001b: 126). The presence of such redundant grammatical elements seems to lack functional motivation as they are not a necessary prerequisite for successful communication and their presence can "presumably, only be explained satisfactorily in historical terms" (Trudgill 1999: 148). Redundancy-based complexity is defined in relation to the communicative necessity and needs of a language user and should therefore be strictly categorised as an agent-related, relative notion of complexity.

Second language acquisition difficulty. This notion of complexity is an agent-related, 'subjective' and thus relative measure which defines complexity in terms of processing and acquisition difficulty as experienced by a (second) language learner: "My thinking was, and is, that 'linguistic complexity', although this [...] is very hard to define or quantify, equates with 'difficulty of learning for adults"' (Trudgill 2001: 371).

Generally, grammatical structures and elements of a language or variety which pose a difficulty to adult learners are considered complex (Kusters 2003, 2008; Szmrecsanyi & Kortmann 2009; Trudgill 2001). Kusters (2003, 2008) defines complexity in relation to an idealised language user, a 'generalized outsider' (Kusters 2008: 9). This outsider is an adult second language learner without any knowledge of the cultural or linguistic background of the target language who is primarily interested in the transmission of information:

> This person speaks a first language, and is not familiar with the second language in question, nor with the customs and background knowledge of the speech community. He or she is primarily interested in using language for communicative purposes [...]. We model this person as a generalized outsider, in order to prevent either facilitating or hampering influences of the first language on acquiring the second language.
>
> (Kusters 2008: 9)

Thus, from the perspective of this outsider all linguistic phenomena are considered complex which (i) are more difficult to acquire for a second language learner than for a first language learner (ii) are functional highly language-specific and (iii) are difficult to perceive and process (Kusters 2003: 6–7, Kusters 2008: 9–10).

Efficiency-based complexity. Another, less well-explored concept of relative complexity is efficiency-based complexity (Hawkins 2004, 2009) or processing difficulty (Seuren & Wekker 1986). Hawkins (2004) introduces a theory based on both language performance and grammar which measures complexity in relation to the communicative efficiency between a speaker and a hearer. He postulates that the most efficient communication, i.e. the communication which requires the least processing effort, is the least complex.

> [...] complexity is a function of the number of formal units and conventionally associated properties that need to be processed in domains relevant for their processing. Efficiency may therefore involve more or less complexity, depending on the proposition to be expressed and the minimum number of properties that must be signalled in order to express it. Crucially, efficiency is an inherently relative notion that compares alternative form-property pairings for expressing the same proposition, and the (most) efficient one is the one that has the lowest overall complexity in on-line processing.
>
> (Hawkins 2004: 25)

According to this metric, complexity is understood as the processing difficulty experienced by a language user. In a nutshell, more processing is more complex.

Transparency-based complexity. Transparency-based complexity is a relative notion of complexity and is also referred to as iconicity (Dammel & Kürschner 2008; Steger & Schneider 2012). Essentially, it boils down to the one-form-one-meaning principle which states that a language is maximally transparent for a language user if one (structural) surface form corresponds to one underlying semantic form (Langacker 1977: 110; Slobin 1977: 176; Seuren & Wekker 1986). This notion to some extent conflates irregularity-based and efficiency-based complexity. In particular, the borders between irregularity-based complexity and transparency are not clear-cut in the literature, and mostly no distinction is made (Trudgill 2004; Szmrecsanyi & Kortmann 2012; Nichols 2013). Inherent linguistic irregularities are often considered as non-transparent while (non-)transparency should merely be viewed as the experience of a language user who is faced with irregularities in a linguistic system. In fact, the two notions could basically be regarded as two sides of the same coin, i.e. one side seen from a theoretical perspective and one side seen from a user perspective. Nonetheless, I propose to view transparency-based complexity as a separate notion of complexity because a substantial amount of research explicitly uses a transparency-based metric.

Steger & Schneider (2012), for instance, use a metric based on various types of iconicity whereby iconic / transparent structures are considered less complex because they are easier to process than non-iconic structures. Specifically, they analyse complement clause constructions in English L2 varieties and find that these varieties—due to language contact and adult second language acquisition—are comparatively more iconic than British English (Steger & Schneider 2012: 187–188). A similar approach to complexity has been proposed by Seuren & Wekker (1986) who investigate the impact of semantic transparency on creole genesis and in relation to second language acquisition. They identify three grammatical strategies which maximize the transparency between surface structures and semantic structures. One of these is the 'simplicity' strategy which "implies that the amount of processing needed to get from semantic analysis to surface structures, and vice versa, is kept to a minimum" (Seuren & Wekker 1986: 66).

From this abundance of complexity metrics and the controversy involved in the definitions of complexity, it can be seen that addressing linguistic complexity is not a simple task. Up to date, no commonly agreed on metric or methodology for assessing either overall or local complexity has been found.

2.1.2. Explaining linguistic complexity

Very much unlike scholarly attempts from the eighteenth and nineteenth century, modern linguistics does no longer accredit linguistic complexity variation to the differing mental capacities of language users or the 'developmental stage' of their cultures. On the contrary, linguistic complexity variation is generally seen as the result of (an interplay of) culturally-historically and geographically conditioned extra-linguistic factors such as language contact or isolation and accompanying second language acquisition as well as the maturity / age of a language, dialect or variety. These factors, on a more user-oriented level, are associated with a number cognitive and communicative constraints such as (online-)processing or transfer. As a matter of fact, all of these factors are interrelated and could be seen as being nested within the three extra-linguistic core dimensions age, geography and history / culture. Figure 2.1.2 illustrates the layering of (some of the) main factors impacting on and explaining linguistic complexity variation.

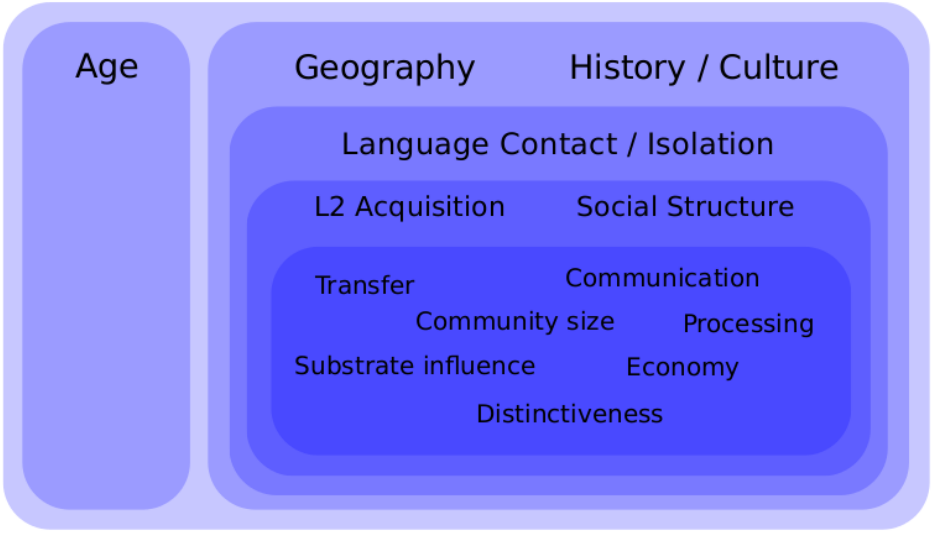

Figure 2.1.: Layering of main interrelated factors impacting on and explaining complexity variation.

The major factor impacting on linguistic complexity is language contact and the resultant (adult) second language acquisition. Situations of language contact, specifically high-contact situations, in which a given language or variety is acquired by a large number of adult language learners encourage and lead to simplification of this language (variety) (Szmrecsanyi & Kortmann 2009; Trudgill 2004, 2009b). Thus, in situations with a high-rate of second language acquisition, languages tend to shed a certain amount

of complexity: "[...] communities involved in large amounts of language contact, to the extent that this contact between adolescents and adults who are beyond the critical threshold for language acquisition, are likely to demonstrate linguistic pidginisation, including simplification, as a result of imperfect language learning" (Trudgill 2004: 306). This point is nicely illustrated by pidgin and creole languages which, originating from situations of extreme language contact, exhibit far less complexity than their source languages (McWhorter 2001b; Trudgill 2004). Language contact, however, does not always lead to simplification but can, under certain circumstances, have quite the opposite effect, namely complexification. Heavy and continuous language contact involving a high rate of child bilingualism often leads to the complexification of the languages in contact (Nichols 1992, 2009; Trudgill 2004):

> It can be concluded that contact among languages fosters complexity, or, put differently, diversity among neighboring languages fosters complexity in each of the languages. Residual zones [, i.e. zones where bi-or multilingualism are common,] are naturally areas of diversity, so we can expect languages in residual zones to exceed the averages of their continents in complexity.
>
> (Nichols 1992: 193)

Trudgill sums the effects of the two types of contact up as follows: "So, long-term contact involving child bilingualism may produce large inventories through borrowing, and adult language contact may produce smaller inventories through imperfect learning, pidginisation, and simplification" (Trudgill 2004: 314). Although Trudgill (2004) explores the size of phoneme inventories in Austronesia, this conclusion can be generalised to other areas of grammar.

A closely related factor, so to speak the twin of language contact, is isolation. Languages in isolated, low-contact situations in which second language acquisition is rare, tend to retain and / or develop complexities (Trudgill 2004). Another factor related to isolation is the social structure of such isolated speech communities. The social structures within these isolated communities are often very tightly-knit, i.e. such communities share a lot of common (cultural) background, which allows them to leave certain structures unmarked on the one hand and to maintain a certain amount of linguistic conservatism on the other hand.

> Small, isolated low-contact communities with tight social network structures are more likely to be able to maintain linguistic norms and ensure the transmission of linguistic complexity from one generation to another. [...] [These communities] will have large amounts of shared information in common and will therefore be able to tolerate lower degrees of linguistic redundancy of certain types.

(Trudgill 2004: 307)

Nichols (2013) successfully approximates the degree of isolation by geographical altitude and shows that while altitude per se does not predict linguistic complexity it determines sociolinguistic aspects such as community size, social network structure and contact which are important factors influencing linguistic complexity.

Kusters (2003, 2008) offers a more culturally-oriented model for the explanation of complexity variation which, however, essentially boils down to the factors contact and isolation. He defines two prototypes of speech communities: the first type are small communities with tight network structures in which language functions as a cultural vehicle, i.e. language does not merely serve as a means of communication but is a means of creating and / or maintaining the identity of the community. Language in these communities is used to express and transmit identities, cultural and religious values, and therefore assumes a large shared background (Kusters 2008: 15). This type of community fits the low-contact, isolated type introduced by Trudgill (2004) and furthers or retains complexities. The second type of community can be large, and often the shared language is not the mother tongue of the speakers. In this type of community, language is primarily a tool of communication. Thus, there is no common cultural ground shared between the members of this community and no values or identities are attached to the language. This type is characterised by high language contact and therefore less likely to maintain or develop linguistic complexity (Kusters 2008: 14–15).

In a similar vein, Lupyan & Dale (2010) (see also Dale & Lupyan 2012) distinguish between exoteric and esoteric languages; languages with a large number of speakers and languages with a comparatively smaller number of speakers. In exoteric communities language serves primarily as a communicative tool, mostly between non-native speakers. They furthermore formulate this in a theoretical framework, the *Linguistic Niche Hypothesis*, to evaluate the interaction between language and social structure (Lupyan & Dale 2010: 2–3). Specfically, Lupyan & Dale (2010) predict the morphological complexity of 2,236 languages together with demographic variables such as speaker population, geographic spread and the number of neighbouring languages using regression models. Languages with a larger number of speakers and higher degrees of language contact were found to use less morphological marking than languages spoken by smaller communities. Instead exoteric languages tend to encode semantic distinctions by lexical means (Lupyan & Dale 2010: 3,6). Be that as it may, however, their Linguistic Niche Hypothesis is merely a re-formulation of Trudgill's (2004) distinction between high-contact and low-contact communities.

Bentz & Winter (2013) adapt the Linguistic Niche Hypothesis to explore the relation between the number of second language learners / speakers, and nominal case marking in a set of 66 typologically stratified languages

by means of regression analysis. Their results confirm that adult second language acquisition leads to a reduction of morphosyntactic complexity and argue that languages not only adapt to the sociolinguistic situation of their speech communities but also to the cognitive constraints of their users (Bentz & Winter 2013: 5–7,19). This is in accordance with a number of studies which explain linguistic complexity variation by cognitive mechanisms such as transfer or substrate influence resulting from second language acquisition (e.g. Huber 2012; Odlin 2012; Siegel 2012).

Finally, the age of a given language or variety has also been suggested to be a predictor for linguistic complexity (McWhorter 2001b). All other things being equal, the complexity of a language increases with increasing age. Even though both, processes of simplification and complexification, do occur in languages, complex, mature patterns habitually evolve (Dahl 2004) or accumulate (McWhorter 2001b; Trudgill 1999) over long periods of time: "The general conclusion was that in older grammars, millennia of grammaticalization and reanalysis have given overt expression to often quite arbitrary slices of semantic space, the results being a great deal of baroque accretion [...] " (McWhorter 2001b: 126).

Table 2.1.: Overview of relevant empirical studies on language complexity since 2000 ordered by year of publication. The linguistic domain, type of complexity metric, explanans for complexity variation and languages or sample size are given. Note that commentaries, reviews and purely theoretical papers not containing original research are not listed.

Study	Domain	Metric	Explanans	Language / Sample
Deutscher (2000)	Syntax	—	Social structure, communicative needs	Akkadian
McWhorter (2001b)	Syntax, phonology, morphology, semantics/pragmatics	Quantitative, redundancy-based	Age	Saramaccan, (Tsez, Lahu, Maori)
Kusters (2003)	Morphology	L2 difficulty	Contact, social structure, community size	Arabic, Quechua, and Swahili varieties, Scandinavian languages
Dahl (2004)	Phonology, morphology, syntax, lexicon	Quantitative, information-theoretic	Age	Various[4]
Moscoso del Prado Martin et al. (2004)	Morphology	Information-theoretic	—	Dutch
Trudgill (2004)	Phonology	Irregularity-based	Contact, social structure, community size	Polynesian
Kettunen et al. (2006)	Morphology, syntax, lexicon	Information-theoretic	—	27 European languages
Shosted (2006)	Phonology, morphology	Quantitative	—	32 languages
Bane (2008)	Morpholgy	Information-theoretic	—	20 languages
Dahl (2008)	Syntax	Quantitative	Age	Siriono
Dammel & Kürschner (2008)	Morphology	Quantitative, transparency-based	—	10 Germanic languages
de Groot (2008)	Morphology, syntax	Quantitative, system complexity	Contact	Hungarian

Continued on next page

[4]Various languages are quoted but none are studied in detail and the list would be too long to present here. Therefore, suffice it to say that Dahl (2004) refers to a wide variety of languages such as English and Swedish but also Guarani or Mandarin Chinese.

Study	Domain	Metric	Explanans	Language / Sample
Fenk-Oczlon & Fenk (2008)	Phonology, semantics, morphology	Quantitative, irregularity-based	Functional economy	8 languages, English, pidgins
Gil (2008)	Morphology, semantics, compositional semantics	—	Contact	10 languages
Juola (2008)	Morphology, syntax	Information-theoretic	—	Historical English varieties, 15 languages
Kusters (2008)	Morphology	L2 difficulty	Contact	Quechua
Lindström (2008)	Morphology, syntax	Quantitative, irregularity-based, L2 difficulty	(Contact)	Kuot, Nlik, Notsi, Madak
McWhorter (2008)	Morphology, phonology, semantics	Quantitative, redundancy-based, irregularity-based	Contact	Riau Indonesian, Tetun Terik, Tetun Diki
Parkvall (2008)	Phonology, morphology, syntax	Quantitative	—	185 languages
Riddle (2008)	Morphology, lexicon	Quantitative, redundancy-based	?	Mandarin Chinese, Thai, Hmong
Sadeniemi et al. (2008)	Morphology, syntax, lexicon	Information-theoretic	—	21 European languages
Sinnemäki (2008)	Morphology, syntax, lexicon	Transparency-based	Economy vs. distinctiveness	50 languages
Bisang (2009)	Pragmatics, morphology	Irregularity-based, transparency-based	Age	East / Southeast Asian languages
Dahl (2009)	Phonology, morphology, syntax	Quantitative	Contact	Elfdalian, Swedish
Gil (2009)	Morphology, syntax, semantics	Quantitative	Age	Riau Indonesian
Hawkins (2009)	Morphology, syntax	Efficiency-based	Processing	English, German, Japanese
Karlsson (2009)	Syntax	Quantitative	Processing (literacy)	Standard Average European Languages

Continued on next page

Study	Domain	Metric	Explanans	Language / Sample
Kortmann & Szmrecsanyi (2009)	Morphology, syntax	Quantitative, redundancy-based, L2 difficulty, irregularity-based	Geography, contact	50 English varieties, pidgins and creoles
Maas (2009)	Morphology, syntax	—	Processing (literacy)	(medieval) German, Latin and some Bible translations
McWhorter (2009)	Morphology, syntax	Redundancy-based	Age	Saramaccan
Miestamo (2009)	Morphology, syntax	Quantitative, implicational hierarchies	Communicative constraints	50 languages
Nichols (2009)	Phonology, Morphology, syntax, lexicon	Quantitative	Contact, population size	68 languages
Progovac (2009)	Morphology, syntax	Quantitative	Age, evolution	English, Serbian
Sinnemäki (2009)	Morphology, syntax	Redundancy-based	Community size	50 languages
Stapert (2009)	Syntax	Quantitative	Age, evolution, processing	English, Piraha
Szmrecsanyi & Kortmann (2009)	Morphology, syntax	Quantitative, redundancy-based, L2 difficulty, irregularity-based	Geography, contact	English varieties
Szmrecsanyi (2009)	Morphology, syntax	Quantitative	Geography, contact, communicative constraints, diachrony	English varieties and registers
Trudgill (2009a)	Morphology	Irregularity-based, redundancy-based	Contact, social structure, community size	English dialects
Trudgill (2009b)	Phonolgy, Morphology, Syntax	Irregularity-based, redundancy-based	Contact	English dialects
Lupyan & Dale (2010)	Morphology, syntax	Quantitative	Contact, community size, communicative constraints	2,236 languages
McWhorter (2012)	Morphology, syntax	Quantitative, redundancy-based, irregularity-based	Age	Saramaccan

Continued on next page

Study	Domain	Metric	Explanans	Language / Sample
Mesthrie (2012)	Morphology, syntax	Quantitative	Contact, SLA, substrate influence	Black South African English, Singapore English
Mühlhäusler (2012)	Morphology, syntax	Quantitative	Social structure, communicative constraints	Pitcairn Norf'k
Odlin (2012)	Lexicon, semantics, syntax (?)	Quantitative	SLA, transfer, fossilisation	L2 English
Han & Lew (2012)	Syntax	L2 difficulty, transparency-based	SLA, transfer, fossilisation	L2 English (L2 Italian)
Huber (2012)	Syntax	Quantitative	Contact, SLA, transfer	Ghanaian English, British English
Steger & Schneider (2012)	Morphology, syntax	Transparency-based	Contact, processing	English L2 varieties
Siegel (2012)	Morphology	Quantitative, irregularity-based, transparency-based	Contact, SLA, transfer	Creoles, pidgins
Bentz & Winter (2013)	Morphology, syntax	Quantitative	SLA	66 languages
Nichols (2013)	Phonology, morphology, syntax, lexicon	Quantitative, transparency-based	Contact, altitude	47 Nakh-Daghestanian languages
Ehret (2014)	Morphology, syntax	Information-theoretic	—	English
Shin (2014)	Morphology	Quantitative	Contact, communicative constraints	Spanish
Siegel et al. (2014)	Morphology, syntax	Quantitative	Contact, SLA	Tok Pisin, Hawai'i Creole, German, Russian, Italian, Spanish, English varieties
Ehret & Szmrecsanyi (2016b)	Morphology, syntax	Information-theoretic	—	Historical English varieties, 10 European languages
Ehret & Szmrecsanyi (2016a)	Morphology, syntax	Information-theoretic	SLA	English learner varieties

2.2. Information theory

2.2.1. Information theory, Shannon entropy and Kolmogorov complexity

Information theory is "the science which deals with the concept 'information', its measurement and its applications" (van der Lubbe 1997: 1). In his landmark paper "A Mathematical Theory of Communication", Shannon (1948) defines *communication* in mathematical terms thereby founding the field which has become known world-wide as information theory. More precisely, he analyses the information content between a message source and a listener and establishes the upper bounds for the efficiency with which messages can be transmitted along a channel (Shannon 1948). The information content of a message is measured in *Shannon entropy*, which quantifies the amount of uncertainty or choice, i.e. entropy, involved in the selection of a message. In order to shed light on the relation between information content and entropy, I will introduce the protagonists of Shannon's theory and outline, bit by bit, how he derives this measure of information. Thereafter, a related metric put forward by and named after the Russian mathematician Kolmogorov (1963, 1965) will be presented and its relevance to the methodology explored in the current work will be illustrated.

In the context of Shannon's theory, "communication" is defined as the transmission of a message from one point A to a point B through a *communication system*. In general, a communication system (see Figure 2.2 below) consists of an *information source*, point A, which produces a message to be sent to a point B, the *destination*. The message is sent through a medium—the *channel*—such as a telephone line. In order to send a message along a channel, the message needs to be transformed into some sort of signal. This is achieved by passing the message to a *transmitter* which, after having transformed the message into a suitable signal, sends it along a channel to a *receiver*. Subsequently, the receiver converts the received signal back to its original form and passes the message on to the intended destination (Shannon 1948: 2).

Shannon discusses three different systems of communication—discrete, continuous and mixed systems. For reasons of relevance, however, I shall only focus on one specific variation of the first type of communication system, namely, the *discrete noiseless system*. A discrete communication system is characterised by messages which are a sequence of discrete, i.e. one after the other, separately produced symbols (Shannon 1948: 3–5). For example, natural (written) languages could be considered such a discrete communication system as each message (word / sentence) consists of a sequence of individually produced symbols (letters). So far, this may sound simple; yet, the process of sending a message from one point to another is not a trivial issue. This is particularly true considering that a given sender would like to send not just one specific message but would like to choose

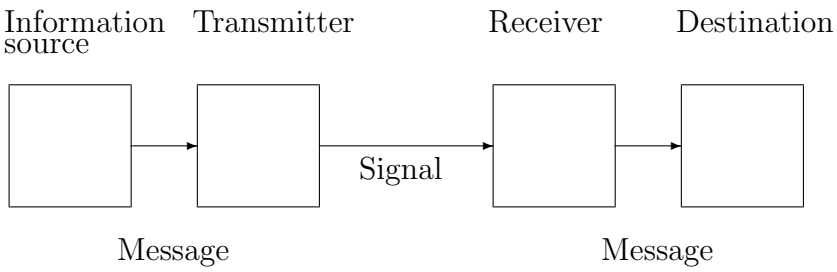

Figure 2.2.: Simplified schematic diagram of a communication system. Adapted from Shannon (1948: 2).

and send any one message (from a set of possible messages).

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. [...]The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design [of the communication system].
>
> (Shannon 1948: 1)

In selecting and sending a message via a discrete communication system as described above, *information* is produced and transmitted along the channel. It is important to note that, in the framework of this theory, the term "information" does not refer to the meaning of an individual message. Rather, information refers to the communicative situation as a whole, i.e. it refers to the amount of freedom in the selection / choice of one message over another from a possible set of messages.

> In fact, two messages, one of which is heavily loaded with meaning and the other of which is pure nonsense, can be exactly equivalent, from the present viewpoint as regards information. [...]To be sure, this word information in communication theory relates not so much as to what you *do* say, as to what you *could* say. That is, information is a measure of one's freedom of choice when one selects a message.
>
> (Weaver 1959: 99–100)

In order to quantify the amount of information which is produced in selecting and sending a message, Shannon derives a measure of information or entropy based on the probabilities of each possible message. A necessary postulate for the development of this quantitative measure of information

is that the set of messages / symbols transmitted in the outlined communication system is finite. In other words, the set of possible messages is restricted to a certain number and this number is known to both the sender and receiver. In terms of language or, as a concrete example, English, a writer (the sender) and a reader (the receiver) both know that the possible set of messages in their communication consists of the 26 letters of the Latin alphabet and not, say, of dots and dashes. The set of messages needs to be finite because "if the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of information produced when one message is chosen from the set, all choices being equally likely" (Shannon 1948: 1).

Shannon chooses a logarithmic base for this measure as the logarithm is—for a variety of mathematical reasons—the most suitable function and many parameters in the physical and engineering sciences such as bandwidth, for instance, are based on a logarithmic scale (Shannon 1948: 1). Furthermore, each of these messages occurs with a certain probability. In the most basic case, the occurrence of each possible message is equally likely. Thus, "we have a set of possible events [or messages or symbols] whose probabilities of occurrence are p_1, p_2, ..., p_n. These probabilities are known but that is all we know concerning which event will occur" (Shannon 1948: 10). The entropy H is the sum of the logarithm of these probabilities:

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

H expresses how likely or predictable a message is. If it is certain which message will be chosen, i.e. the probability of all messages but one is zero, the entropy equals zero ($H = 0$). This means that, as the outcome is known, no information is transmitted. In all other cases the entropy has a positive value ($H > 0$). If all messages are equally likely to be chosen, the entropy H is maximal. This means any outcome is informative as it is not certain (Shannon 1948: 10–11). Thus, the information content of a message is directly related to its probability of occurrence, or its unpredictability: a message is informative if it is not known in advance, predictable or expected but conveys something surprising / new. In a nutshell, Shannon entropy calculates the information contained in a message in relation to its unpredictability.

As an illustration of how the entropy / information is calculated, consider an event x with two possibilities of choice, i.e. if there are, for instance, two possible messages in a set. Each of these possibilities has a probability which is known: if the first choice has a probability of p, then the second has the probability q which is $1 - p$.[5] The entropy or information of this event x

[5]The probability of q must be $1 - p$ as in probability theory, the sum of the probability of all events / possibilities, or in this case messages, is always defined as 1.

is then the sum of the logarithms of p and q: $H(x) = -(p \log p + q \log q)$. In Figure 2.3 which illustrates this example event, we see that the entropy is zero if the probability of p is either zero or one; if $p = 0$ q occurs with certainty thus there is zero entropy. Likewise, there is no uncertainty / entropy if $p = 1$ as we know that choice p occurs. The entropy reaches its maximum value if both choices p and q are equally likely, i.e. both choices have a probability of 0.5.

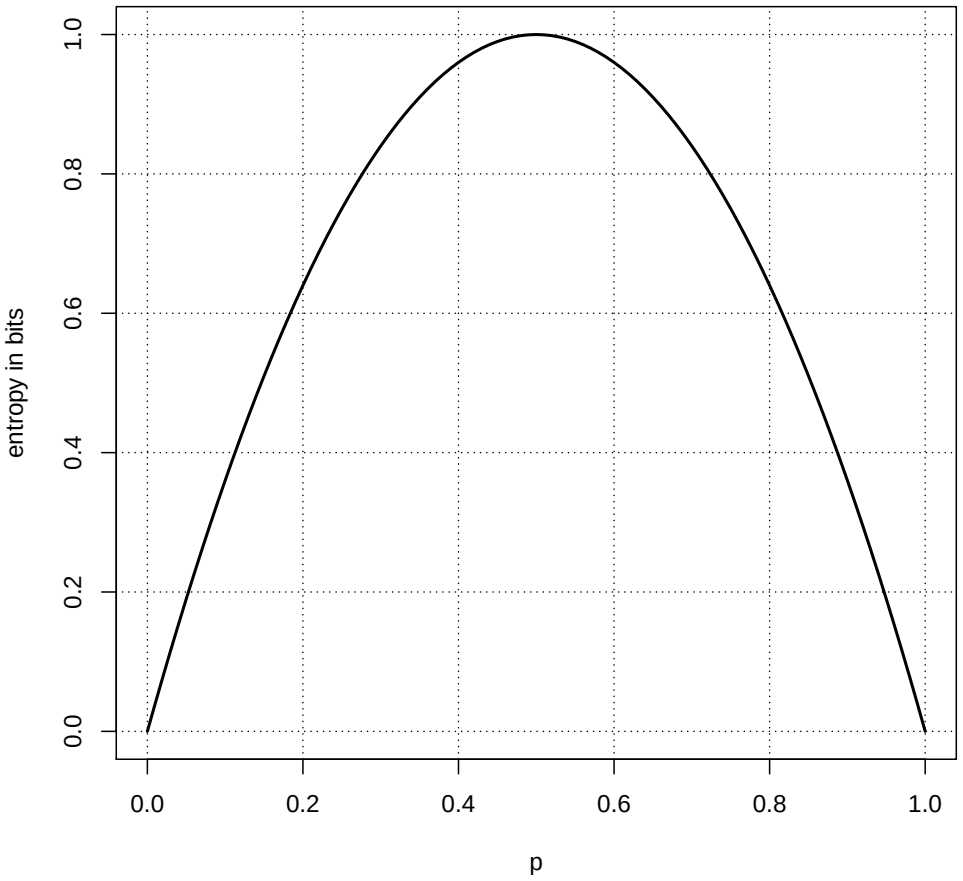

Figure 2.3.: Entropy / information of a two-choice situation. Abscissa indexes the probability of the event p, ordinate the amount of entropy in bits.

Even though Shannon's theory has found application in modern data encoding and compression—the idea is that events with high probabilities have shorter encodings than events with low probabilities (Shannon 1948: 17; see also MacKay 2003; Li & Vitányi 1997)—Shannon entropy is an "ensemble notion" (Li & Vitányi 1997: 65), i.e. information is always measured in relation to a set of possibilities and their probabilities. As such, however, it cannot measure the information content of an individual message

independent of the probabilities in the set. Thus, in order to measure the information content of individual objects, "information" needs to be defined differently. Precisely, information needs to be defined in absolute terms, i.e. it must refer to the information inherent in the object alone and not depend on external factors.

> The most natural approach to defining the quantity of information is clearly to define it in relation to the individual object (be it Homer's *Odyssey* or a particular type of dodo) rather than in relation to a set of objects from which the individual object may be selected. To do so, one could define the quantity of information in an object in terms of the number of bits required to describe it. A description of an object is evidently only useful if we can reconstruct the object from this description. We aim at something different from C.E. Shannon's theory of communication [...]. Our task is to widen the limited set of alternatives until it is universal. We aim at a notion of 'absolute' information of individual objects, that is the information that by itself describes the object completely.
>
> (Li & Vitányi 1997: 93)

In short, the information content of an object is defined as the description of the object from which it can be reconstructed. Yet an object can have several descriptions and not every complete description of an object can be considered to be a measure of its information content or complexity. While all of these descriptions may be complete, they may vary in their length such that some descriptions may be longer and some shorter. Intuition dictates that longer descriptions are considered more complex than shorter ones. The idea is therefore that "from all descriptions of an object we can take the length of the shortest description as a measure of the object's complexity. It is natural to call an object 'simple' if it has at least one short description, and to call it 'complex' if all of its descriptions are long" (Li & Vitányi 1997: 1).

Kolmogorov complexity, also known as *descriptive complexity*, *algorithmic complexity* or *algorithmic information content*, is based on exactly this assumption: it measures the information content or complexity of an object as the length of the shortest possible description of this object. This quantity is absolute, i.e. it is a property of the object alone and does not depend on, for instance, the probabilities in a set of messages (Li & Vitányi 1997: 48). Technically speaking, the complexity K of a given object x is measured by the length of the shortest description $|d|$ of x which is required to (re)generate x:

$$K(x) = |d(x)|$$

For illustrative purposes, assume that we have two strings of symbols (see example (1) below) which are the objects whose complexity we want to

measure. Both strings consist of the same number of symbols, yet string (1-a) can be compressed to the expression 5×cd counting four symbols whereas the shortest description of string (1-b) is the string itself. Measuring the complexity of string (1-a) and string (1-b) according to the length of their shortest possible description, string (1-a) is obviously less complex than string (1-b).

(1) a. cdcdcdcdcd (10 symbols) → 5×cd (4 symbols)
b. c4gh39aby7 (10 symbols) → c4gh39aby7 (10 symbols)

In what has become known as the *Invariance Theorem* Kolmogorov (1965) has shown that his measure of complexity is independent of the description method used. The difference between the length of two descriptions d_1 and d_2 which are produced by two different specification methods / description languages D_1 and D_2 is bounded independently of the input. This is another way of saying that the difference in length between descriptions of different specification methods is negligible and the complexity of an object x under any specification method is therefore invariant (Li & Vitányi 1997: 96–97; 100–101). Moreover, Kolmogorov complexity can be shown to be a suitable measure of information as it is quantitatively related to Shannon's classic notion of information. In fact, Kolmogorov complexity is asymptotically equal to Shannon entropy (Li & Vitányi 1997: 522–525).

For mathematically non-trivial reasons which are related to the *Halting Problem*, Kolmogorov complexity cannot be effectively calculated (Kolmogorov 1965; Li & Vitányi 1997). However, it can be approximated and its upper bounds computed by adaptive entropy estimation methods which are employed by modern file compression programs. In fact, "the Kolmogorov complexity of a file is essenstially the length of the ultimate compressed version of the file" (Li et al. 2004: 3252; see also Ziv & Lempel 1977).

To sum up, Shannon's groundbreaking theory of communication has established the field of information theory and introduced a first quantitative measure of information, Shannon entropy. A related measure of information is Kolmogorov complexity which permits, in contrast to Shannon entropy, the independent measurement of the complexity of individual objects. Kolmogorov complexity, while itself uncomputable, can be approximated by compression algorithms as implemented in modern file compression programs.

2.2.2. Information-theoretic complexity

In Section 2.1 various linguistic complexity metrics have been introduced, among them the largely unexplored notion of information-theoretic complexity. Information-theoretic complexity has its roots, as the name suggests, in information theory. More precisely, it is based on and approximates the non-computable notion of Kolmogorov complexity. Yet, there is

no uniform expression of information-theoretic complexity. In fact, there is a variety of different, Kolmogorov-based metrics and methodologies. In the following I will present several of these measures and conclude with giving a concise definition of what—in the context of this work—information-theoretic complexity is. For the sake of completeness I notice that there are a few non-Kolmogorov based, information-theoretic measures in the complexity literature (Moscoso del Prado Martin et al. 2004; Moscoso del Prado Martin 2011: e.g.). These measures are, however, not relevant to the current work, which focuses exclusively on Kolmogorov-based metrics of complexity, and will be only briefly addressed for illustrative purposes.

Moscoso del Prado Martin (2011) adapts an information-theoretic metric known as *effective complexity* to assess the relation between functional syntactic / semantic information and the morphological, viz. inflectional, complexity of six European languages drawing on data from the *Europarl corpus*, a collection of transcripts from sessions of the European Parliament. Effective complexity is based on Shannon entropy and, in simplified terms, measures the complexity of an object as the length of a short description of the object's regularities (Gell-Mann & Lloyd 1996: 49). Furthermore, effective complexity, as it is based on Shannon entropy, is a probabilistic notion (for details see Section 2.2.1 above). In other words, it considers the probability and distribution of these regularities in relation to a set of objects, and not just in the individual object itself (Gell-Mann & Lloyd 1996: 48–49). Be that as it may, Moscoso del Prado Martin (2011) finds that, firstly, the presence of inflections reduces the overall complexity of a language. Secondly, the presence (or absence) of word order information impacts on the morphological complexity in the analysed language samples: when word order was left intact the difference in morphological complexity between the six samples was negligible. When word order was randomized, on the other hand, the samples differed in their morphological complexity such that English ranked low in morphological complexity while the Romance languages (French, Italian, Spanish) ranked high (Moscoso del Prado Martin 2011: 3528).

A second notion of information-theoretic complexity which is also based on Shannon entropy is explored in Moscoso del Prado Martin et al. (2004). In order to predict response latencies in the morphological processing of Dutch words as sampled in the *CELEX Lexical Database*, the informational complexity of each word in the dataset is measured in terms of Shannon entropy. Specifically, the measure assumes that the surface frequency of a given word is more or less equal to the amount of information necessary to encode this word in a lexicon (Moscoso del Prado Martin et al. 2004: 2,7). Multiple regression models are used to predict the correlation between response latencies and the information residual of a word vis-à-vis more traditional type-token counts (Moscoso del Prado Martin et al. 2004). The results show that the information residual of a word is a reliable predictor

for the morphological processing cost involved in the recognition of Dutch words (Moscoso del Prado Martin et al. 2004: 22).

Let us now turn to information-theoretic measures which are based on Kolmogorov complexity. Kolmogorov-based measurements of linguistic complexity were first proposed by Juola (1998) who successfully demonstrated how algorithms which were originally designed for data compression can be used to measure linguistic complexity. These programs work on the assumption that (text) strings always exhibit—at least to some extent—structural regularities as well as redundancies which can be reduced, i.e. compressed. Algorithms of the Lempel-Ziv family such as, for instance, `gzip` compress new text strings on the basis of previously seen and "memorised" strings making use of this structural redundancy. Technically speaking, they "[...]employ the concept of encoding future segments of the source-output [for example a given text string] via maximum-length copying from a buffer containing the recent past output"(Ziv & Lempel 1977: 337). Simply put, the program "loads" a certain amount of text and "stores" it in a temporary "lexicon". On the basis of this lexicon of memorised strings it can "recognise" newly encountered text (sub)strings and compress them by eliminating redundancy.[6] The amount of information thus measured in a given text string is essentially a measure of the (structural) redundancy in this string.

Juola (1998) measures morphological complexity—defined as the information in a text which is expressed through inflectional endings—by using the open-source compression program `gzip` . The idea is to measure linguistic complexity via the information content in text samples where a higher amount of information is taken to equal higher linguistic complexity of the respective language sample (Juola 1998). For this purpose he uses translations of the same text—here the entire Bible in Maori, Russian, English, Dutch, French and Finnish—assuming that the informativeness of all texts across these different languages should be roughly the same if all languages are equally complex (Juola 1998: 209). Morphological complexity in these text samples is addressed by altering the information at the morphological level, i.e. the regularity of inflectional endings, prior to measuring. The suffix *–ing* in English, for example, is considered a morphological regularity as it mostly signals a present participle and is likely to be preceded by strings such as *we are*, *he is* etc. By changing these regularities through numeric type-substitution, the prediction of particular word forms on the basis of other word forms is rendered more difficult, i.e. the text becomes less compressible. Simply put, each token of a word type is replaced by a random number so that morphological structures are no longer recognisable. For instance, if *singing* and *eating* are replaced by 14 and 3258 the

[6]Strictly speaking, compression is achieved by back-referencing redundant (sub)strings with length-distance pairs, i.e. the length of the copied sequence and the distance (in the buffer) to the previous identical sequence (Ziv & Lempel 1977: 337).

tokens are no longer identifiable as present participles and hence the tokens no longer exhibit any morphological regularity. This means that the morphological information and hence the morphological complexity is increased (Juola 1998: 209–210). Juola (1998) argues that such a degradation process and resulting increase in complexity should have a stronger effect on the compressibility of less morphologically complex languages than on languages with an already complex morphological system due to the fact that the prediction of word forms in the latter is already difficult. The comparison of the file sizes of the original unaltered samples with the degraded samples yields compression ratios which approximate the morphological complexity of the given sample (Juola 1998: 209–210). Returning to the Bible texts mentioned above, the complexity hierarchy for the morphological level according to their compression ratios is (in increasing order of morphological complexity) Maori, English, Dutch, French, Russian, Finnish (Juola 1998: 211). Part of this ranking can be confirmed and is in line with previous studies—Nichols (1992), for example, also finds that English is less morphologically complex than French and Russian—lending more credibility to the compression method. Instead of substituting morphological regularities, Juola (2008) achieves distortion by random deletion. In this experiment 24 Bible versions, among them nine English translations, are analysed. At the morphological level distortion is achieved by random deletion of 10% of the characters. At the syntactic level 10% of all word tokens are deleted thus destroying syntactic relations (Juola 2008: 101). The rationale is that the compressibility of syntactically simple languages, i.e. languages with free word order, will not be as badly affected as the compressibility of syntactically complex languages, i.e. languages with relatively fixed word order, as the former lack complex inter-word dependencies. The results of this experiment tie in neatly with Juola's previous findings (1998) and indicate a trade-off between complexity at the morphological and syntactic level, i.e. morphologically complex languages exhibit low syntactic complexity whereas syntactically complex languages exhibit low morphological complexity (Juola 2008: 104).

The compression technique, i.e. measuring linguistic complexity by using file compression programs, as introduced by Juola (1998), was further explored in two papers which both analyse the 21 official languages of the European Union (Kettunen et al. 2006; Sadeniemi et al. 2008). Both papers draw on a corpus consisting of translations / texts of the European Union Constitution in 21 languages: Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Slovak, Slovenian, Spanish and Swedish (Kettunen et al. 2006; Sadeniemi et al. 2008). The data is modified at the morphological level, similar to Juola (1998), by numeric type-substitution. The syntactic level is manipulated by random shuffling of the word order within each sentence thereby maintaining the order of the

sentences as a whole (Kettunen et al. 2006: 101; Sadeniemi et al. 2008: 193–194). In contrast to Juola (1998), the open source compression program `bzip2` is used to compress the text samples. Subsequently, the complexity of the samples is analysed on the basis of the compressed file sizes of the original and manipulated samples. On the whole, the results are as expected and dovetail with previous findings. Minor differences in the complexity order are observable, mainly among members of the Baltic and Slavic languages (Kettunen et al. 2006: 105–107; Sadeniemi et al. 2008: 194–200).

Ehret (2014) adapts the Juola-style compression technique (Juola 1998) to measure the complexity of linguistic features in a detailed fashion. Specifically, the paper explores the contribution of inflectional morphs (for example *–ing* or *–ed*) and functional constructions (such as progressive or passive) to the morphological and syntactic complexity of a mixed-genre corpus consisting of *Alice's Adventures in Wonderland*, the Gospel of Mark and newspaper texts. The results show that the presence of more marker types (morphs) increases the morphological complexity in the corpus while the syntactic complexity is reduced. Furthermore it is demonstrated that the morphological ranking of the morphs analysed coincides with the morpheme order acquisition reported in second language acquisition research (Ehret 2014: 55–58). More generally, the paper fills a gap in algorithmic complexity research by showing how compression algorithms can be utilised to address morphology and syntax in a very detailed manner—previous research in this line has mainly focused on measuring complexity in a more global manner (Ehret 2014: 63–64).

Further explorations of Kolmogorov measurements include a paper by Li et al. (2004) who propose a universal, unsupervised distance metric measuring the (dis)similarity between two objects using general standard algorithms for compression. Their measure is based on the pairwise comparison of Kolmogorov complexity measurements. Specifically, the distance between two objects x and y is calculated by taking the ratio of the compressed length of y if x serves as auxiliary input (i.e. x comes first in a joined file consisting of x and y) and the compressed length of y (Li et al. 2004: 3254). It is important to note that the total length of the objects (files) compared does not exceed the size of the compression algorithm's buffer. In order to illustrate the universality of the metric, Li et al. (2004) present two case studies from two unrelated fields of research: firstly, they compute a mitochondrial phylogeny tree tracing the evolutionary history of mitochondria and, secondly, they construct a language family tree for 52 Euro-Asian languages. For the linguistic case study they draw on a parallel corpus of the *Universal Declaration of Human Rights* and use `gzip` as compressor. On the basis of their similarity metric a distance matrix is calculated which, in turn, serves as input for the language tree construction. The resulting language tree successfully distinguishes between the

main language families (Romance, Celtic, Germanic, Finno-Ugric, Slavic, Baltic and Altaic). Minor misclassifications on the micro level—English, for instance, is classified as a Romance language—can be explained by the latinised vocabulary of the English version of the *Universal Declaration of Human Rights* on the one hand and by the fact that contemporary corpora / languages exhibit linguistic traits which are not inherited but stem from language contact (Li et al. 2004: 3260). Still, the similarity metric yields convincing results. Cilibrasi & Vitányi (2005) further develop this metric using a novel hierarchical clustering method for the calculation of their phylogenetic trees and deliver convincing results demonstrating the universality of their metric in two ways. On a methodological level, they expand their study by not only using dictionary compressors (like `gzip`) as well as statistical and block sorting compressors (like `PPMZ` and `bzip2`), but also by applying the metric to numerous diverse areas such as virology, music, literature or astronomy (Cilibrasi & Vitányi 2005: 1).

Kettunen et al. (2006) and Sadeniemi et al. (2008), inspired by Cilibrasi & Vitányi (2005), measure the difference / similarity in complexity between two languages L_1 and L_2. The assumption is relatively straightforward: if two languages are similiar the compression algorithm should manage the transition in a concatenated text file from L_1 to L_2 well and the sample should compress accordingly. If, on the other hand, two languages are very different, compressibility should suffer, i.e. be comparatively worse. However, this approach seems to be problematic and the results are not particularly intuitive: most of the Romance languages cluster together but Greek and Estonian should not cluster with Slavic languages. Neither should Hungarian and Maltese be next to French (Kettunen et al. 2006: 107–108).

A third, rather different, approach to measure linguistic complexity by approximating Kolmogorov complexity was proposed by Bane (2008). He is interested in the grammar of a language and uses a software package (*Linguistica*) that constructs models to best predict the data. The idea is to measure the complexity of a grammar, in this case on the morphological level, by finding the shortest possible description of the grammatical information that allows the construction of a language's morphology (Bane 2008: 71–72). In detail, the software describes morphological models on the basis of stems, affixes and "signatures" which give a specification of the possible affixes a given stem can take. To illustrate, the signature for the English verb *wait* would contain the possible affixes that can attach to the stem *wait*, i.e. *Ø* , *ing*, *s* and *ed*. The simplest model which accurately describes the input corpus is taken as the smallest total description length of the grammar of the language to approximate the complexity of this grammar. Grammars with a simple morphology should have many stems but few affixes and signatures, while morphologically complex grammars have more affixes and signatures than stems. The morphological complexity of a grammar is therefore calculated by dividing the description length

of affixes and stems by the total description length of the model (Bane 2008: 72–73). This method approaches Kolmogorov complexity rather indirectly and relies on a—even if machine-generated—categorisation of grammatical features, namely stems and affixes as well as their distributional details. For this reason, the method is, unlike the Kolmogorov measurements presented above, not entirely objective. In a case study, the Bible in twenty languages, among them creoles and pidgins (Danish, Dutch, Bislama, English, French, German, Haitian Creole, Hungarian, Icelandic, Italian, Kituba, Latin, Maori, Nigerian Pidgin, Papiementu, Solomon Pijin, Spanish, Swedish, Tok Pisin, and Vietnamese) is analysed. Creoles and pidgins, together with Vietnamese exhibit low morphological complexity while Latin and Hungarian exhibit the highest morphological complexity (Bane 2008: 73). Compared to the type-token ratio, morphological complexity in the sample languages increases with an increasing number of types and decreases with a decreasing number of tokens. All in all, Bane's (2008) hierarchy is in line with what one would expect and ties in with other research (e.g. Juola (1998)).

In the context of this work, information-theoretic complexity is defined as a Kolmogorov-based, unsupervised, algorithmic measure of complexity which is approximated by utilising compression algorithms of the Lempel-Ziv family. Due to the nature of these compression programs, information-theoretic complexity is largely based on structural (ir)regularity and redundancy. In the following chapters information-theoretic complexity will be further explored and defined in linguistic terms by applying and expanding the Juola-style compression methodology. Furthermore, the workings of compression algorithms will be subjected to an in-depth analysis.

3. Experimenting with the compression technique[1]

This chapter is dedicated to the validation and extension of the Juola-style compression technique and presents several experiments in which the reach and limits of the methodology are explored. Applying the compression technique to different data types, I demonstrate that Kolmogorov-based complexity measurements yield linguistically interpretable results, because they provide complexity rankings that match our intuitions—e.g. West Saxon should be morphologically more complex than Present-day English—and are in line with what more orthodox complexity notions would lead one to expect. For instance, Bakker (1998) measures syntactic complexity in terms of word order flexbility and finds that English is less flexible than Finnish. Furthermore, I will show that the method is not restricted to parallel text databases but also works on semi-parallel and non-parallel corpus data.

3.1. Parallel texts

3.1.1. Method and data

The use of file compression programs for measuring linguistic complexity has to date been limited to parallel text corpora, i.e. translational equivalents of the same text in different languages. Such parallel text databases—originally used in historical linguistic studies and more recently in computational approaches to machine translation—have become quite popular in typological research (Auwera et al. 2005; Cysouw & Wälchli 2007; Dahl 2004). Parallel text databases, while still being usage-based, facilitate the comparability across different languages and language varieties due to the fact that differences in propositional content can be ruled out: "Direct comparability of concrete examples across languages is a strong point of the parallel text method. In the ideal case the same domains, instantiated in the same examples, are represented in the same textual environment with the same degree of emphasis in the same register"(Wälchli 2007: 131–132). Furthermore, analyses based on parallel texts permit the generalisation of the results beyond the individual texts or documents studied. This is an-

[1] A partial summary of this chapter has appeared as Ehret & Szmrecsanyi (2016b).

other way of saying that the complexity of languages as a whole can be inferred from the complexity of a parallel sample of these languages. The reason for this is that the complexity measured in any given text sample or document, be it *Harry Potter* or *Hamlet*, should be the same if all languages were equally complex, i.e. all languages should use the same amount of complexity to convey the same meaning (propositional content). Any observed variation in the complexity of parallel text samples, then, should be attributable to the differing levels of complexity of the languages as the propositional content is the same. The classic database for parallel text studies is, due to its availability in a vast number of languages, the Bible (Cysouw et al. 2007; Dahl 2007; de Vries 2007). In this vein, I set the stage by applying the compression technique to the Gospel of Mark in several historical varieties of English and six other languages listed below.

English varieties:

- West Saxon (approx. 10$^{\text{th}}$ century [from Bright 1905])
- Wycliffe's Bible (14$^{\text{th}}$ century [1395])
- The Douay-Rheims Bible (16$^{\text{th}}$ century [1582])
- The King James Version (17$^{\text{th}}$ century [1611])
- Webster's Revision (19$^{\text{th}}$ century [1833])
- Young's Literal Translation (19$^{\text{th}}$ century [1862])
- The Darby Bible (19$^{\text{th}}$ century [1867])
- The American Standard Version (20$^{\text{th}}$ century [1901])
- The Bible in Basic English (20$^{\text{th}}$ century [1941]), using mostly 850 Basic English words and simplified grammar (Ogden 1934, 1942)
- The English Standard Version (21$^{\text{st}}$ century [2001])

Other languages:

- Esperanto (Esperanto Londona Biblio, 20$^{\text{th}}$ century [1926])
- Finnish (Pyhä Raamattu, 20$^{\text{th}}$ century [1992])
- French (Ostervald, 20$^{\text{th}}$ century [1996 revision])
- German (Schlachter, "Miniaturbibel", 20$^{\text{th}}$ century [1951 revision])
- Hungarian (Vizsoly Bible [a.k.a. Károli Bible], 16$^{\text{th}}$ century)

– Latin (Vulgata Clementina, 4thcentury)

Table 3.1 lists the number of words and sentences for each version of the Gospel of Mark. Note that, generally, corpus size is not a crucial factor when working with parallel text databases because, as mentioned above, the propositional content of the components of such corpora is identical. Yet, even with parallel text databases, a minimum corpus size is required: while the Gospel of Mark (counting 14,370 words, English Standard Version) is sufficiently large, the Lord's Prayer (counting 52 words, Matthew 6: 5–14, English Standard Version), for instance, is not.[2]

Table 3.1.: Number of words and sentences in the Gospel of Mark.

Bible version	**Words**	**Sentences**
English varieties:		
American Standard	15,043	874
Basic English	16,461	868
Darby	15,185	898
Douay Rheims	15,036	909
English Standard	14,402	826
King James	15,186	703
Websters	15,232	853
West Saxon	16,014	928
Wycliffe	19,172	843
Young's Literal	15,524	947
Other varieties:		
Esperanto	13,045	876
Finnish	10,507	1,040
French	15,712	867
German	14,114	867
Hungarian	11,779	844
Latin	10,545	877
Total	232,957	14,020

Each of these Bible texts was saved in a separate text file and all formatting and punctuation including (verse) numbers were removed in `R`.[3] The

[2]Precursory experiments with the Lord's Prayer and the compression technique had to be aborted due to data sparsity.

[3]R 2.14.0 (R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0).

early versions of the Gospel of Mark were furthermore manually screened for orthographic variance and, where applicable, variants of the same form were removed by replacing them with the most frequent variant. In the West Saxon text, for instance, *broðer* spelled with an "e" occurs once whereas *broðor* with an "o" occurs 16 times in the original text. Thus, *broðer* being the less frequent form was replaced by the more frequent *broðor*. In total 101 pairs or, in some cases, triplets were corrected in the West Saxon text. In the King James version, Webster's and Wycliffe only one pair was replaced whereas two pairs were replaced in Young's Literal translation (for a complete list of the variants see Table 3.2). This procedure ensures that the presence of more than one spelling version for the same form does not cause higher incompressibility and thus increased morphological complexity in the subsequent measurement.

Table 3.2.: Spelling variance in the Gospel of Mark.

Bible version	**Spelling variants**
King James	oft — often
Webster's	shouldest — shouldst
Wycliffe	borne — born
Young's Literal	honor — honour sware — swear
West Saxon	adrifð — adrifþ afyrð — afyrþ alyfð — alyfþ anddetende — andettende angin — angyn anwald — anweald aset — asett belimpð — belimpþ beoð — beoþ bescofen — besceofen bethsaida — bethzaida bigspel — bigspell bið — biþ bringað — bringaþ broðer — broðor bysmeriað — bysmeriaþ cumað — cumaþ cwæðon — cwædon cymð — cymþm deofolseocnyssa — deofulseocnessa — deofolseocnessa deð — deþ

Continued on next page

Bible version	Spelling variants
West Saxon	
	doð – doþ
	drincð — drincþ
	dweligaþ — dweligeaþ
	dysegað — dysegaþ
	ecnesse — ecnysse
	eorþam — eorþan
	feo — feoh
	flæsce — flæse
	forgeafe — forgefe
	forgyfaþ — forgifaþ
	forwurþað — forwurðaþ
	fullað — fullaþ
	furðon — furþon
	fylidge — fyligde
	fæstað — fæstaþ
	gaæþ — gaþ
	gebidaþ — gebiddaþ
	gebrodðra — gebroðra
	gebroþru — gebroðru
	gebundene — gebundenne
	gecyrede — gecyrrede
	gegearwiað — gegearwiaþ
	gehyrað — gehyraþ
	gelyfað — gelyfaþ
	gemette — gemete
	gemunde — gemynde
	genealæcð — genelæcþ
	gesceasfte — gesceafte
	geswutelaþ – geswutelod
	gesylþ — gesylt
	gewearð — gewearþ
	gewurðaþ — gewurþað
	godspel — godspell
	godspelles — godspellys
	habbað — habbaþ
	hreofnes — hreofnys
	hriðigende — hriþigende
	hweðer — hweþer
	hwæþer — hwæðer
	hyrligum — hyrlingum
	hyrsumiað — hyrsumiaþ
	hæfð — hæfþ
	hæland — hælend
	hælyndes — hælendes
	iacobun — iacobum
	menegu — menigu
	metað — metaþ
	middre — midre
	minne — mine
	moðor — modor

Continued on next page

Bible version	**Spelling variants**
West Saxon	
	muðan — muþan
	nabbað — nabbaþ
	nane — nanne
	net — nett
	næfð — næfþ
	ongean — ongen — ongan
	sceall — sceal
	scep — scip — scyp
	scype — scipe
	secð — secþ
	seocnyssa — seocnessa
	soplice — soðlice — soþliice — soþlice
	spreæc — spræc
	spycð — spycþ
	sunnu — sunu
	swyðram — swyðran
	symbel — symble
	syþþan — syððan
	sæwð — sæwþ
	templ — tempel
	tobrycð — tobrycþ
	towearde — toweardre
	towurpe — towyrpþ
	unclæne — unclænne
	uncnytte — uncytte
	ungeleaffulnesse — ungeleaffulnysse
	warniað — warniaþ
	wearp — wearð — wearþ
	wife — wif
	winð — winþ
	winter — wintra — wintre
	witlodlice — witodice — witodlice
	wundrodon — wundredon
	yrmðe — yrmþe
	yrnþ — yrmþe
	ðas — ðað — ðat
	ðonne — ðone
	ðænne — ðæne
	þone — þonne
	þrysmiað — þrysmiaþ

Methodologically, I utilise the open source compression program `gzip` [4] to approximate Kolmogorov complexity and to assess linguistic complexity on the overall, syntactic and morphological level. A word on the understanding and definition of these complexity notions: In the context of this work, overall complexity coincides with Miestamo's concept of global complexity

[4] gzip (GNU zip), Version 1.2.4. URL `http://www.gzip.org/`

(2008: 29–32) which comprises the complexity of all levels of a language and thus refers to the complexity of a language as a whole. Morphological complexity corresponds to the degree of structural (ir)regularity of word forms a given language exhibits. More structural (ir)regularity is considered more morphologically complex; the morphological complexity axis is therefore negatively poled. Syntactic complexity, on the other hand, is specified in terms of word order (rules) such that maximally simple syntax is essentially defined as maximally free word order. Hence, the polarity of the syntactic complexity axis is set so that fixed word order, i.e. more word order rules and less variation of syntactic patterns, counts as complex (see also Section 7.2).

In this spirit, the overall complexity of the text samples is assessed by obtaining two measurements for each text file analysed: the file size in bytes before compression, and the file size in bytes after compression. These file sizes are directly associated, i.e. the bigger an uncompressed text file is to start with, the bigger is the resulting compressed text file. In order to eliminate the trivial correlation between the original uncompressed file sizes and the compressed file sizes of the samples, the two values are subjected to linear regression. The resulting *adjusted overall complexity scores* (regression residuals, in bytes) are a measure of the left-over variance between the language samples and are taken as indicators of the overall complexity of a given sample. Bigger adjusted complexity scores can be equated with higher informativeness of a given text sample and thus indicate higher levels of Kolmogorov complexity.

Complexity at the morphological and syntactic level can be addressed by manipulating the information at the respective linguistic level in each text file prior to compression. Largely following Juola (2008), morphological distortion is achieved by random deletion of 10% of the orthographic characters in each text file. Through this procedure new word forms are created and, at the same time, morphological regularity is compromised. Subsequently, the distorted samples are compressed in order to determine how well or badly the compression program deals with the distortion. Morphologically complex languages exhibit overall a relatively large amount of word forms in any case. Hence, distortion should not compromise them as much as morphologically simple languages, in which distortion creates proportionally more random noise and thus entropy / complexity. Comparatively worse compression ratios thus signify low morphological complexity. Distortion at the syntactic level is accomplished by randomly deleting 10% of all orthographically transcribed word tokens in each sample. This procedure is assumed to have little impact on languages with relatively simple syntax—defined as free word order—as they lack between-word interdependencies that could be compromised. Syntactically complex languages with many word order rules, however, should be greatly affected as word order regularities are distorted and compromised. In the Basic English text sample, for example,

the auxiliary sequence *would have been* (1-a) occurs twice. This sequence could presumably be altered to *would Ø been* (1-b) through distortion. In this case the compression algorithm would encounter two hapax legomenon patterns—instead of encountering one pattern twice. This leads to uncompressible entropy and compromises compression efficiency. To make a long story short, comparatively bad compression ratios after syntactic distortion indicate high syntactic complexity.

(1) a. It **would have been** well for that man if he had never been given birth.
b. It **would Ø been** well for that man if he had never been given birth. [Mark 14: 21 [Basic English]]

On a more technical note, the size in bytes of the original undistorted file and the size in bytes of the syntactically / morphologically distorted file are taken. On the basis of these values I calculate two complexity quotients: the *morphological complexity score*, defined as $-\frac{m}{c}$, where m is the compressed file size after morphological distortion and c is the compressed file size before distortion; and the *syntactic complexity score*, defined as $\frac{s}{c}$, where s is the compressed file size after syntactic distortion and c the file size before distortion.

3.1.2. The Gospel of Mark

Proceeding as described above, the file sizes in bytes before and after compression are established for each text file. I then calculate adjusted overall complexity scores for all language samples and obtain a hierarchy of overall complexity (Figure 3.1). West Saxon, Hungarian, Finnish, Latin, German and French are (in decreasing order) rather complex whereas Esperanto and all English texts after 1066 are rather simple. Note that there is an interaction between the overall and morphological complexity measures such that exceptionally high morphological complexity (e.g. in Hungarian) tends to be reflected in high overall complexity scores. This is due to the algorithmic nature of the measure which is based on structural surface redundancy (for more details see Chapter 5).

The analysis of morphological and syntactic complexity yields equally intuitive results. Thus in Figure 3.2, languages which are morphologically complex but syntactically simple cluster in the top left quadrant: West Saxon, Finnish, Latin and Hungarian exhibit the most complex morphology. All the English varieties—apart from West Saxon—as well as French are morphologically simple but syntactically complex and are scattered across the bottom right quadrant. Basic English, in the very bottom left part of the plot, is the morphologically least complex but syntactically most complex sample. German and Esperanto cover the middle ground and seem to be balanced in regard to morphological versus syntactic complexity.

In this Bible sample, morphological complexity trades off against syn-

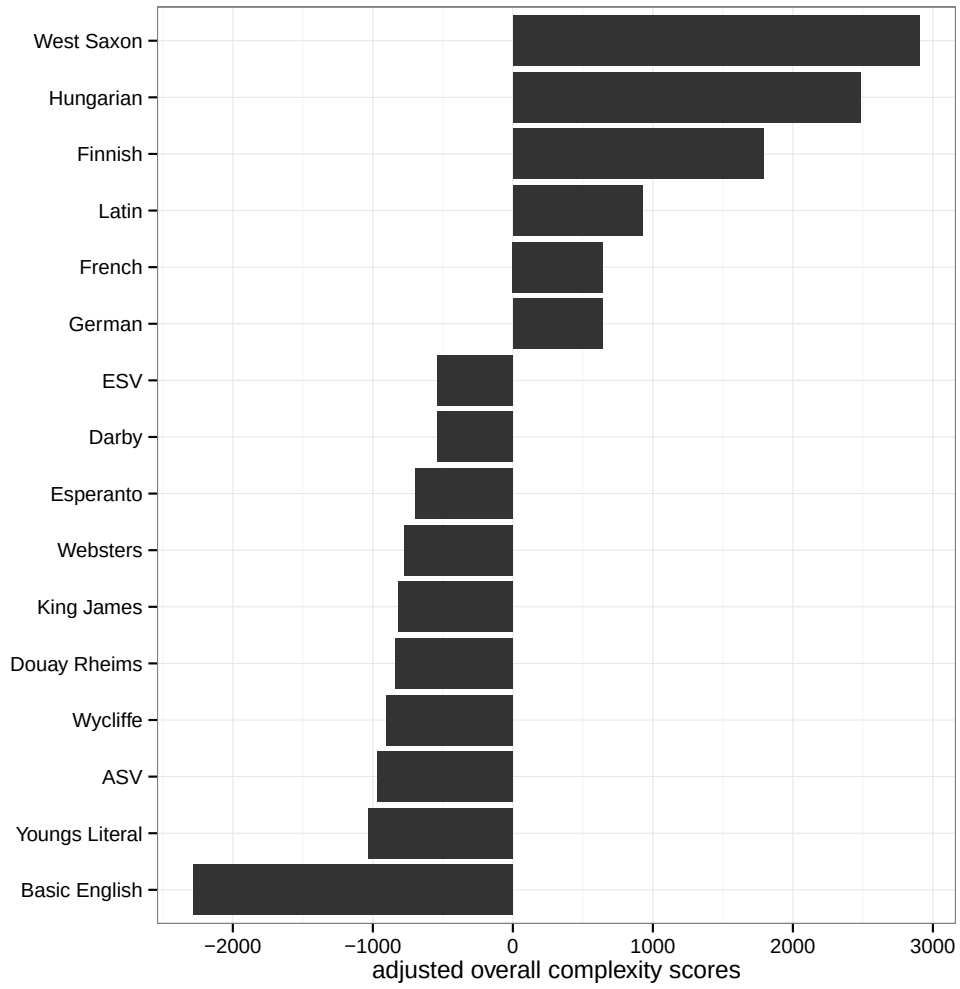

Figure 3.1.: Overall complexity hierarchy in the Gospel of Mark database. Negative residuals indicate below-average complexity; positive residuals indicate above-average complexity.

tactic complexity and vice versa. A negative correlation between morphological complexity and syntactic complexity is particularly prominent when focusing on the English varieties; with a Pearson's correlation coefficient of $r = -0.92$, $p = 0.0000$ the correlation between the complexity scores indicates a textbook-style trade-off.

The workings of the compression technique will be illustrated with an example passage from Mark 1: 8–9 in West Saxon (classified as a morphologically complex but syntactically simple language) and Basic English (classified as a morphologically simple but syntactically complex language). In terms of morphology (see Table 3.3), nine different segmentable inflected word tokens (*fullige*, *wæter-e*, *full-aþ*, *Halg-um*, *Gast-e*, *dag-um*, *ge-full-od*, *Iordan-e*, *Iohann-e*) can be counted in the West Saxon version whereas only three tokens (2 x *giv-en*, *day-s*) and two types (*giv-en*, *day-s*) can be counted

Table 3.3.: Segmentable inflected word tokens in Mark 1: 8–9.

West Saxon	**Basic English**
[8] Ic fullig-e eow on wæter-e; he eow full-aþ on Halg-um Gast-e.	[8] I have giv-en you baptism with water, but he will give you baptism with the Holy Spirit.
[9] And on ðam dag-um, come se Hælend fram Nazareth Galilee, and wæs ge-full-od on Iordan-e fram Iohann-e.	[9] And it came about in those day-s, that Jesus came from Nazareth of Galilee, and was giv-en baptism by John in the Jordan.

Table 3.4.: Word order patterns in Mark 1: 8–9.

West Saxon	**Basic English**
[8] $[\text{Ic}]_{\text{subject}}$ $[\text{fullige}]_{\text{verb}}$ $[\text{eow}]_{\text{object}}$ $[\text{on wætere}]_{\text{adverbial;}}$ $[\text{he}]_{\text{subject}}$ $[\text{eow}]_{\text{object}}$ $[\text{fullaþ}]_{\text{verb}}$ $[\text{on Halgum Gaste}]_{\text{adverbial.}}$	[8]$_{[\text{I}]\text{subject}}$ $[\text{have given}]_{\text{verb}}$ $[\text{you}]_{\text{object}}$ $[\text{baptism}]_{\text{object}}$ $[\text{with water}]_{\text{adverbial,}}$ but $[\text{he}]_{\text{subject}}$ $[\text{will give}]_{\text{verb}}$ $[\text{you}]_{\text{object}}$ $[\text{baptism}]_{\text{object}}$ $[\text{with the Holy Spirit}]_{\text{adverbial.}}$
[9]$_{\text{And}}$ $[\text{on ðam dagum}]_{\text{adverbial,}}$ $[\text{come}]_{\text{verb}}$ $[\text{se Hælend}]_{\text{subject}}$ $[\text{fram Nazareth Galilee}]_{\text{adverbial,}}$ and $[\text{wæs gefullod}]_{\text{verb}}$ $[\text{on Iordane}]_{\text{adverbial}}$ $[\text{fram Iohanne}]_{\text{adverbial.}}$	[9]$_{\text{And}}$ $[\text{it}]_{\text{subject}}$ $[\text{came about}]_{\text{verb}}$ $[\text{in those days}]_{\text{adverbial,}}$ that $[\text{Jesus}]_{\text{subject}}$ $[\text{came}]_{\text{verb}}$ $[\text{from Nazareth of Galilee}]_{\text{adverbial,}}$ and $[\text{was given}]_{\text{verb}}$ $[\text{baptism}]_{\text{object}}$ $[\text{by John in the Jordan}]_{\text{adverbial.}}$

in the Basic English version. In short, the compression algorithm encounters more patterns in the West Saxon than in the Basic English version. For this reason, the West Saxon text is less compressible on the morphological plane than the Basic English text.

Let us now turn to syntax (Table 3.4), the West Saxon version features four different word order patterns and thus possesses a rather flexible syntax. In the Basic English version, on the other hand, word order is relatively rigid (i.e. complex) as the pattern subject-verb dominates throughout the passage. Therefore, Basic English is classified as a syntactically complex language—in contrast to West Saxon—as it has many word order rules to break.

I will now briefly focus on historical drifts as measured by the compression technique in the English Bible translations. It is a well-known fact that in the course of its history the English language has changed from a rather synthetic language which uses many inflections to encode grammatical in-

formation into a rather analytic language which relies on word order and function words to convey grammatical information instead. In Figure 3.3 which plots real time drifts in the history of English, this textbook story is nicely depicted: we can see that the Kolmogorov complexity measurements of the Bible samples clearly suggest a morphological simplification and syntactic complexification over time, some outliers notwithstanding.

In this experiment with parallel texts, I have demonstrated that compression algorithms such as implemented in `gzip` can be utilised to measure linguistic complexity on an overall, morphological and syntactic level. Furthermore, it could be shown that the compression technique captures both intra-linguistic and cross-linguistic complexity variation fairly well and the obtained results dovetail with intuitive complexity assessments as well as previous complexity research (Bakker 1998; Nichols 1992). For instance, my ranking of syntactic complexity is largely congruent with a ranking reported by Bakker (1998) who measures syntactic complexity in terms of flexibility (e.g. Finnish and German are less complex than French and Present-day English).

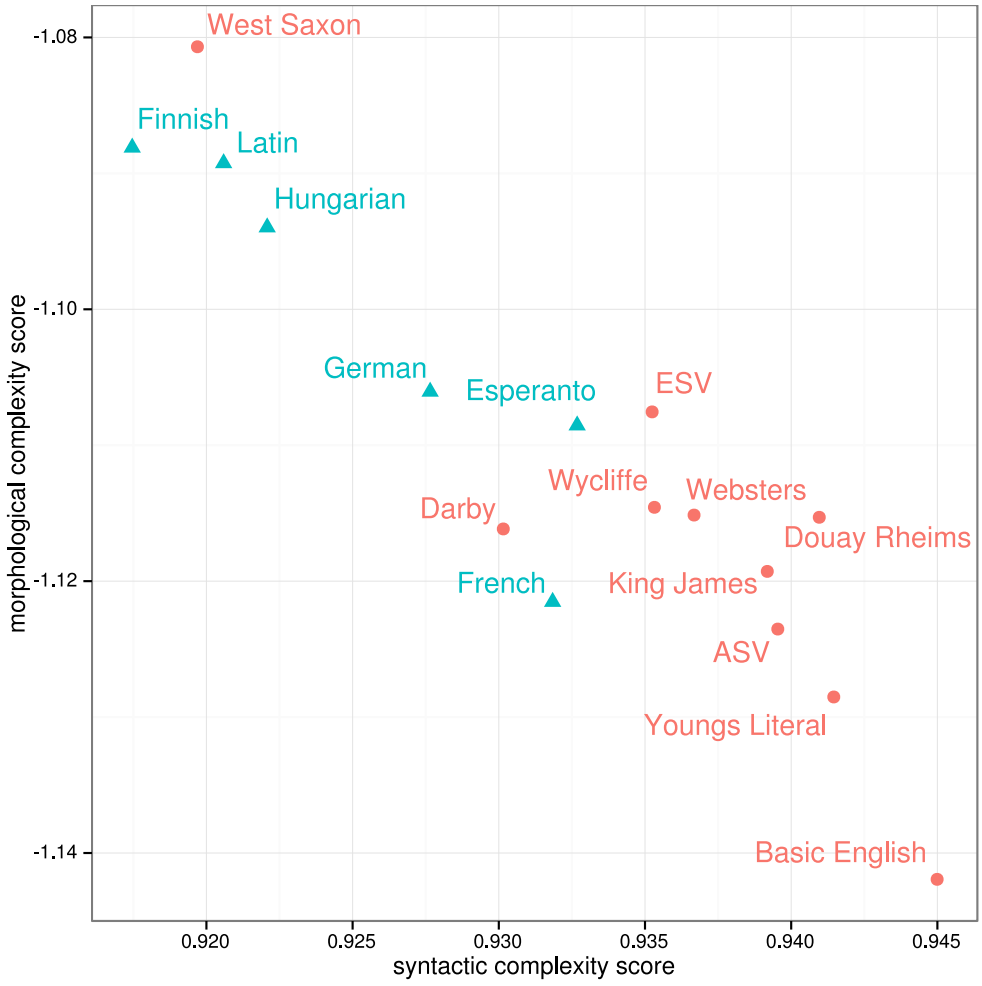

Figure 3.2.: Morphological complexity by syntactic complexity in the Gospel of Mark database. Abscissa indexes increased syntactic complexity, ordinate indexes increased morphological complexity. English versions are represented by a red circle, non-English versions by a blue triangle.

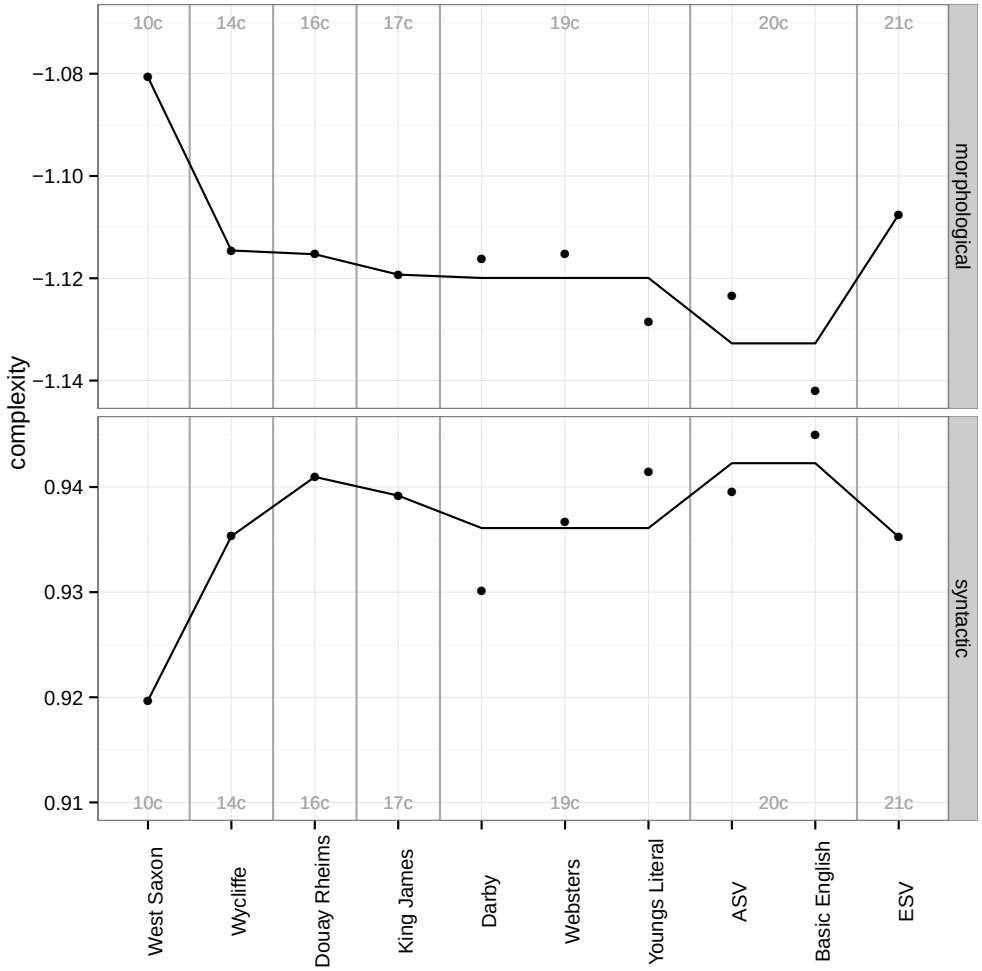

Figure 3.3.: Real time drifts in English: morphological (upper plot) and syntactic complexity (lower plot) in the Gospel of Mark database. Abscissa arranges Bible translations chronologically. The mean morphological and syntactic complexity are given where multiple Bible samples per period are available.

3.2. Parallel, semi-parallel and non-parallel texts

In this section I will further explore the compression technique introduced above and demonstrate that it need not be limited to parallel text corpora but can also be applied to non-parallel text samples as there is no theoretical reason for which only parallel corpora should be used (Juola 1998: 211), and the compression ratios yielded should be (con)text independent. To furnish a case study, I draw on two datasets

1. a parallel and,—after permutation wizardry, semi-parallel—corpus of *Alice's Adventures in Wonderland* in nine languages (Dutch, English, Finnish, French, German, Hungarian, Italian, Romanian, and Spanish) and,

2. a non-parallel sample of newspaper texts covering the same nine European languages (Dutch, English, Finnish, French, German, Hungarian, Italian, Romanian, and Spanish).

This experiment is set up in two steps. Firstly, I measure and subsequently compare linguistic complexity in the parallel corpus of *Alice's Adventures in Wonderland* and a re-sampled semi-parallel version of the same corpus. Secondly, linguistic complexity in non-parallel newspaper texts will be measured and the results compared to the complexity hierarchy obtained from the parallel Alice corpus.

3.2.1. Method and data

In a first step, a parallel corpus of a literary text, i.e. *Alice's Adventures in Wonderland* by Lewis Carroll, is sampled in nine European languages chosen from Germanic, Romance and Finno-Ugric languages which use the Latin alphabet and are frequently utilised as test cases in the complexity literature (Bakker 1998; Juola 1998; Kettunen et al. 2006; Sadeniemi et al. 2008): Dutch, English, Finnish, French, German, Hungarian, Italian, Romanian and Spanish. Table 3.5 lists the number of words and sentences in the Alice database.
Next, I compile a non-parallel corpus of newspaper texts on several contemporary topics[5] in the same nine languages as the Alice corpus (i.e. Dutch, English, Finnish, French, German, Hungarian, Italian, Romanian and Spanish). The topics were chosen according to their availability across the nine languages and cover the European currency crisis, the political situation in Tunisia and the Democratic Republic of Congo, the death of Kim il Jong, the nuclear crisis in Iran and the pending elections in Russia. Thus, all articles dealing with these topics were automatically identified by their HTML

[5]The texts were all retrieved between December 2011 and February 2012.

Table 3.5.: Number of words and sentences in the Alice database.

Alice version	Words	Sentences
Dutch	28,897	1,866
English	26,446	1,625
Finnish	18,572	1,996
French	25,327	2,064
German	26,309	1,659
Hungarian	19,517	2,081
Italian	24,709	1,878
Romanian	23,870	1,952
Spanish	27,128	2,115

topic tag and retrieved from the following online newspapers using custom-made `R` web scrapers (see Appendix B.1 for the basic scraper code).

- Dutch: Volkskrant (`http://www.volkskrant.nl/`)
- English: The Guardian (`http://www.guardian.co.uk/`)
- Finnish: Iltasanomat (`http://www.iltasanomat.fi`), Helsinki Sanomat (`http://www.hs.fi/`)
- French: Le Figaro (`http://www.lefigaro.fr/`)
- German: Die Welt (`www.welt.de`)
- Hungarian: HvG (`http://hvg.hu/`), Nepszava (`http://www.nepszava.hu`)
- Italian: La repubblica (`http://www.repubblica.it/`)
- Romanian: Adevarul (`http://www.adevarul.ro/`)
- Spanish: ABC (`http://www.abc.es`)

Note, however, that the subsequent analysis will only focus on two sample corpora of news articles. One corpus containing articles dealing with the 'Euro crisis' and 'Congo' and a second corpus containing articles dealing with the 'Euro crisis', 'Congo' and 'Tunisia'. I chose these two and three topic-corpora respectively, because not all of the topics retrieved yielded satisfying results. Due to the vast number of articles and the span of languages covered, a manual control of each article's topic was not feasible. For this reason, some of the news sources might substantially differ—probably

depending also on the political relations / interests among the respective countries—in the topics published under the same topic tag.

On a more technical note, all texts of the Alice corpus as well as the newspaper corpus were saved as text files from which all punctuation, numbers or any other non-alphabetical characters such as, for example, bullet points were removed in `R` and subsequently subjected to manual screening. The newspaper corpus was also screened for multiply occurring identical articles or passages which were, where necessary, deleted as they would inflate compressibility. In case of the newspaper corpus, the constant variable "number of sentences" was introduced to determine equally sized samples across languages and topics. By choosing the same number of sentences instead of, for instance, words or characters, syntactic interdependencies remain intact. Thus, for each topic the same number of sentences in each of the languages is sampled. The number of sentences, in turn, is determined by the language sample with the smallest amount of sentences available for a given topic—not all newspapers sample the same amount of text / articles on a given topic.[6] Table 3.6 shows the number of sentences per newspaper corpus.

Table 3.6.: Number of sentences by topic for the newspaper corpora "Euro-Congo" and "Euro-Congo-Tunisia".

Corpus	**Topic**	**Sentences**	**Total**
Euro-Congo	Euro-Crisis	417	782
	Congo	311	
Euro-Congo-Tunisia	Euro-Congo	417	1,248
	Congo	311	
	Tunisia	466	

The methodology used in this experiment is essentially identical to the method described in Section 3.1.1. In order to measure overall complexity in the parallel and non-parallel corpus, the file sizes in bytes of each text file before and after compression are established. Subsequently, the adjusted complexity scores which indicate the overall complexity of each language sample are calculated by subjecting the file sizes to linear regression. Next, we address syntactic and morphological complexity. Let us briefly rehearse how this is achieved: syntactic distortion is performed by deleting 10% of all word tokens in each text file prior to compression. This procedure leads to the disruption of word order regularities and should greatly affect syntactically complex languages, i.e. languages with relatively strict word order rules. The morphological information is manipulated by deletion of

[6] It is for this reason that I had to tap into two online newspapers for Finnish and Hungarian while one source delivered enough data for each topic in the other languages.

10% of all orthographic characters in each text file thereby creating new word forms, i.e. random noise / entropy. This noise compromises compressibility of morphologically simple languages which, overall, have fewer word forms than morphologically complex languages. Concretely, I apply a multiple distortion and compression script to the complete corpora which implements the methodology as outlined above but allows for multiple iterations (the full script is provided in Appendix B.2). Even though I have shown in the Bible experiment that simple distortion and compression yields linguistically meaningful results, the critical observer might claim that the results achieved by simple random deletion are not statistically sound but a product of coincidence. Therefore, I will take multiple measuring points in order to ensure that my findings are statistically robust. For further illustration, in the process of random deletion, any character or word token of a given text file could be modified. This means, however, that depending on which character or word precisely was subject to deletion, the impact of the deletion on complexity might vary. Consider example (2), which illustrates the procedure of random syntactic distortion. (2-a) is the unaltered sentence that will be subjected to random deletion. In this example the distortion script is supposed to delete two words at random. While both in (2-b) and (2-c) two words were deleted, the impact of the deletion differs greatly: (2-b) is still syntactically intact, whereas (2-c) has been rendered incomprehensible because syntax is compromised badly.

(2) a. There was a table set out under a tree in front of the house and the march hare and the hatter were having tea at it [...].
b. There was a table set Ø under a tree in front of the house and the Ø hare and the hatter were having tea at it [...].
c. There was a table set out under a tree in front Ø the house and the march hare and the hatter were Ø tea at it [...].
[Alice]

In terms of complexity, compression of neither (2-a) nor (2-b) in isolation would reflect the actual complexity of the sentence. However, taking the average of several measuring points, the actual complexity of the string can be approximated. Turning back to the actual analysis, I thus apply multiple distortion and compression with $N = 1,000$ iterations to each file in the parallel and non-parallel corpus. Every iteration of the script returns the compressed file sizes for each language sample before and after syntactic/morphological distortion. On the basis of these file sizes the *average morphological complexity score* and the *average syntactic complexity score* are calculated. More precisely, the average complexity scores are obtained by taking the mean of the total number of ratios from each of the measuring points ($N = 1,000$) for morphological and syntactic complexity respectively. In short, the average complexity score is the mean of $N = 1,000$ morphological / syntactic complexity ratios. In order to take stock of the dispersion

across the individual data points the standard deviation is calculated. The values are given in Table 3.7 for the parallel Alice corpus and Tables 3.8 and 3.9 for the non-parallel newspaper corpora. What, then, does the standard deviation tell us? The standard deviation of the average syntactic complexity score for English in the parallel Alice corpus, for example, is 0.0016 and the average syntactic complexity score is 0.9276. This means that most values in this dataset fall between $0.9276 - 0.0016 = 0.926$ and $0.9276 + 0.0016 = 0.9292$. In other words, 68% of all measuring points (assuming a normal distribution) fall within a range of 0.926 and 0.9292. The mean thus seems to be a good representation of the actual complexity across the different measuring points.

Finally, a semi-parallel corpus of *Alice's Adventures in Wonderland* is created by means of permutation. Before every iteration of the multiple distortion and compression script, I randomly sample 10% of the total number of sentences from the parallel Alice database (by incorporating a random sampling function into the distortion script, see Appendix B.3). These samples vary in terms of their propositional content due to the process of multiple random permutation, i.e. the new database is no longer parallel but semi-parallel. Applying the script with $N = 1,000$ iterations, I obtain $1,000$ measuring points for the compressed and uncompressed file sizes before and after syntactic / morphological distortion of each permutated language sample. On the basis of these values, the average morphological complexity score and the average syntactic complexity score are subsequently computed and their standard deviations (see Table 3.10) are calculated as described above. Variation between the individual measuring points is negligible and the average morphological / syntactic complexity scores are a fit proxy for the actual complexity in my text samples.

Subsequently, the *average overall complexity score* is obtained by calculating regression residuals of the mean compressed file sizes (dependent variable) and the mean uncompressed file sizes (independent variable). Intra-sample dispersion is accounted for by calculating the variation coefficient as described above. Table 3.11 shows the dispersion of the compressed and uncompressed file sizes.

Table 3.7.: Average morphological and syntactic complexity scores and their standard deviations in the parallel Alice corpus by language.

Language	**Morphological score**	**Standard deviation**	**Syntactic score**	**Standard deviation**
Dutch	-1.0827	0.0013	0.9190	0.0014
English	-1.1068	0.0015	0.9276	0.0016
Finnish	-1.0587	0.0014	0.9112	0.0015
French	-1.1049	0.0015	0.9253	0.0015
German	-1.0930	0.0014	0.9180	0.0014
Hungarian	-1.0816	0.0015	0.9158	0.0016
Italian	-1.0947	0.0013	0.9213	0.0016
Romanian	-1.0710	0.0015	0.9177	0.0015
Spanish	-1.1040	0.0014	0.9239	0.0015
Swedish	-1.1002	0.0015	0.9234	0.0015

Table 3.8.: Average morphological and syntactic complexity scores and their standard deviations in the Euro-Congo news corpus by language.

Language	Morphological score	Standard deviation	Syntactic score	Standard deviation
Dutch	-1.0932	0.0019	0.9208	0.0020
English	-1.1290	0.0018	0.9347	0.0020
Finnish	-1.0813	0.0018	0.9127	0.0019
French	-1.1432	0.0018	0.9318	0.0018
German	-1.1026	0.0019	0.9212	0.0022
Hungarian	-1.0904	0.0016	0.9154	0.0018
Italian	-1.0786	0.0015	0.9168	0.0016
Romanian	-1.1157	0.0018	0.9281	0.0019
Spanish	-1.1248	0.0019	0.9320	0.0021

Table 3.9.: Average morphological and syntactic complexity scores and their standard deviations in the Euro-Congo-Tunisia news corpus by language.

Language	**Morphological score**	**Standard deviation**	**Syntactic score**	**Standard deviation**
Dutch	-1.0999	0.0016	0.9205	0.0017
English	-1.1196	0.0014	0.9284	0.0015
Finnish	-1.0869	0.0015	0.9119	0.0015
French	-1.1559	0.0014	0.9313	0.0016
German	-1.0951	0.0014	0.9173	0.0016
Hungarian	-1.0983	0.0012	0.9153	0.0014
Italian	-1.0805	0.0012	0.9162	0.0013
Romanian	-1.1153	0.0015	0.9256	0.0016
Spanish	-1.1209	0.0014	0.9271	0.0015

Table 3.10.: Average morphological and syntactic complexity scores and their standard deviations in the semi-parallel Alice corpus by language.

Language	**Morphological score**	**Standard deviation**	**Syntactic score**	**Standard deviation**
Dutch	-1.0051	0.0057	0.9149	0.0044
English	-1.0180	0.0060	0.9191	0.0043
Finnish	-0.9814	0.0053	0.9112	0.0048
French	-1.0028	0.0060	0.9172	0.0047
German	-1.0177	0.0055	0.9153	0.0041
Hungarian	-0.9894	0.0057	0.9135	0.0054
Italian	-1.0034	0.0061	0.9156	0.0045
Romanian	-0.9876	0.0056	0.9135	0.0044
Spanish	-1.0017	0.0062	0.9171	0.0048
Swedish	-1.0065	0.0063	0.9163	0.0048

Table 3.11.: Mean uncompressed and compressed file sizes (in bytes) and their standard deviations in the semi-parallel Alice corpus by language.

Language	Uncompressed	Standard deviation	Compressed	Standard deviation
Dutch	13,559	914	5,662	350
English	13,605	1,055	5,551	398
Finnish	11,071	657	4,752	265
French	11,651	718	5,015	281
German	15,406	1,131	6,280	427
Hungarian	10,575	580	4,857	247
Italian	12,457	904	5,310	352
Romanian	11,952	868	5,216	347
Spanish	11,628	766	4,983	299
Swedish	10,951	575	4,532	220

3.2.2. Alice's Adventures in Wonderland

This section focuses on the parallel and semi-parallel Alice corpora. Specifically, I will compare the complexity rankings obtained from the parallel corpus to the rankings obtained from the semi-parallel corpus and establish whether the results correlate positively, i.e. whether the compression technique yields consistent results for both parallel and semi-parallel texts.

First, I will compare the complexity of the two corpora on the overall level. In order to obtain overall complexity hierarchies, the adjusted complexity scores for the parallel Alice corpus and the average overall complexity scores for the semi-parallel corpus are calculated as described above. The overall complexity ranking in the parallel corpus (Figure 3.4) is in decreasing order of complexity, Hungarian, Romanian, Dutch, Finnish, German, Italian, Spanish, French and English.

On the whole, these results are as I would expect them to be: Hungarian exhibits the highest overall complexity in the sample whereas English exhibits the lowest complexity. Most of the Romance languages, i.e. Italian, Spanish and French, cluster together in the less complex left area of the plot. Romanian, while being a Romance language, is ranking next to Hungarian. Finnish in the right, complex area and German, in the left less complex area of the plot, cover the middle ground. In fact, only Dutch, which is third on the complexity hierarchy in this sample, exhibits surprisingly high overall complexity. Comparing this ranking to the results of the semi-parallel corpus (Figure 3.5) I find that Hungarian still exhibits the highest overall complexity. English can still be found on the less complex left area of the plot and counts among the less complex languages, yet Spanish is the least complex language in the semi-parallel sample. In the ranking of overall complexity of the other languages I observe minor shifts: Italian, German and Dutch as well as French are (in decreasing order) complex whereas Romanian and Finnish have become less complex and can now be found in the left area of the plot. Thus, the complete complexity hierarchy in decreasing order is: Hungarian, Italian, German, Dutch, French, Romanian, Finnish, English and Spanish. In order to assess the similarity and the extent to which the two overall complexity hierarchies correspond, I perform the *Spearman Rank Correlation Test* for ordinal ranked data. The overall correlation of the two complexity hierarchies is with $r = 0.5$ ($p = 0.09$) moderate but significant. The reason for the only moderate correlation between the two overall complexity rankings is the loss of content alignment which was achieved through permutation and the resulting shifts in complexity. Strictly speaking, the parallel and semi-parallel Alice corpora are in terms of content two different databases.

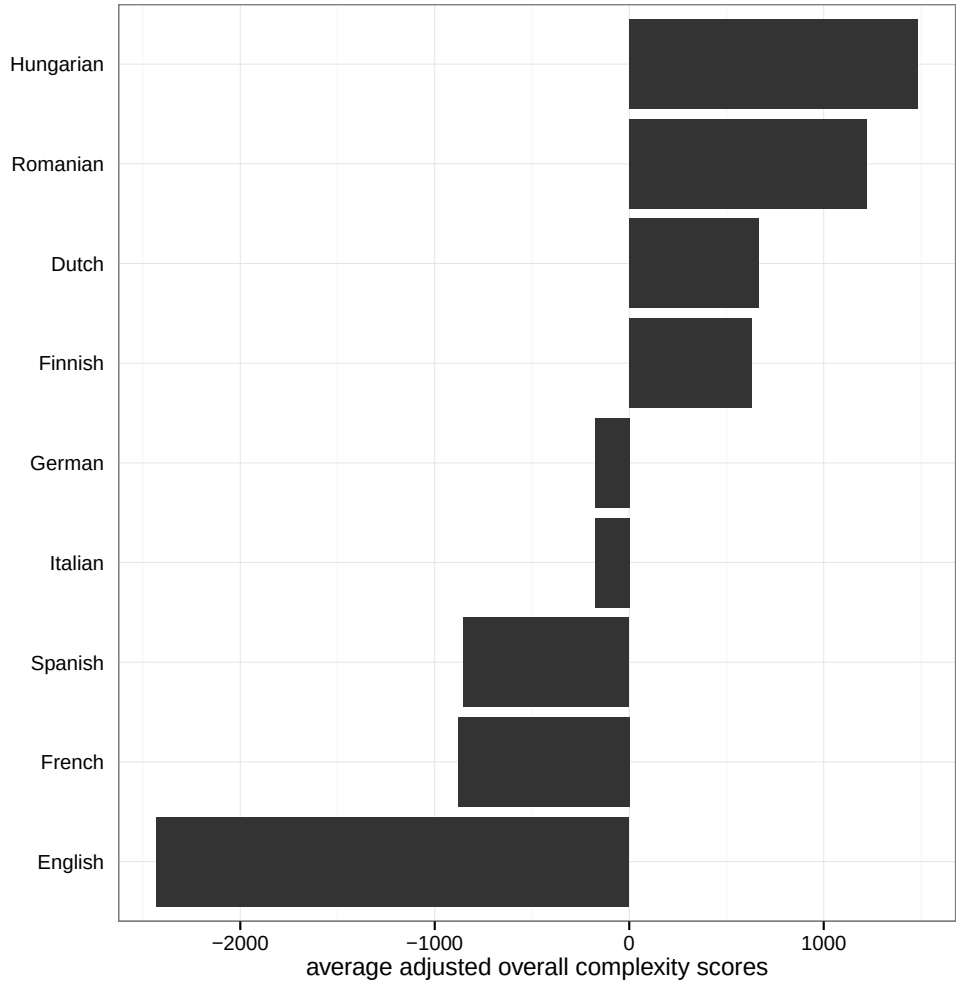

Figure 3.4.: Overall complexity hierarchy of the parallel Alice corpus. Negative residuals indicate below-average complexity; positive residuals indicate above-average complexity.

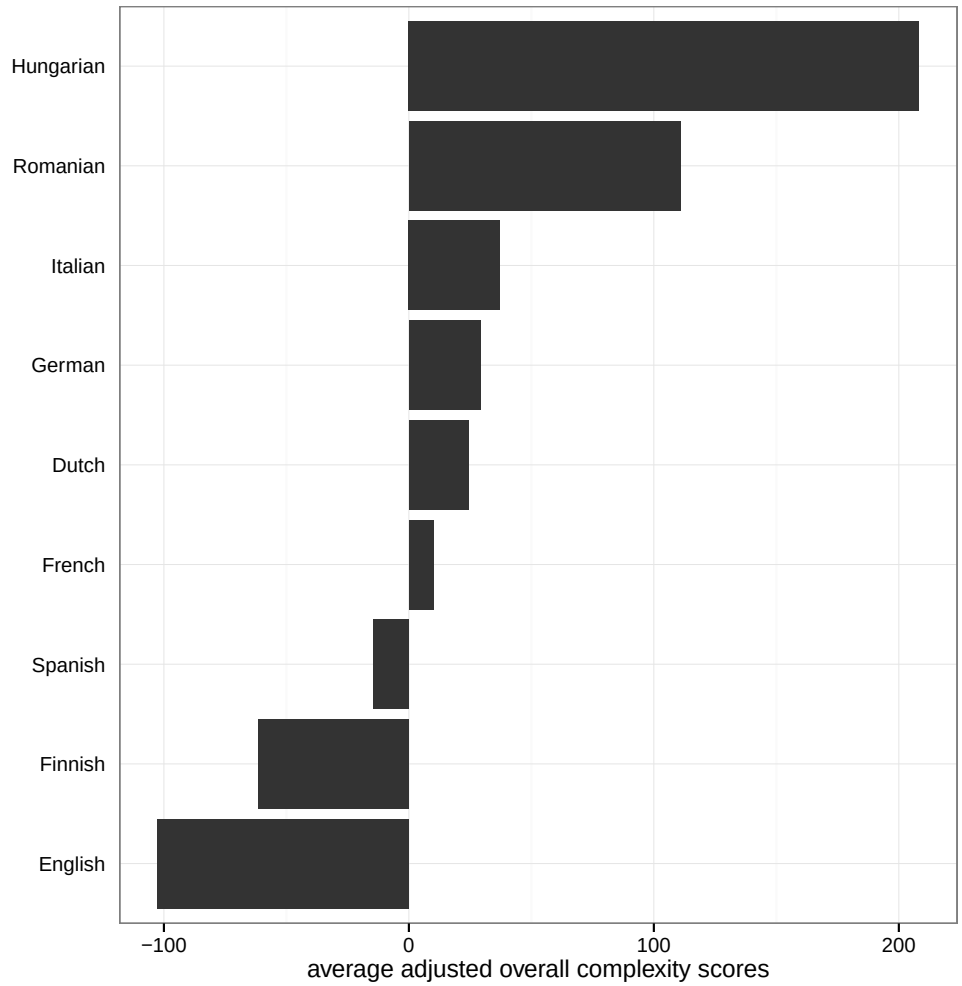

Figure 3.5.: Overall complexity hierarchy of the semi-parallel Alice corpus. Negative residuals indicate below-average complexity; positive residuals indicate above-average complexity.

Turning to morphological and syntactic complexity, I compute the average morphological complexity score and the average syntactic complexity score for both the parallel and semi-parallel corpus. The analysis of the parallel Alice corpus (Figure 3.6) dovetails with intuitions: Finnish, which according to the overall complexity score is the morphologically most complex but syntactically most simple language in the sample, is positioned in the extreme top left quadrant while morphologically simple but syntactically complex languages—in this sample Spanish, French and English—cluster in the bottom right quadrant of the plot. The more balanced languages in the sample are scattered across the middle field of the plot: Romanian, Hungarian and Dutch are morphologically relatively more complex than syntactically while German and Italian seem to be well balanced.

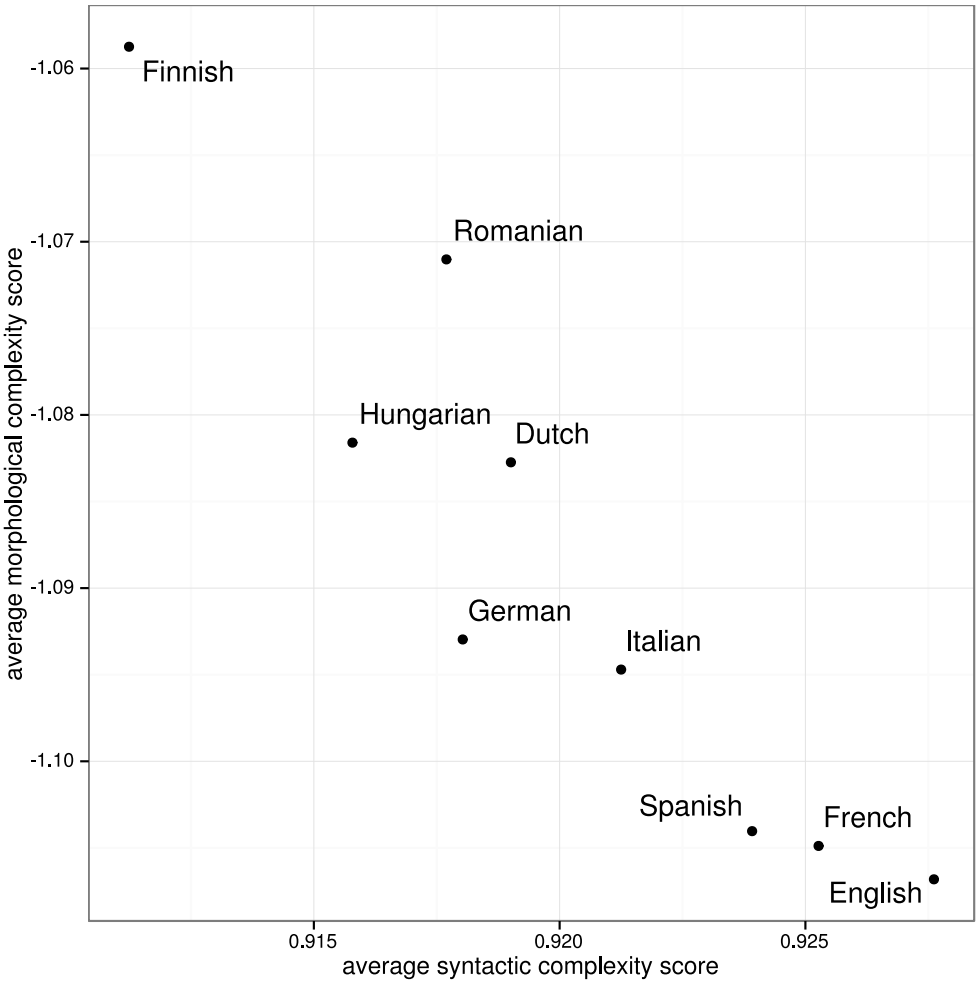

Figure 3.6.: Morphological by syntactic complexity in the parallel Alice corpus. Abscissa indexes increased syntactic complexity, ordinate indexes increased morphological complexity.

In the semi-parallel corpus (Figure 3.7) the nine languages seem to be sim-

ilarly distributed: Finnish and Hungarian, both located in the (extreme) top left quadrant, are clearly the morphologically most complex and syntactically most simple languages in the sample. English, in the bottom right quadrant, is the syntactically most complex but morphologically most simple language. Between these poles of either extreme morphological or syntactic complexity, the comparatively "balanced" languages, Dutch, Italian, Spanish, French and Romanian, are grouped in the centre of the plot. German, which exhibits medium syntactic complexity and low morphological complexity, is positioned at the bottom centre. In both datasets, a very significant trade-off between morphological and syntactic complexity can be observed: the negative correlation coefficient for the parallel corpus (Pearson's $r = -0.93, p = 0.0002$) and the semi-parallel corpus (Pearson's $r = -0.79, p = 0.006$) indicate that most languages in this sample trade off morphological for syntactic complexity. As a standalone fact, this finding is interesting insofar as it seems to strengthen the claim of the equal complexity hypothesis which postulates that complexity in one linguistic sub-domain is counterbalanced by simplicity in another sub-domain. Yet, this is just one piece in a bigger puzzle: the results of the overall complexity measurement obtained by the compression technique clearly suggest that some languages are more complex than others. Furthermore, in order to measure an overall complexity trade-off, all levels of a language (i.e. pragmatics, phonology, etc.) would need to be compared, not only morphology and syntax (Deutscher 2009). Be that as it may, despite very minor shifts among the balanced languages, the results of the morphosyntactic complexity analysis in the parallel and semi-parallel Alice corpus seem to be rather congruent. In other words, the compression technique can be effectively used with both parallel and semi-parallel texts. These findings are backed up statistically by conducting a Spearman's correlation test of the average syntactic and morphological complexity scores, respectively, of the two corpora. Overall, I observe high correlation between the two datasets: at the syntactic level Spearman's rho with $r = 0.82$ ($p = 0.005$) indicates very high correlation whereas correlation at the morphological level is slightly lower but still high ($r = 0.73, p = 0.02$).

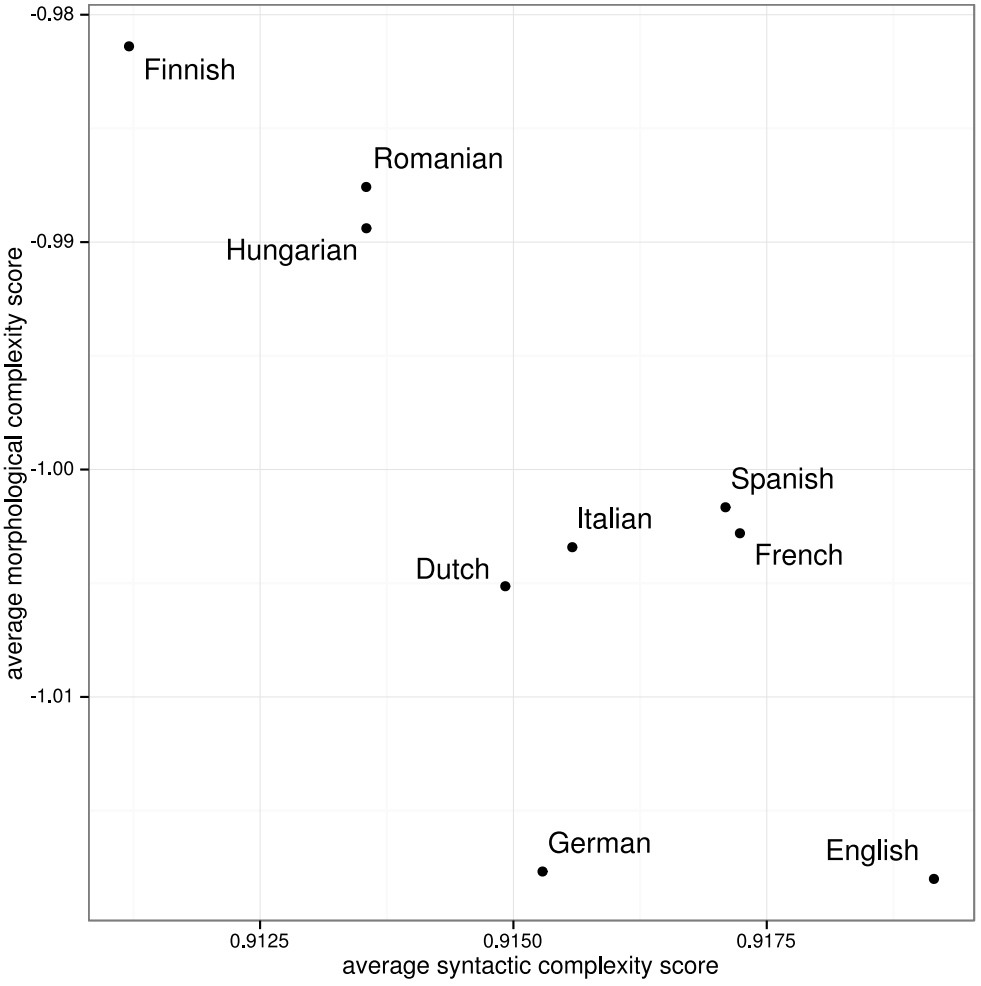

Figure 3.7.: Morphological by syntactic complexity in the semi-parallel Alice corpus. Abscissa indexes increased syntactic complexity, ordinate indexes increased morphological complexity.

In order to further validate these findings, I compare the results obtained by the compression technique to results from more traditional research. Bakker (1998) investigates syntactic complexity on the basis of flexibility in word order patterns. He defines flexibility in terms of the variability and number of word order patterns in a language and assigns values between 0 and 1 for syntactic flexibility. The more word-order patterns a given language exhibits, the more flexible is a language. Values close to zero indicate less flexibility and thus increased syntactic complexity (Bakker 1998: 387). This is another way of saying that flexible languages are syntactically simple while inflexible languages are syntactically complex. In Table 3.12 below, the syntactic complexity ranking based on Bakker's flexibility values is compared to the syntactic complexity rankings established by the Kolmogorov-based syntactic complexity scores.

On a statistical level, I compare Bakker's ranking with the rankings in the parallel and semi-parallel corpus by means of Spearman's correlation test. The syntactic complexity order of the languages in the parallel Alice corpus very highly correlates with Bakker's ranking ($r = 0.75, p = 0.02$). Spearman's rho of correlation for the syntactic complexity ranking in the semi-parallel Alice corpus indicates with $r = 0.43$ ($p = 0.1$) a moderate but significant correlation.

To sum up, the overall rankings of complexity of the parallel and semi-parallel Alice corpus are fairly similar and statistically correlate to a significant extent. On a morphological and syntactic plane, the results from the two corpora are highly congruent and are in line with complexity rankings reported in traditional research. Interestingly, dislocations in the complexity hierarchies between the parallel and semi-parallel corpus are particularly prominent among the morphosyntactically "balanced" languages such as German or Italian. These languages are, as measured by the compression technique, neither extremely syntactically complex nor extremely morphologically complex but seem to express grammatical information equally both with syntax and morphology. My results therefore suggest that the propositional content—which might decisively influence the choice between morphologically versus syntactically encoded information—should be a more important factor for successful complexity measurement of balanced languages than for "extreme" languages such as English or Finnish.

Table 3.12.: Comparison of syntactic complexity ranking of the nine languages according to Bakker's flexibility index and syntactic complexity scores in the parallel and semi-parallel Alice corpora. Note that Bakker's analysis does not include Hungarian.

Bakker's ranking	**Flexibility index**	**Parallel Alice**	**Syntactic complexity score**	**Semi-parallel Alice**	**Syntactic complexity score**
1. French	0.1	1. English	0.928	1. English	0.919
2. Spanish	0.3	3. French	0.925	2. French	0.917
3. Italian	0.3	3. Spanish	0.924	3. Romanian	0.917
4. English	0.4	4. Italian	0.921	4. Spanish	0.916
5. German	0.4	5. Dutch	0.919	5. Italian	0.916
6. Dutch	0.4	6. German	0.918	6. German	0.915
7. Romanian	0.5	7. Romanian	0.918	7. Dutch	0.915
8. Finnish	0.6	8. Hungarian	0.916	8. Hungarian	0.914
9. Hungarian	NA	9. Finnish	0.911	9. Finnish	0.911

3.2.3. Newspaper texts

After having explored Kolmogorov complexity measurements in parallel and semi-parallel texts in Section 3.2.2, I will now apply the compression technique to genuinely non-parallel texts. Non-parallel texts are neither translational equivalents nor re-sampled parts of parallel texts but, in the context of this chapter, non-parallel texts are independently composed texts on a given identical topic. Thus, complexity rankings of two non-parallel newspaper corpora will be discussed and compared to the rankings of the parallel Alice corpus as well as to the complexity hierarchy established by Bakker (1998).

First, the three-topic newspaper corpus on the European currency crisis, and the political situation in Congo and Tunisia, will be discussed. The overall complexity is assessed by obtaining two measurements for each language file: the file size before compression and the file size after compression. Subsequently, I calculate adjusted overall complexity scores by applying linear regression to the two measurements. The resulting overall complexity hierarchy (Figure 3.8) in this corpus is as follows: Italian is overall the most complex language in this sample, followed closely by Hungarian and—at a considerable distance—by German. Finnish, English, Romanian, Spanish, Dutch and Spanish are, in this order, simple. These results are rather different from the overall hierarchy which I obtained for the parallel corpus. The ranking in the parallel Alice corpus is, in decreasing order of complexity, Hungarian, Romanian, Dutch, Finnish, German, Italian, Spanish, French and English. Particularly prominent is the difference in complexity observed for the Italian and English language samples. In the non-parallel corpus Italian ranks highest in complexity while it is rather simple in the parallel corpus. English, formerly the simplest language in the parallel corpus data, now exhibits even more complexity than Romanian. All in all, I observe considerable dislocations regarding overall complexity in the non-parallel Euro-Congo-Tunisia dataset. Performing Spearman's correlation test in order to statistically back-up these observations, the correlation between the parallel and non-parallel ranking is low ($r = 0.26, p = 0.25$).

Next, I analyse overall complexity in the two-topic newspaper corpus which samples texts on the topics 'Euro crisis' and 'Congo' only. The adjusted overall complexity scores are calculated as described above and yield the following hierarchy of overall complexity (Figure 3.9): Hungarian is rated the most complex language in this dataset, closely followed by Italian and, in decreasing order of complexity, Finnish, German, Romanian, Dutch, Spanish, English and French. Although Italian is still surprisingly and unproportionally complex, the other languages in this dataset behave roughly as one would expect. Hungarian and Finnish are relatively complex whereas French, Spanish and English are relatively simple. Complexity dislocations can mainly be observed among languages which are, according to Kolmogo-

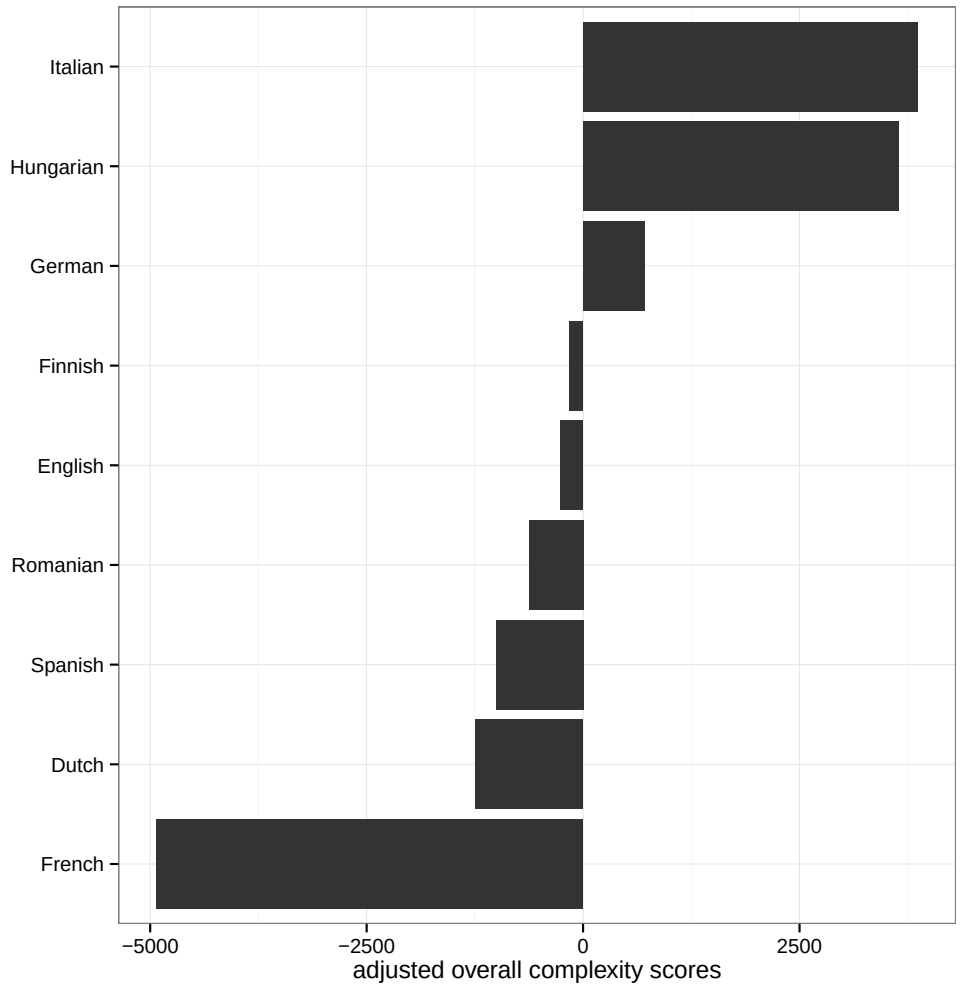

Figure 3.8.: Overall complexity hierarchy in the Euro-Congo-Tunisia newspaper corpus. Negative residuals indicate below-average complexity; positive residuals indicate above-average complexity.

rov measurements, "balanced" between morphological and syntactic complexity such as German, Dutch or Italian. Despite the fact that such dislocations among balanced languages seem to be common (cf. Chapter 3.2.2), it does not sufficiently explain the extremely high morphological complexity exhibited by the Italian outlier. Considering the fact that propositional content plays an important role in cross-linguistic complexity analyses, a lack of homogeneity in the composition of this particular subcorpus is likely at fault. Be that as it may, comparing this ranking to the hierarchy of the parallel Alice corpus, I find that the correlation is moderate (Spearman's $r = 0.63, p = 0.04$).

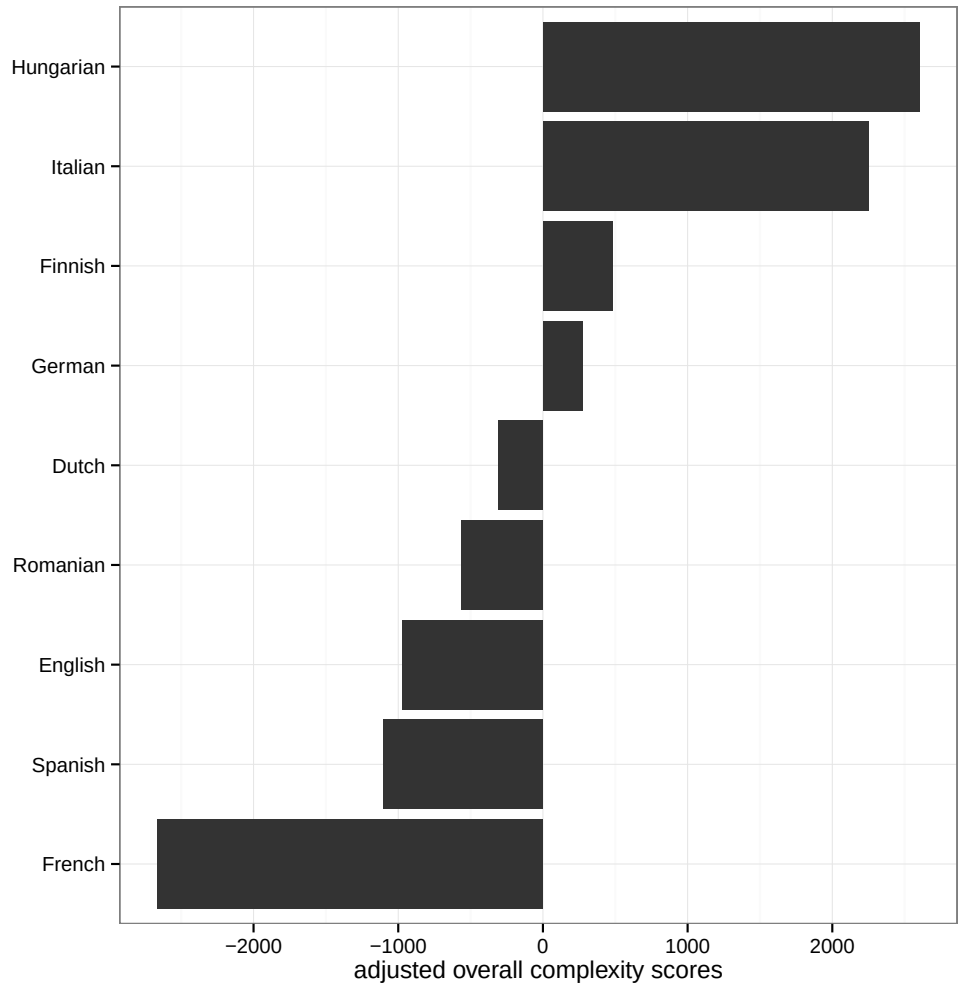

Figure 3.9.: Overall complexity hierarchy in the Euro-Congo newspaper corpus. Negative residuals indicate below-average complexity; positive residuals indicate above-average complexity.

I will now turn to the analysis of morphological and syntactic complexity in the two newspaper corpora. In the three-topic dataset, Finnish and, surprisingly, Italian, located in the extreme left quadrant of the plot (Figure 3.10), are the morphologically most complex and syntactically most simple languages. Hungarian, German and Dutch cluster in the upper middle field and are morphologically more complex than syntactically whereas Romanian, Spanish and English cluster in the lower middle field and are syntactically more complex than morphologically. French, which is positioned in the bottom right quadrant of the plot, is syntactically complex but morphologically simple. Despite the Italian outlier, the results for syntactic complexity in the Euro-Congo-Tunisia corpus correlate very highly with the results of the parallel Alice corpus (Spearman's $r = 0.83, p = 0.004$). Complexity on the morphological level correlates only moderately (Spearman's

$r = 0.55, p = 0.067$).

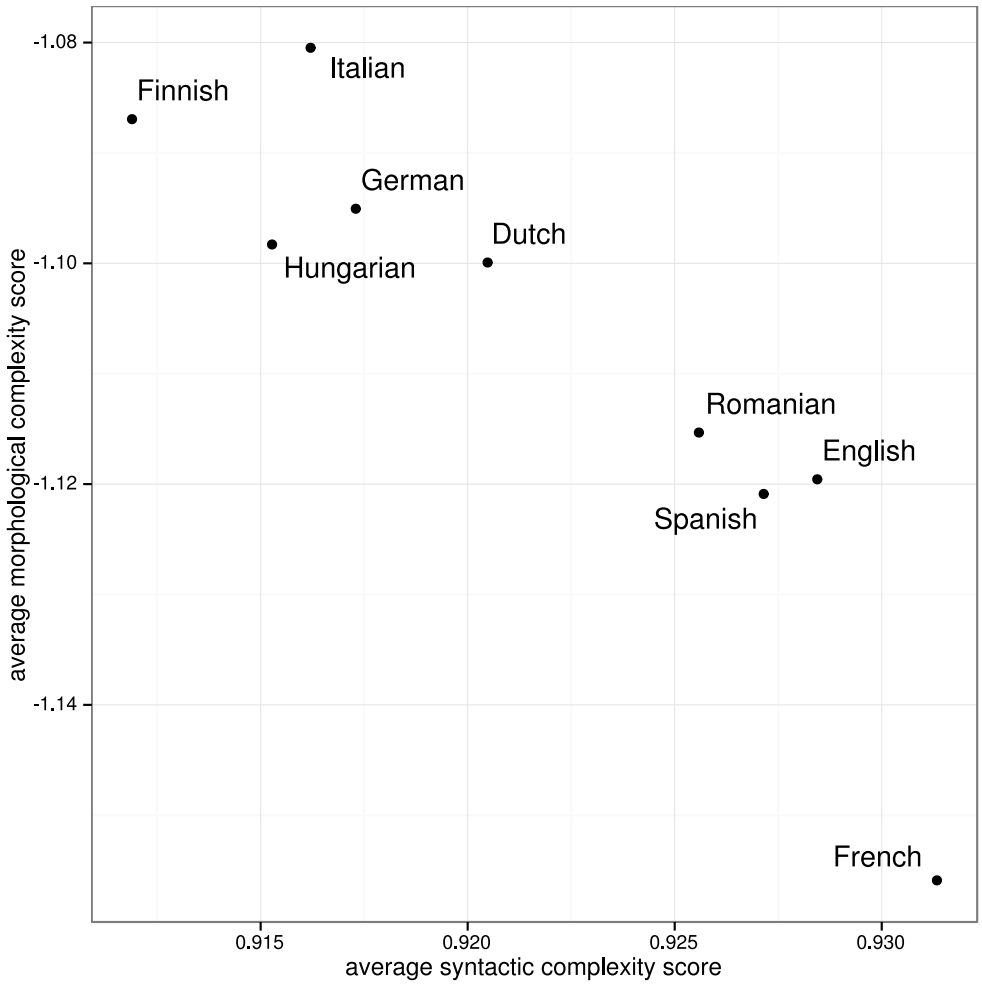

Figure 3.10.: Morphological by syntactic complexity in the non-parallel Euro-Congo-Tunisia newspaper corpus. Abscissa indexes increased syntactic complextiy, ordinate indexes increased morphological complexity.

The analysis of morphological and syntactic complexity in the two-topic Euro-Congo dataset is shown in Figure 3.11. The distribution is similar to the three-topic dataset. Morphologically complex but syntactically simple languages are grouped in the top left quadrant: Finnish, Italian, and Hungarian. Dutch, German and Romanian are scattered across the whole middle area of the plot. The syntactically complex but morphologically simple languages Spanish, English and French are positioned in the bottom right quadrant. Apart from the Italian data point, which, again, is an outlier, these results tie in neatly with my previous research of syntactic and morphological complexity in the parallel Alice corpus. These findings are statistically backed-up by conducting a Spearman's correlation

test between the non-parallel Euro-Congo corpus and the parallel Alice corpus. At the syntactic level the correlation between the two datasets is very high ($r = 0.82, p = 0.005$), yet at the morphological level the two datasets correlate only moderately but still significantly ($r = 0.63, p = 0.038$). As in the other datasets, I observe a significant trade-off between morphological and syntactic complexity (Pearson's correlation coefficient: Euro-Congo $r = -0.94, p = 0.00007846$; Euro-Congo-Tunisia $r = -0.90, p = 0.0004$) in both newspaper corpora.

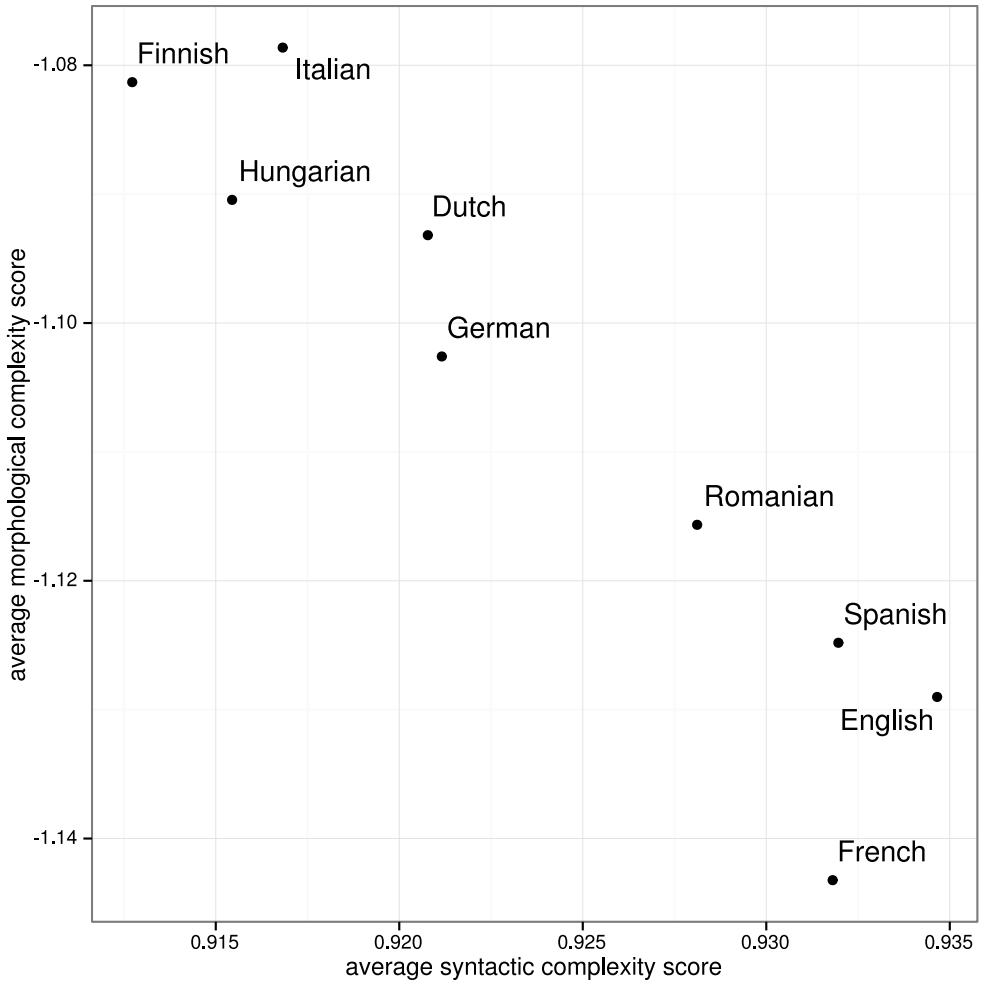

Figure 3.11.: Morphological by syntactic complexity in the non-parallel Euro-Congo newspaper corpus. Abscissa indexes increased syntactic complextiy, ordinate indexes increased morphological complexity.

In order to further validate the results of the non-parallel newspaper corpora, I compare the syntactic complexity scores of each dataset to the complexity ranking established by Bakker (1998). Table 3.13 shows the ranking of the sample languages according to Bakker and according to the Kolmogorov complexity metric for both newspaper corpora. Statistically,

the correlation between the Euro-Congo-Tunisia dataset and Bakker's flexibility ranking is moderate (Spearman's $r = 0.39, p = 0.17$). The ranking in the Euro-Congo dataset correlates (Spearman's $r = 0.36, p = 0.19$) moderately.

In summary, the compression technique works fairly well for measuring syntactic and morphological complexity in non-parallel texts. The comparison of the Euro-Congo-Tunisia and Euro-Congo newspaper datasets with the parallel Alice corpus shows that syntactic complexity measurements correlate very highly while morphological complexity measurements correlate only moderately. Measuring overall complexity in non-parallel texts is possible, yet the technique works only moderately well as correlation with the parallel Alice corpus has shown. On the whole, the differences between the two news datasets—the two-topic corpus yields overall slightly better results than the three-topic dataset—and the Italian outlier suggest that the propositional content of the texts analysed plays an important role. This means that, while the texts need not be translational equivalents, a certain amount of content control needs to be applied in the composition of the corpus to ensure homogeneity within and comparability across the corpus components. Randomly chosen texts cannot be successfully assessed with the compression technique and compared across languages.

Table 3.13.: Comparison of syntactic complexity ranking of the nine languages according to Bakker's flexibility index and syntactic complexity scores in the two newspaper datasets. Note that Bakker's analysis does not include Hungarian.

Bakker's ranking	**Flexibility index**	**Euro-Congo-Tunisia**	**Syntactic complexity score**	**Euro-Congo**	**Syntactic complexity score**
1. French	0.1	1. French	0.931	1. English	0.935
2. Spanish	0.3	2. English	0.928	2. Spanish	0.932
3. Italian	0.3	3. Spanish	0.927	3. French	0.932
4. English	0.4	4. Romanian	0.926	4. Romanian	0.928
5. German	0.4	5. Dutch	0.920	5. German	0.921
6. Dutch	0.4	6. German	0.917	6. Dutch	0.921
7. Romanian	0.5	7. Italian	0.916	7. Italian	0.917
8. Finnish	0.6	8. Hungarian	0.915	8. Hungarian	0.915
9. Hungarian	NA	9. Finnish	0.912	9. Finnish	0.913

3.3. Summary

This chapter explored and substantially extended the Juola-style compression technique (Juola 2008, 1998) by conducting several experiments. First, I corroborated prior applications of the compression technique by measuring overall, syntactic and morphological complexity in a parallel corpus of the Gospel of Mark covering six languages and a handful of historical varieties of English. Breaking new ground, a statistically more robust version of the compression technique was then extended to the analysis of semi-parallel and non-parallel corpora of literary and newspaper writing in nine European languages.

The complexity measurements obtained with the compression technique yield linguistically meaningful results which are in line with previous complexity rankings reported in more traditional research (Bakker 1998; Juola 2008; Nichols 2009). All in all, it captures both intra-linguistic as well as cross-linguistic complexity variation quite well. Section 3.1 verified that compression algorithms of the Lempel-Ziv family accurately assess linguistic complexity of different languages on the overall, morphological and syntactic level. Even intra-linguistic complexity developments can be algorithmically captured as I trace the diachronic change of English from early West-Saxon, a morphologically complex language rich in inflections, to present-day English, a syntactically complex language heavily relying on syntax to convey meaning. A follow-up experiment, which uses a refined, statistically more robust version of the Juola-style compression technique, has shown that Kolmogorov measurements need not be limited to parallel-text databases but can be successfully applied to semi-parallel and—to some extent—non-parallel texts. Morphological and syntactic complexity can be measured well in semi- and non-parallel texts, outliers notwithstanding. Particularly the measurement of balanced languages, i.e. languages which encode grammatical information to roughly equal parts through syntax and morphology, seems to be sensitive to the propositional content and composition of semi- and non-parallel copora. Put differently, in balanced languages, the propositional content seems to influence the choice between encoding information syntactically or morphologically. In short, the measurement of overall complexity in non-parallel texts is, in principle, possible but its success depends to a large extent on the propositional content of the texts and is thus subject to textual variation. Content control of non-parallel text databases is therefore crucial and randomly chosen texts cannot be used for cross-linguistic complexity analyses with the compression technique. To conclude, the compression technique was shown to be a radically objective and economical shortcut to measure linguistic complexity on the overall, morphological and syntactic tier.

4. Excursion: Targeted file manipulation[1]

So far the compression technique has been shown to yield linguistically meaningful results when measuring the overall, syntactic and morphological complexity of parallel, semi-parallel and non-parallel corpora. This chapter builds on and extends the work presented in Ehret (2014) in order to demonstrate how algorithms can be utilised to measure morphological and syntactic complexity in a detailed fashion. In this spirit, a new flavour of the compression technique to measure Kolmogorov complexity of specific linguistic features—*targeted file manipulation*—is introduced. To be more precise, the degree to which individual morphological markers such as *–ing*, and functional constructions such as progressive (*be* + verb–*ing*) contribute to the syntactic and morphological complexity in three different texts will be analysed. In contrast to Ehret (2014) who applies targeted file manipulation to a mixed-genre corpus of literary writing, scripture and newspaper texts (for a brief summary of the paper see Section 2.2.2), this chapter measures the features' intertextual variation in terms of Kolmogorov complexity by separately analysing the three text types. On an interpretational plane, their textual complexity on the syntactic and morphological level will be inferred.

4.1. Method and data

In order to measure the contribution of morphological markers and constructions to complexity in English texts, and to further explore the possibility of assessing detailed morphological and syntactic complexity with compression algorithms, I draw up a set of $N = 10$ high-frequency morphosyntactic features comprising:

(i) morphological markers: *–ing*, *–ed*, genitive *'s*, plural *–s* and third person singular *–s*;

(ii) and a handful of functional constructions: progressive aspect *be* + verb–*ing*, perfect aspect *have* + verb past participle, passive voice *be* + verb past participle and the future markers *will* and *going to*.

[1] A partial summary of this chapter has appeared as Ehret (2014).

The feature set is restricted to the most frequent (common) markers in the English verb and noun phrase (Biber et al. 1999; Johansson & Hofland 1989) as well as functional constructions encoding tense, aspect and voice. The analysis of other feature areas such as modals, adjectives or pronouns is outside the scope of this pilot study and is reserved for future research. Table 4.1 provides the text frequency of each feature per text type. Note that for reasons of operationalisation, no distinction is made between inflectional and non-inflectional occurrences of morphological markers (*he is singing* vs. *he hates singing*) in this study.

Table 4.1.: Text frequency of morphological markers and constructions per text type.

Feature	**Text frequency**			
	Alice	Mark	Euro-Congo	Total
–ing	336	244	316	896
–ed	390	424	550	1,364
Genitive *'s*	25	23	124	172
Plural *–s*	326	535	974	1,835
3rd ps sg *–s*	188	290	321	799
going to	8	1	4	13
Passive	59	123	229	411
Perfect	121	110	127	358
Progressive	96	87	71	254
will	68	109	89	266

The effect of these morphological markers and constructions on the complexity of texts is analysed in samples of Carroll's *Alice's Adventures in Wonderland*, the Gospel of Mark in the English Standard Version (ESV) and the Euro-Congo newspaper corpus, thus covering three distinct text types: literary writing, religious writing / scripture and newspaper texts. Each text type is analysed separately in order to gauge the variation of the features in terms of Kolmogorov complexity across the different text types. The comparability of the measurements across the three texts is ensured by sampling an equal amount of data from each text, roughly 14,000 words, in full sentences so that syntactic structures remain intact. The number of sentences and words for each text type are provided in Table 4.2.

Table 4.2.: Number of sentences and words per text genre.

Text	Text type	Sentences	Words
Alice	literary	792	14,010
Mark	religious	807	14,009
Euro-Congo	newspaper	604	14,007
Total		22,203	42,026

Methodologically, targeted file manipulation is used to measure the contribution of these morphological markers and constructions to the morphological and syntactic complexity in the three texts. Essentially, targeted file manipulation removes specific target structures from a text. The idea is to assess the contribution of these target structures to the morphological and syntactic complexity in text samples by comparing the complexity of the manipulated samples and the unmanipulated samples. In other words, I compare the morphological / syntactic complexity of a given text sample from which a certain feature was removed to the morphological / syntactic complexity of the original text sample, including the specific feature. To this end, targeted file manipulation, specifically systematic removal, is combined with the compression technique, i.e. multiple random distortion and subsequent compression. On a more technical note, I remove one feature at a time from each of the three texts and obtain a set of feature-manipulated text samples. These manipulated texts, and the original intact version of each text, are then subjected to random distortion and compression. For each text sample the average morphological and the average syntactic complexity scores are calculated. Based on these scores, the morphological and syntactic complexity of the feature-manipulated texts, and the complexity of the original texts, can be compared. The difference in complexity between the manipulated and original texts indicates the amount of morphological / syntactic complexity that an individual feature contributes to the original text. Note that targeted file manipulation is a text-based method, that is to say the precise amount of a feature's contribution to the morphological and syntactic complexity of a given text varies according to the morphological and syntactic complexity of the original text, respectively. The extent of this text-dependent variation will be assessed in the following section.

On an interpretational level, the morphological and syntactic complexity of each feature is inferred from the amount of complexity it contributes to the original text. Since this complexity is as already mentioned, to some extent text-dependent, I will refer to it as *textual complexity*. A feature that increases the complexity of the original text is considered complex while a feature that decreases the complexity of the original text is considered simple (less complex).

The targeted manipulation of the above listed morphological markers and constructions is implemented as follows. First, the texts are annotated with part-of-speech tags using the `Stanford CoreNLP tool` (Toutanova et al. 2003). The part-of-speech tags permit the automatic manipulation of the morphological markers and facilitate the manual coding of the functional constructions. Generally, each feature is identified and manipulated in such a way that the texts are altered as little as possible. Technically speaking, the markers were automatically deleted, i.e. conjugated verbs and inflected nouns were replaced by their lemma with the help of a `python` [2] script which identifies the endings on the basis of their part-of-speech tags (see Appendix C for the full script). For instance, the corresponding part-of-speech tag for *–ing* is VBG. Examples (1)–(5) illustrate how the manipulation of morphological markers was implemented.

(1) a. Alice was [**beginning**]$_{\text{ing}}$ to get very tired of [**sitting**]$_{\text{ing}}$ by her sister on the bank and of having nothing to do: [...].
b. Alice was **begin** to get very tired of **sit** by her sister on the bank and of having nothing to do: [...].
[ALICE]

(2) a. [...]John [**appeared**]$_{\text{ed}}$ baptizing in the wilderness and proclaiming a baptism of repentance for the forgiveness of sins.
b. [...]John **appear** baptizing in the wilderness and proclaiming a baptism of repentance for the forgiveness of sins.
[MARK]

(3) a. [...]and was surprised to see that she had put on one of the [**Rabbit's**]$_{\text{genitive s}}$ little white kid gloves while she was talking.
b. [...]and was surprised to see that she had put on one of the **Rabbit** little white kid gloves while she was talking.
[ALICE]

(4) a. [...]which sparked [**clashes**]$_{\text{plural s}}$ between angry [**demonstrators**]$_{\text{plural s}}$ and police, according to witnesses.
b. [...]which sparked **clash** between angry **demonstrator** and police, according to witnesses.
[EURO-CONGO]

(5) a. If the Lisbon treaty is reopened, Cameron has to tread carefully between Tory backbenchers [...]and most of the rest of the EU, who are wary of getting bogged down in a row about what Britain [**wants**]$_{\text{3rd person s}}$
b. If the Lisbon treaty is reopened, Cameron has to tread carefully between Tory backbenchers [...]and most of the rest of the EU, who are wary of getting bogged down in a row about what Britain **want**
[EURO-CONGO]

Adapted from Ehret (2014)

[2] Python Software Foundation. Python Language Reference, version 3. URL `http://www.python.org`

The future markers *going to* and *will* were deleted as illustrated in examples (6)–(7). The manipulation of the other functional constructions was manually conducted, as not every present participle ending in *–ing* and annotated as VBG is part of a progressive construction. The progressive, passive and perfect were thus manually identified and manipulated by deleting the auxiliary *be* / *have* and replacing the main verb with its lemma (see examples (8)–(10)). Construction manipulation is implemented as lemma-substitution because it does not introduce new irregularity and complexity. For example, replacing verbal constructions such as *was singing* with irregular past tense forms (*sang*) instead of the lemma (*sing*) would add irregularity and thus unproportionally increase the complexity of the text.

(6) a. And, as you might like to try the thing yourself, some winter day, I [**will**]$_{\text{will}}$ tell you how the Dodo managed it.
b. And, as you might like to try the thing yourself, some winter day, I tell you how the Dodo managed it.
[ALICE]

(7) a. And he did not want anyone to know for he was teaching his disciples, saying to them, the son of man [**is going to**]$_{\text{going to}}$ be delivered into the hands of men and they kill him.
b. And he did not want anyone to know for he was teaching his disciples, saying to them, the son of man be delivered into the hands of men and they kill him.
[MARK]

(8) a. Alice [**was beginning**]$_{\text{progressive}}$ to get very tired of sitting by her sister on the bank and of having nothing to do: [. . .].
b. Alice **begin** to get very tired of sit by her sister on the bank and of having nothing to do: [. . .].
[ALICE]

(9) a. A further 110 people [**were arrested**]$_{\text{passive}}$ on suspicion of affray.
b. A further 110 people **arrest** on suspicion of affray.
[EURO-CONGO]

(10) a. More than 140 people [**have been**]$_{\text{perfect}}$ arrested at a protest in central London over the bitterly contested elections in the Democratic Republic of Congo.
b. More than 140 people **be** arrested at a protest in central london over the bitterly contested elections in the Democratic Republic of Congo.
[EURO-CONGO]

Adapted from Ehret (2014)

Constructions with negative contractions such as *won't*, *hasn't* or *isn't* were manipulated by deleting the constructions and replacing *n't* with the negative particle *not* (11).

(11) a. Oh! [**won't**]$_{\text{will negative contraction}}$ she be savage if I've kept her waiting!

b. Oh! **not** she be savage if I've kept her waiting!
[Alice]

Adapted from Ehret (2014)

Subsequently, all part-of-speech tags were removed and the plain texts, feature-manipulated and originals, were treated with the compression technique described in the previous chapters. Each text is morphologically and syntactically distorted by removing 10% of all orthographically transcribed characters / word tokens prior to compression. Through morphological distortion new word forms are created, i.e. morphological complexity is increased, which affects the compression performance of languages with an overall simpler morphology. Syntactic distortion leads to the collapsing of word-order regularities and highly affects syntactically complex languages. The distortion and compression script is implemented with $N = 1000$ iterations in order to obtain statistically robust results which are not due to mere coincidence. The script returns the compressed file sizes of each text before and after syntactic / morphological distortion for every iteration. Based on these file sizes, the average syntactic complexity score and the average morphological complexity score are calculated, specifically the mean of $N = 1000$ morphological / syntactic complexity ratios is taken. The intra-sample variation, i.e. the dispersion between the measurements taken in the different iterations, in the three texts (Table 4.3 and Table 4.4) is very small as indicated by the low standard deviations. This means that, statistically, the mean syntactic and mean morphological complexity scores approximate and reflect the average complexity of the texts well.

On a statistical sidenote, the measurements which are presented in the following sections are calculated from compressed file sizes in bytes (as described above). The differences between the values of the average morphological and syntactic complexity scores across manipulated texts and their originals often seem to be of statistical insignificance. However, this is not the case as differences between compressed file sizes are very small to start with (Juola 1998). I calculate *Tukey's honestly significant difference test* (Tukey's HSD) for all pairs of the morphological and syntactic complexity scores in the three texts. Tukey's HSD is a statistical siginificance test that allows for multiple comparisons in a single step and is based on an *ANOVA* (Analysis of Variance) table. It calculates whether differences between two means are statistically significant, returning the difference between the two means, a p-value and the upper and lower bounds of the confidence intervals (Baayen 2008: 106–107). Nonetheless, the statistical significance of the measurements is not of primary interest to this analysis. Suffice it to say that Tukey's HSD demonstrates that even minimal differences between morphological / syntactic complexity scores can be of statistical significance; this means that measurements obtained by compression are statistically ro-

bust and replicable. The tables with the full statistics of Tukey's HSD are provided in Appendix A.

Table 4.3.: Average morphological and syntactic complexity scores and their standard deviations by text and morph.

Morph	**Alice**		**Mark**		**Euro-Congo**	
	Morphological score	Syntactic score	Morphological score	Syntactic score	Morphological score	Syntactic score
Original	0.00194	0.00214	0.00217	0.00232	0.00201	0.00229
–ing	0.00199	0.00217	0.00217	0.00247	0.00208	0.00227
–ed	0.00201	0.00221	0.00221	0.00248	0.00211	0.0023
Genitive *'s*	0.00196	0.00229	0.00229	0.00238	0.00212	0.00233
Plural *–s*	0.00191	0.00224	0.00224	0.00243	0.00207	0.0023
3rd ps sg *–s*	0.00196	0.00217	0.00217	0.0025	0.00211	0.0023

Table 4.4.: Average morphological and syntactic complexity scores and their standard deviations by text and construction.

Construction	**Alice**		**Mark**		**Euro-Congo**	
	Morphological score	Syntactic score	Morphological score	Syntactic score	Morphological score	Syntactic score
Original	0.002	0.00216	0.00226	0.00232	0.00212	0.00231
going to	0.00204	0.00212	0.00226	0.00228	0.00208	0.00223
Passive	0.00207	0.00217	0.00225	0.00242	0.00207	0.00235
Perfect	0.00202	0.00207	0.00219	0.00237	0.00211	0.0023
Progressive	0.00194	0.00214	0.00222	0.00247	0.00208	0.00233
will	0.00199	0.00218	0.00230	0.00236	0.00204	0.00239

4.2. Analysing morphological markers

Having set the stage by outlining the methodology, I proceed to analyse the effect of the morphological markers *–ing*, *–ed*, genitive *'s*, plural *–s* and third person singular *–s* on the complexity of the three texts at the syntactic and morphological level. First, the variation of textual marker complexity across the three texts is statistically assessed. This is followed by a detailed discussion and comparison of their textual complexity in the three texts.

The extent to which the effect of morphological markers on the complexity in the three different texts varies is assessed by pairwise correlation of the complexity scores obtained as described in Section 4.1. Specifically, Spearman's rank correlation rho is used to measure the similarity between the rankings of the average morphological and syntactic complexity scores of the marker-manipulated texts in Alice, Mark and the Euro-Congo newspaper corpus. The correlation of the syntactic complexity scores of the marker-manipulated texts (Table 4.5) is moderate to negative. While the syntactic ranking of the marker-manipulated texts in Alice and Mark correlates moderately well, the correlation between the other pairs, Alice and Euro-Congo and Mark and Euro-Congo, is negative. The morphological complexity scores (Table 4.6) exhibit moderate correlation for the Alice–Euro-Congo pair but very low correlation for Alice–Mark and Mark–Euro-Congo. In short, the statistical similarity between the textual complexity of morphological markers across the three different texts is moderate.

Figures 4.1–4.3 present the original Alice, Mark and Euro-Congo news texts as well as their marker-manipulated texts according to morphological and syntactic complexity. In all three plots, the originals are located in the top left quadrant and are the morphologically most complex but syntactically most simple text. The marker-manipulated texts are spread across the right middle to lower right half of the plots, i.e. they exhibit lower morphological but higher syntactic complexity than the originals. Thus, in the big-picture perspective, all morphological markers increase the morphological complexity of the texts analysed but decrease their syntactic complexity. The former observation is not unexpected and is in line with the "more is

Table 4.5.: Correlation of syntactic complexity scores for marker-manipulated texts across the three texts.

	Syntactic correlation	
	Mark	Euro-Congo
Alice	$rho = 0.5$ $p = 0.225$	$rho = -0.9$ $p = 0.99$
Mark		$rho = -0.08$ $p = 0.96$

Table 4.6.: Correlation of morphological complexity scores for marker-manipulated texts across the three texts.

	Morphological correlation	
	Mark	Euro-Congo
Alice	$rho = -0.1$ $p = 0.61$	$rho = 0.6$ $p = 0.18$
Mark		$rho = 0.3$ $p = 0.34$

more complex" motto (Arends 2001: 180): the more morphological marker types / distinctions a text contains the more morphologically complex it is overall (see, for instance, Arends 2001; McWhorter 2001a, 2012; Shosted 2006). The fact that morphological markers decrease syntactic complexity seems surprising. Yet, some markers like *–ing* are part of morphosyntactic patterns such as the progressive aspect. They are therefore likely to be preceded by a form of the verb *be*. In very simplified terms, the *ing*-marker could be said to facilitate the prediction of progressive patterns and thus decrease syntactic complexity.

Let us turn to a more detailed comparison of textual marker complexity in the three texts. On the morphological level, the textual complexity of each morphological marker is inferred from its contribution to the morphological complexity in the original. Technically speaking, the difference between the average morphological complexity score of the marker-manipulated texts and the original text is taken. In this context, markers that increase the morphological complexity of the original are considered information-theoretically more complex than markers that decrease the morphological complexity in the original. In Mark, for instance, the text without third person singular *–s*, exhibits the lowest morphological complexity, i.e. it increases the complexity of the original text. The *ing*-marker, in comparison, increases the morphological complexity of the original to a much smaller degree. Therefore, third person singular *–s* is information-theoretically more complex on the morphological level than *–ing*. Table 4.7 gives an overview of the morphological markers ranked to their textual complexity on the morphological level. In Alice, the ranking of the markers is in increasing order of morphological complexity: *–ed*, genitive *'s*, plural *–s*, third person singular *–s* and *–ing*. Comparing the morphological complexity of the markers in Alice to their complexity in the other two texts, *–ing* appears to be most affected by variation. In Alice the text without *–ing* is the most morphologically complex, in the Euro-Congo news text it is positioned in the middlefield and in Mark it is the least complex text. Plural *–s* is similarly affected by intertextual variation, though to a lesser extent: it comes third in the ranking in Alice and Mark but first in the Euro-Congo corpus, and is

thus less complex in the news text than in the other two texts. Apart from these deviations, a "core complexity ranking" of the morphological markers which is identical across all three texts can be established: *–ed*, genitive *'s* and third person singular *–s* are in increasing order morphologically complex independent of the text.

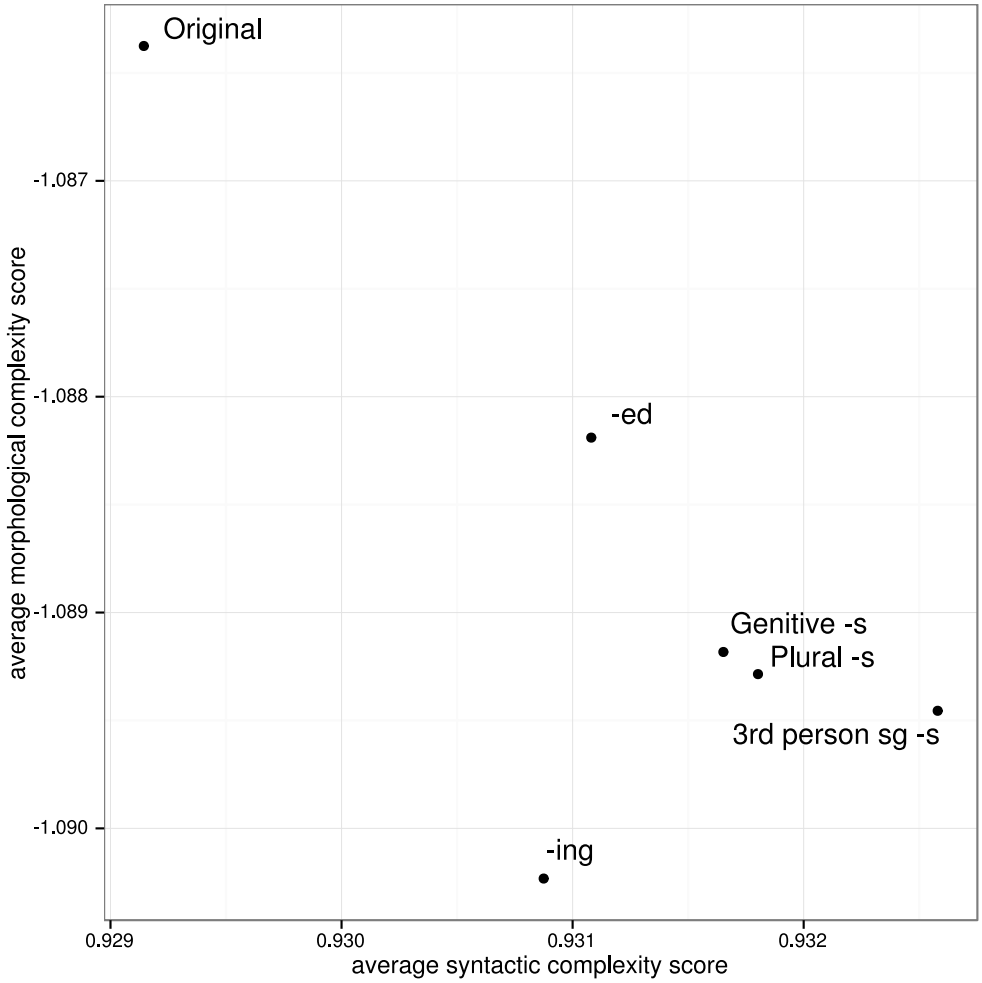

Figure 4.1.: Morphological by syntactic complexity of marker-manipulated texts and original Alice. Abscissa indexes increased syntactic complextiy, ordinate indexes increased morphological complexity.

The textual complexity of the morphological markers on the syntactic level is likewise inferred from their contribution to the complexity of the original. In syntactic terms, textual complexity is the amount of complexity a given morph reduces in the original. Take for example the *ed*-marker: the original Alice text (with *ed*) is syntactically less complex than the text without *ed*. Its presence therefore decreases the syntactic complexity of the original. Consequently, it follows that markers which decrease the syntactic complexity of the original are considered information-theoretically simple.

Table 4.7.: Morphological ranking of morphological markers in Alice, Mark and the Euro-Congo news corpus. Ranking is given in increasing order of morphological complexity.

Alice	**Mark**	**Euro-Congo**
–ed	*ing*	Plural *–s*
Genitive *–s*	*ed*	*–ed*
Plural *–s*	Plural *–s*	Genitive *'s*
3rd ps sg *'s*	Genitive *'s*	*–ing*
–ing	3rd ps sg *–s*	3rd ps sg *–s*

The ranking of the markers in Alice according to their textual complexity on the syntactic level is in increasing order of simplicity: *–ing*, *–ed*, genitive *'s*, plural *–s* and third person singular *–s*. Thus, the three *s*-marker reduce the syntactic complexity of the original more than *–ed* and *–ing*. The ranking in the Gospel of Mark comes closest to Alice such that the *s*-markers tend to be more syntactically simple than the other manipulated texts. In the Euro-Congo news corpus however, *–ed* and *–ing* decrease the syntactic complexity of the original to a greater degree than the *s*-markers.

Table 4.8.: Syntactic ranking of morphological markers in Alice, Mark and the Euro-Congo news corpus. Ranking is given in increasing order of syntactic simplicity.

Alice	**Mark**	**Euro-Congo**
–ing	*ing*	3rd ps sg *–s*
–ed	Plural *–s*	Genitive *'s*
Genitive *–'s*	*–ed*	Plural *–s*
Plural *s*	3rd ps sg *s*	*–ed*
3rd ps sg *–s*	Genitive *–'s*	*–ing*

Finally, Pearson's correlation coefficient is used for testing whether the complexity of the marker-manipulated texts is related to the token frequency of the morphological markers. On the morphological level, marker frequency and the complexity of the marker-manipulated texts does essentially not correlate. Consequently, more frequent markers do not contribute more complexity to the texts than infrequent markers. Pearson's correlation coefficient for marker frequency and complexity in Alice and Mark is very low (Alice $r = 0.14$, $p = 0.41$, Mark $r = 0.27$, $p = 0.33$). In the Euro-Congo news texts, the correlation is moderately high (Pearson's $r = 0.63$, $p = 0.13$). This means that in the newspaper genre, there is a slight, but not significant, trend for morphological complexity to increase with increasing marker frequency. The correlation between occurrence frequency of the

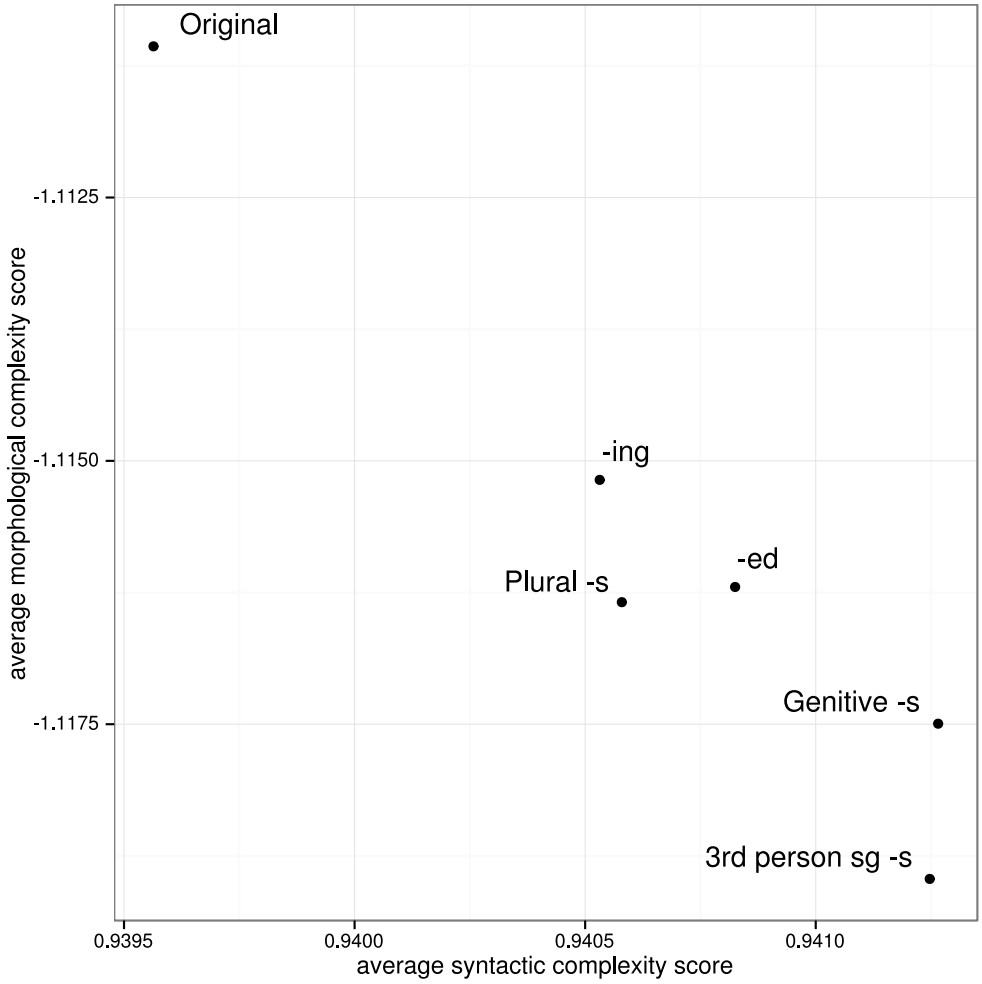

Figure 4.2.: Morphological by syntactic complexity of marker-manipulated texts and original Mark. Abscissa indexes increased syntactic complextiy, ordinate indexes increased morphological complexity.

markers and syntactic complexity of the marker-manipulated texts is negative (Pearson's correlation coefficient: Alice $r = -0.49$, $p = 0.8$, Mark $r = -0.61$, $p = 0.87$, Euro-Congo $r = -0.02$, $p = 0.51$).

In short, despite the low statistical similarity between the complexity scores of the marker-manipulated texts, the general complexity trend of the morphological markers is very similar. This is another way of saying that the exact measurements vary depending on the original text but that, on the whole, the textual complexity of the markers relative to the complexity of the original text is similar. Apart from one exception, there is no to little positive correlation between the complexity of marker-manipulated texts and their occurrence frequency.

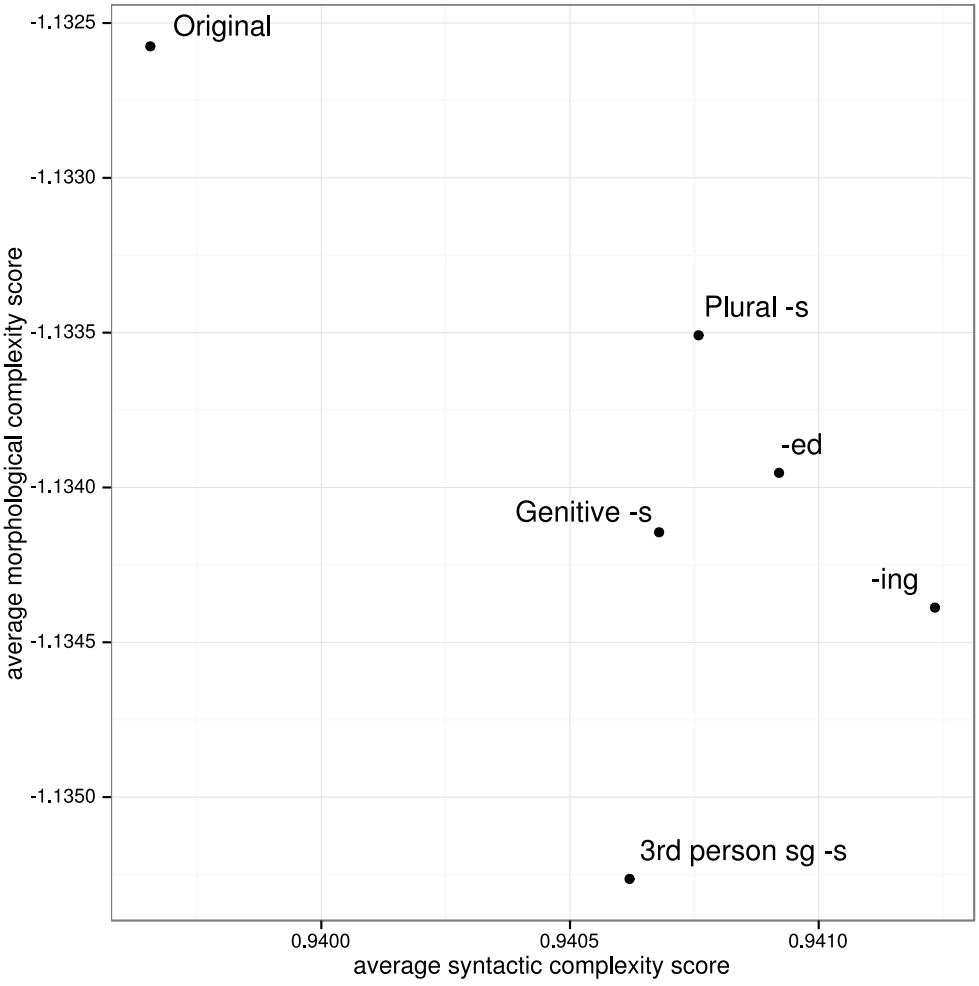

Figure 4.3.: Morphological by syntactic complexity of marker-manipulated texts and original Euro-Congo news corpus. Abscissa indexes increased syntactic complextiy, ordinate indexes increased morphological complexity.

4.3. Analysing constructions

This section analyses the five functional constructions progressive aspect *be* + verb–*ing* perfect aspect *have* + verb past participle passive voice *be* + verb past participle, and the future markers *will* and *going to* and their effect on the morphological and syntactic complexity in Alice, Mark and the Euro-Congo news corpus.

The effect of constructions on the complexity of the texts is statistically compared, and the degree of variation measured by pairwise calculation of Spearman's rank correlation rho for the average syntactic and morphological complexity scores of the construction-manipulated texts in Alice, Mark and the Euro-Congo newspaper corpus. On the syntactic level, construction-

manipulated texts correlate moderately to very highly (Table 4.9). The correlation of syntactic complexity scores between Alice and Mark is very high so that the effect of the constructions on the syntactic complexity in these two texts is virtually identical. The morphological complexity scores generally correlate highly, with the exception of the only moderate correlation between Alice and the Euro-Congo corpus (Table 4.10).

Table 4.9.: Correlation of syntactic complexity scores for construction-manipulated texts across the three text.

	Syntactic correlation	
	Mark	Euro-Congo
Alice	$rho = 0.9$ $p = 0.042$	$rho = 0.7$ $p = 0.12$
Mark		$rho = 0.6$ $p = 0.18$

Figures 4.4–4.6 plot the construction-manipulated text and the originals by syntactic and morphological complexity. The original texts, positioned in the top right quadrant of the plots, are (almost) the morphologically most complex and syntactically most simple texts. The texts without perfect, passive and progressive constructions exhibit substantially less morphological and syntactic complexity than the originals. Therefore, the presence of these constructions increases both the morphological and syntactic complexity in the texts. This is expected since all functional constructions—future markers excluded—affect both word order regularities and word forms. For instance, the perfect construction consists of two discrete components which are morphologically marked, namely a form of the auxiliary verb *have* and a verb marked as past participle. The syntactic sequence 'auxiliary *have* + past participle form', simply put, signals a passive construction. Thus, the construction *have laugh-ed* cuts across syntax and morphology. The texts without the future markers orbit around the originals across all three

Table 4.10.: Correlation of morphological complexity scores for construction-manipulated texts across the three texts.

	Morphological correlation	
	Mark	Euro-Congo
Alice	$rho = 0.8$ $p = 0.67$	$rho = 0.5$ $p = 0.23$
Mark		$rho = 0.7$ $p = 0.012$

texts. The complexity of the texts without *going to* is almost identical to the complexity of Alice, Mark and the Euro-Congo news texts and thus its presence or absence hardly affects the complexity of the originals. The future marker *will*, on the other hand, is the only construction which slightly decreases the morphological complexity in the originals but hardly affects their syntactic complexity at all. Therefore, I conclude that future markers in general hardly affect the complexity of the texts as morphological and syntactic structures remain largely intact even without the markers.

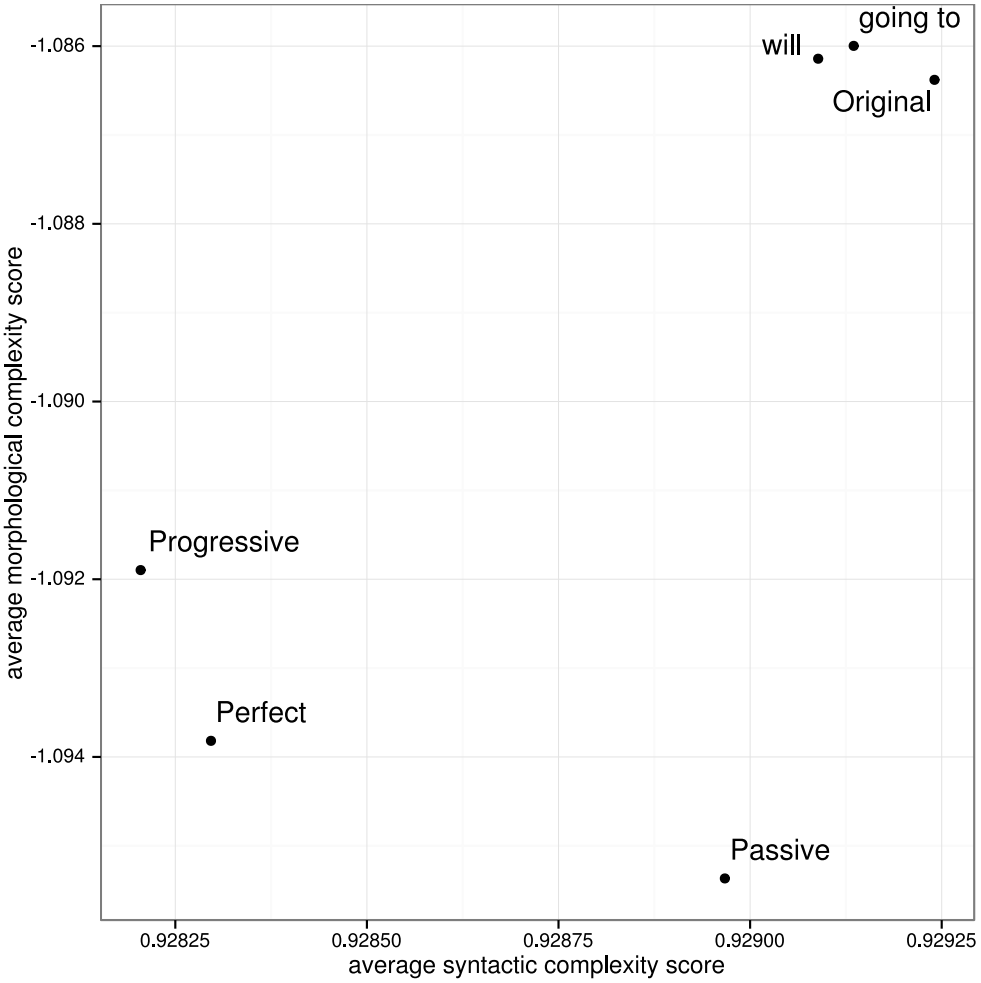

Figure 4.4.: Morphological by syntactic complexity of construction-manipulated texts and original Alice. Abscissa indexes increased syntactic complextiy, ordinate indexes increased morphological complexity.

In order to compare the effect of the constructions on the complexity in the three different texts more closely, their textual complexity on the morphological and syntactic level is inferred by taking the difference in morphological / syntactic complexity of each construction-manipulated text and

their original. Since all constructions increase the morphological and syntactic complexity in the original, textual complexity indicates the amount of morphological / syntactic complexity a given construction adds to the original text. Constructions that contribute more morphological and syntactic complexity, respectively, are considered comparatively more information-theoretically complex. In Table 4.11 and Table 4.12 the ranking of the constructions according to their textual complexity is given. In regard to morphological complexity, the ranking of the constructions progressive, perfect and passive are very similar in Alice and Mark. Progressive seems to be less complex than the other two constructions. This dovetails with intuitions as the *ing*-participle is more regular than past participles, which come in different forms (e.g. *sung*, *tak-en*, lingtokencall-ed). Irregularity, then, is known to increase complexity (see e.g. McWhorter (2001b), McWhorter (2012)). Only in the Euro-Congo news text is progressive more complex than the other two constructions. The two future markers exhibit the least textual complexity but their order varies such that in Alice *going to* is less complex than *will* while their order is inversed in Mark and the Euro-Congo corpus. The differences in the ranking of the constructions according to their textual complexity on the syntactic level is less pronounced such that only the order of the passive, perfect and progressive constructions varies. In Alice and Mark, the progressive construction exhibits the highest textual complexity while the most syntactically complex construction in the Euro-Congo news text is perfect.

Table 4.11.: Morphological ranking of constructions in Alice, Mark and the Euro-Congo news corpus. Ranking is given in increasing order of morphological complexity.

Alice	**Mark**	**Euro-Congo**
going to	*will*	*will*
will	*going to*	*going to*
Progressive	Progressive	Passive
Perfect	Passive	Perfect
Passive	Perfect	Progressive

Pearson's correlation coefficient reveals that the effect of constructions on text complexity is not sensitive to frequency effects. In all three texts the correlation between frequency of occurrence and the contribution of the constructions to text complexity is negative both for syntactic (Alice $r = -0.84$, $p = 0.96$, Mark $r = -0.59$, $p = -0.85$, Euro-Congo $r = -0.74$, $p = 0.92$) and morphological complexity (Alice $r = -0.59$, $p = 0.85$, Mark $r = -0.4$, $p = 0.75$, Euro-Congo $r = -0.46$, $p = 0.78$). In other words, the syntactic and morphological complexity of the texts analysed in this section does not depend on, or is influenced by, the construction's frequency in the

Table 4.12.: Syntactic ranking of constructions in Alice, Mark and the Euro-Congo news corpus. Ranking is given in increasing order of syntactic complexity.

Alice	**Mark**	**Euro-Congo**
going to	*going to*	*going to*
will	*will*	*will*
Passive	Perfect	Progressive
Perfect	Passive	Passive
Progressive	Progressive	Perfect

text.

In a nutshell, the overall contribution of constructions to text complexity does not vary substantially across genre and general tendencies regarding morphological and syntactic complexity are largely congruent. Progressive, passive and perfect constructions increase morphological and syntactic complexity in all three texts. The two future markers affect complexity to a much lesser extent than the other constructions. This suggests that analytical invariant markers are less complex than inflectional markers due to their regularity and transparency (Szmrecsanyi & Kortmann 2012; Nichols 2009; Trudgill 2004). Lastly, the complexity contributions of the constructions is not influenced by their frequency in the texts.

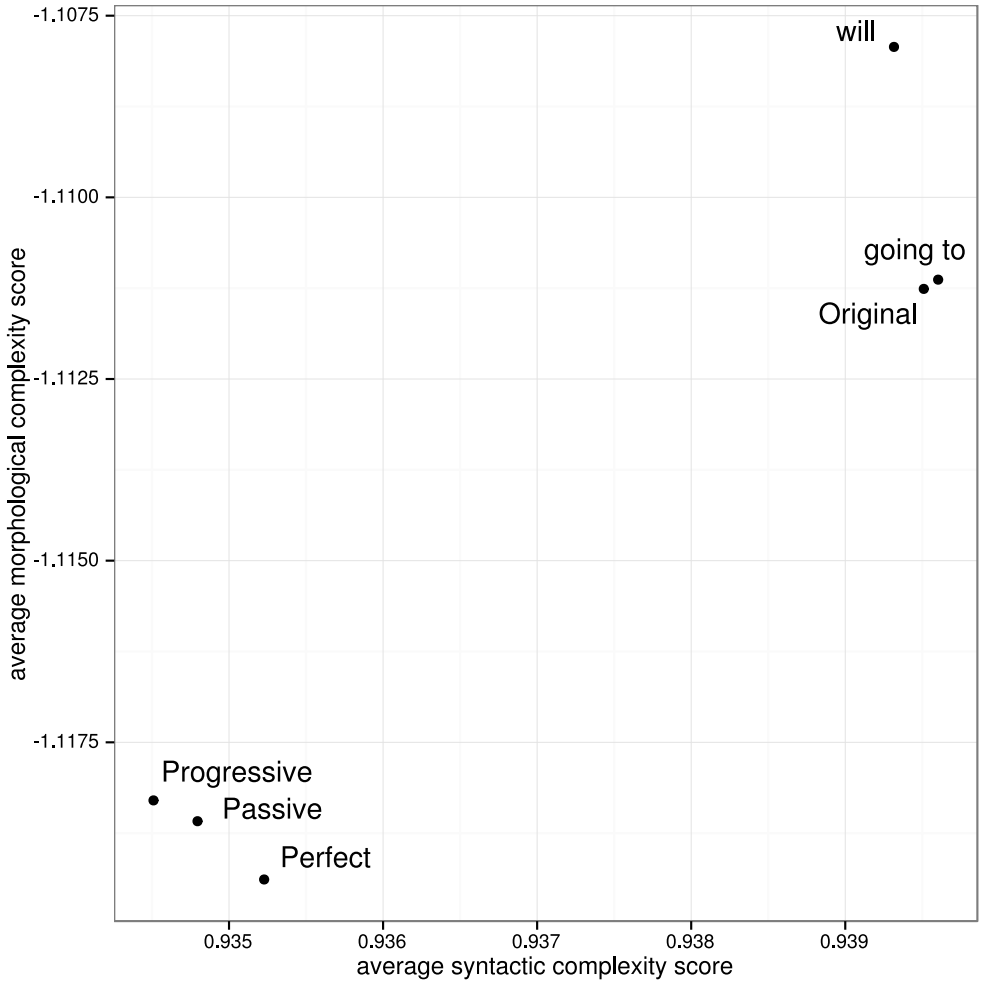

Figure 4.5.: Morphological by syntactic complexity of construction-manipulated texts and original Mark. Abscissa indexes increased syntactic complextiy, ordinate indexes increased morphological complexity.

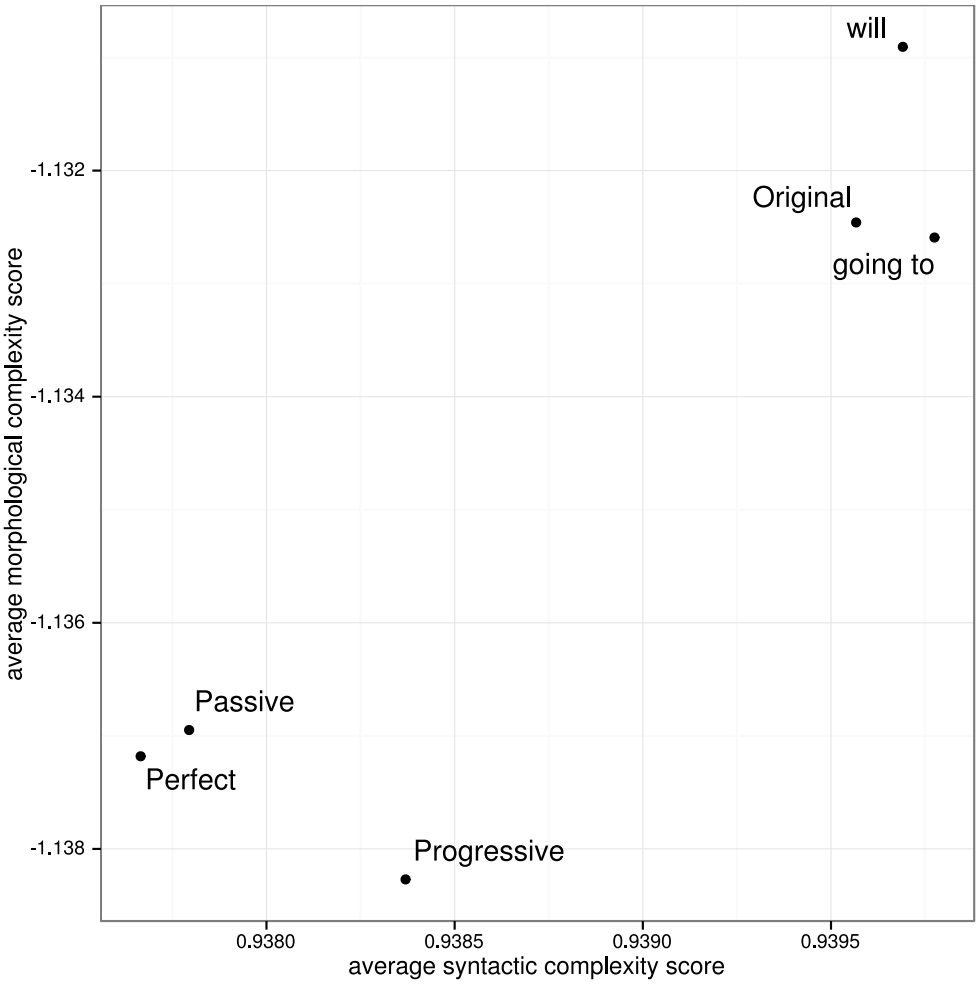

Figure 4.6.: Morphological by syntactic complexity of construction-manipulated texts and original Euro-Congo news corpus. Abscissa indexes increased syntactic complextiy, ordinate indexes increased morphological complexity.

4.4. Summary

In this chapter, targeted file manipulation, a method which permits the detailed analysis of morphological and syntactic complexity in texts, was introduced. Extending previous work in this field (Ehret 2014), the contribution of specific morphological markers and constructions to the morphological and syntactic complexity in three different texts was analysed with a special focus on intertextual variation. On an interpretational level, the textual complexity of these features was derived from their complexity contribution to the text.

Although there is intertextual variation regarding the exact amount of complexity a given feature contributes to a text, general trends can be reliably assessed and hold across different texts. In plain English, the textual complexity of the features assessed is, relative to the complexity of the original text, very similar. Generally, the presence of more morphological marker types increases the morphological complexity of the texts. On the other hand, the presence of more marker types facilitates the algorithmic predicition of syntactic patterns. All the constructions analysed—with the exception of the future markers *going to* and *will*—increase morphological but decrease syntactic complexity in literary, religious and newspaper writing. This indicates that higher amounts of morphological markers / inflections generate higher amounts of morphological complexity. Invariant grammatical markers, on the other hand, increase simplicity.

These findings are of threefold importance:

(i) The measurements provide evidence for the effectiveness of targeted file manipulation and bolster the somewhat unorthodox approach of the compression technique because they correspond to previous measurements and metrics of complexity (Arends 2001; McWhorter 2001a, 2012; Kusters 2008; Szmrecsanyi & Kortmann 2009; Trudgill 2004).

(ii) I demonstrate that the complexity measured with targeted file manipulation is largely text-independent. This is of relevance for and validates the results reported in Ehret (2014) who measures information-theoretic complexity of morphological markers and constructions in a mixed-genre corpus.

(iii) The findings throw light upon algorithmically measured complexity because they establish that the algorithm is sensitive to and capable of capturing the (ir)regularity of morphosyntactic structures. The following chapter will pursue this topic in more depth.

To conclude, targeted manipulation can serve as a powerful diagnostic for identifying general complexity trends of specific linguistic structures in written texts.

5. Exploring compression algorithms

5.1. Comparing and interpreting algorithmic complexity

This section describes and interprets algorithmically recognised strings on the basis of `gzip` 's lexicon output and aims at defining information-theoretic complexity in linguistic terms. To this end, the lexicon output of *Alice's Adventures in Wonderland* will be subjected to an in-depth analysis. Moreover, the impact of syntactic and morphological distortion on compressed strings will be discussed by comparing the original Alice lexicon output to the lexica of a syntactically and a morphologically distorted version of Alice.

5.1.1. Interpreting compressed strings

The `gzip` lexicon output is obtained by configuring the `gzip` source code and constructing a version of the algorithm which is usually used for debugging, i.e. finding errors in the code of the algorithm. Code listing 5.1 provides the commands used to build the debug version—henceforth referred to as `dgzip` .[1]

```
apt-get source gzip
cd gzip-*
./configure
make +=-DDEBUG
```

Code listing 5.1: How to build the `gzip` debug version.

In the next step, the relevant text is compressed by invoking the original `gzip` algorithm. To retrieve the lexicon of compressed strings, the compressed text is then piped to the configured `dgzip` with a call for verbose decompression (see Code listing 5.2).

[1] All commands listed in this chapter were implemented on Debian GNU/Linux, Version 7. URL `http://www.debian.org`

```
inputtext > gzip -f | dgzip -d -v -v -f
```

Code listing 5.2: How to retrieve the lexicon output for a given input text.

For ease of use, `dgzip` is saved as a separate program and incorporated in a shell script (Code listing 5.3) which takes a given input text file, removes all punctuation (using the shell script listed in Appendix D) and subsequently retrieves the lexicon output as described above.

```
if test $# -ne 1
then
        echo >&2 "Syntax: $0 <filename>"
        exit 1
fi

cat "$1" |
rmpunc |
dgzip -f 2>/dev/null |
dgzip -d -f -v -v 2>&1 1>/dev/null |
sed '1,/^$/d' |
head -n 1 |
sed 's/\\/\n\\/g'
```

Code listing 5.3: Shell script to generate a line-by-line lexicon output.

The shell script returns a line-by-line output consisting of back-referenced sequences, their length and distance to the preceding identical sequence as well as literal (text) sequences. The minimum length for referenced sequences is three characters including spaces, so that the lexicon does not contain any back-referenced sequences shorter than 3 symbols (Salomon 2007: 230–240). Code listing 5.4 provides the first twenty lines of the lexicon output from the original Alice text. The first line of the output contains no compressed sequences as the algorithm has yet to encounter strings on whose basis the text of the first line can be compressed. The subsequent entries start with a backslash and the length-distance pair in square brackets which is immediately followed by the compressed sequence of the specified length. For example, the second entry *\[29,4]ing by her* is to be interpreted as follows: the first integer enclosed in square brackets indicates the distance in characters (including spaces) to the previously encountered identical string (in the search buffer) on whose basis the referenced string (in the look-ahead buffer) is compressed. The second integer in square brackets indicates the length of the compressed string. The referenced compressed string in

line two was first encountered 29 characters (including spaces) before, and counts 4 characters (including spaces). Thus, the compressed sequence in this example is *ing␣*. Note that spaces are part of compressed sequences.[2]

```
alice was beginning to get very tired of sitt
\[29,4]ing by her
\[15,3] sist
\[7,3]er on the bank an
\[41,5]d of hav
\[40,4]ing noth
\[77,7]ing to do
\[40,3] on
\[102,3]ce or tw
\[111,4]ice s
\[51,3]he had peep
\[94,3]ed in
\[37,3]to
\[71,5]the book
\[94,12] her sister
\[151,4]was read
\[120,5]ing but it
\[55,5] had no pictures
\[84,4] or con
\[171,3]versations
```

Code listing 5.4: First twenty lines of the lexicon output of the original *Alice's Adventures in Wonderland.*

The lexicon of compressed strings of the original Alice text contains a total of 16,991 entries including the first line, yet only 11,683 unique strings. In the parlance of linguistics, one could thus say that the lexicon counts 16,991 tokens but only 11,683 types.[3] Table 5.1 lists the number of unique strings, and their frequency of compression, i.e. how often a given string was recognised and compressed. In the original Alice lexicon compression frequencies range from fourteen to one. The strings can be grouped according to their compression frequency: high-frequency strings, frequent strings, rare strings and very rare strings. Among the fifteen high-frequency strings are sequences such as *very␣* (compression frequency = 16) and *ing␣* (com-

[2]In this chapter and throughout this work, spaces at the beginning and end of compressed strings are represented by an open box '␣'. Spaces within compressed strings are represented by themselves.

[3]Note that unless otherwise stated the terms '(raw) frequency' and 'string frequency' always refer to the token frequency of strings. 'Compression frequency' on the other hand is the number of times a given type of string was compressed.

pression frequency = 14). The group of frequent strings consists of 171 unique entries and contains, for instance, the sequences *was␣* (compression frequency = 12) and *uch␣* (as in *s-uch*) (compression frequency = 6). In contrast to the former two categories, strings with a compression frequency smaller or equal to five, i.e. rare strings, count roughly twice as many unique entries as the category 'frequent'. Examples of rare strings are *king␣* (compression frequency = 5) and *uddenly␣* (compression frequency = 4). Strings with a compression frequency smaller or equal to two constitute the largest group in the lexicon with a total of 10,586 strings. Example sequences of the category 'very rare' are *␣herself␣* (compression frequency = 2) and *down down down␣* (compression frequency = 1). In other words, over 90 percent of all compressed strings fall into the category 'very rare' and occur only once or twice while only about two percent of all strings fall into the categories highly frequent and frequent. This means that the number of strings decreases with increasing compression frequency. In fact, the distribution of strings strongly resembles a *Zipfian distribution* (see Figure 5.1 for visualisation). This is interesting but not unexpected as word frequencies in natural human languages are known to follow *Zipf's law* and the compressed strings were created from a natural English text. Zipf's law states that the frequency of a word decreases exponentially to its frequency rank. Thus, the probability of word occurrences in a language sets out high but gradually decreases. This is another way of saying that, in human languages, only a small number of words occur very frequently, while the majority of words occur rarely (Zipf 1935, 1949; Cancho & Solé 2003). This is also true for the frequency distribution of compressed strings.

The length of compressed sequences ranges from 3 (the minimum length required by the algorithm for back-referencing strings) characters to 148 characters (including spaces) (Table 5.2). Yet, 85 percent of all strings have a length of three to ten characters, 13 percent of the strings count between eleven and eighteen characters and only a very small percentage of strings exceeds this count. In fact, strings longer than 37 characters occur only once or twice. Thus, a significant trade-off between the length and number of compressed strings can be observed (Pearson's correlation coefficient $r = -0.47$, $p = 0.0005$). This is another way of saying that the number of compressed strings decreases with increasing string length so that, for example, for the highest character count (length = 148) only one string exists (example (1)).

(1) ome and join the dance will you won't you will you won't you will you join the dance will you won't you will you won't you won't you join the dance␣

Table 5.1.: Distribution of unique strings in the original Alice lexicon according to frequency of compression. The first column categorises the strings according to their frequency.

Category	Unique strings	Compression frequency
highly frequent	4	16
	4	15
	5	14
	2	13
frequent	10	12
	21	11
	18	10
	9	9
	23	8
	34	7
	56	6
rare	127	5
	248	4
	536	3
very rare	1,580	2
	9,006	1

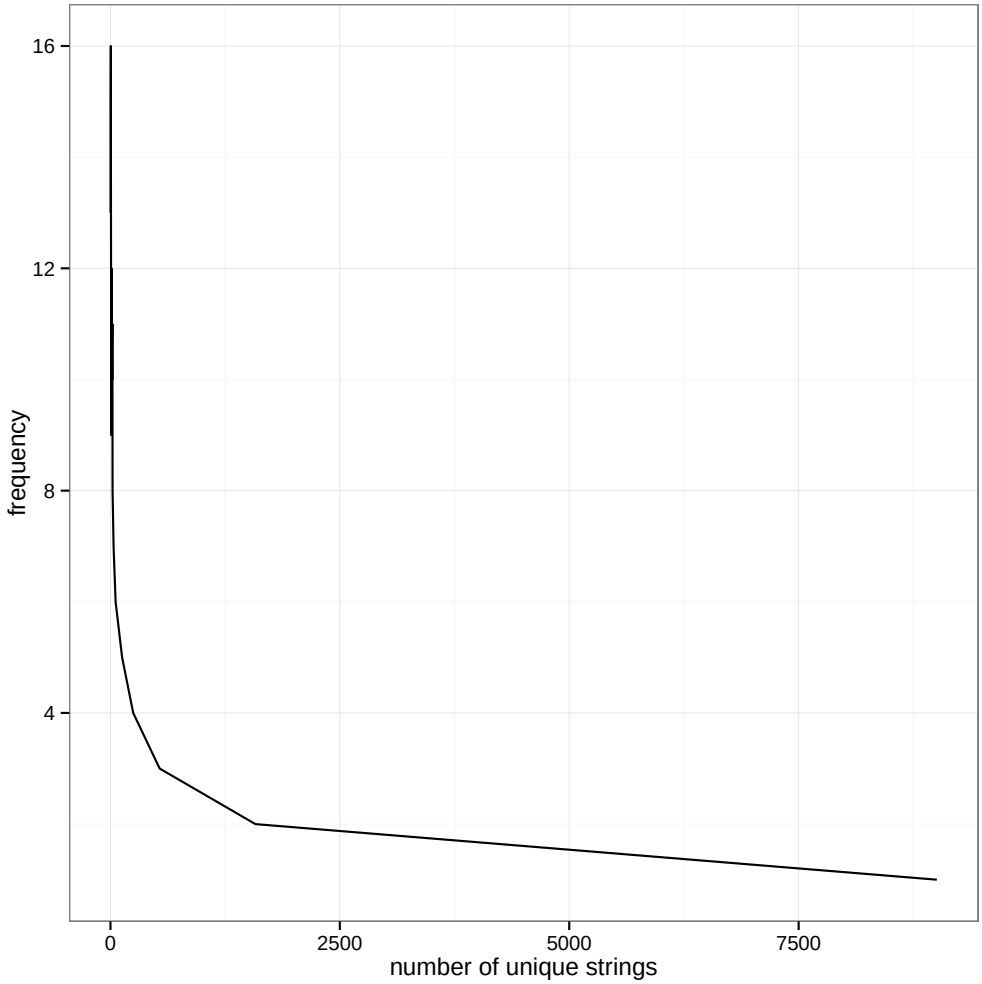

Figure 5.1.: Distribution of unique strings in the lexicon of the original Alice text. Abscissa indexes number of unique strings, ordinate indexes increased compression frequency.

Table 5.2.: Length of compressed strings by number of compressed strings in the original Alice lexicon. Length is given in characters including white space.

Length	Strings
3	826
4	1,961
5	2,700
6	2,465
7	2,228
8	1,940
9	1,408
10	985
11	582
12	467
13	316
14	279
15	175
16	147
17	112
18	100
19	57
20	59
21	28
22	44
23	20
24	14
25	12
26	9
27	16
28	4
29	4
30	2
31	3
32	2
33	5
34	3
37	2
39	2
40	1
41	1
46	1
47	1
50	2
51	1
53	1
54	2
55	1
57	1
85	1
148	1

In order to determine the nature of compressed strings and thus gain a better understanding of information-theoretic complexity, every string in the original Alice lexicon output was manually analysed and, subsequently, the strings were manually annotated for linguistic, and non-linguistic, category. For this purpose a six-fold coding scheme was developed consulting *The Longman Grammar of Spoken and Written English* (Biber et al. 1999) on word class categories. The coding scheme gives a detailed (linguistic) description of compressed sequences while at the same time allowing for subsequent semi-automatic annotation of the distorted lexica (for more detail on distorted lexica refer to Section 5.1.2). Therefore, words of multiple word class membership were either subsumed under a macro-category or, in cases where this is not possible, assigned a default membership. For example, nouns and verbs were subsumed under the macro-category 'lexical'. Thus, the resulting coding scheme essentially distinguishes between two major word classes (lexical and functional), other linguistically meaningful units, mixed strings and random chunks:

(i) *Lexical.* Lexical words include nouns, verbs, adjectives and adverbs (Biber et al. 1999: 62–66) as well as *to*-infinitives (e.g. *to see*) and established phrasal verbs (e.g. *make out*). Examples from the original Alice lexicon are *hedgehogs*, *considering* and *dreadfully*. For practical reasons, auxiliary forms of the verbs *have*, *be* and *do* and the borderline cases *ought to*, *used to* and *have to* were included in this category.

(ii) *Functional.* This category comprises prepositions, determiners, pronouns, coordinators, subordinators, numerals, the negator *not*, adverbial particles and wh-words as well as modal verbs (Biber et al. 1999: 69–91). Inserts were also subsumed under this category—despite the fact that they constitute an independent, if somewhat ambiguous, class of words (for a discussion see Biber et al. (1999: 56–57))—because the greetings and response words ocurring in the lexicon (e.g. *yes*, *please*) are more or less a closed word class (Biber et al. 1999: 56). Furthermore, semi-determiners (e.g. *same*, *such*) as well as quantifiers (e.g. *every*) and subordinators (e.g. *yet*) with multiple word class membership were by default coded as 'functional'.

(iii) *Other.* This category includes word segments, endings and linguistically meaningful "chunks". Specifically, it contains noun suffixes such as *-ment* or *-ity* (for a complete list refer to Biber et al. (1999: 321)), genitive *'s*, verb endings such as *-ing* or *-ed*, adjectival / adverbial endings such as *-ly* or *-est*. Parts of contractions such as *'ll* or *'ve* and the endings *herSELF* and *forWARD(S)*, as well as *conclusION* and *greenISH* were counted to this category. Furthermore, any of the above forms plus one or more intact pattern from category (i) and (ii) were also coded as 'other' (e.g. *'s no use*␣).

(iv) *Phrasal.* Phrasal patterns are defined as multi-word strings. They include combinations and phrases of two or more intact words (e.g. *do cats eat bats*␣, *there was nothing*␣, *her sister*␣) as well as contractions (e.g. *that's*␣, *can't*␣). Note that 'phrases' as defined here are not identical with prosodic or grammatical phrases as the algorithm lacks knowledge of sentence boundaries or other syntactic units. Phrasal patterns may therefore be combinations of words that, in the original text, belong to different sentences / syntactic units and were formerly separated by punctuation marks such as *child said the*␣ and *gryphon we*, or cut-off word sequences such as *you ever*␣ or *the best*␣.

(v) *Mixed.* Mixed strings contain at least one intact pattern from categories (i), (ii) or (iii) which are mixed with random chunks such as, for instance, *the b* or *abbit was*␣.

(vi) *Random.* This category consists of random chunks and nonsensical phrases such as *cks*␣ or *ich w*.

On a binary scale, I can thus distinguish between linguistically meaningful strings (Example (2)), i.e. strings which belong to categories (i) to (iv), and random, linguistically not meaningful strings (Example (3)), i.e. strings from category (v) and (vi).

(2) a. their␣
b. looked anxiously␣
c. ␣opportunity␣
d. 'd better␣

(3) a. s to f
b. dance t
c. gree␣
d. omet

Most strings in the original Alice lexicon belong to the mixed category and twenty percent are random chunks (Table 5.3). The categories 'phrasal' and 'lexical' make up twenty and fifteen percent of the compressed strings, respectively, while function words and suffixes (category 'other') are the two smallest categories. In other words, roughly half of the compressed strings in the original Alice lexicon are linguistically meaningful units while the other half consists of more or less random strings. It is important to note that the number of strings per category does, of course, depend on the linguistic composition of the text which serves as input for the algorithm. A lexicon based on a different text can, for example, contain more strings of the category 'lexical' or less strings of the category 'phrasal' than the Alice lexicon. It goes without saying that the exact matches and compressed sequences are also strictly text-bound and determined by the input text. The Alice lexicon contains phrases like ␣*beautiful soup*␣ and lexical words such

as *uglification*␣ which will not occur in a lexicon created from the Gospel of Mark or a text on the crisis in Iraq. However, the lexicon analysis of Alice demonstrates that lexicon-based compression algorithms such as `gzip` do capture recurring linguistic structures. Needless to say, this does not mean that compression algorithms possess any kind of linguistic knowledge about the structures they encounter.

Table 5.3.: Raw frequency and percentage of compressed strings by linguistic category in the original Alice lexicon. Note that the percentages were rounded down to the nearest integer and therefore sum up to 99 percent in total instead of 100.

Linguistic category	**Raw frequency**	**Percentage**
Functional	913	5
Lexical	2,558	15
Mixed	6,170	36
Other	175	1
Phrasal	3,730	22
Random	3,445	20
Total	16,991	99

Having established that compression algorithms do indeed capture linguistic structures, the question remains why not every instance of a linguistic structure is compressed. Put differently, what motivates `gzip` to sometimes 'recognise' linguistic structures and sometimes not? For example, why is a given linguistic structure such as *ing* compressed only once if it occurs 965 times in total in the text? The answer is very simple and related to the workings of the compression algorithm. Lexicon-based compression algorithms of the Lempel-Ziv family—to which `gzip` belongs—achieve compression by back-referencing redundant strings with the length of the copied string and the distance to the previous, identical string in the search buffer (which serves as referent). In this process, the algorithm tries to find a matching string of maximum length (Ziv & Lempel 1977: 377; Salomon 2007), i.e. the algorithm is greedy. This is another way of saying that the algorithm—agnostic about form-meaning relationships—will choose longer sequences over shorter sequences no matter whether these sequences, from a linguistic point of view, are meaningful or not. Example (4) illustrates maximum length compression of compressed strings containing *ing*. The text in bold represents the text stored in the search buffer of the algorithm. The umarked rest represents the text in the look-ahead buffer, which is to be compressed on the basis of the search buffer content. The actual referent strings and their matches are enclosed in square brackets. In the process of compression, then, the algorithm is looking for the longest match to any sequence of letters and spaces stored in the search buffer. In example (4-a) the

search buffer is *alice was beginning to get very tired*. The longest possible match in the look-ahead buffer is the sequence *ing␣* counting four characters. In example (4-b) the first possible match is *[ing␣]* of length four. Yet, the next possible match *[ing to the␣]* counts eleven characters. Hence, the algorithm compresses the second match because it is the sequence of maximum length.

(4) a. **alice was beginn[ing␣]$_{\text{referent}}$ to get very tired** of sitt[ing␣]$_{\text{match}}$by her sister on the bank [...]
b. **and began bow[ing to the]$_{\text{referent}}$king the queen** [...]and then turn[ing to the]$_{\text{match}}$rose-tree

In the example above, both matched (compressed) strings are linguistically meaningful patterns and in both strings the sequence *ing* is a verbal suffix. However, the algorithm distinguishes neither between different functions / usages of a given pattern—for example gerund vs. present participle—nor does it distinguish between linguistically meaningful patterns and structurally identical, non-meaningful patterns such as in example (5) where the algorithm matches *beginnING* with *nothING*.

(5) **alice was beginn[ing to]$_{\text{referent}}$get very tired** [...]and of having noth[ing to]$_{\text{match}}$do

Furthermore, the greedy behaviour of the algorithm leads to the compression of linguistically nonsensical strings such as in example (6).

(6) **she found herself fa[lling]$_{\text{referent}}$down a very deep well** [...]to drop the jar for fear of ki[lling]$_{\text{match}}$somebody

In summary, `gzip` compresses any sequence of characters and white space that is a match of maximum length for a given sequence of characters and white space in the search buffer. In many cases, these sequences coincide with linguistically interpretable (surface) structures such as, for instance, suffixes, verbs, nouns or whole phrases, but the algorithm does not systematically select structures or possesses any knowledge of the structures it compresses.[4] This is reflected in the fact that the algorithm also compresses nonsensical strings or strings which, on the surface, resemble linguistically meaningful structures. In short, algorithmic compression is based on the form of structures, not on their function and meaning. Information-theoretic complexity must therefore be defined as a measure of structural surface redundancy.

This definition comes with an important implication: information-theoretic Kolmogorov-based complexity tends to favour morphological complexity

[4]The algorithm could theoretically be rewritten to be less agnostic and able to recognise linguistic units. This would presumably affect the objectivity of the method because the algorithm would be apriorily told which strings to compress and which strings to ignore.

because algorithmic compression as described here is based on structural redundancy. Structural redundancy, on the other hand, is closely linked to morphological complexity which, in this work, is defined as the complexity related to the structural (ir)regularity of word forms. As a consequence, compression algorithms are not ideal for measuring the overall complexity of languages. Thus, algorithms like `gzip` should be used with caution and users should bear in mind that algorithmically measured overall complexity is largely a function of morphological complexity.

5.1.2. Comparing compressed strings

After having described `gzip` 's lexicon output on the basis of *Alice's Adventures in Wonderland*, I will now describe the impact of morphological and syntactic distortion on the distribution, frequency and nature of compressed strings. For this purpose, `gzip` 's lexicon output of a morphologically distorted and a syntactically distorted version of Alice will be extracted and analysed.

Before discussing the distribution and frequency of compressed strings in the distorted lexica, let us briefly recapitulate how distortion is implemented: morphological distortion is achieved by random deletion of 10% of all orthographically transcribed characters in the text. This is assumed to increase the amount of word forms in the text and thus increase morphological complexity. Syntactic distortion is implemented as random deletion of 10% of all orthographically transcribed words. This leads to the disruption of word-order interdependencies and compromises syntax. The morphologically distorted lexicon is therefore expected to contain more unique strings than the original Alice lexicon while the syntactically distorted lexicon should contain less unique strings. Table 5.4 illustrates how distortion affects morphology and syntax in Alice by providing an example passage from a morphologically distorted Alice text and a syntactically distorted Alice text.

Table 5.4.: Distorted passages from *Alice's Adventures in Wonderland*. Morphologically distorted tokens are marked in bold. The zero symbol 'Ø ' indicates where a token was deleted through syntactic distortion.

Distortion	**Text**
Morphological	alice was **egining** to get very tired of sitting by her **ist** on the bank **an** of **havig** nothing to do once or **wice** she had **pped** into the book her sister was **rading** but it had **n picures** or **conversatons** in it and what is the **se** of a **boo** thought **ali** without pictures or conversation
Syntactic	alice was beginning to get very Ø of sitting by her sister on the bank and of having nothing to do once or twice she had peeped into Ø book her sister was Ø but it had no pictures or conversations Ø it and what is Ø use of a book thought alice without pictures or conversation

Let us first turn to the description of the morphologically distorted lexicon. Code listing 5.5 shows the first twenty lines of the morphologically distorted Alice lexicon illustrating how morphological distortion impacts on the structure of compressed strings in the lexicon. In line two, for example, the original string *her sister* was transformed to *her ist*. Thus, the subsequent occurrence of the original string is not compressed because the pattern is not contained in the search buffer. For this reason, the algorithm—unlike in the original lexicon where the complete sequence ␣*her sister*␣ (length = 12) is back-referenced—compresses three shorter sequences ␣*her*␣ (length = 5), *ist* (length = 3) and *er*␣ (length = 3). In general terms, three conclusions can be drawn from this example. Firstly, the morphologically distorted lexicon contains more unique strings. Secondly, if it contains more unique strings compression frequencies are lower (because unique strings are compressed only once), and thirdly, it contains more short strings than the original lexicon. On a linguistic level, the morphologically distorted lexicon should also contain more random strings than the original lexicon.

```
alice was egining to get very tired of sitt
\[29,4]ing by her ist on the bank an
\[37,4] of havig noth
\[72,7]ing to do
\[38,3] on
```

```
\[95,3]ce or w
\[103,4]ice s
\[48,3]he had pp
\[86,3]ed in
\[34,3]to
\[66,5]the book
\[86,5] her s
\[87,3]ist
\[7,3]er
\[141,4]was rad
\[111,5]ing but it
\[52,5] had n picures
\[78,4] or con
\[160,3]versatons
\[73,3] in
```

Code listing 5.5: First twenty lines of the lexicon output of a morphologically distorted version of Alice.

The statistics of the morphologically distorted lexicon back these observations. The morphologically distorted lexicon consists of 18,452 entries and 13,346 unique strings, i.e. it does indeed contain more unique strings than the original Alice lexicon with 11,683 unique strings and a total of 16,991 entries. This confirms the assumption that morphological distortion leads to the creation of new word forms such as *picures* or *aisy-cha* and, ultimately, more random noise which is difficult to compress. In terms of Kolmogorov complexity this means that the lexicon of the morphologically distorted text is longer than the lexicon of the original text.

An overview of the compression frequency of unique strings (Table 5.5) show that the range of compression frequencies in the morphologically distorted lexicon is similar to the original Alice lexicon—ranging from fifteen to one—and the number of strings gradually declines with increasing compression frequency, i.e. string distribution in the morphologically distorted lexicon generally follows Zipf's law. However, the distorted lexicon contains less high-frequency and frequent strings than the original lexicon despite the fact that it contains overall more strings. Most notably, it contains only one high-frequency string (*very␣*, compression frequency = 15) while the original Alice lexicon counts fifteen high-frequency strings. The category 'frequent' counts 150 strings as opposed to 171 in the original lexicon, while both lexica count roughly the same number of rare strings (morphologically distorted = 913, original = 911). As a consequence, the category 'very rare' counts with 12,281 strings, considerably more strings than in the original Alice lexicon (very rare = 10, 586). In other words, morphological distortion which consumes recurrent patterns leads to lower compression frequencies

Table 5.5.: Distribution of unique strings in the morphologically distorted Alice lexicon according to frequency of compression. The first column categorises the strings according to their frequency.

Category	Unique strings	Compression frequency
highly frequent	1	15
	2	12
	10	11
	12	10
frequent	15	9
	17	8
	35	7
	59	6
	101	5
rare	228	4
	584	3
very rare	1,862	2
	10,419	1

and an increased number of unique strings.

String lengths in the morphologically distorted lexicon range from 3 to 24 characters including spaces (Table 5.6). In comparison to the original Alice lexicon, where the maximum string length counts 148 characters, the upper limit is with a length of 24 characters much lower. Overall, 95 percent of all strings have a length between three and ten characters. Thus, the morphologically distorted lexicon contains 10 percent more short sequences than the original Alice lexicon. This is statistically reflected in a very high negative correlation between string length and string frequency: Pearson's correlation coefficient $r = -0.78$ is very highly significant ($p = 0.0000081$).

Table 5.6.: Length of compressed strings by number of compressed strings in the morphologically distorted Alice lexicon. Length is given in characters including white space.

Length	Strings
3	1,260
4	3,386
5	3,869
6	3,250
7	2,460
8	1,638
9	982
10	595
11	333
12	238
13	152
14	94
15	48
16	46
17	32
18	25
19	12
20	15
21	6
22	3
23	3
24	4

As mentioned in the beginning of this chapter, the syntactically distorted lexicon should be shorter, i.e. it should contain less entries and unique strings, than the original lexicon due to the deletion of 10% of all word tokens in the text. Furthermore, word-order interdependencies should be compromised through syntactic distortion and as a result uncompressible noise should be created. To exemplify syntactic distortion and its effect on the lexicon, the first twenty lines of the syntactically distorted Alice lexicon output are provided in Code listing 5.6. In the first line the word *tired* was deleted from the text. As a consequence, the pattern *ed* in *peepED* is not in the buffer and is thus not back-referenced. In contrast to the morphologically distorted lexicon, where the algorithm compressed several shorter sequences instead of the original pattern, the algorithm omits the sequence altogether. Thus, overall less lexicon entries are created and the syntactically distorted lexicon consists of only 15,827 entries and 11,041 unique strings.

```
alice was beginning to get very of sitt
\[23,4]ing by her
\[15,3] sist
\[7,3]er on the bank and
\[41,4] of hav
\[40,4]ing noth
\[71,7]ing to do
\[40,3] on
\[96,3]ce or tw
\[105,4]ice s
\[51,3]he had peeped in
\[37,3]to book
\[90,12] her sister
\[141,5]was but it
\[43,5] had no pictures
\[72,4] or con
\[153,3]versations
\[36,4] it
\[125,4]and wha
\[48,3]t is use
```

Code listing 5.6: First twenty lines of the lexicon output of a syntactically distorted version of Alice.

The compression frequency ranging from eighteen to one as well as the distribution of unique strings in the syntactically distorted lexicon (Table 5.7) is very similar to the distribution and frequency of strings in the original Alice lexicon and also follows Zipf's law. The syntactically distorted lexicon counts twelve high-frequency strings (original = 15) and 173 frequent strings (original = 173). As the syntactically distorted lexicon contains overall less strings, the number of strings in the categories 'rare' and 'very rare' is with 784 and 10,071 strings, respectively, lower than in the original lexicon (rare = 911, very rare = 10,586).

String lengths in the syntactically distorted lexicon range from 3 to 47 characters including spaces (Table 5.8). Although the maximum string length is larger than in the morphologically distorted lexicon, it is still considerably lower than in the original Alice lexicon (maximum length = 148). However, 87 percent of the strings in the syntactically distorted lexicon count between three and ten characters, 12 percent count between eleven and eighteen characters and only about 1 percent is equal to or longer than 19 characters. In other words, the distribution of strings according to their length is virtually identical to the distribution of strings in the original lexicon. Furthermore, I observe the typical negative trade-off between string

Table 5.7.: Distribution of unique strings in the syntactically distorted Alice lexicon according to frequency of compression. The first column categorises the strings according to their frequencies.

Category	Unique strings	Compression frequency
highly frequent	1	20
	1	18
	1	16
	2	15
	3	14
	4	13
rare	9	12
	10	11
	15	10
	15	9
	27	8
	42	7
	55	6
rare	108	5
	199	4
	477	3
very rare	1,466	2
	8,605	1

length and frequency (Pearson's r $= -0.7$, $p = 0.0000032$).

All in all, the syntactically distorted lexicon is almost identical to the original Alice lexicon in regard to compression frequency and distribution of compressed strings. The only major difference seems to lie in the total number of lexicon entries and unique strings which is, as expected, lower in the syntactically distorted lexicon. Thus, while syntactic distortion does lead to the disruption of word-order interdependencies (see Table 5.4 above), the creation of random noise in the text is not as overtly reflected in the syntactically distorted lexicon as it is in the morphologically distorted lexicon.

Table 5.8.: Length of compressed strings by number of compressed strings in the syntactically distorted Alice lexicon. Length is given in characters including white space.

Length	Strings
3	776
4	1,902
5	2,626
6	2,382
7	2,159
8	1,797
9	1,286
10	876
11	539
12	399
13	282
14	221
15	132
16	112
17	87
18	59
19	43
20	42
21	34
22	24
23	10
24	8
25	4
27	11
28	2
29	3
31	1
33	1
34	2
35	1
39	2
40	1
47	1

Let us now focus on the linguistic interpretation of compressed strings in the morphologically and syntactically distorted lexica. On the basis of the coding scheme for linguistic categories described in Section 5.1.1, all compressed strings in the distorted lexica were semi-automatically annotated using the programming language and statistics package `R` .[5] Specifically, the annotated lexicon of the original Alice text was used as reference dictionary for the coding of identical sequences occurring in the distorted lexica. Sequences which did not occur in the original lexicon were manually annotated according to this categorisation scheme. Suspect sequences which could potentially belong to more than one category were manually disambiguated. Sequences such as *ing* or *thin*, for instance, can either be linguistically meaningful units (e.g. *havING*, *thin*) or nonsensical strings such as in *nothING* or *THINk*. In the case of the syntactically distorted lexicon, the category 'phrasal' was subjected to manual verification in order to eliminate syntactically distorted phrases from which words were deleted in the process of distortion. These "junk" phrases were subsumed under the category 'mixed' because their components (words) are intact and linguistically meaningful despite the fact that syntax is corrupted. Example (7) gives two such corrupted phrases (in bold) along with their context in the original text. The zero symbol 'Ø ' marks where a word was deleted in the process of syntactic distortion.

(7) a. Then she looked at **the Ø of** the well and noticed that they were filled with cupboards and book-shelves.

b. [. . .]yet you finished the goose with the bones and the beak, pray **how Ø you** manage to do it?

Figure 5.2 displays the percentage of compressed strings for each of the six linguistic categories in the distorted lexica and, for comparison, in the original Alice lexicon. The raw frequencies of strings per linguistic category are provided in Table 5.9. The proportion of random strings in the morphologically distorted lexicon is roughly twice as much as in the original and the syntactically distorted lexica. The morphologically distorted lexicon contains about half the amount of phrasal strings, and also less lexical and fucntional strings than the other two lexica. The percentage of strings per category in the syntactically distorted lexicon is virtually identical to the percentage of strings in the original lexicon.

The effect of morphological distortion on the nature of compressed strings is obvious and as expected: through morphological distortion, new "word forms" such as *lizad* and *shoed* are created which can no longer be considered linguistically meaningful. Consequently, the lexicon contains more random strings than the original lexicon. Interestingly, the percentage of

[5] R 2.15.1. R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

mixed strings remains stable and is about the same size as in the other two lexica. This is probably due to the fact that previously linguistically meaningful strings are either distorted to mixed strings or random strings, while strings that are linguistically non-meaningful to begin with will remain in the former two categories. Thus, overall, we observe a shift from linguistically meaningful strings to mixed strings to random strings, resulting in an increased number of random strings in total.

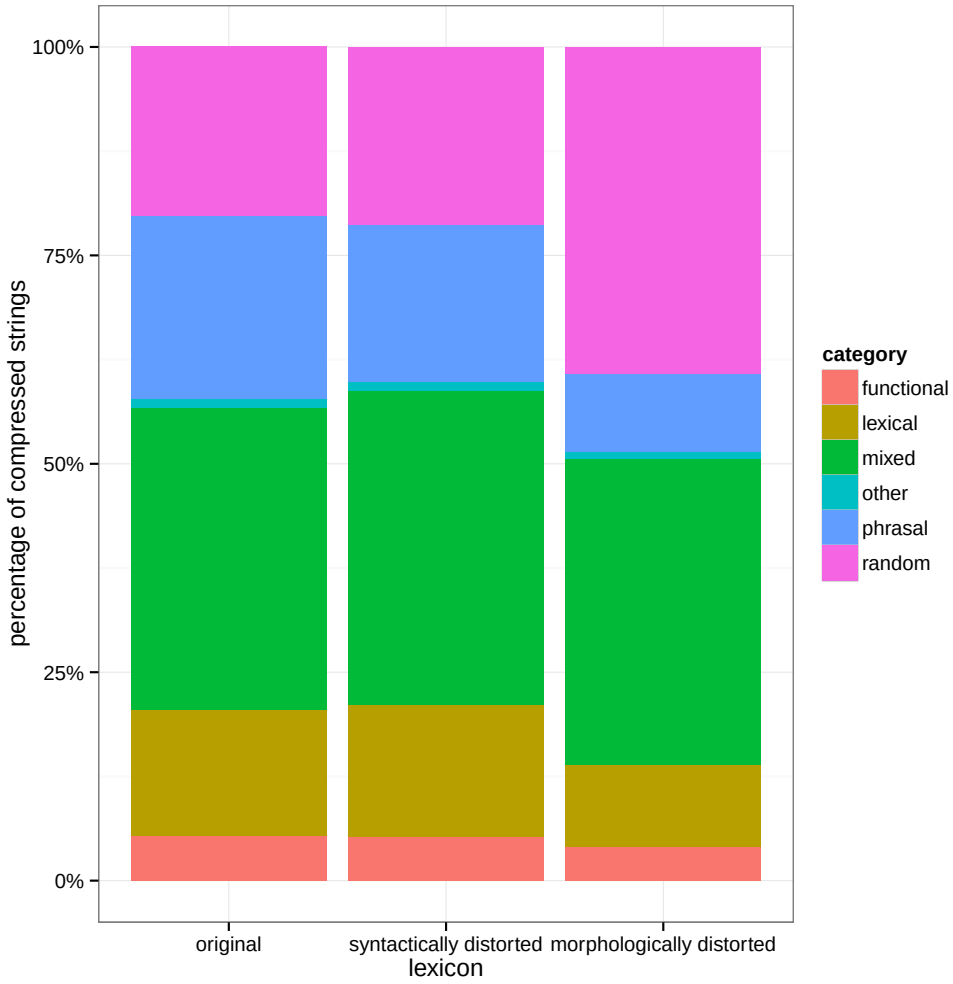

Figure 5.2.: Percentage of compressed strings according to linguistic category in the three Alice lexica.

Again, the effect of syntactic distortion on the lexicon output is more subtle than the impact of morphological distortion. In fact, the percentages of strings per category are, as noted above, almost identical to the original lexicon. The syntactically distorted lexicon does not contain more random strings than the original lexicon but, while containing less strings in total, the percentage of linguistically interpretable strings remains near-identical. As a matter of fact, the number of strings per category varies only between

two and three percent across the two lexica. The question, then, is to which extent syntactic distortion affects the nature of compressed strings. Or, in other words, what happens to inter-word dependencies which should be compromised through distortion? In order to shed light on this matter, the number of syntactically distorted phrases is retrieved from the mixed category. In total the lexicon contains only 101 junk phrases, which represents merely about 1% of the total strings in the syntactically distorted lexicon. Thus, while syntactic interdependencies are corrupted through distortion, the distribution of compressed strings on the linguistic level is not influenced in a major way. Rather, syntactic distortion, as has been discussed in Section 5.1.1, leads to the omission and reduction of compressed strings.

To sum up, morphological and syntactic distortion affect `gzip` 's lexicon output very differently. While morphological distortion leads, as expected, to an increased number of random / unique strings in the lexicon, the impact of syntactic distortion is very subtle so that the syntactically distorted lexicon is near-identical to the original lexicon. Nevertheless, the analysis of the distorted lexica illustrated the process of distortion and demonstrated that it affects the compressibility of texts as intended.

Table 5.9.: Raw frequencies of compressed strings by linguistic category in the three Alice lexica.

Linguistic category	**Original**	**Syntactically distorted**	**Morphologically distorted**
Functional	913	835	749
Lexical	2,558	2,495	1,826
Mixed	6,170	5,970	6,759
Other	175	162	165
Phrasal	3,730	2,999	1,722
Random	3,445	3,364	7,230

5.2. Summary

This chapter took a peek inside the black box of `gzip` 's algorithm in order to gain understanding of the compression technique and define Kolmogorov-based information-theoretic complexity in linguistic terms. This was accomplished by extracting and analysing `gzip` 's lexicon, a line-by-line output of compressed text sequences, for *Alice's Adventures in Wonderland*.

The first section of this chapter gave a detailed description of the distribution of compressed strings in the lexicon of the original Alice text. Every entry in the lexicon was manually analysed and annotated according to linguistic and non-linguistic categories such as lexical or functional words, other linguistically interpretable sequences, phrasal sequences, random non-linguistically meaningful sequences or mixed sequences (containing both meaningful and random sequences). The lexicon analysis revealed that the algorithm compresses both linguistically meaningful and random strings because it does—as expected—not possess any linguistic knowledge. Rather, `gzip` works on the form and structure of (ir)regularities in a text. Thus, Kolmogorov-based information-theoretic complexity is essentially a measure of structural surface redundancy. This implies that the methodology is morphology-sensitive and slightly tends to favour morphological complexity as very high structural redundancy can result in low overall complexity and vice versa. The compression technique is therefore not an ideal tool for measuring overall complexity and should be applied with caution.

In the second section of this chapter, the impact of distortion on compressed strings was assessed by analysing two distorted lexica of *Alice's Adventures in Wonderland*: a syntactically and a morphologically distorted version of Alice were described, annotated and compared to the original Alice lexicon. The survey of string distribution in the distorted lexica confirmed that morphological distortion does indeed lead to an increased amount of "word forms" in the morphologically distorted text and results in an increased number of unique strings which are difficult to compress. In the syntactically distorted text, syntactic inter-dependencies and word order regularities were compromised through distortion as intended. Strings which were previously (i.e. in the undistorted Alice version) added to the lexicon were no longer recognised by the algorithm and were thus omitted in the syntactically distorted lexicon. On a linguistic level, the morphologically distorted lexicon contains a higher percentage of linguistically non-meaningful strings than the original Alice lexicon while the syntactically distorted lexicon contains less strings altogether.

To sum things up, compression algorithms do not intentionally measure or count linguistic features because they do not possess any knowledge of form-function pairings. Instead it was shown that Kolmogorov-based information-theoretic complexity is a measure of structural surface redundancy and based on the recurrence of orthographic character sequences.

Moreover, it was established that the process of morphological and syntactic distortion affect text compressibility as intended.

6. Case studies

This chapter is, ultimately, concerned with the applicability of algorithmic measurements to naturalistic corpus resources and the extent to which intra-linguistic complexity variation can be algorithmically approximated. Thus, two case studies are presented in which the compression technique is applied to naturalistic corpus data. Specifically, the compression technique will be used to assess intra-linguistic variability in English in terms of overall, syntactic and morphological complexity. Empirically, I study two text corpora, the *British National Corpus* (BNC) and the *International Corpus of Learner English* (ICLE). The first case study assesses the complexity variation across different written registers of British English. The results show that more formal registers are overall, and morphologically, more complex than less formal registers, which tend to be syntactically more complex. The second case study takes an interest in the complexity of learner essays produced by students with different levels of instructional exposure in English and, in a small subset of the data, measures the complexity of the learner essays across national varieties. All other things being equal, the amount of instruction in English is shown to be an indicator for the writing proficiency of the learners. In fact, the amount of instructional exposure positively correlates with the complexity of the learners' essays, i.e. texts produced by more advanced learners are overall more complex than texts produced by less advanced learners. Although mother tongue seems to influence the complexity of learner essays, the observed relationship between Kolmogorov complexity measurements and the amount of instruction received in English is robust across different mother tongue backgrounds.

6.1. Assessing complexity variation in British English registers

6.1.1. Method and data

This section draws on data from the *British National Corpus* (BNC World Edition). The BNC is a general, synchronic corpus of standard British English and samples a variety of different spoken and written registers amounting to 100 million words in total (Aston & Burnard 1998). The corpus is fully part-of-speech annotated and comes in SGML (*Standard General Markup Language*) format. This study is restricted to the written compon-

ent of the BNC because, in contrast to the spoken component, sentence boundaries are clearly marked.[1] In this case study the presence of marked sentence boundaries is important for the generation of equally sized samples, and implementation of the compression technique (for more details see below). The written part counts approximately 90 million words and comprises 46 different registers such as academic writing from the social sciences (W_ac_soc_science), school essays (W_essay_sch) or poetry (W_fict_poetry). These registers fall into eighteen macro-registers (Aston & Burnard 1998). Table 6.1 lists the macro-registers in the BNC World Edition which were used in this analysis. Note that the three different newspaper types (broadsheet, tabloid and other) are listed and analysed as separate registers.

Table 6.1.: Overview of the number of texts per written macro-registers and the newspaper micro-registers in the BNC World Edition. The first column gives the class code by which the individual registers can be identified in the corpus.

Class code	**Register**	**Texts**
W_ac	academic writing	501
W_essay	essay	11
W_fict	fiction	464
W_letters	letters	17
W_newsp_brdsht	broadsheet newspapers	340
W_newsp_tabloid	tabloid newspapers	6
W_newsp_other	other newspapers	140
W_non_ac	non-academic writing	534
W_pop_lore	popular lore	211
W_religion	religion	35
W_admin	administrative writing	12
W_advert	advertisements	60
W_biography	biography	100
W_commerce	commerce	112
W_email	email	7
W_misc	miscellaneous writing	500
W_news_script	news scripts	32
W_hansard	hansard	4
W_institut_doc	institutional documents	43
W_instructional	instructional writing	15

All annotation and mark-up was removed from the data and the files

[1] This does not mean that spoken data cannot be assessed with the compression technique. Any kind of language, be it written or spoken, can be used as input for the compression technique as long as it comes in machine-readable form and an appropriate format—in this case study, the technique requires clearly marked (e.g. by a fullstop) sentence boundaries.

were converted to plain text format using several shell scripts (see Appendix D). On the basis of their class code—an SGML tag which identifies the individual registers—all text files per register were retrieved and subsequently merged. The current dataset thus consists of one text file per (macro-)register, amounting to twenty text files.

Methodologically, the open source compression program `gzip` [2] is used to measure linguistic complexity at the overall, syntactic and morphological tier. Largely following Juola (2008), the relative informativeness of the text samples is approximated with a compression algorithm and taken as indicator of a given sample's complexity. The compression technique is implemented with $N = 1,000$ iterations using random sampling. In plain English, the script randomly samples 10% of the sentences per text / macro-register for each iteration. By sampling sentences rather than words, syntactic interdependencies remain intact. This serves two purposes: first, the samples analysed per register are of the same, constant size, which ensures the comparability across the different text samples. Second, random samples of a constant size—especially if they are taken from differently sized populations—should be more representative of the entire population than a fixed sample covering only a certain amount of the text. Keeping these parameters constant, any observed variability in complexity should be attributable to the factor 'register'.

Overall complexity is measured through multiple compression. For each register and iteration two measurements are obtained; the file size in bytes before compression, and the file size in bytes after compression. The mean compressed file size and the mean uncompressed file size is then calculated for each register. Subsequently, the correlation between these two file sizes is eliminated through linear regression analysis with the compressed file sizes as dependent variable, and the uncompressed file sizes as independent variable. This yields the average adjusted overall complexity score (regression residuals) which serves as indicator of the overall complexity of a given text sample. Larger scores can be equated with higher informativeness of a given text sample and thus higher overall complexity. The mean uncompressed and compressed file sizes from which the average adjusted overall complexity scores were calculated, as well as their standard deviations, are presented in Table 6.2.

Morphological and syntactic complexity are addressed by distorting the respective level of information in a given text sample and subsequently measuring the impact on the compressibility of the sample. Let us briefly rehearse how distortion is implemented. Morphological distortion is achieved by random deletion of 10% of all orthographically transcribed characters in a text file, leading to the creation of new word forms, and thus random noise. Texts which have a large number of different word forms to start with should be less affected by distortion than texts which have a

[2] gzip (GNU zip), Version 1.2.4. URL `http://www.gzip.org/`

comparatively smaller number of word forms. In the latter, morphological distortion should create comparatively more random noise, i.e. it should increase the morphological complexity in the sample. Comparatively worse compression ratios thus indicate lower morphological complexity. Syntactic distortion is accomplished by random deletion of 10% of all orthographically transcribed word tokens in a text sample. This should greatly affect syntactically complex texts—syntactic complexity is defined in terms of word order whereas maximum flexibility is equated with low complexity—because it disrupts word order regularities. Comparatively bad compression ratios after syntactic distortion indicate high syntactic complexity.

As described above, the distortion and compression loop (see Appendix B.2) is applied with random sampling and $N = 1,000$ iterations. For each iteration of the loop, the script returns the morphological and syntactic complexity score of each text sample. The morphological complexity score is defined as $-\frac{m}{c}$, where m is the compressed file size after morphological distortion and c the original compressed file size. The syntactic complexity score is defined as $\frac{s}{c}$, where s is the compressed file size after syntactic distortion and c the file size before distortion. Finally, I obtain the average morphological and average syntactic complexity score for each sample by taking the mean of the total number of measuring points (Table 6.3).

Table 6.2.: Mean uncompressed and compressed file sizes (in bytes) and their standard deviations by register.

Register	Uncompressed	Standard deviation	Compressed	Standard deviation
Academic	72245	2490	48256	22082
Administrative	73970	2726	47301	24355
Advertisements	55669	1967	37682	16387
Biography	53126	1990	36259	15379
Broadsheet newspapers	60929	2135	41564	17657
Commerce	72155	2392	47622	22384
Email	45963	1579	30781	13905
Essays	56774	1976	38180	16956
Fiction	29556	1355	20263	8501
Hansard	58521	2237	38250	18484
Institutional documents	77839	2531	50885	24614
Instructional	65380	2166	43786	19700
Letters	65044	2183	42804	20263
Miscellaneous	61476	2147	41820	18027
Non-academic	65873	2326	44574	19597
Other newspapers	56928	1974	38872	16510
Popular lore	55449	1879	38130	15794
Religion	59217	1960	39757	17786
News scripts	61155	1906	40901	18415
Tabloid newspapers	44259	1566	30482	12635

Table 6.3.: Average morphological and syntactic complexity scores and their standard deviations by register.

Register	Morphological score	Standard deviation	Syntactic score	Standard deviation
Academic	-0.9774	0.0636	0.9066	0.0063
Administrative	-1.0104	0.0943	0.9112	0.0103
Advertisements	-0.9695	0.0589	0.9069	0.0066
Biography	-0.9612	0.0497	0.9063	0.0060
Broadsheet newspapers	-0.9640	0.0519	0.9059	0.0057
Commerce	-0.9852	0.0710	0.9072	0.0068
Email	-0.9719	0.0611	0.9086	0.0081
Essays	-0.9712	0.0579	0.9072	0.0069
Fiction	-0.9505	0.0459	0.9069	0.0069
Hansard	-0.9910	0.0774	0.9106	0.0099
Institutional documents	-0.9943	0.0792	0.9078	0.0074
Instructional	-0.9743	0.0606	0.9074	0.0070
Letters	-0.9911	0.0758	0.9110	0.0103
Miscellaneous	-0.9648	0.0529	0.9059	0.0058
Non-academic	-0.9694	0.0568	0.9060	0.0058
Other newspapers	-0.9636	0.0519	0.9061	0.0059
Popular lore	-0.9577	0.0469	0.9058	0.0055
Religion	-0.9710	0.0579	0.9069	0.0065
News scripts	-0.9762	0.0635	0.9088	0.0083
Tabloid newspapers	-0.9572	0.0476	0.9063	0.0061

6.1.2. Register variability

This section discusses register variability in terms of overall, morphological and syntactic complexity comparing the twenty written BNC registers introduced above.

An overall complexity hierarchy of the registers is presented in Figure 6.1. The extreme cases on the complex side of the spectrum are broadsheet newspapers and popular lore. Other above average complex registers are, in decreasing order of overall complexity: miscellaneous, non-academic, other newspapers, academic, biography, instructional, advertisement, tabloid newspapers, religion. The registers essays and news scripts are of average overall complexity as indicated by the average adjusted overall complexity score close to zero. The extreme cases on the "simple" side of the spectrum are administrative and hansard. Below-average complex registers are, in decreasing order of complexity: commerce, institutional documents, letters, email and fiction. Generally, less formal registers, i.e. registers which are relatively close to spoken language such as email or fiction, are less complex than more formal registers such as newspapers which are known to be subject to strict editorial and economy-driven constraints. In assessing grammatical variation in written and spoken texts Biber (1988) establishes several dimensions along which texts typically vary. One of these dimensions is labelled 'informational versus involved production' and refers to a highly informational, abstract style that is the result of planned and edited production on one end of the dimension and, on the other end, to an involved emotional style that is produced in interaction (Biber 1988: 104–108). Using the parlance of Biber (1988) then, I find that more involved registers are less complex than informational-abstract registers. In particular the three newspaper registers lend themselves well for illustrating complexity variation along the formality / register dimension: the most formal broadsheet newspapers are the most complex register, other regional newspapers are less complex and tabloid newspapers, which are known for their involved style, are the least complex newspaper register. In this vein, the registers fiction, letters and email are below-average complex while the newspaper registers, academic writing and biography are above-average complex. However, the ranking is not flawless as it is unexpected that administrative writing, hansards and institutional documents—all rather formal registers—are less complex than, for instance, popular lore.

Figure 6.2 shows the BNC registers according to morphological and syntactic complexity. For the sake of readibility, the registers were categorised into "more informational" and "more involved" registers according to the scale presented in Biber (1988: 128). Administrative writing is the syntactically most complex but morphologically least complex register. As noted before, morphological complexity seems to interact with overall complexity in that it is more strongly reflected in the overall measure than

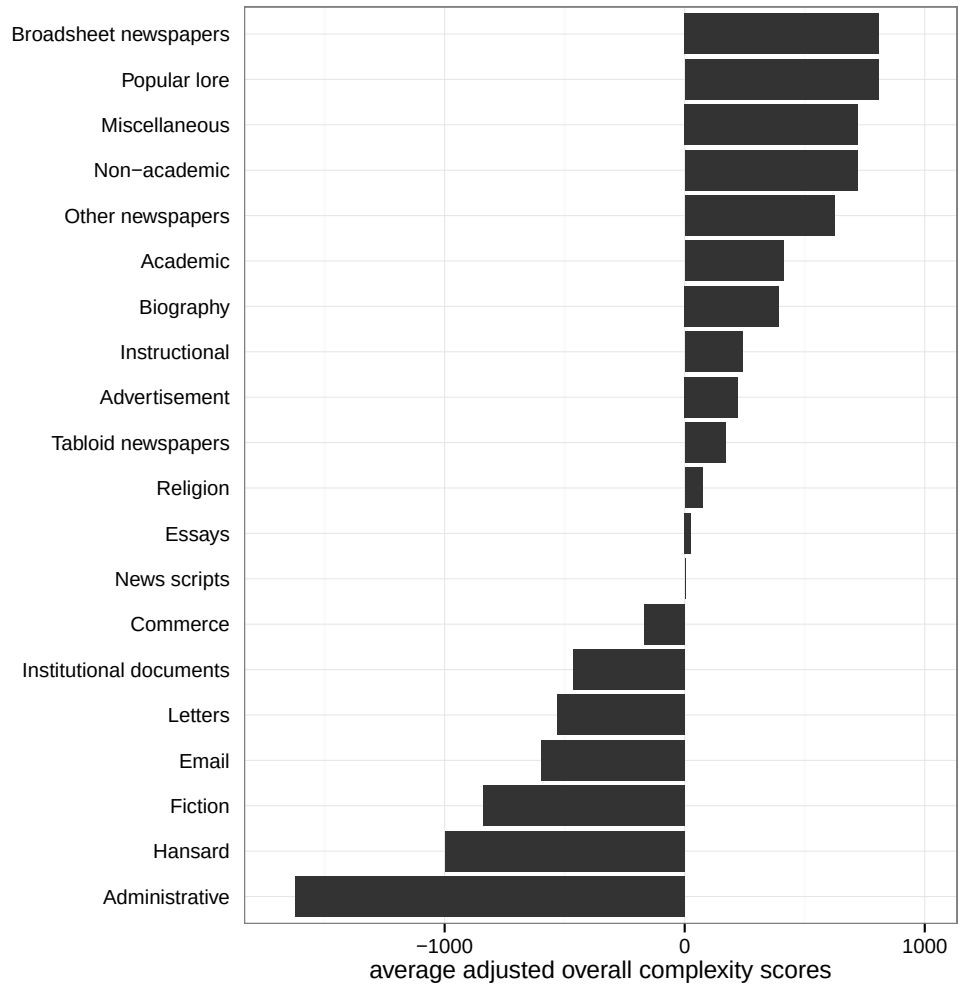

Figure 6.1.: Overall complexity hierarchy of written BNC registers. Negative residuals indicate below-average complexity; positive residuals indicate above-average complexity.

syntactic complexity. This is due to the nature of the algorithm which essentially relies on structural (surface) redundancy for compression and thus somewhat favours morphological complexity. Hence, administrative writing, which loads low on the morphological dimension, is the overall least complex sample. Fiction is the morphologically most complex register and popular lore is the syntactically least complex register. Tabloid newspapers, biography, other newspapers, broadsheet newspapers, miscellaneous and non-academic writing cluster together in the top left quadrant; they are roughly of the same morphological and syntactic complexity with a tendency to above-average complexity on the morphological tier and below-average complexity on the syntactic tier. Most of the other registers are scattered across the left upper centre of the plot and exhibit average morphological and syntactic complexity. The registers commerce, and notably institutional

documents are of below-average morphological complexity. Both hansard and letters are very low in morphological complexity but high in syntactic complexity. Furthermore, in this sample of registers, morphological complexity significantly trades off against syntactic complexity and vice versa (Pearson's correlation coefficient $r = -0.78$, $p = 0.0001$).

Szmrecsanyi (2009) investigates intra-linguistic variablity in terms of analyticity and syntheticity, two measures which have a long tradition in linguistic typology and are closely linked to the current complexity debate. Analyticity basically refers to the number of unbound grammatical markers in a text sample while syntheticity refers to the number of bound grammatical markers in a text sample (Szmrecsanyi 2009: 319–320, 322). Szmrecsanyi (2009) analyses, inter alia, the variability of the written and spoken macro-registers in the BNC and reports that high analyticity indices correlate with involved production while abstract informational registers tend to be more synthetic. Although analyticity and syntheticity are not directly related to Kolmogorov complexity measurements, and are in fact two very different things, I will nevertheless assess to which extent the current results correspond to Szmrecsanyi's findings. Analyticity refers to free (unbound) structures and will therefore be compared to syntactic complexity, which is basically a measure of word order flexibility and could be said to refer to unbound structures (words). Syntheticity could be said to roughly correspond to morphological complexity in that both measures refer to bound grammatical markers. Thus, testing this against the algorithmic register analysis, I correlate my results with the indices reported in Szmrecsanyi (2009: 333). Note that the separately analysed newspaper registers are excluded. Pearson's correlation coefficient for analyticity and the average syntactic complexity scores is positive but low ($r = 0.34$, $p = 0.09$). The correlation between syntheticity and the average morphological complexity scores is likewise positive but very low ($r = 0.28$, $p = 0.14$). While the correlation between the two measures is very low—which is not surprising since they are not directly related and hence comparable—it is positive. In regard to the relationship between the formality of the register and complexity, my findings therefore tie in with Szmrecsanyi's (2009) findings and suggest that more formal / abstract-informational registers tend to be morphologically more complex than informal / involved registers which tend to be syntactically more complex (e.g. newspapers vs. emails, biography vs. letters).

In sum, written British English registers vary in overall and morphological complexity such that less formal / involved registers are less complex than formal / abstract-informational registers. On the other hand, less formal registers have a tendency to be more syntactically complex than formal registers. On the whole, these Kolmogorov-based results are in line with previous findings (e.g. Szmrecsanyi (2009)), outliers notwithstanding.

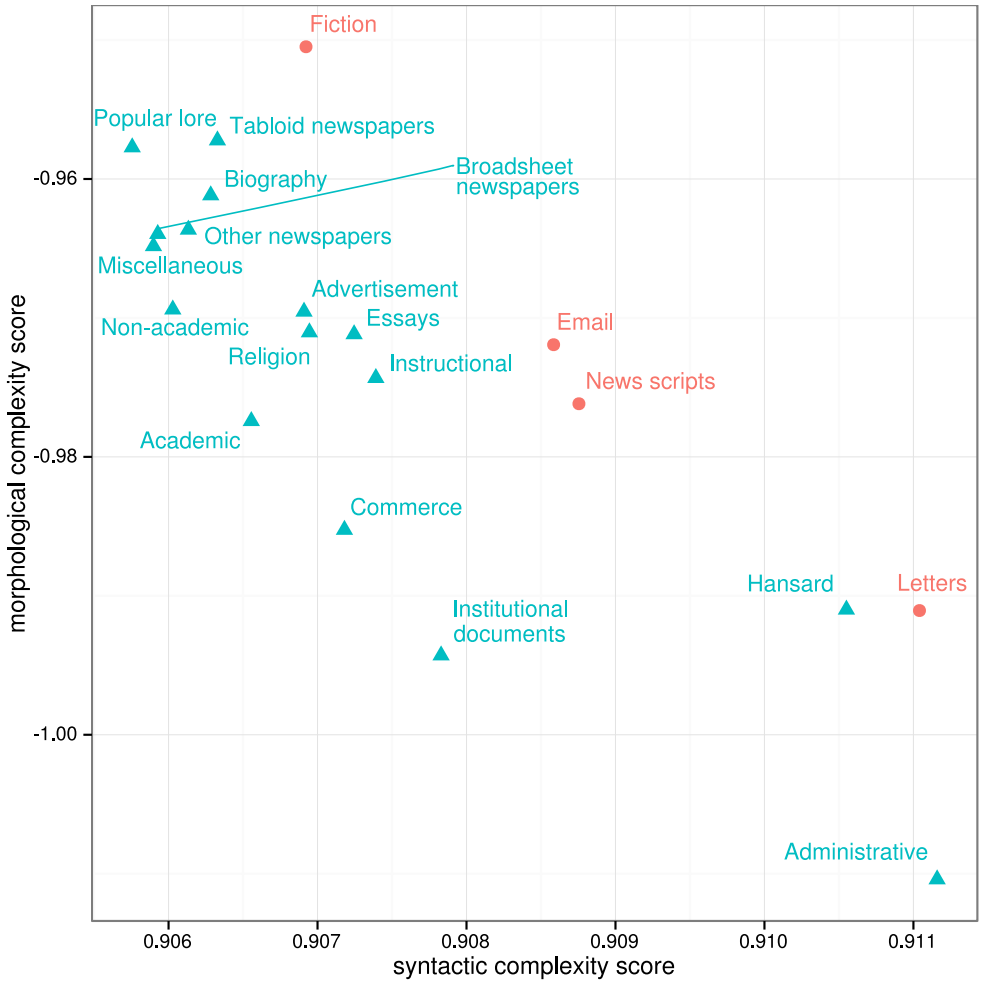

Figure 6.2.: Morphological complexity by syntactic complexity of written BNC registers. More informational registers are represented by blue triangles, more involved registers by red filled circles. Abscissa indexes increased syntactic complexity, ordinate indexes increased morphological complexity.

6.2. Measuring complexity in learner English[3]

6.2.1. Method and data

This section investigates data from the *International Corpus of Learner English* (ICLE, version 1). ICLE, first published in 2002, is a corpus of written learner English (English as a foreign language) and contains both argumentative and literary essays composed by higher to intermediate advanced learners from 11 different mother tongue backgrounds: Bulgarian, Czech, Dutch, Finnish, French, German, Polish, Russian, Spanish and Swedish. The corpus comprises approximately 200,000 words per national variety and counts 2.5 million words in total. All national components of the corpus were designed according to the same guidelines and provide extensive data on learner and task variables as listed below (Granger et al. 2002).

Learner variables:

- age
- gender
- mother tongue
- regional provenance (e.g. Belgian Dutch vs. Netherlandic Dutch)
- knowledge of other foreign languages
- time spent in an English speaking country
- time spent studying English at school
- time spent studying English at university

Task variables:

- topic
- length
- argumentative vs. literary essay
- timed vs. untimed essay
- exam conditions vs. use of reference tools

[3] A partial summary of this chapter has appeared as Ehret & Szmrecsanyi (2016a).

This study takes an interest in the complexity of learner language as represented by the essays sampled in ICLE. In particular, the relationship between the complexity of the learner essays and the amount of previous instruction in English is explored. For this reason, the data sampled in ICLE is categorised into four different groups according to the amount of instructional exposure, i.e. the number of years of instruction in English at school and university. Furthermore, the dataset is restricted to argumentative essays, which constitute the largest part of the data, because content control is a crucial factor for the successful application of the compression technique. In the following a brief account of the categorisation procedure and guidelines is given.

First, the number of texts for every combination of years spent studying English at school and university was surveyed. Table 6.4 shows the number of argumentative essays according to the number of years the learners have spent studying English at school and university. In some cases, very little data exists or the existing data stems from learners who have not attended university. For example, there are only four texts from learners who have studied English for one year at school but have not studied English at university at all (Table 6.4, first row). In order to obtain a representative sample for each learner group, such borderline cases were excluded; the range of years at school was restricted to 4–9 years and the range of years at university to 1–5 years.

Table 6.4.: Number of argumentative essays according to years of studying English at school and university in ICLE.

School	Texts	University	Texts
1	4	—	—
2	30	—	—
3	70	—	—
4	503	1	169
5	328	2	686
6	375	3	650
7	365	4	543
8	399	5	195
9	420	6	33
10	351	7	4
11	48	—	—
12+	82	—	—

Second, on the basis of the number of years spent studying English at school and university, six groups of learners / instructional exposure were determined in such a way that the groups overlap as little as possible. Table 6.5 lists the six learner groups which are used in this analysis. The

Table 6.5.: Learner groups by years of instruction in English at school and university. Total number of years, number of texts, words and sentences are provided for each group.

Group	School	University	Total	Texts	Words	Sentences
I	4-6	1-2	5-8	340	230,054	12,531
II	4-6	3	7-9	345	238,590	13,644
III	4-6	4-5	8-11	464	303,233	16,792
IV	7-9	1-2	8-11	533	335,091	17,285
V	7-9	3	10-12	262	171,762	9,328
VI	7-9	4-5	11-14	253	169,237	8,765

most advanced groups—the groups sampling essays from learners with the highest amount of instructional exposure in English—are groups VI and V, while groups I and II are the groups with the least amount of instructional exposure in English. Groups III and IV represent both intermediate levels of instructional exposure and have received the same amount of instruction in English, yet to different parts at school and university. The current dataset thus consists of six text files which can be taken to represent learner groups of different proficiency levels (see also Bestgen & Granger 2014: 30).

On a methodological plane, the open source `gzip` algorithm is used to approximate the relative informativeness in the text samples and thereby measuring their overall, morphological and syntactic complexity. More precisely, the Juola-style compression technique is applied with $N = 1,000$ iterations and implemented with random sampling. In other words, in each iteration of the distortion and compression script 10% of the sentences per text / learner group are randomly sampled. I sample the same number of sentences rather than, for instance, the same number of words; because in this manner syntactic interdependencies remain intact. Random sampling ensures the comparability of the different texts because it keeps sample size constant. Furthermore, random samples of a constant size that are taken from differently sized texts are more representative than fixed-size samples which cover only a certain part of a text. Thus, all other things being equal, any variability in the complexity between the different texts should be due to the different level of instructional exposure of the learners. In order to avoid a confounding of variables, the influence of the variable 'national background' on the complexity of the learner groups will be addressed in Section 6.2.3.

The technical details of the method used in this case study are essentially identical to the method described in Section 6.1.1 above. In order to measure the overall complexity of learner groups in ICLE, the file sizes in bytes of each text file before and after compression are established for each iteration. Thereafter, the adjusted complexity scores, which indicate

the overall complexity of each text sample, are calculated by subjecting the file sizes to linear regression. Finally, the mean of the total number of iterations ($N = 1,000$) yields the average adjusted overall complexity score for a given text sample. Comparatively larger scores indicate higher overall complexity of a text sample. Table 6.6 presents the mean uncompressed and compressed file sizes on whose basis the average adjusted overall complexity scores were calculated, and their standard deviations by learner group.

Next, I address syntactic and morphological complexity. Syntactic distortion is performed by deletion of 10% of all word tokens in each text file prior to compression. This procedure leads to the disruption of word order regularities and should greatly affect syntactically complex texts, i.e. texts with a comparatively fixed word order. The morphological information is manipulated by deletion of 10% of all orthographic characters in each text file thereby creating new word forms. This compromises the compressibility of morphologically simple texts which, on the whole, have fewer word forms than morphologically complex texts. Applying the distortion and compression loop with $N = 1,000$ iterations, the morphological and syntactic complexity score for each text is calculated (for details refer to Section 6.1.1 above) and returned for each iteration of the script. The mean of the total number of measuring points for morphological and syntactic complexity yields the average morphological and syntactic complexity score, respectively (Table 6.7).

Table 6.6.: Mean uncompressed and compressed file sizes (in bytes) and their standard deviations by learner group.

Group	Uncompressed	Standard deviation	Compressed	Standard deviation
I	93212	2485	34407	1236
II	89416	2410	33182	1202
III	92971	2507	34435	1248
IV	96803	2568	35837	1294
V	90997	2412	34289	1237
VI	95769	2587	36189	1329

Table 6.7.: Average morphological and syntactic complexity scores and their standard deviations by learner group.

Group	Morphological score	Standard deviation	Syntactic score	Standard deviation
I	-1.0539	0.0026	0.9143	0.0019
II	-1.0529	0.0027	0.9149	0.0020
III	-1.0536	0.0026	0.9145	0.0019
IV	-1.0523	0.0026	0.9141	0.0019
V	-1.0463	0.0027	0.9139	0.0019
VI	-1.0452	0.0025	0.9135	0.0019

6.2.2. Complexity variation in learner essays

In this section, the complexity of the six groups of learners with different amounts of instructional exposure—I, II, III, IV, V, and VI—on the overall, syntactic and morphological tier is discussed.

In terms of overall complexity, the results obtained through compression match with our expectations in that essays of more advanced learners are more complex than essays of less advanced learners (Figure 6.3). In other words, overall complexity correlates positively with the amount of instructional exposure in English, and by inference, with proficiency (Pearson's correlation coefficient $r = 0.85$, $p = 0.034$). To be more specific, texts IV and V are overall the most complex texts. These texts stem from learners who studied English for ten to fourteen years in total and therefore represent the group with the highest amount of instructional exposure in this dataset. Text I, on the other hand, is overall the least complex text and was produced by learners who have studied English for about five to eight years in total. The texts II, III and IV are, in decreasing order of overall complexity, below average complex. This ranking is somewhat unexpected because learners of the levels III and IV should be more advanced, and hence the texts more complex, than learners of group II who have spent less time studying English. Yet, the overall complexity hierarchy of learner groups generally conforms to the natural progression from less complex production to more complex production.

Let us turn to morphological and syntactic complexity (Figure 6.4). The texts produced by the learners with the highest amount of instructional exposure, texts V and VI, are by far the most complex texts in regard to morphological complexity. Yet, in regard to syntactic complexity they are the least complex texts. All other texts are morphologically considerably less complex but syntactically more complex: text II is the syntactically most complex text followed, in decreasing order of syntactic complexity, by the texts III, I and IV. The morphologically most simple text is text I, which represents the group with the least instructional exposure in this study. In fact, morphological complexity highly correlates with the amount of instructional exposure of the learners: the two-sided Pearson's correlation coefficient for morphological complexity and learner group is positive at $r = 0.89$ ($p = 0.018$).[4] Put in other words, increasing amounts of instructional exposure in English lead to more morphological complexity but less syntactic complexity. This trade-off is statistically confirmed by a negative correlation between morphological and syntactic complexity (Pearson's correlation coefficient $r = -0.82$, $p = 0.02$). All other things being equal, the amount of instructional exposure received by the ICLE learners is taken as

[4]The correlation between syntactic complexity and instructional exposure is, on the other hand, negative (Pearson's correlation coefficient $r = -0.83$, $p = 0.042$) indicating the decrease of syntactic complexity with increasing amounts of instructional exposure.

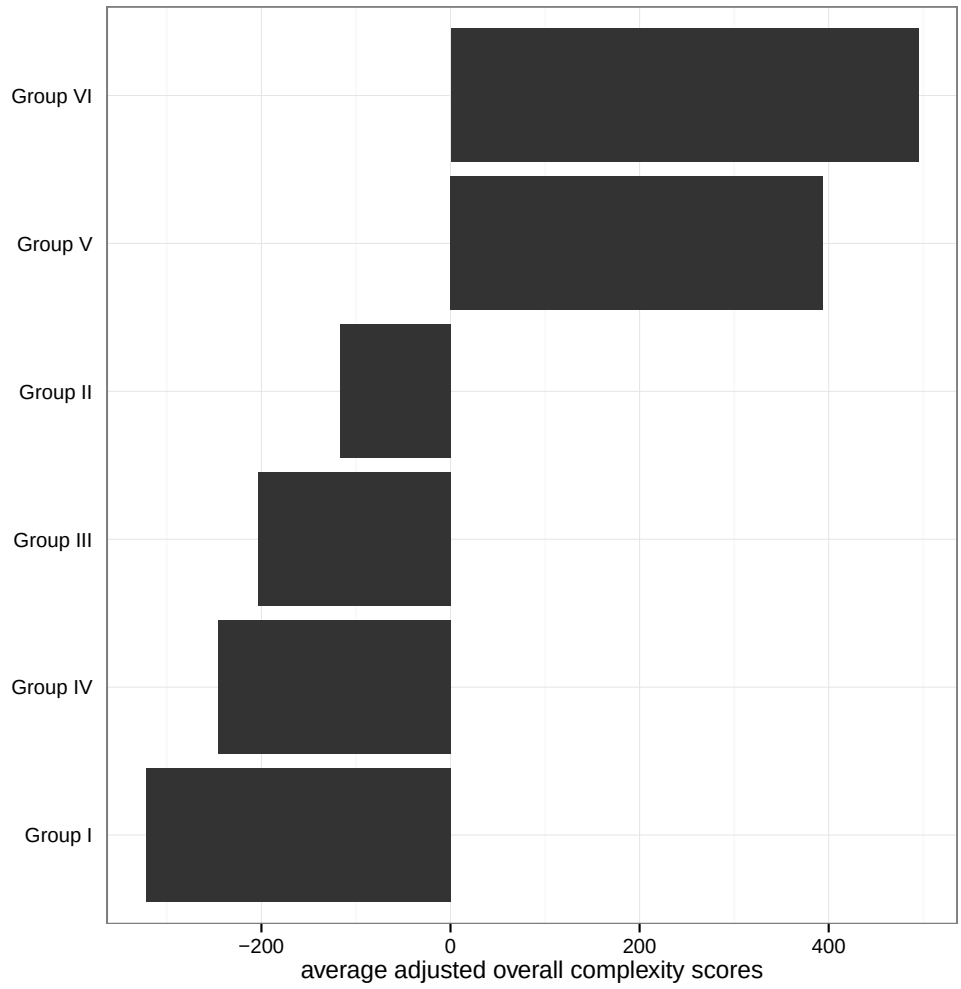

Figure 6.3.: Overall complexity hierarchy of learner groups in ICLE. Negative residuals indicate below-average complexity; positive residuals indicate above-average complexity.

a proxy for their (writing) proficiency in English.

The decrease of syntactic complexity in more advanced writing may seem surprising. However, it can be explained; firstly, by the fact that more proficient learners are more likely to use different word order patterns such as inversion of the type *never have we been more suprised* versus *we have never been more surprised*, than less advanced learners who should prefer basic SVO syntax. The use of more variable—and thus comparatively freer—word order leads to lower syntactic complexity because this work defines maximally simple syntax as maximally free word order (see Section 7). Secondly, my findings can be seen as further evidence along the lines of Biber et al. (2011) who show that the measure of complexity, namely the degree of clausal embedding, commonly used in writing development studies, does not at all capture the complexity of advanced writing proficiency

(Biber et al. 2011: 10–12; see also Biber & Gray 2010). On the contrary, Biber et al. (2011) find that clausal embedding is a feature of conversational language which is acquired in early stages of language development. Later stages of proficiency, instead, are marked by a higher degree of phrasal complexity and greater range of lexico-grammatical combinations such as finite complement clauses (e.g. *I think that [. . .]*) (Biber et al. 2011: 29–32). How then, is this related to morphological and syntactic complexity as measured with the compression technique? Clausal embedding is concerned with the degree of subordination and thus with syntactic complexity while it could be argued that an increasing use of different lexico-grammatical patterns increases morphological complexity. For instance, according to Biber et al. the majority of *that*-clauses in spoken language occur with only four different verbs (Biber et al. 2011: 31). In the context of Kolmogorov complexity, this means that a text with few verbal patterns is easily compressible and hence morphologically simple. A text with many different verbal patterns, on the other hand, should be more difficult to compress and thus morphologically complex. In short, the current study supports the finding that higher complexity in writing is not necessarily accompanied by higher syntactic complexity. As an aside, this tallies with the BNC results reported in Chapter 6.1 above, according to which more informal (i.e. oral) registers are syntactically more complex than formal (i.e. written) registers.

Finally, the complexity measurements presented in this section seem to systematically correlate with measures more commonly used in second language acquisition research such as lexical diversity or noun phrase density (Ehret & Szmrecsanyi 2016a). Ehret & Szmrecsanyi (2016a) calculate seven different SLA measures for the six learner groups and correlate them with the Kolmogorov measurements for overall, morphological and syntactic complexity. These calculations were conducted using two online tools (`http://corpora.lancs.ac.uk/vocab/analyse/morph.phpandhttp://cohmetrix.com`). Both tools come with a text size limit so that the calculations were based on random samples of sentences from the six texts analysed in this section. Table 6.8 shows an example of the correlation coefficients for each of the SLA measures calculated. Overall and morphological complexity best correspond to morphological verb complexity (MCI_{verbs}; in simplified terms a measure of morphological variation based on word types (Pallotti 2015: 121–122)), noun phrase density (DRNP) and number of noun phrase modifiers (SYNNP). This observation yet again shows that overall and morphological Kolmogorov complexity are somewhat related (see Chapter 5 and Chapter 7). The syntactic complexity scores are negatively correlated with the number of noun phrase modifiers (SYNNP)—less noun phrase modifiers predict more rigid word order—and morphological verb complexity (MCI_{verbs}) (for more details refer to Ehret & Szmrecsanyi (2016a)).

Table 6.8.: Pearson correlation coefficients between SLA measures and Kolmogorov measures. Significance codes: *significant at $p < 0.05$ under Bonferroni correction. MCI_{verbs}/MCI_{nouns}: morphological complexity indices (Brezina & Pallotti 2015; Pallotti 2015); LDTTRc: lexical diversity, type-token ratio (content words); SYNLE: left embededness, words before main verb (mean); SYNNP: number of modifiers per noun phrase (mean); DRNP: noun phrase density, incidence; DRPVAL: agentless passive voice density incidence.

SLA measure	**Overall complexity score**	**Morphological complexity score**	**Syntactic complexity score**
MCI_{verbs}	* 0.96	0.92	-0.61
MCI_{nouns}	0.55	0.43	0.14
LDTTRc	0.53	0.58	-0.40
SYNLE	0.24	0.23	0.13
SYNNP	0.83	0.88	-0.85
DRNP	0.78	0.69	-0.28
DRPVAL	-0.07	0.14	0.04

(Adapted from Ehret & Szmrecsanyi (2016a))

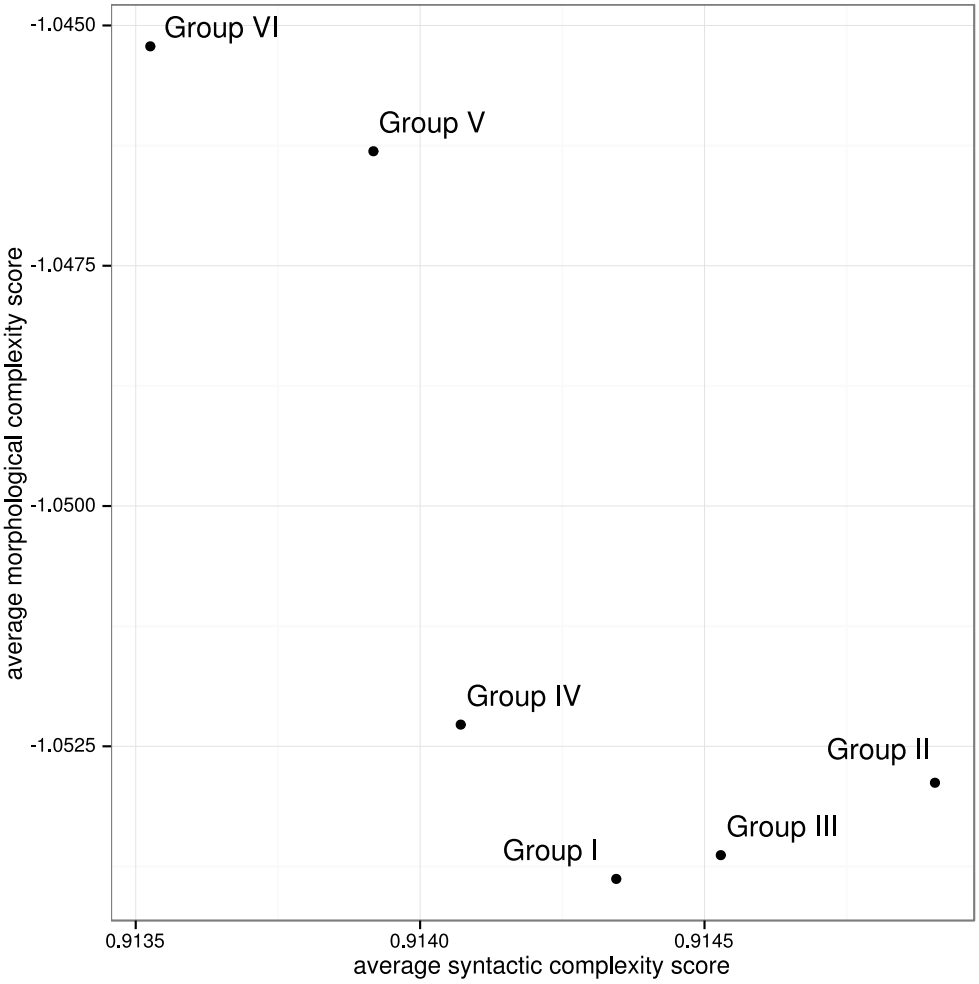

Figure 6.4.: Morphological complexity by syntactic complexity of learner groups in ICLE. Abscissa indexes increased syntactic complexity, ordinate indexes increased morphological complexity.

6.2.3. Does mother tongue matter?

The previous section analysed the complexity of six different learner groups, i.e. groups with different levels of instructional exposure, drawing on essays written by learners with different national backgrounds. This section seeks to elucidate the influence of the factor 'national variety' or 'mother tongue background' on the overall, syntactic and morphological complexity of these essays. In this section, I address the question of whether the complexity measured in these essays significantly varies between learners of different mother tongue backgrounds or whether their complexity is invariable across the national varieties sampled in ICLE. To this end, I analyse two subsets of the current dataset: a subset of the six groups across four different national varieties (German, French, Italian and Spanish) and a subset of the six

groups in one national variety (German).

Table 6.9 lists the number of texts, words and sentences per learner group for the national components German, French, Italian and Spanish. The national components of ICLE are relatively small to start with—they only comprise 200,000 words—so that the amount of data which is available for each national variety per group is very little. For some of the groups merely two texts (roughly around 1300 words) exist, and for the French IV level no data at all is available. In particular the French, Italian and Spanish subsets are extremely unbalanced in regard to the distribution of data across the six groups of learners. Due to this distributional imbalance, I first analyse the complete subset consisting of four national varieties, and afterwards as a control group, the complete German subset which is the distributionally most balanced sample in the set. The German subset is also the largest national component and the amount of data per learner group ranges from approximately 11,000 to 50,000 words.

Methodologically, the compression technique is separately applied to each subset with $N = 1,000$ iterations and randomly samples 10% of the sentences in each text per iteration. Thereby, the maximum size of the sample is determined by the number of sentences contained in the smallest text. To illustrate, in the four nationality subset the smallest text counts 54 sentences. Thus the script randomly samples 54 sentences from each text in the subset and keeps a random 10% of these sentences, in this case 5 sentences per iteration (as the script rounds non-integer values to the next smaller integer).[5] As the implementation of the compression technique and the calculation of the complexity scores are identical to the procedure described in Sections 6.1.1 and 6.2.1, the details will not be repeated here. Suffice it to say that the average overall complexity scores are calculated based on the compressed and uncompressed file sizes of each text sample. Furthermore, the average morphological and average syntactic complexity scores are calculated on the basis of the morphologically / syntactically distorted file sizes and undistorted file sizes respectively. The tables with the full statistics are provided in Appendix A.2.

Let us first turn to the overall complexity hierarchies. If mother tongue does not impact on the overall complexity of the six groups, the texts should be ranked according to the amount of instructional exposure received in English, i.e. the group with the highest amount of instructional exposure should be more complex than the group with the least amount of instructional exposure. However, in the four nationality subset (Figure 6.5), the texts seem to cluster according to the national background of the learners rather than according to learner group. The overall picture is rather noisy and there is no clear hierarchy of the six learner groups across the different

[5]This results in a very small amount of data per learner group and the representativity of these samples is questionable. This is a major caveat and the results of the four nationality study should therefore be considered tentative.

Table 6.9.: Learner groups according to national background. Number of texts, words and sentences are provided for each group. Note that for French learners no texts are available for group IV.

Nationality	Group	Texts	Words	Sentences
German	I	25	11,490	595
	II	22	11,922	635
	III	24	13,259	697
	IV	107	51,589	2,753
	V	80	37,010	1,981
	VI	85	50,868	2,471
French	I	2	1,344	100
	II	77	46,651	2,742
	III	134	80,505	4,423
	IV	—	—	—
	V	4	2,244	138
	VI	3	1,674	114
Italian	I	3	1,503	54
	II	6	3,093	117
	III	27	14,543	634
	IV	2	980	66
	V	9	5,166	232
	VI	34	20,064	906
Spanish	I	14	8,869	437
	II	2	1,341	60
	III	6	3,885	188
	IV	86	55,494	2,806
	V	6	3,642	181
	VI	28	16,744	858

mother tongue backgrounds. The German and Italian texts are overall the most complex texts while the Spanish and French texts range from slightly above-average complexity to below-average complexity; French and Spanish learners seem to produce overall less complex texts than German and Italian learners. To be more specific, the German groups VI and V—representing the most advanced groups—are the two most complex texts followed by Italian III, German III, German IV and Italian VI and V. On the "simple" end of the spectrum, Italian I, French I and Spanish II and V are the least complex texts. Apart from the extremely low complexity of the Spanish group V, essays written by more advanced learners tend to be more complex than essays written by less advanced learners. Nevertheless, no clear ranking according to the amount of instructional exposure, neither within nor across national varieties, is discernable in the mid-range. In particular the ranking of the French and Spanish groups does not seem to follow any pattern. This is not surprising, however, as these two datasets are particularly unbalanced in regard to data quantity and distribution across the different learner groups. Therefore, these results are merely provisional in nature.

In the German national subset (Figure 6.6), the learner groups are ordered in textbook-style fashion in decreasing order of complexity from the most advanced group to the least advanced group. As in the complete ICLE study, increasing overall complexity highly correlates with instructional exposure and, by inference, with increasing proficiency (Pearson's correlation coefficient $r = 0.96$, $p = 0.001$). Thus, the texts VI and V which were produced by the most advanced learners are the most complex texts, while text I which stems from the least advanced learners is by far the least complex text. Moreover, the complexity of all six groups (I, II, III, IV, V, VI) strictly increases with increasing amounts of instructional exposure. On the other hand, in the complete ICLE study presented in the previous section, the less advanced group II is more complex than the more advanced groups III and IV. This could well be an effect of the mixing of various national backgrounds. In order to statistically compare these results, I correlate the overall complexity scores of the German national subset and the complete ICLE study.[6] Pearson's correlation coefficient indicates with $r = 0.87$ ($p = 0.012$) that the ranking of the six groups of instructional exposure in terms of overall complexity across the two studies highly correlates.

In regard to morphological and syntactic complexity, the results of the two national subsets seem to confirm the findings of the complete ICLE study. In the four nationality subset (Figure 6.7), national clusters seem to emerge which suggest that mother tongue background and nationality matters. However, within the national clusters, I observe that texts produced by

[6]The scores of the four nationality subset are not correlated because firstly, the results are somewhat problematic due to data sparsity and, secondly, the differing number of observations does not permit a direct comparison between the studies.

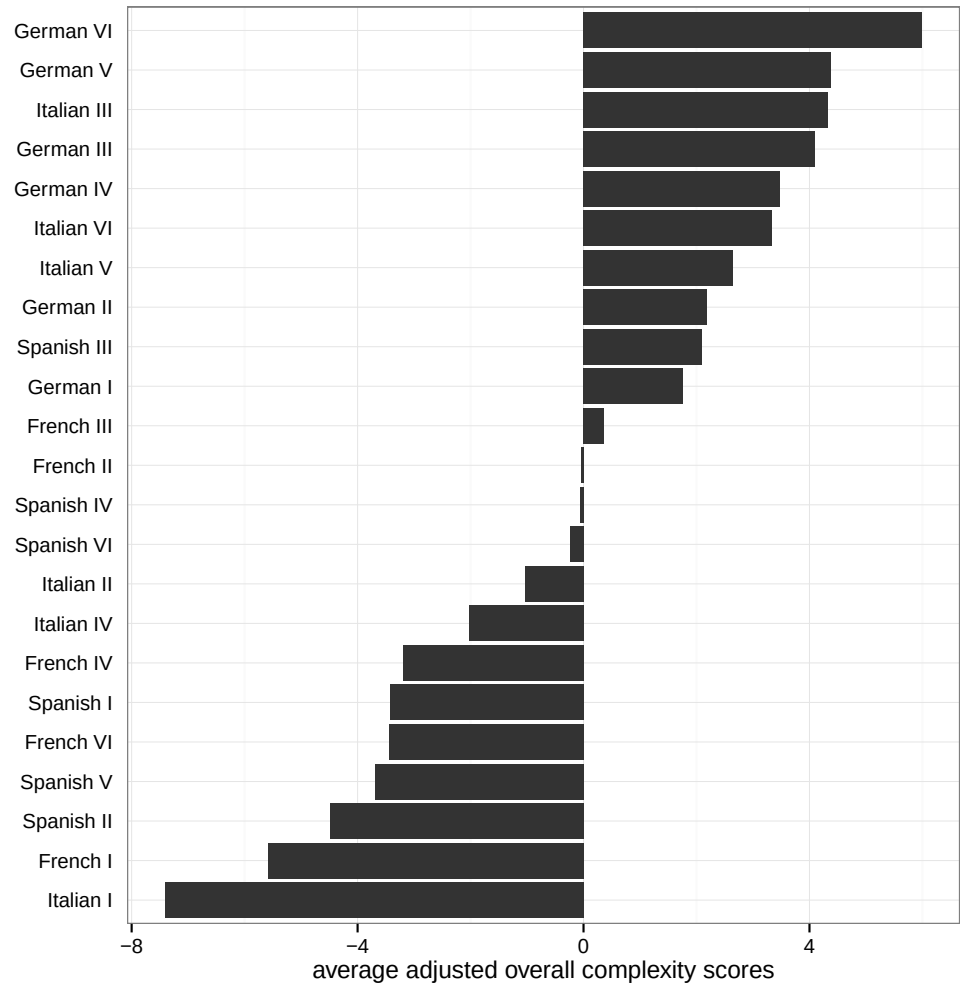

Figure 6.5.: Overall complexity hierarchy of ICLE learner groups by four different mother tongue backgrounds. Negative residuals indicate below-average complexity; positive residuals indicate above-average complexity.

more advanced learners tend to be morphologically more complex than texts produced by less advanced learners. Syntactically, there is a tendency for texts produced by less advanced learners to be more complex than texts produced by more advanced learners. Yet, these findings have to be taken with a grain of salt because in terms of morphological complexity, these results are not as clear-cut and seem to be flawed by the distributional imbalance of the data. The German cluster, for example, is the largest dataset and is morphologically the most complex cluster while the Italian I and II and Spanish II and V texts are the morphologically least complex texts and also happen to be some of the smallest samples. Further research with a much larger dataset is needed to investigate the cross-linguistic (in)variability of morphological and syntactic learner language complexity.

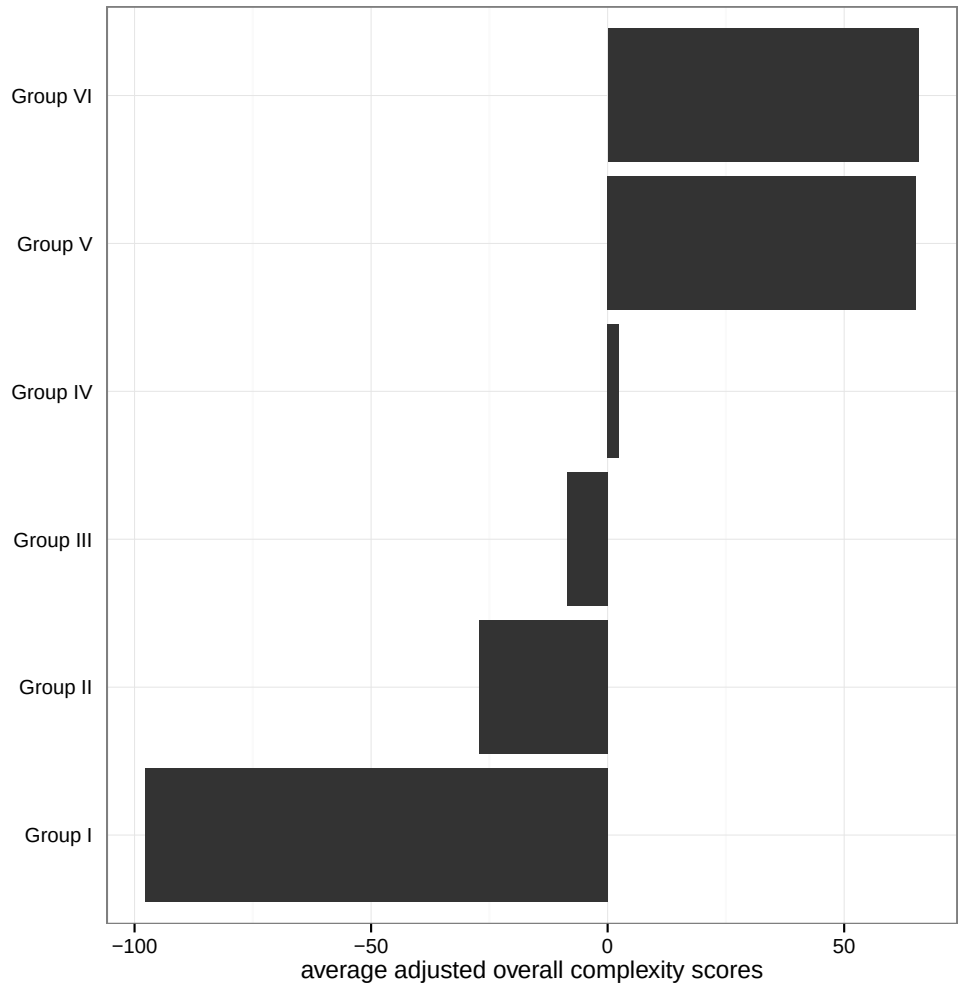

Figure 6.6.: Overall complexity hierarchy of German learner groups in ICLE. Negative residuals indicate below-average complexity; positive residuals indicate above-average complexity.

In the German national subset, the results in regard to the morphological and syntactic complexity of the six groups of instructional exposure dovetail with the findings of the complete ICLE study (Figure 6.8). The morphological complexity of the texts increases with higher amounts of instructional exposure while syntactic complexity decreases. Taking instructional exposure as a proxy for (writing) proficiency, this is another way of saying that advanced proficiency leads to more morphological complexity but less syntactic complexity in written learner production. Specifically, text I, which was produced by learners of the lowest level of proficiency, is the syntactically most complex but morphologically least complex text. The texts produced by learners with a medium amount of instructional exposure in English are scattered across the centre of the plot and are of average morphological and syntactic complexity. Texts VI and V, the texts which

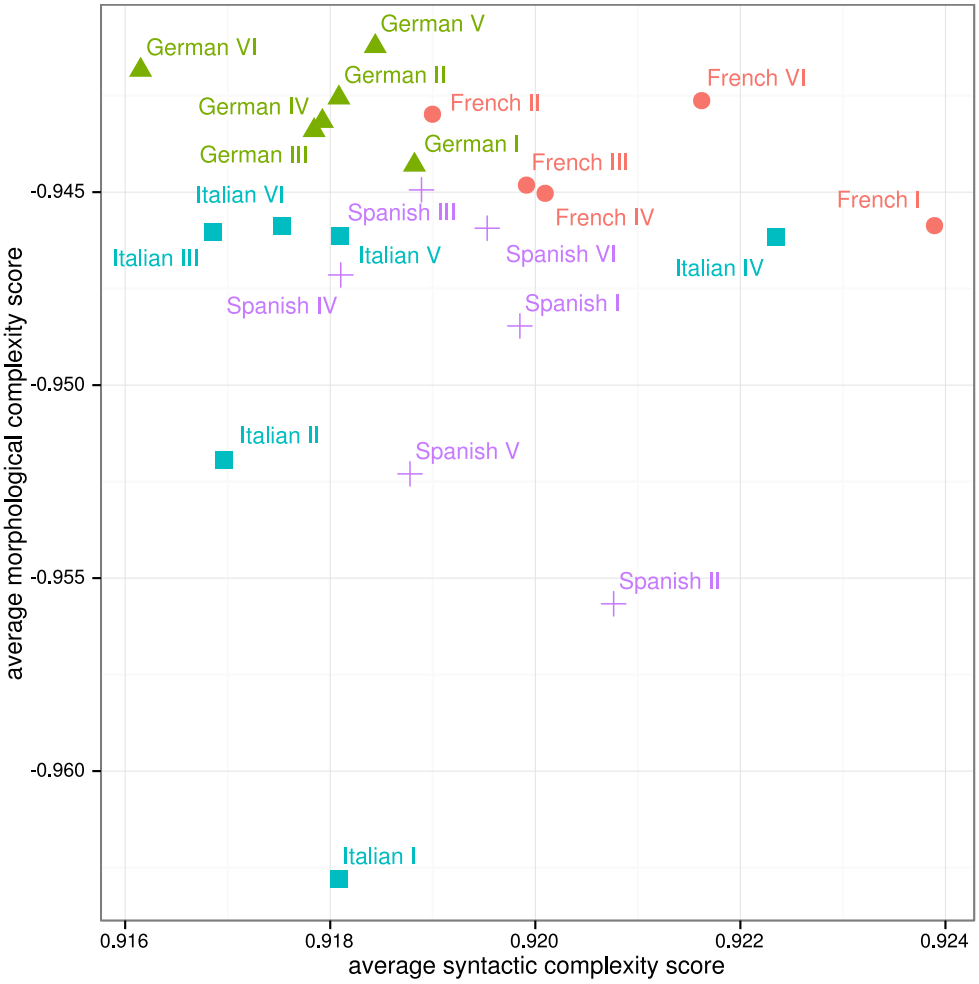

Figure 6.7.: Morphological complexity by syntactic complexity of ICLE learner groups by four different mother tongue backgrounds. Abscissa indexes increased syntactic complexity, ordinate indexes increased morphological complexity.

were produced by learners who have received the highest amount of instruction in English and can therefore be considered the most proficient learners in this set, are the morphologically most complex but syntactically least complex texts. Thus, morphological complexity trades off against syntactic complexity (Pearson's correlation coefficient $r = 0.97$, $p = 0.0007$). The statistical comparison of the morphological and syntactic complexity scores of the German national subset and the complete ICLE study indicates a very high correlation on the morphological level (Pearson's correlation coefficient $r = 0.81$, $p = 0.024$) and a high correlation on the syntactic level (Pearson's correlation coefficient $r = 0.72$, $p = 0.054$).

To sum up, the analysis of the German national subset demonstrates that the overall complexity of written learner production increases with

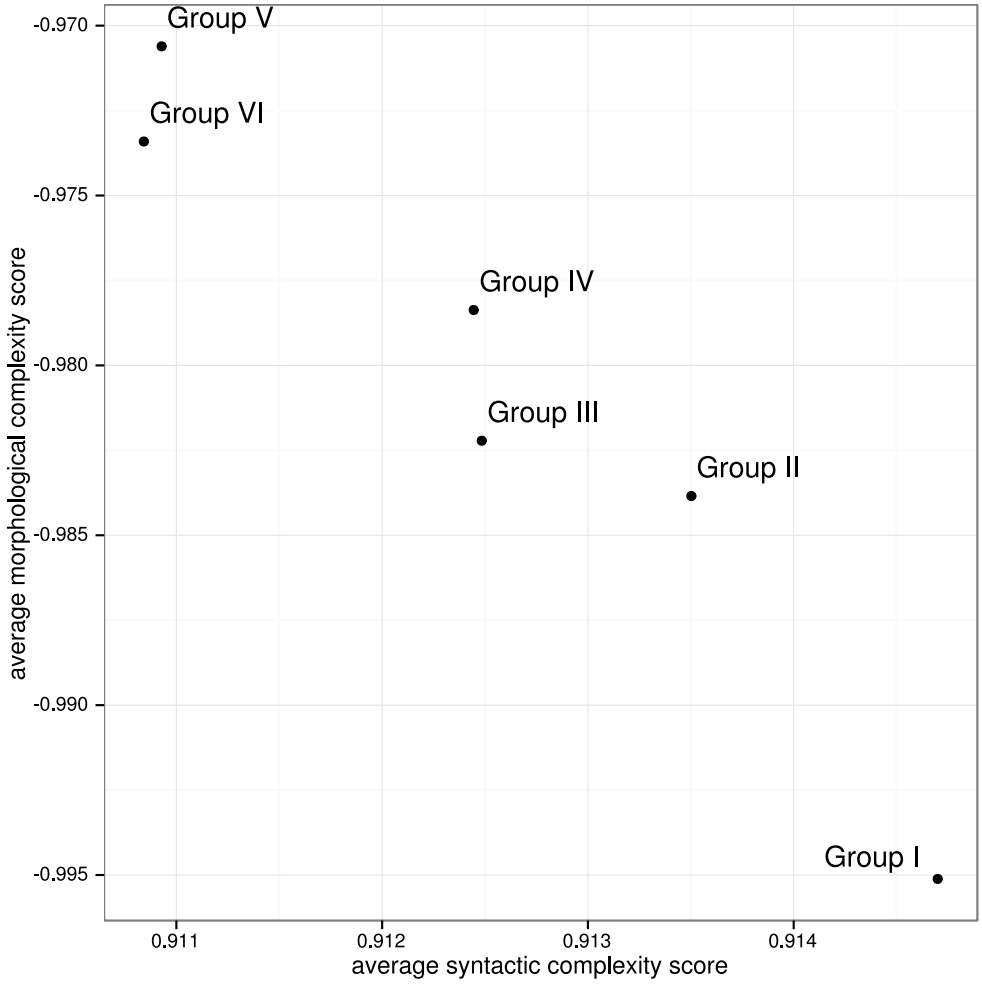

Figure 6.8.: Morphological complexity by syntactic complexity of German learner groups in ICLE. Abscissa indexes increased syntactic complexity, ordinate indexes increased morphological complexity.

increasing proficiency. The results of the four nationality subset seem to exhibit the same tendency within each national subgroup but are clearly flawed by data sparsity. Needless to say, further exploration with a larger dataset and learners with more different mother tongue backgrounds is required. Furthermore, in both subsets higher amounts of instructional exposure lead to an increase in morphological complexity and a decrease in syntactic complexity. These observations hold across different national varieties and confirm the measurements of the complete ICLE dataset which does not distinguish between the mother tongue backgrounds of the learners. In short, while mother tongue background does to some extent matter, the relationship between instructional exposure and complexity is quite robust across different mother tongue backgrounds.

6.3. Summary

In this chapter intra-linguistic variability in regard to overall, morphological and syntactic complexity was analysed in two different datasets drawn from the *British National Corpus* and the *International Corpus of Learner English*. In more general terms, this chapter was concerned with the applicability of the compression technique to naturalistic corpus resources.

In a first case study, the complexity variation across different written registers of British English as sampled in the BNC was analysed. I find that more formal registers are overall more complex than less formal registers, i.e. I observe variation along Biber's (1988) informational vs. involved dimension. Furthermore, morphological complexity positively correlates with formality of the register such that more formal / abstract-informational registers are more morphologically complex than less formal registers. On the other hand, more informal / involved registers tend towards more syntactic complexity.

In a second case study, the relationship between complexity and the amount of instructional exposure in ICLE was assessed. The measurements indicate that higher amounts of instructional exposure in English—which is taken as a proxy for writing proficiency in this study—leads to higher overall and morphological complexity in the written production of learners. Writing produced by less advanced learners, on the other hand, is marked by higher syntactic complexity and lower morphological complexity. Furthermore, Kolmogorov complexity measures of learner language are systematically correlated with more traditional SLA metrics (cf. Ehret & Szmrecsanyi (2016a)).

The analysis of two small subsets of essays split up according to national variety suggests that mother tongue background influences production complexity but that the relationship between instructional exposure and complexity, reported for ICLE as a whole is quite stable across different mother tongue backgrounds. For reasons of data availability and distribution these results should be treated with caution. Further research with a much extended dataset in terms of size is needed to shed more light on the (in)variance of cross-linguistic learner language complexity. The ICLE study moreover emphasizes that the success of the compression technique—as most quantitative methodologies—relies on the quality of the input. In other words, the technique requires not only a large quantity of data but these data should be equally distributed across the analysed samples to ensure the reliability, comparability and representativity of the measurements.

All in all, this chapter demonstrated that the compression technique can be successfully applied to naturalistic corpus resources, and yields results which tie in with findings obtained by more traditional means, while at the same time being more economical and objective.

7. Summary and Conclusion

This chapter starts with a short overview of the research context. After this a summary of the results, both linguistic and methodological, will be provided and the research questions posed in the introduction addressed. I will conclude with a discussion of the advantages and drawbacks of the compression technique, and give an outlook on future research.

Before diving into the subject matter, let us briefly recount the theoretical background of, and motivation for this work. Situated at the intersection of information theory and research on linguistic complexity (in)variation, the major focus of this work was the exploration, development, and advancement of the compression technique. My research was motivated by the recent interest in the definition and measurement of linguistic complexity, and the lack of a metric which is both objective and economical. The majority of previous studies on linguistic complexity is either based on empirically expensive evidence, i.e. evidence which is labour-intensive to obtain, and therefore difficult to replicate; or it is selective and hence subjective in nature because the methodologies used require the *a priori* definition, categorisation and selection of features serving as input for analyses.

In contrast to these "traditional" methods, the compression technique combines radical objectivity and economy with a usage-based holistic approach to measuring linguistic complexity in corpora. The basic idea of the compression technique is to measure linguistic complexity with compression algorithms. In technical terms, compression algorithms use adaptive entropy estimation methods to approximate Kolmogorov complexity, and thus measure the complexity of a given text sample by the length of the shortest possible description of this text sample. Consequently, string (1-a) is less complex than string (1-b) (see Example (1) below). Although both strings count the same number of orthographic characters, the length of their shortest possible description differs, such that string (1-a) can be compressed to the expression 5×cd counting four symbols, whereas the shortest description of string (1-b) is the string itself.

(1) a. cdcdcdcdcd (10 symbols) → 5×cd (4 symbols)
b. c4gh39aby7 (10 symbols) → c4gh39aby7 (10 symbols)

Generally, better compression rates of a given text sample indicate lower Kolmogorov complexity and thus lower linguistic complexity. Kolmogorov-based information-theoretic complexity is an absolute metric of complexity

and is essentially a measure of structural surface redundancy, i.e. the repetition of orthographic character sequences.

7.1. Results

This section provides summaries of the empirical chapters and presents the results in light of the research questions posed in the introduction.

(i) **Can compression algorithms be applied to data other than parallel copora?**

In previous algorithmic complexity research, the compression technique was only used with parallel text databases, which are basically translational equivalents of one text in different languages. Such parallel corpora have become popular in typological research (e.g. Auwera et al. 2005; Cysouw & Wälchli 2007), because they facilitate the comparability across different languages as observed differences due to the propositonal content of the texts can be ruled out (Wälchli 2007). I explored this technical limitation of the compression technique by applying it to several different data types; parallel, semi-parallel, and genuinely non-parallel texts. Furthermore, in two case studies the compression technique was applied to naturalistic corpora (see research question (iv) below).

Chapter 3 first established the validity of the Juola-style compression technique by measuring overall, morphological, and syntactic complexity in a parallel database of the Gospel of Mark in some historical varieties of English and a few other languages (Esperanto, Finnish, French, German, Hungarian, and Latin). The rankings dovetail with intuitions and rankings reported in the literature (Bakker 1998; Nichols 2009). In terms of overall complexity, for instance, Hungarian and Finnish are comparatively complex while Esperanto and the English varieties (exlcuding West Saxon) are comparatively simple. In terms of morphological and syntactic complexity, West Saxon, Finnish, and Latin are the morphologically most complex and syntactically least complex texts. In contrast, Basic English is the morphologically most simple and syntactically most complex text. Moreover, the measurements depict the historical drift of English from a rather morphologically complex language to a language that relies heavily on syntax to convey grammatical information. For example, the West Saxon text exhibits more morphological but less syntactic complexity than the English texts after 1066, such as the sample from the English Standard Version.

Having thus validated the methodology, its applicability to new data types was tested. Specifically, I utilised the compression technique

with a parallel and semi-parallel corpus of *Alice's Adventures in Wonderland* spanning nine European languages (Dutch, English, Finnish, French, German, Hungarian, Italian, Romanian, and Spanish), and two non-parallel datasets of newspaper articles in the same nine languages. In order to assess the degree to which algorithmic measurements are influenced by the propositional content of the samples, the complexity rankings of the semi-parallel Alice corpus and the newspaper corpora were statistically compared to the ranking obtained for the parallel Alice corpus, which served as control group. The syntactic rankings of all copora were furthermore correlated with a hierarchy of syntactic complexity reported in Bakker (1998).

The similarity between the parallel and semi-parallel Alice dataset was moderate in terms of overall complexity, but very high in terms of morphological and syntactic complexity. Dislocations in the complexity hierarchies were particularly notable among languages which neither tend to extreme morphological (such as Hungarian), nor extreme syntactic (such as English) complexity. This suggests that the propositional content might play a role in the measurement of balanced languages (such as German) in which content seems to influence the choice between morphologically and syntactically encoded information. The statistical comparison between the parallel Alice corpus and the two newspaper datasets ranged from low to moderate in terms of overall complexity, and from moderate to high in terms of morphosyntactic complexity. The performance differences of the compression technique on the two newspaper datasets—which sample articles on two and three different topics respectively—confirms that the propositional content of texts has an effect on the measurements.

Thus, the findings showed that the compression technique can in principle be applied to non-parallel data. Yet, the propositional content of the texts can have an influence on the results, whereby the degree of parallelity impacts more on the measurement of overall complexity than morphological and syntactic complexity. Content control of non-parallel corpora is therefore crucial for obtaining reliable measurements with the compression technique, especially in comparative studies. In a word, randomly chosen texts cannot be used as input.

This section also introduced a statistically more robust version of the compression technique which takes multiple measurements over several iterations of a compression and distortion loop, instead of relying on a single measure (as does the Juola-style compression technique (Juola 2008)).

In summary, I find that the algorithmically obtained complexity rankings are in tandem with the literature (Bakker 1998; Nichols 2009). Therefore, compression algorithms can be used with different data

types, i.e. semi-parallel and non-parallel texts, if content is controlled for. The applicability of the technique to large-scale naturalistic corpora will be discussed below.

(ii) **Can compression algorithms measure the complexity of specific linguistic features?**

The compression technique is an inherently holistic methodology which assesses the overall, morphological and syntactic complexity of texts as a whole, i.e. from a bird's eye perspective. In Chapter 4 the classic compression technique was combined with the systematic removal of specific target structures—a method I refer to as targeted file manipulation—in order to measure morphological and syntactic complexity in a detailed fashion. In this spirit, the contribution of a handful of morphological markers (e.g. *–ing*) and functional constructions (e.g. progressive aspect *be* + verb–*ing*) to the complexity in three different texts was measured, and their textual complexity was inferred. The chapter extended and complemented a first exploration of targeted file manipulation in a mixed-genre corpus of the same texts (Ehret 2014) by analysing to which extent the measurements are text-dependent.

I found that generally, the presence of more morphological marker types generates more morphological complexity in the texts, while their presence enhances the algorithmic prediction of syntactic patterns. The constructions which were analysed (progressive, passive and perfect) increase information-theoretically measured morphological complexity in the texts but decrease syntactic complexity. Invariant markers such as the future markers *will* and *going to*, in contrast, hardly affected the complexity of the texts. In fact, they were found to increase simplicity. From a linguistic perspective, these results were unsurprising as they are in accordance with the literature (Arends 2001; McWhorter 2001a, 2012; Kusters 2008; Szmrecsanyi & Kortmann 2009; Trudgill 2004). From a methodological perspective, however, this was taken as evidence for the effectiveness of the compression technique, and demonstrated that compression algorithms can be used to measure morphological and syntactic complexity in a detailed fashion. Furthermore, I demonstrated that the measurements obtained through targeted file manipulation are largely text-independent. Although the precise amount of algorithmic complexity measured depends to some extent on the complexity of a given text, the general complexity trends hold across different texts.

(iii) **What do compression algorithms, linguistically speaking, actually measure?**

One of the major objectives of this work was to understand the work-

ings of the compression algorithm that underlies the compression technique, and thus provide a linguistic definition of Kolmogorov-based information-theoretic complexity.

Chapter 5 therefore looked into the black box of the algorithm by analysing `gzip` 's lexicon output for *Alice's Adventures in Wonderland*, a line-by-line output of text sequences which the algorithm recognised, and on whose basis the input text was compressed. The strings in the lexicon output were manually annotated according to linguistic and non-linguistic categories such as lexical or functional words, other linguistically interpretable sequences, phrasal sequences, random non-linguistically meaningful sequences or mixed sequences (containing both meaningful and random sequences). Example (2) provides example sequences for each category. Spaces are part of compressed sequences and are represented by "␣" at the end or beginning of a string.

(2)
a. Lexical ␣opportunity␣
b. Functional her␣
c. Other ing␣
d. Phrasal do cats eat bats␣
e. Mixed dance t
f. Random omet

The in-depth analysis of the annotated lexicon revealed that about half of the compressed strings are linguistically meaningful units while the other half is made up of more or less random strings which cannot be interpreted in a linguistically meaningful way. Thus, I find that compression algorithms do indeed capture recurring linguistic structures, and compressed strings often coincide with linguistically meaningful units such as verbs or suffixes. However, the algorithm does not possess any linguistic knowledge and does not prefer linguistic structures over random patterns. For this reason, the algorithm also compresses random strings, or strings which on the surface resemble linguistic units. This means that the compression algorithm works on the form of structures, not on their function or meaning. Consequently, compression algorithms measure structural surface redundancy.

This characteristic of Kolmogorov-based information-theoretic complexity implies that the methodology has a slight tendency to favour morphological complexity, and should be used with caution when measuring overall complexity in text samples. In plain English, the method heavily relies on, and is sensitive to structural (ir)regularities and redundancy. Therefore, very high structural redundancy can result in low overall complexity and vice versa.

In order to assess the impact of distortion on the nature of com-

pressed strings, the lexicon output of a morphologically and a syntactically distorted version of *Alice's Adventures in Wonderland* were semi-automatically annotated for the (non-)linguistic categories introduced above, and compared to the lexicon output of the original Alice text. This analysis showed that the compression technique and the morphological and syntactic distortion of texts, which is necessary for the measurement of morphological and syntactic complexity respectively, works as intended.

Let us rehearse how distortion was implemented. The idea was to address morphological and syntactic complexity by distorting the information at the respective level in the texts prior to compression, and then measuring the impact of distortion on the compressibility of the texts. To be more precise, morphological distortion was implemented as random deletion of orthographically transcribed characters. This is supposed to create random noise and thus increase the morphological complexity in the distorted texts. Text samples with a large variety of word forms should not be as badly affected as text samples with a relatively small amount of word forms in which distortion creates comparatively more random noise. Therefore, the compression algorithm should perform comparatively worse on text samples with low morphological complexity. Syntactic distortion on the other hand, was implemented as random deletion of word forms, thus disrupting word order regularities. The impact of syntactic distortion on text samples with a relatively simple syntax—defined here as free word order—should not be as great as on text samples with complex syntax, i.e. fixed word order. It follows that the performance of the algorithm should be comparatively worse on text samples with high syntactic complexity.

The survey of string distribution in the distorted lexica confirmed first, that morphological distortion does indeed lead to an increased amount of "word forms" in the distorted text resulting in an increased number of unique strings in the lexicon which are difficult to compress. Second, syntactic inter-dependencies and word order regularities in the text are compromised through distortion as intended. This is reflected in the fact that strings which the algorithm added to the lexicon of the original Alice text were not recognised, and thus omitted in the syntactically distorted lexicon. On a linguistic level, the morphologically distorted lexicon contains a higher percentage of linguistically non-meaningful strings than the original Alice lexicon, while the syntactically distorted lexicon contains less strings altogether.

In short, compression algorithms do not intentionally measure or count linguistic features because they do not possess any knowledge of form-function pairings. Rather, Kolmogorov-based information-theoretic

complexity is a measure of structural surface redundancy.

(iv) **How well do compression algorithms capture intra-linguistic, i.e. within language, complexity variation in naturalistic corpora?**

The red thread running through all studies presented in this work is the question of the compression technique's applicability to various data types, among others its applicability to naturalistic corpus resources. In the spirit of the Freiburg School of complexity research (Kortmann & Szmrecsanyi 2009; Szmrecsanyi & Kortmann 2009), which focuses on analysing variation within varieties or dialects of a language (Kortmann & Szmrecsanyi 2012: 14), its applicability was tested by measuring intra-linguistic complexity variation in two large-scale corpora of English. Specifically, a random sampling technique was used to approximate the overall, morphological and syntactic complexity in written British English registers and learner essays produced by students who received different amounts of instruction in English.

The study on register variability drew on data from the *British National Corpus* (BNC) and analysed twenty written registers ranging from newspapers and biography to emails and administrative writing. I found that in terms of overall complexity less formal registers are less complex than more formal registers, outliers notwithstanding. Particularly the overall complexity ranking of the three newspaper registers (broadsheet, regional and tabloid) depicted this trend in a text-book style fashion (see Figure 7.1 for illustration).

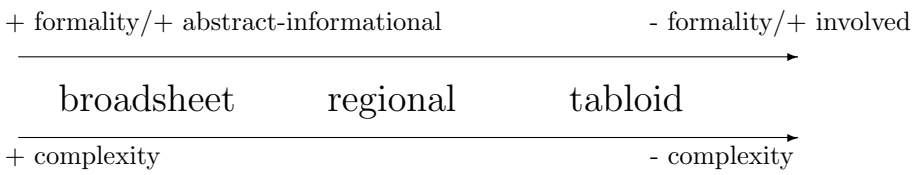

Figure 7.1.: Overall complexity ranking of newspaper registers in the BNC.

In terms of morphological and syntactic complexity, less formal registers tend to be more syntactically complex but less morphologically complex than formal registers. Altogether, these findings are in line with the register variation described by Biber (1988) along the involved and abstract-informational dimensions, i.e. more involved registers should be less complex while abstract-informational registers should be more complex, and roughly with Szmrecsanyi (2009).

On a methodological plane, the results of the BNC study furthermore indicated that the reliance of the compression technique on surface

redundancy influences the overall complexity scores of some registers. For instance, the low overall complexity score for administrative writing is most likely a result of the low morphological complexity score of this register.

The second analysis is based on the *International Corpus of Learner English* (ICLE) which samples essays produced by students with different levels of instructional exposure in English. The complexity variation of these texts at the overall, morphological and syntactic level was measured and the influence of the learners' national background, i.e. mother tongue, was evaluated for two subsets of the corpus. Keeping all other things equal, the amount of instructional exposure in English is considered a proxy for the learners' level of proficiency in English. Higher amounts of instructional exposure in English lead learners to produce texts with higher overall and morphological complexity. On the other hand, texts produced by less advanced learners are comparatively syntactically more complex. These findings seemed to be largely independent of the learners' national background: although mother tongue does lead to differences in complexity between essays of learners with different national backgrounds, the relationship between the amount of instruction in English and essay complexity is overall quite robust. Kolmogorov complexity in learner essays generally increases with increased instructional exposure. Furthermore, Kolmogorov measurements of learner language correlate with commonly used SLA complexity measures in a systematic way. Nevertheless, further research with larger datasets is required to back-up my findings, because as with most quantitative methodologies, the reliability, comparability and representativity of compression measurements depend on the quality of the input data. Put in other words, data sparsity and distribution are crucial factors impacting on the quality of the results obtained by compression.

Concludingly, I can thus say that the compression technique can be succesfully applied to naturalistic corpora, and considering data availability and distribution, is a reliable means for measuring intra-linguistic complexity. Moreover, the compression technique could potentially serve as diagnostic tool for language proficiency in Second Language Acquisition testing (e.g. TOEFL or IELTS testing).

7.2. Discussion

After having presented the results of my research, this section offers a more general discussion of the methodological aspects and features of the compression technique, and considers its advantages and drawbacks. I conclude with outlining future applications and research with the compression tech-

nique.

Objectivity and economy. In contrast to previous measures and methodologies used in complexity research, the use of compression algorithms allows for unparalleled objectivity and economy. Firstly, the compression technique is radically objective and unsupervised because compression algorithms are totally agnostic about form-function pairings, i.e. they do not possess any linguistic knowledge or understanding of the text they compress. This property is often quoted as a major flaw of the methodology. However, it is this "flaw" which makes the compression technique a radically objective tool for measuring linguistic complexity in texts. Secondly, the compression technique is an economical means for measuring linguistic complexity because it can be easily implemented and applied to any orthographically transcribed text database—assuming a certain amount of content control is executed. In fact, the compression technique does not rely on empirically expensive evidence (e.g. elaborately coded data, manually selected features etc.) and results can therefore be replicated with comparatively little effort.

Text-dependence. This brings us to the next point: the compression technique is a strictly text-based methodology. Referring to Cysouw & Good (2013: 338) who have recently coined the term *doculect* to refer to a variety represented by, and documented in a given text source, the compression technique could be said to measure complexity in doculects. This is another way of saying that the compression technique and the resulting measurements are to some extent text-dependent. Again, this characteristic can be considered both an advantage and a drawback. On the one hand, text-dependence is a drawback because the technique depends on orthographic transcription conventions—variant spellings of the same word inflate complexity because the algorithm encounters two different strings (e.g. *neighbourhood* and *neighborhood*) which cannot be compressed, instead of one string twice (2 × *neighbourhood*). On the other hand, the technique depends on the content of the texts. While orthography is an issue that can be easily addressed, for instance by normalising spelling variation in historical texts[1], content can only be controlled for to a limited degree, particularly in large copora as the content needs to be manually screened or selected. Content control, i.e. the control or screening of the propositional content of texts, is an important factor for the successful application of the compression technique to non-parallel and, to a lesser extent, to naturalistic corpora. This is because content

[1]Spelling is normalised by replacing two variant forms (e.g. *broðor* and *broðer*) with one form (e.g. *broðer*) throughout a text.

control ensures that the measurements obtained are robust, representative and reliable. Therefore, the compression technique cannot be used with randomly chosen texts. In other words, the components of non-parallel (cross-linguistic) corpora should be comparable in terms of their propositional content or topic. Additionaly, extra-linguistic variables such as sample size / text length should be held constant to ensure comparability (more details on sample size are given under the topic 'Data preparation' below).

Furthermore, text-dependence plays a role in targeted manipulation, the measurement of specific features with compression algorithms. Measurements obtained through targeted manipulation depend on the complexity of the texts in which a certain feature is measured. Still, it was shown that the trends of the measurements hold across different texts.

The fact that the compression technique is text-based is also one of its major advantages and inherent features. For one, it is usage-based because the measurements are based on written or spoken texts which were produced by language users, and thus permit the study of naturally occurring language phenomena in their usage context. Another advantage is that the compression technique is holistic because it works directly on the data (texts) and its input is not restricted to, for instance, a selection of certain features.

Data availability. It is a well known fact that the output quality of quantitative methodologies largely depends on the quality of the input. This is also valid for the compression technique. The complexity measurements obtained through compression are more robust and representative, if they are based on larger datasets (i.e. at least 10,000 words of running text). Furthermore, the data should be equally distributed across the samples measured and compared, i.e. the samples should be of the same size. Data availability therefore needs to be considered when using the compression technique, especially when applying it to other than parallel data.[2]

Data preparation. Another factor which goes hand in hand with data quality is data preparation. Raw texts cannot serve as input for the compression technique but need to be carefully prepared lest the compressibility of the texts is compromised, and their complexity inflated: non-textual information of any kind such as corpus metadata, markup or XML codes as well as punctuation and non-alphabetical charac-

[2]In parallel corpora content differences can be ruled out. Therefore, it can be safely assumed that any observed differences in the data do not stem from data distribution or size but from the samples' complexity. However, a minimum sample size of a few thousand words is still required.

ters (e.g. numbers, special UTF-8 characters or byte order marks which can be found in data retrieved from the web) and superfluous (double or multiple) whitespaces need to be removed. Depending on the database, orthography has to be normalised, i.e. in historical texts, for instance, spelling variants of the same form (e.g. *broðor* versus *broðer*) need to be replaced by one form (e.g. *broðer*) as they affect the compressibility of the texts.

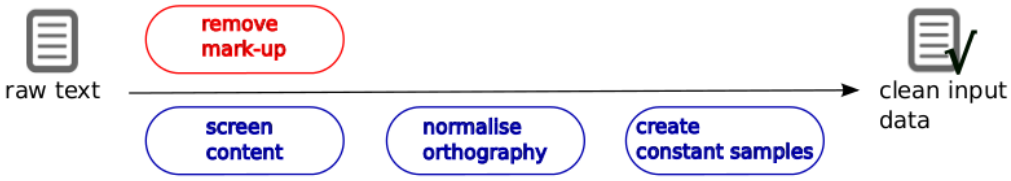

Figure 7.2.: Analytical pipeline for the creation of clean input data for the compression technique. Steps required with every data type are marked in red.

In addition to these factors, sample size and content control have to be considered when working on non-parallel data. As mentioned above, non-parallel text samples can only be compared if they are of the same size—because the propositional content of such texts is not identical and hence constant. If non-parallel samples are not of the same size, observed differences cannot be reliably attributed to a difference in complexity. Be reminded that the complexity measurements obtained with the compression technique are related to the text size and the compressibility of the text samples. For this reason, I have introduced the constant variable "number of sentences". By sampling the same number of sentences per text sample as opposed to words or characters, the sample size is held constant and syntactic interdepencies are left intact at the same time. Last but not least, content control is a crucial factor for the successful application of the compression technique to non-parallel text databases. As this point has already been stressed elsewhere (see 'Text-dependence' above), suffice it to say that random texts cannot be used as input for the compression technique but the propositional content should be similar. Figure 7.2 depicts the analytical pipeline for turning raw texts into appropriate input data for the compression technique.

Structural redundancy. This work determined Kolmogorov-based information-theoretic complexity as a measure of structural surface redundancy. In this context, structural surface redundancy refers to the recurrence of orthographically transcribed character sequences in a text. This definition implies that the complexity measured with the

compression technique has a slight tendency of favouring morphological complexity, i.e. it is morphology-sensitive, as morphological complexity and structural redundancy are somewhat correlated. This was reflected in some of the overall complexity measurements of the Gospel of Mark (see Chapter 3, Section 3.1.2) and the BNC register analysis (see Chapter 6, Section 6.1). In other words, overall complexity measurements can be affected by large amounts of structural redundancy in the texts such that texts with high structural redundancy exhibit high overall complexity. While this is not a major caveat, researchers need to be aware of this characteristic when using the compression technique for measuring overall complexity in corpora.

Complexity axes. Last but not least, a word on the polarity of the axes of morphological and syntactic complexity is in order. In this work, syntactic complexity was defined in terms of word order (rules), that is to say maximally free word order (more variation of word order patterns and less word order rules) is defined as maximally simple. This might seem counterintuitive because Kolmogorov complexity is related to the predictability of sequences in a text and fixed word order (patterns) should be more predictable than free word order. Simply put, fixed word order should be Kolmogorov-simple. However, Kolmogorov-based syntactic complexity is an "inverted" or "indirect" measure of syntactic complexity as it is measured through distortion. Syntactic distortion is basically the deletion of words in a text and permits the subsequent measurement of its impact on the predictability in this text. If the text is comparatively less predictable after distortion, the text is considered syntactically complex. Hence, fixed word order must be defined as Kolmogorov-complex. Consequently, the axis of syntactic complexity in the graphs is poled such that fixed word order counts as complex and free word order counts as simple.

The polarity of the morphological complexity axis, on the contrary, is arbitrarily set such that more morphological (structural) (ir)regularity counts as more morphologically complex. This definition is in accordance with definitions of morphological complexity customarily used in typological research (e.g. McWhorter 2001b; Szmrecsanyi & Kortmann 2009).

Finally, what do we gain by using compression algorithms? The compression technique is an objective and economical tool for measuring linguistic complexity in texts. It thus constitutes an independent analysis tool in complexity research that forgoes the labour-intensive, manual selection of more traditional metrics. Yet, it could also be employed as a short-cut or complementary diagnostic tool to gauge and / or confirm general complexity trends before applying more orthodox methods. Moreover, the compression

technique could be used to measure language performance: In second language acquisition contexts where proficiency needs to be tested and assessed such as IELTS (International English Language Testing System) or TOEFL (Test Of English as a Foreign Language) the compression technique could potentially serve as diagnostic tool for assessing language proficiency.

Yet, the full potential of the compression technique introduced in this work is not yet fully explored and awaits further study, both in terms of the range of its applicability and in terms of the implementation of the methodology itself.

The current work has left plenty of lose ends to be tied up: Future research should take a closer look at intra-linguistic complexity variation in learner varieties in order to establish whether the findings based on ICLE (see Section 6.2) can be confirmed when tapping other (larger) databases of learner English. Specifically, the question of whether the mother tongue background of learners influences the complexity of learner English is in need of further examination. The possibility of measuring historical complexity variation and change was very briefly discussed in Section 3.1 where different historical English Bible texts were compared, and the shift of English from a morphologically rich language to a language that encodes grammatical information predominantly through syntax was depicted. This area of study deserves much more attention in future algorithmic complexity research, especially as large-scale historical copora of English such as ARCHER (*A Representative Corpus of Historical English Registers*) or COHA (*Corpus of Historical American English*) are available and lend themselves well for analyses with the compression technique. Researchers could, for example, analyse complexity changes in historical English registers comparing British and American English, or might even be able to trace colloquialisation, the approximation of written language to spoken language (for a detailed discussion see Hundt & Mair 1999), by means of algorithmic complexity measurements. It has been reported that spoken language is characterised by high degrees of clausal embedding, which could be said to very roughly correspond to syntactic complexity, while written language is rather characterised by higher degrees of phrasal complexity and lexical density, which could both be said to correspond to morphological complexity (Biber et al. 2011: 29–32; see also Biber 1988: 113–114). In the context of algorithmic complexity research, it should be possible to prove a colloquialisation of written language by measuring a shift from morphological complexity to syntactic complexity, or at least by measuring an increase in syntactic complexity in written texts over time. Furthermore, the compression technique might also find application in measuring complexity in child language acquisition data and spoken language corpora. Research into geographical complexity variation in English is currently being prepared by the author. More generally, algorithmic complexity variation in languages and varieties other than English is completely underresearched and should be investigat-

ed.

Apart from extending algorithmic complexity research to different fields of linguistics, the method itself can also be extended. This work, and much of the previous research in this field, has focused on measuring overall, morphological and syntactic complexity—apart from Juola (2008) who also addressed pragmatic complexity. However, other sub domains of language might also be measurable with compression algorithms. Phonological and phonetic complexity, for example, could be addressed by compressing orthographically transcribed phonological and phonetic transcripts. Syntactic complexity might also be measured in more elaborate ways by compressing syntactic tree structures or syntactic annotations. It could also be interesting to make compression algorithms less "ignorant", i.e. `gzip` , for instance, could be equipped with linguistic knowledge so that the algorithm would only compress linguistically meaningful units such as words. This could be achieved by restricting the compressible sequences to words listed in a dictionary that is fed to the algorithm during compression. Such an algorithm would of course no longer be universally applicable and objective, as it would be tuned to a certain language and possess *a priori* knowledge of linguistically meaningful sequences.

To conclude, this work has provided a first in-depth linguistic analysis of compression algorithms and their application in corpus-based complexity research. All in all, compression algorithms provide an easy, efficient and economic way for measuring linguistic complexity.

Bibliography

Adrian Akmajian, Richard A. Demers, Ann K. Farmer, and Robert M. Harnish. *Linguistics: An Introduction to Language and Communication*. MIT Press, Cambridge, 4th edition, 1997.

Jacques Arends. Simple grammars, complex languages. *Linguistic Typology*, 5(2/3): 180–182, 2001.

Guy Aston and Lou Burnard. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press, Edinburgh, 1998.

Johan van Auwera, Eva Schalley, and Jan Nuyts. Epistemic possibility in a Slavonic parallel corpus – a pilot study. In Björn Hansen and Petr Karlik, editors, *Modality in Slavonic languages. New perspectives*, pages 201–217. Otto Sagner, München, 2005.

R. Harald Baayen. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press, Cambridge, 2008.

Dik Bakker. Flexibility and consistency in word order patterns in the languages of Europe. In Anna Siewierska, editor, *Constituent order in the languages of Europe*, pages 384–419. Mouton de Gruyter, Berlin, 1998.

Max Bane. Quantifying and measuring morphological complexity. In Charles B. Chang and Hannah J. Haynie, editors, *Proceedings of the 26th West Coast Conference on Formal Linguistics*, pages 69–76. Cascadilla Proceedings Project, Somerville, MA, 2008.

Heike Behrens. Usage-based and emergentist approaches to language acquisition. *Linguistics*, 47(2):383–411, 2009.

Christian Bentz and Bodo Winter. Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change*, 3:1–27, 2013.

Yves Bestgen and Sylviane Granger. Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, 26:28–41, 2014.

Douglas Biber. *Variation across Speech and Writing*. Cambridge University Press, Cambridge, 1988.

Douglas Biber and Bethany Gray. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, 9: 2–20, 2010.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. *Longman grammar of spoken and written English*. Longman, Harlow, 1999.

Douglas Biber, Bethany Gray, and Kornwipa Poonpon. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45(1):5–35, 2011.

Derek Bickerton. *Language and Human Behaviour*. University of Washington Press, Seattle, 1995.

Walter Bisang. On the evolution of complexity: sometimes less is more in East and mainland Southeast Asia. In Geoffrey Sampson, David Gil, and Peter Trudgill, editors, *Language Complexity as an Evolving Variable*, pages 34–49. Oxford University Press, Oxford, 2009.

Vaclav Brezina and Gabriele Pallotti. Morphological complexity tool, available from http://corpora.lancs.ac.uk/vocab/analyse_morph.php. 2015.

Joan Bybee. From usage to grammar: the mind's response to repetition. *Language*, 82: 711–733, 2006.

Joan Bybee. *Language, usage and cognition*. Cambridge University Press, Cambridge, 2010.

Ferrer i Ramon Cancho and Ricard V. Solé. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America*, 100(3):788–791, 2003.

Rudi Cilibrasi and Paul M. B Vitányi. Clustering by compression. *IEEE Transactions on Information Theory*, 51(4):1523–1545, April 2005.

David Crystal. *The Cambridge Encyclopedia of Language*. Cambridge University Press, Cambridge, 1987.

Peter W. Culicover. *Grammar and Complexity: Language at the Intersection of Competence and Performance. 1st ed. Oxford Linguistics*. Oxford University Press, Oxford, 2013.

Michael Cysouw and Jeff Good. Languoid, doculect and glossonym: Formalizing the notion 'language'. *Language Documentation and Conservation*, 7:331–359, 2013.

Michael Cysouw and Bernhard Wälchli. Parallel texts: using translational equivalents in linguistic typology. *Language Typology and Universals*, 60(2):95–99, 2007.

Michael Cysouw, Chris Biemann, and Matthias Ongyerth. Using Strong's Numbers in the Bible to test an automatic alignment of parallel texts. *Language Typology and Universals*, 60(2):158–171, 2007.

Östen Dahl. *The growth and maintenance of linguistic complexity*. John Benjamins, Amsterdam/Philadelphia, 2004.

Östen Dahl. From questionnaires to parallel corpora in typology. *Language Typology and Universals*, 60(2):172–181, 2007.

Östen Dahl. Grammatical resources and linguistic complexity: Sirionó as a language without NP coordination. In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors, *Language Complexity: Typology, Contact, Change*, pages 153–164. John Benjamins, Amsterdam/Philadelphia, 2008.

Östen Dahl. Testing the assumption of complexity invariance: the case of Elfdalian and Swedish. In Geoffrey Sampson, David Gil, and Peter Trudgill, editors, *Language Complexity as an Evolving Variable*, pages 50–63. Oxford University Press, Oxford, 2009.

Rick Dale and Gary Lupyan. Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in Complex Systems*, 15(03n04):1150017/1–1150017/16, 2012.

Antje Dammel and Sebastian Kürschner. Complexity in nominal plural allomorphy: A contrastive survey of ten Germanic languages. In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors, *Language Complexity: Typology, Contact, Change*, pages 243–262. John Benjamins, Amsterdam/Philadelphia, 2008.

Casper de Groot. Morphological complexity as a parameter of linguistic typology: Hungarian as a contact language. In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors, *Language Complexity: Typology, Contact, Change*, pages 191–215. John Benjamins, Amsterdam/Philadelphia, 2008.

Lourens de Vries. Some remarks on the use of Bible translations as parallel texts in linguistic research. *Language Typology and Universals*, 60(2):148–157, 2007.

Guy Deutscher. *Syntactic Change in Akkadian*. Oxford University Press, Oxford/New York, 2000.

Guy Deutscher. "Overall complexity": a wild goose chase? In Geoffrey Sampson, David Gil, and Peter Trudgill, editors, *Language Complexity as an Evolving Variable*, pages 243–251. Oxford University Press, Oxford, 2009.

John Edwards. *Multilingualism*. Penguin, London, 1994.

Katharina Ehret. Kolmogorov complexity of morphs and constructions in English. *Linguistic Issues in Language Technology*, 2(11):43–71, 2014.

Katharina Ehret and Benedikt Szmrecsanyi. Compressing learner language: an information-theoretic measure of complexity in SLA production data. *Second Language Research (Sage OnlineFirst)*, 2016a.

Katharina Ehret and Benedikt Szmrecsanyi. An information-theoretic approach to assess linguistic complexity. In Raffaela Baechler and Guido Seiler, editors, *Complexity, Isolation, and Variation*, pages 71–94. Walter de Gruyter, Berlin/Boston, 2016b.

Nick C. Ellis. Emergentism, connectionism and language learning. *Language Learning*, 48:631–664, 1998.

Gertraud Fenk-Oczlon and August Fenk. Complexity trade-offs between subsystems of language. In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors, *Language Complexity: Typology, Contact, Change*, pages 43–65. John Benjamins, Amsterdam/Philadelphia, 2008.

Benjamin W. Fortson. *Indo-European language and culture: an introduction*. Blackwell, Oxford, 2004.

Murray Gell-Mann and Seth Lloyd. Information measures, effective complexity, and total information. *Complexity*, 2(1):44–52, 1996.

David Gil. How complex are isolating languages? In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors, *Language Complexity: Typology, Contact, Change*, pages 109–132. John Benjamins, Amsterdam/Philadelphia, 2008.

David Gil. How much grammar does it take to sail a boat? In Geoffrey Sampson, David Gil, and Peter Trudgill, editors, *Language Complexity as an Evolving Variable*, pages 19–33. Oxford University Press, Oxford, 2009.

Sylviane Granger, Estelle Dagneaux, and Fanny Meunier, editors. *The International Corpus of Learner English: Handbook and CD-ROM*. Presses universitaires de Louvain, Louvain-la-Neuve, 2002.

Zhao Hong Han and Wai Man Lew. Acquisitional complexity: What defies complete acquisition in Second Language Acquisition. In Bernd Kortmann and Benedikt Szmrecsanyi, editors, *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, Linguae & Litterae, pages 192–217. Walter de Gruyter, Berlin/Boston, 2012.

John A. Hawkins. *Efficiency and Complexity in Grammars*. Oxford University Press, Oxford/New York, 2004.

John A. Hawkins. An efficiency theory of complexity and related phenomena. In Geoffrey Sampson, David Gil, and Peter Trudgill, editors, *Language Complexity as an Evolving Variable*, pages 252–268. Oxford University Press, Oxford, 2009.

Charles Francis Hockett. *A course in modern linguistics*. Macmillan, New York, 1958.

Magnus Huber. Syntactic and variational complexity in British and Ghanaian English. Relative clause formation in the written parts of the International Corpus of English. In Bernd Kortmann and Benedikt Szmrecsanyi, editors, *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, Linguae & Litterae, pages 218–242. Walter de Gruyter, Berlin/Boston, 2012.

Marianne Hundt and Christian Mair. 'Agile' and 'uptight' genres: the corpus-based approach to language change in progress. *International Journal of Corpus Linguistics*, 4:221–242, 1999.

Stig Johansson and Knut Hofland. *Frequency analysis of English vocabulary and grammar: based on the LOB corpus*. Clarendon, Oxford, 1989.

Patrick Juola. Measuring linguistic complexity: the morphological tier. *Journal of Quantitative Linguistics*, 5(3):206–213, 1998.

Patrick Juola. Assessing linguistic complexity. In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors, *Language Complexity: Typology, Contact, Change*, pages 89–107. John Benjamins, Amsterdam/Philadelphia, 2008.

Juhani Järvikivi, Pirita Pyykkönen-Klauck, and Matti Laine. Words & constructions: Language complexity in linguistics and psychology. *Special Issue of The Mental Lexicon*, 9(2), 2014.

Fred Karlsson. Origin and maintenance of clausal embedding complexity. In Geoffrey Sampson, David Gil, and Peter Trudgill, editors, *Language Complexity as an Evolving Variable*, pages 193–202. Oxford University Press, Oxford, 2009.

Kimmo Kettunen, Markus Sadeniemi, Tiina Lindh-Knuutila, and Timo Honkela. Analysis of EU languages through text compression. In Tapio Salakoski, Filip Ginter, Sampo Pyysalo, and Tapio Pahikkala, editors, *Advances in Natural Language Processing*, number 4139 in Lecture Notes in Artificial Intelligence, pages 99–109. Springer-Verlag, Berlin/Heidelberg, 2006.

Andrej Kolmogorov. On tables of random numbers. *Sankhya*, 25:369–375, 1963.

Andrej N. Kolmogorov. Three approaches to the quantitative definition of information. *Problemy Peredachi Informatsii*, 1(1):3–11, 1965.

Bernd Kortmann and Benedikt Szmrecsanyi. World Englishes between simplification and complexification. In Lucia Siebers and Thomas Hoffmann, editors, *World Englishes - Problems, Properties and Prospects: selected papers from the 13th IAWE conference*, pages 265–285. John Benjamins, Amsterdam/Philadelphia, 2009.

Bernd Kortmann and Benedikt Szmrecsanyi, editors. *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Lingua & Litterae. Walter de Gruyter, Berlin/Boston, 2012.

Wouter Kusters. *Linguistic Complexity: The Influence of Social Change on Verbal Inflection*. LOT, Utrecht, 2003.

Wouter Kusters. Complexity in linguistic theory, language learning and language change. In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors, *Language Complexity: Typology, Contact, Change*, pages 3–21. John Benjamins, Amsterdam/Philadelphia, 2008.

Ronald Langacker. Syntactic re-analysis. In Charles Li, editor, *Mechanisms of Syntactic Change*, pages 57–139. University of Texas Press, Austin, 1977.

Ronald Langacker. *Foundations of cognitive grammar, vol. 1: Theoretical prerequisites*. Stanford University Press, Stanford, CA, 1987.

Ronald Langacker. A usage-based model. In Brygida Rudzka-Ostyn, editor, *Topics in cognitive linguistics*, pages 127–161. John Benjamins, Amsterdam/Philadelphia, 1988.

Diane Larsen-Freeman. An ESL index of development. *TESOL Quarterly*, 12(4):439–448, 1978.

Diane Larsen-Freeman. The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, 27(4): 590–619, 2006.

John Leavitt. *Linguistic Relativities: Language diversity and modern thought*. Cambridge University Press, Cambridge, 2011.

Ming Li and Paul M. B Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer-Verlag, New York, 1997.

Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul M. B Vitányi. The similarity metric. *IEEE Transactions on Information Theory*, 50(12):3250–3264, 2004.

Eva Lindström. Language complexity and interlinguistic difficulty. In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors, *Language Complexity: Typology, Contact, Change*, pages 439–448. John Benjamins, Amsterdam/Philadelphia, 2008.

Gary Lupyan and Rick Dale. Language structure is partly determined by social structure. *PLoS ONE*, 5(1):1–10, 2010.

Utz Maas. Orality versus literacy as a dimension of complexity. In Geoffrey Sampson, David Gil, and Peter Trudgill, editors, *Language Complexity as an Evolving Variable*. Oxford University Press, Oxford, 2009.

David J. C MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, Cambridge, 2003.

John McWhorter. What people ask David Gil and why: Rejoinder to the replies. *Linguistic Typology*, 5(2/3), 2001a.

John McWhorter. The world's simplest grammars are creole grammars. *Linguistic Typology*, 6:125–166, 2001b.

John McWhorter. The rest of the story: Restoring pidginization to creole genesis theory. *Journal of Pidgin and Creole Languages*, 17(1):1–48, 2002.

John McWhorter. Why does a language undress? Strange cases in Indonesia. In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors, *Language Complexity: Typology, Contact, Change*, pages 167–190. John Benjamins, Amsterdam/Philadelphia, 2008.

John McWhorter. Oh noo!: A bewilderingly multifunctional Saramaccan word teaches us how a creole language develops complexity. In Geoffrey Sampson, David Gil, and Peter Trudgill, editors, *Language Complexity as an Evolving Variable*. Oxford University Press, Oxford, 2009.

John McWhorter. Complexity hotspot: The copula in Saramaccan and its implications. In Bernd Kortmann and Benedikt Szmrecsanyi, editors, *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, Linguae & Litterae, pages 243–246. Walter de Gruyter, Berlin/Boston, 2012.

Rajend Mesthrie. Deletions, antideletions and complexity theory, with special reference to Black South African and Singaporean English. In Bernd Kortmann and Benedikt Szmrecsanyi, editors, *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, Linguae & Litterae, pages 90–100. Walter de Gruyter, Berlin/Boston, 2012.

Matti Miestamo. On the feasibility of complexity metrics. In Krista Kerge and Maria-Maren Sepper, editors, *FinEst Linguistics. Proceedings of the Annual Finnish and Estonian Conference of Linguistics [Publications of the Department of Estonian of Tallinn University 8]*, pages 11–26. Tallinna Ülikooli Kirjastus, Tallinn, 2006.

Matti Miestamo. Grammatical complexity in a cross-linguistic perspective. In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors, *Language Complexity: Typology, Contact, Change*, pages 23–41. John Benjamins, Amsterdam/Philadelphia, 2008.

Matti Miestamo. Implicational hierarchies and grammatical complexity. In Geoffrey Sampson, David Gil, and Peter Trudgill, editors, *Language Complexity as an Evolving Variable*, pages 80–97. Oxford University Press, Oxford, 2009.

Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors. *Language Complexity: Typology, Contact, Change*. John Benjamins, Amsterdam/Philadelphia, 2008.

Fermin Moscoso del Prado Martin. The mirage of morphological complexity. In *Proc. of the 33rd Annual Conference of the Cognitive Science Society*, pages 3524–529, 2011.

Fermin Moscoso del Prado Martin, Aleksandar Kostic, and R. Harald Baayen. Putting the bits together: an information theoretical perspective on morphological processing. *Cognition*, 94(1):1–18, 2004.

Peter Mühlhäusler. *Pidginization and simplification of language*. Pacific Linguistics, Series B No. 26. Australian National University, Canberra, 1974.

Peter Mühlhäusler. The complexity of the personal and possessive pronoun system of Norf'k. In Bernd Kortmann and Benedikt Szmrecsanyi, editors, *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, Linguae & Litterae, pages 101–126. Walter de Gruyter, Berlin/Boston, 2012.

Frederick J. Newmeyer and Laurel B. Preston, editors. *Measuring Grammatical Complexity*. Oxford University Press, New York, 2014.

Johanna Nichols. *Linguistic Diversity in Space and Time*. University of Chicago Press, Chicago, IL, 1992.

Johanna Nichols. Linguistic complexity: a comprehensive definition and survey. In Geoffrey Sampson, David Gil, and Peter Trudgill, editors, *Language Complexity as an Evolving Variable*, pages 64–79. Oxford University Press, Oxford, 2009.

Johanna Nichols. The vertical archipelago: Adding the third dimension to linguistic geography. In Peter Auer, Martin Hilpert, Anja Stukenbrock, and Benedikt Szmrecsanyi, editors, *Space in language and linguistics: geographical, interactional, and cognitive perspectives*. Walter de Gruyter, Berlin/Boston, 2013.

Terence Odlin. Nothing will come of nothing. In Bernd Kortmann and Benedikt Szmrecsanyi, editors, *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, Linguae & Litterae, pages 62–89. Walter de Gruyter, Berlin/Boston, 2012.

William O'Grady, Michael Dobrovolsky, and Mark Aronoff. *Contemporary Linguistics: An Introduction*. St. Martin's Press, New York, 3rd edition, 1997.

Lourdes Ortega. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24:492–518, 2003.

Lourdes Ortega. Interlanguage complexity: A construct in search of theoretical renewal. In Bernd Kortmann and Benedikt Szmrecsanyi, editors, *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, Linguae & Litterae, pages 127–155. Walter de Gruyter, Berlin/Boston, 2012.

Gabriele Pallotti. A simple view of linguistic complexity. *Second Language Research*, 31 (1):117–134, 2015.

Mikael Parkvall. The simplicity of creoles in a cross-linguistic perspective. In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors, *Language Complexity: Typology, Contact, Change*, pages 265–285. John Benjamins, Amsterdam/Philadelphia, 2008.

Ljiljana Progovac. Layering of grammar: Vestiges of protosyntax in present-day languages. In Geoffrey Sampson, David Gil, and Peter Trudgill, editors, *Language Complexity as an Evolving Variable*. Oxford University Press, Oxford, 2009.

Elizabeth M. Riddle. Complexity in isolating languages: Lexical elaboration versus grammatical economy. In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors, *Language Complexity: Typology, Contact, Change*, pages 134–151. John Benjamins, Amsterdam/Philadelphia, 2008.

Markus Sadeniemi, Kimmo Kettunen, Tiina Lindh-Knuutila, and Timo Honkela. Complexity of European Union languages: A comparative approach. *Journal of Quantitative Linguistics*, 15(2):185–211, 2008.

David Salomon. *Data Compression. The complete Reference.* Springer-Verlag, London, 4 edition, 2007.

Geoffrey Sampson. A linguistic axiom challenged. In Geoffrey Sampson, David Gil, and Peter Trudgill, editors, *Language Complexity as an Evolving Variable*, pages 1–18. Oxford University Press, Oxford, 2009.

Geoffrey Sampson, David Gil, and Peter Trudgill, editors. *Language Complexity as an Evolving Variable*. Oxford University Press, Oxford, 2009.

August Wilhelm von Schlegel. *Observations sur la langue et littérature provençale*. Librairie Grecque-Latine-Allemande, Paris, 1818.

Friedrich von Schlegel. *Über die Weisheit und Sprache der Indier*. Mohr und Zimmer, Heidelberg, 1808.

Pieter Seuren and Herman Wekker. Semantic transparency as a factor in creole genesis. In Pieter Muysken and Norval Smith, editors, *Substrata versus Universals in Creole Genesis*, pages 57–70. John Benjamins, Amsterdam/Philadelphia, 1986.

Claude E Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

Naomi Lapidus Shin. Grammatical complexification in Spanish in New York: 3sg pronoun expression and verbal ambiguity. *Language Variation and Change*, 26:1–28, 2014.

Ryan K. Shosted. Correlating complexity: A typological approach. *Journal of Linguistic Typology*, 10:1–40, 2006.

Jeff Siegel. Accounting for analyticity in creoles. In Bernd Kortmann and Benedikt Szmrecsanyi, editors, *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, Linguae & Litterae, pages 35–61. Walter de Gruyter, Berlin/Boston, 2012.

Jeff Siegel, Benedikt Szmrecsanyi, and Bernd Kortmann. Measuring analyticity and syntheticity in creoles. *Journal of Pidgin and Creole Languages*, 29(1):49–85, 2014.

Kaius Sinnemäki. Complexity trade-offs in core argument marking. In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors, *Language Complexity: Typology, Contact, Change*, pages 67–88. John Benjamins, Amsterdam/Philadelphia, 2008.

Kaius Sinnemäki. Complexity in core argument marking and population size. In Geoffrey Sampson, David Gil, and Peter Trudgill, editors, *Language Complexity as an Evolving Variable*. Oxford University Press, Oxford, 2009.

Dan Slobin. Language change in childhood and in history. In John Macnamara, editor, *Language Learning and Thought*, pages 185–221. Academic Press, London, 1977.

Eugénie Stapert. Universals in language or cognition? Evidence from English language acquisition and from Pirahã. In Geoffrey Sampson, David Gil, and Peter Trudgill, editors, *Language Complexity as an Evolving Variable*. Oxford University Press, Oxford, 2009.

Maria Steger and Edgar W. Schneider. Complexity as a function of iconicity: The case of complement clause constructions in New Englishes. In Bernd Kortmann and Benedikt Szmrecsanyi, editors, *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, Linguae & Litterae, pages 156–191. Walter de Gruyter, Berlin/Boston, 2012.

Benedikt Szmrecsanyi. Typological parameters of intralingual variability: Grammatical analyticity versus syntheticity in varieties of English. *Language Variation and Change*, 21(3):319–353, 2009.

Benedikt Szmrecsanyi and Bernd Kortmann. Between simplification and complexification: non-standard varieties of English around the world. In Geoffrey Sampson, David Gil, and Peter Trudgill, editors, *Language Complexity as an Evolving Variable*, pages 64–79. Oxford University Press, Oxford, 2009.

Benedikt Szmrecsanyi and Bernd Kortmann. Introduction: Linguistic complexity, second language acquisition, indigenization, contact. In Benedikt Szmrecsanyi and Bernd Kortmann, editors, *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, Lingua & Litterae, pages 6–34. Walter de Gruyter, Berlin/Boston, 2012.

Michael Tomasello. *Constructing a language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, Cambridge, MA/London, UK, 2003.

Kristina Toutanova, Dan Klein, Chirstopher Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 252–259, 2003.

Peter Trudgill. Language contact and the function of linguistic gender. *Poznan Studies in Contemporary Linguistics*, 35:133–152, 1999.

Peter Trudgill. Contact and simplification: Historical baggage and directionality in linguistic change. *Linguistic Typology*, 5(2/3):371–374, 2001.

Peter Trudgill. Linguistic and social typology: The Austronesian migrations and phoneme inventories. *Linguistic Typology*, 8:305–320, 2004.

Peter Trudgill. Sociolinguistic typology and complexification. In Geoffrey Sampson, David Gil, and Peter Trudgill, editors, *Language Complexity as an Evolving Variable*, pages 98–109. Oxford University Press, Oxford, 2009a.

Peter Trudgill. Vernacular universals and the sociolinguistic typology of English dialects. In Markku Filppula, Juhani Klemola, and Heli Paulasto, editors, *Vernacular universals and language contacts : evidence from varieties of English and beyond*, pages 304–322. Routledge, New York, 2009b.

J. C. A. van der Lubbe. *Information theory*. Cambridge University Press, Cambridge [England]; New York, 1997.

Wilhelm von Humboldt. *Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluß auf die geistige Entwickelung des Menschengeschlechts*. Dümmler, Berlin, 1836.

Wilhelm von Humboldt. Ueber das Entstehen der grammatischen Formen, und ihren Einfluss auf die Ideenentwicklung. In Jürgen Trabant, editor, *Über die Sprache*, pages 52–81. Francke, Tübingen/Basel, 1994.

Warren Weaver. Recent contributions to the Mathematical Theory of Communication. In Clauce E. Shannon and Warren Weaver, editors, *The Mathematical Theory of Communication*, pages 94–117. The University of Illinois Press, Urbana, 8th edition, 1959.

Rulon Wells. Meaning and use. *Word*, 10:235–250, 1954.

Bernhard Wälchli. Advantages and disadvantages of using parallel texts in typological investigations. *Language Typology and Universals*, 60(2):118–134, 2007.

George Kingsley Zipf. *The psycho-biology of language: An introduction to dynamic philology*. Houghton-Mifflin, Boston, 1935.

George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press, Cambridge, MA, 1949.

Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, IT-23(3):337–343, 1977.

Appendices

A. Tables

A.1. Targeted file manipulation: Tukey's HSD tables

Table A.1.: Tukey's HSD for average syntactic and morphological complexity scores of inflectional morphs in Alice. The statistics include the difference between the means, the lower and upper bounds of the 95% confidence interval and the adjusted p-value per pair.

	Syntactic complexity				**Morphological complexity**			
Pair	Difference	Lower bound	Upper bound	p-value	Difference	Lower bound	Upper bound	p-value
orig-nns	0.0021	0.0027	3.160543e-03	0.0000	-0.0027	-0.0029	-2.366911e-03	0.0000
pos-nns	0.0001	-0.0001	3.529221e-04	0.8520	-0.0001	-0.0004	1.413095e-04	0.6855
vbd-nns	0.0011	0.0008	1.346065e-03	0.0000	-0.0007	-0.0010	-4.304618e-04	0.0000
vbg-nns	-0.0009	-0.0012	-6.968026e-04	0.0000	-0.0009	-0.0012	-6.364014e-04	0.0000
vbz-nns	-0.0002	-0.0004	8.046900e-05	0.3811	0.0008	0.0005	1.068852e-03	0.0000
pos-orig	-0.0028	-0.0031	-2.557288e-03	0.0000	0.0025	0.0022	2.799324e-03	0.0000
vbd-orig	-0.0018	-0.0021	-1.564145e-03	0.0000	0.0019	0.0016	2.227552e-03	0.0000
vbg-orig	-0.0039	-0.0041	-3.607012e-03	0.0000	0.0017	0.0014	2.021613e-03	0.0000
vbz-orig	-0.0031	-0.0033	-2.829741e-03	0.0000	0.0034	0.0031	3.726866e-03	0.0000
vbd-pos	0.0010	0.0007	1.243476e-03	0.0000	-0.0006	-0.0009	-2.806683e-04	0.0000
vbg-pos	-0.0010	-0.0013	-7.993919e-04	0.0000	-0.0008	-0.0011	-4.866079e-04	0.0000
vbz-pos	-0.0003	-0.0005	-2.212031e-05	0.0237	0.0009	0.0006	1.218645e-03	0.0000
vbg-vbd	-0.0020	-0.0023	-1.792535e-03	0.0000	-0.0002	-0.0005	8.516341e-05	0.3328
vbz-vbd	-0.0013	-0.0015	-1.015263e-03	0.0000	0.0015	0.0012	1.790417e-03	0.0000
vbz-vbg	0.0008	0.0005	1.027604e-03	0.0000	0.0017	0.0014	1.996356e-03	0.0000

Table A.2.: Tukey's HSD for average syntactic and morphological complexity scores of inflectional morphs in Mark. The statistics include the difference between the means, the lower and upper bounds of the 95% confidence interval and the adjusted *p*-value per pair.

	Syntactic complexity				**Morphological complexity**			
Pair	Difference	Lower bound	Upper bound	p-value	Difference	Lower bound	Upper bound	p-value
orig-nns	-1.015648e-03	-1.325289e-03	-7.060073e-04	0.0000	0.0053	0.0050	0.0056	0.0000
pos-nns	6.859423e-04	3.763015e-04	9.955832e-04	0.0000	-0.0012	-0.0014	-0.0009	0.0000
vbd-nns	2.456531e-04	-6.398774e-05	5.552939e-04	0.2102	0.0001	-0.0001	0.0004	0.6886
vbg-nns	-4.800972e-05	-3.576505e-04	2.616311e-04	0.9979	0.0012	0.0009	0.0014	0.0000
vbz-nns	6.677108e-04	3.580700e-04	9.773516e-04	0.0000	-0.0026	-0.0029	-0.0023	0.0000
pos-orig	1.701590e-03	1.391950e-03	2.011231e-03	0.0000	-0.0064	-0.0067	-0.0061	0.0000
vbd-orig	1.261301e-03	9.516604e-04	1.570942e-03	0.0000	-0.0051	-0.0054	-0.0049	0.0000
vbg-orig	9.676384e-04	6.579976e-04	1.277279e-03	0.0000	-0.0041	-0.0044	-0.0038	0.0000
vbz-orig	1.683359e-03	1.373718e-03	1.993000e-03	0.0000	-0.0079	-0.0082	-0.0076	0.0000
vbd-pos	-4.402893e-04	-7.499301e-04	-1.306484e-04	0.0007	0.0013	0.0010	0.0016	0.0000
vbg-pos	-7.339521e-04	-1.043593e-03	-4.243112e-04	0.0000	0.0023	0.0020	0.0026	0.0000
vbz-pos	-1.823156e-05	-3.278724e-04	2.914093e-04	1	-0.0015	-0.0018	-0.0012	0.0000
vbg-vbd	-2.936628e-04	-6.033036e-04	1.597802e-05	0.0746	0.0010	0.0007	0.0013	0.0000
vbz-vbd	4.220577e-04	1.124169e-04	7.316985e-04	0.0014	-0.0028	-0.0031	-0.0025	0.0000
vbz-vbg	7.157205e-04	4.060797e-04	1.025361e-03	0.0000	-0.0038	-0.0041	-0.0035	0.0000

Table A.3.: Tukey's HSD for average syntactic and morphological complexity scores of inflectional morphs in the Euro-Congo news corpus. The statistics include the difference between the means, the lower and upper bounds of the 95% confidence interval and the adjusted *p*-value per pair.

	Syntactic complexity				**Morphological complexity**			
Pair	Difference	Lower bound	Upper bound	p-value	Difference	Lower bound	Upper bound	p-value
orig-nns	-1.102333e-03	-1.395361e-03	-8.093053e-04	0.0000	0.0009	6.674229e-04	1.199121e-03	0.0000
pos-nns	-7.934222e-05	-3.723702e-04	2.136858e-04	0.9722	-0.0006	-9.019250e-04	-3.702273e-04	0.0000
vbd-nns	1.618292e-04	-1.311988e-04	4.548572e-04	0.6156	-0.0004	-7.101264e-04	-1.784288e-04	0.0000
vbg-nns	4.752911e-04	1.822631e-04	7.683191e-04	0.0001	-0.0009	-1.145225e-03	-6.135274e-04	0.0000
vbz-nns	-1.387653e-04	-4.317933e-04	1.542627e-04	0.7569	-0.0018	-2.021488e-03	-1.489790e-03	0.0000
pos-orig	1.022991e-03	7.299630e-04	1.316019e-03	0.0000	-0.0016	-1.835197e-03	-1.303499e-03	0.0000
vbd-orig	1.264162e-03	9.711344e-04	1.557190e-03	0.0000	-0.0014	-1.643398e-03	-1.111701e-03	0.0000
vbg-orig	1.577624e-03	1.284596e-03	1.870652e-03	0.0000	-0.0018	-2.078497e-03	-1.546799e-03	0.0000
vbz-orig	9.635679e-04	6.705399e-04	1.256596e-03	0.0000	-0.0027	-2.954760e-03	-2.423062e-03	0.0000
vbd-pos	2.411714e-04	-5.185660e-05	5.341994e-04	0.1759	0.0002	-7.405031e-05	4.576474e-04	0.3105
vbg-pos	5.546333e-04	2.616053e-04	8.476613e-04	0.0000	-0.0002	-5.091490e-04	2.254869e-05	0.0952
vbz-pos	-5.942310e-05	-3.524511e-04	2.336049e-04	0.9925	-0.0011	-1.385412e-03	-8.537141e-04	0.0000
vbg-vbd	3.134619e-04	2.043392e-05	6.064899e-04	0.0279	-0.0004	-7.009475e-04	-1.692498e-04	0.0000
vbz-vbd	-3.005945e-04	-5.936225e-04	-7.566508e-06	0.0405	-0.0013	-1.577210e-03	-1.045513e-03	0.0000
vbz-vbg	-6.140564e-04	-9.070844e-04	-3.210284e-04	0.0000	-0.0009	-1.142112e-03	-6.104139e-04	0.0000

Table A.4.: Tukey's HSD for average syntactic and morphological complexity scores of functional constructions in Alice. The statistics include the difference between the means, the lower and upper bounds of the 95% confidence interval and the adjusted *p*-value per pair.

	Syntactic complexity				**Morphological complexity**			
Pair	Difference	Lower bound	Upper bound	p-value	Difference	Lower bound	Upper bound	p-value
orig-going to	1.054373e-04	-0.0002	3.784596e-04	0.8812	-0.0004	-6.384077e-04	-0.0001	0.0003
passive-going to	-1.681030e-04	-0.0004	1.049193e-04	0.4952	-0.0094	-9.626466e-03	-0.0091	0.0000
perfect-going to	-8.386248e-04	-0.0011	-5.656024e-04	0.0000	-0.0078	-8.077666e-03	-0.0076	0.0000
progressive-going to	-9.304051e-04	-0.0012	-6.573828e-04	0.0000	-0.0059	-6.156115e-03	-0.0056	0.0000
will-going to	-4.635413e-05	-0.0003	2.266682e-04	0.9967	-0.0001	-4.008645e-04	0.0001	0.5908
passive-orig	-2.735403e-04	-0.0005	-5.179468e-07	0.0492	-0.009	-9.244145e-03	-0.0087	0.0000
perfect-orig	-9.440620e-04	-0.0012	-6.710397e-04	0.0000	-0.0074	-7.695346e-03	-0.0072	0.0000
progressive-orig	-1.035842e-03	-0.0013	-7.628201e-04	0.0000	-0.0055	-5.773794e-03	-0.0053	0.0000
will-orig	-1.517914e-04	-0.0004	1.212309e-04	0.6086	0.0002	-1.854369e-05	0.0005	0.0871
perfect-passive	-6.705218e-04	-0.0009	-3.974994e-04	0.0000	0.0015	1.292713e-03	0.0018	0.0000
progressive-passive	-7.623021e-04	-0.0010	-4.892798e-04	0.0000	0.0035	3.214264e-03	0.0037	0.0000
will-passive	1.217489e-04	-0.0002	3.947712e-04	0.8008	0.0092	8.969515e-03	0.0095	0.0000
progressive-perfect	-9.178036e-05	-0.0004	1.812420e-04	0.9309	0.0019	1.665465e-03	0.0022	0.0000
will-perfect	7.922706e-04	0.0005	1.065293e-03	0.0000	0.0077	7.420715e-03	0.0079	0.0000
will-progressive	8.840510e-04	0.0006	1.157073e-03	0.0000	0.0058	5.499164e-03	0.0060	0.0000

Table A.5.: Tukey's HSD for average syntactic and morphological complexity scores of functional constructions in Mark. The statistics include the difference between the means, the lower and upper bounds of the 95% confidence interval and the adjusted *p*-value per pair.

	Syntactic complexity				**Morphological complexity**			
Pair	Difference	Lower bound	Upper bound	p-value	Difference	Lower bound	Upper bound	p-value
orig-going to	-9.326373e-05	-0.0004	2.109630e-04	0.953	-0.0001	-4.128106e-04	0.0002	0.8023
passive-going to	-4.807439e-03	-0.0051	-4.503212e-03	0.0000	-0.0075	-7.738493e-03	-0.0072	0.0000
perfect-going to	-4.374950e-03	-0.0047	-4.070723e-03	0.0000	-0.0083	-8.540493e-03	-0.008	0.0000
progressive-going to	-5.093141e-03	-0.0054	-4.788914e-03	0.0000	-0.0072	-7.452161e-03	-0.0069	0.0000
will-going to	-2.867043e-04	-0.0006	1.752244e-05	0.0781	0.0032	2.917329e-03	0.0035	0.0000
passive-orig	-4.714175e-03	-0.0050	-4.409948e-03	0.0000	-0.0073	-7.611371e-03	-0.0070	0.0000
perfect-orig	-4.281686e-03	-0.0046	-3.977459e-03	0.0000	-0.0081	-8.413371e-03	-0.0078	0.0000
progressive-orig	-4.999877e-03	-0.0053	-4.695650e-03	0.0000	-0.0070	-7.325039e-03	-0.0068	0.0000
will-orig	-1.934406e-04	-0.0005	1.107862e-04	0.4576	0.0033	3.044451e-03	0.0036	0.0000
perfect-passive	4.324893e-04	0.0001	7.367161e-04	0.0007	-0.0008	-1.087689e-03	-0.0005	0.0000
progressive-passive	-2.857017e-04	-0.0006	1.852505e-05	0.08	0.0003	6.431250e-07	0.0006	0.0491
will-passive	4.520735e-03	0.0042	4.824961e-03	0.0000	0.0107	1.037013e-02	0.0109	0.0000
progressive-perfect	-7.181910e-04	-0.0010	-4.139643e-04	0.0000	0.0011	8.026436e-04	0.0014	0.0000
will-perfect	4.088245e-03	0.0038	4.392472e-03	0.0000	0.0115	1.117213e-02	0.0117	0.0000
will-progressive	4.806436e-03	0.0045	5.110663e-03	0.0000	0.0104	1.008380e-02	0.0107	0.0000

Table A.6.: Tukey's HSD for average syntactic and morphological complexity scores of functional constructions in the Euro-Congo news corpus. The statistics include the difference between the means, the lower and upper bounds of the 95% confidence interval and the adjusted *p*-value per pair.

	Syntactic complexity				**Morphological complexity**			
Pair	Difference	Lower bound	Upper bound	p-value	Difference	Lower bound	Upper bound	p-value
orig-going to	-2.081583e-04	-0.0005	8.759936e-05	0.3387	0.0001	-0.0001	3.992015e-04	0.7090
passive-going to	-1.980488e-03	-0.0023	-1.684731e-03	0.0000	-0.0044	-0.0046	-4.090248e-03	0.0000
perfect-going to	-2.109328e-03	-0.0024	-1.813570e-03	0.0000	-0.0046	-0.0049	-4.322478e-03	0.0000
progressive-going to	-1.405383e-03	-0.0017	-1.109626e-03	0.0000	-0.0057	-0.0059	-5.411680e-03	0.0000
will-going to	-8.417147e-05	-0.00038	2.115862e-04	0.9656	0.0017	0.0014	1.952679e-03	0.0000
passive-orig	-1.772330e-03	-0.0021	-1.476572e-03	0.0000	-0.0045	-0.0048	-4.223571e-03	0.0000
perfect-orig	-1.901170e-03	-0.0022	-1.605412e-03	0.0000	-0.0047	-0.005	-4.455801e-03	0.0000
progressive-orig	-1.197225e-03	-0.0015	-9.014674e-04	0.0000	-0.0058	-0.0061	-5.545002e-03	0.0000
will-orig	1.239868e-04	-0.0002	4.197444e-04	0.8394	0.0016	0.0013	1.819356e-03	0.0000
perfect-passive	-1.288396e-04	-0.0004	1.669180e-04	0.8162	-0.0002	-0.0005	3.364861e-05	0.1276
progressive-passive	5.751050e-04	0.0003	8.708626e-04	0.0001	-0.0013	-0.0016	-1.055553e-03	0.0000
will-passive	1.896317e-03	0.0016	2.192074e-03	0.0000	0.0060	0.0058	6.308805e-03	0.0000
progressive-perfect	7.039446e-04	0.0004	9.997022e-04	0.0000	-0.0011	-0.0014	-8.233231e-04	0.0000
will-perfect	2.025156e-03	0.0017	2.320914e-03	0.0000	0.0063	0.0060	6.541035e-03	0.0000
will-progressive	1.321212e-03	0.0010	1.616969e-03	0.0000	0.0074	0.0071	7.630237e-03	0.0000

A.2. Case studies: Standard deviation in the ICLE national subsets

Table A.7.: Average morphological and syntactic complexity scores and their standard deviations by national background and learner group.

Nationality	Group	Morphological score	Standard deviation	Syntactic score	Standard deviation
German	I	-0.9443	0.0118	0.9188	0.0160
	II	-0.9426	0.0107	0.9181	0.0161
	III	-0.9434	0.0107	0.9178	0.0169
	IV	-0.9432	0.0107	0.9179	0.0167
	V	-0.9412	0.0099	0.9184	0.0168
	VI	-0.9418	0.0098	0.9162	0.0163
French	I	-0.9459	0.0123	0.9239	0.0177
	II	-0.9430	0.0110	0.9190	0.0174
	III	-0.9448	0.0110	0.9199	0.0169
	V	-0.9450	0.0114	0.9201	0.0182
	VI	-0.9426	0.0111	0.9216	0.0183
Italian	I	-0.9628	0.0147	0.9181	0.0144
	II	-0.9519	0.0121	0.9170	0.0152
	III	-0.9460	0.0105	0.9169	0.0151
	IV	-0.9462	0.0122	0.9223	0.0182
	V	-0.9461	0.0115	0.9181	0.0150
	VI	-0.9459	0.0107	0.9175	0.0154
Spanish	I	-0.9485	0.0128	0.9198	0.0166
	II	-0.9557	0.0142	0.9208	0.0157
	III	-0.9449	0.0112	0.9189	0.0166
	IV	-0.9471	0.0117	0.9181	0.0163
	V	-0.9523	0.0136	0.9188	0.0158
	VI	-0.9459	0.0111	0.9195	0.0172

Table A.8.: Average morphological and syntactic complexity scores and their standard deviations by learner group in the German ICLE subset.

Group	Morphological score	Standard deviation	Syntactic score	Standard deviation
I	-0.9951	0.0079	0.9147	0.0059
II	-0.9838	0.0067	0.9135	0.0061
III	-0.9822	0.0069	0.9125	0.0061
IV	-0.9784	0.0068	0.9124	0.0062
V	-0.9706	0.0064	0.9109	0.0060
VI	-0.9734	0.0061	0.9108	0.0059

B. R scripts

B.1. Basic web scraper

```
library(RCurl)
library(XML)
library(stringr)


#write to file
writeout <- function (contents, filename = "
   PATH_TO_FILE") {

  output <- paste(sapply(contents, paste,
     collapse="\n"), collapse="\n\n")

  write(output, file=filename, append = TRUE,
     sep = "\n\n")
}


#clean up whitespace
trim <- function(x) gsub("^\\s+|\\s+$", "", x)


#repair incomplete links
fixlinks = function (links) {
      links = paste("RELEVANT_LINK", links,
         sep="")
}

#extract text content form a given webpage
getcontent <- function(url) {
        Sys.sleep(5.0)

  webpage <- getURL(url)
```

```
        webpage <- readLines(tc <-
           textConnection(webpage)); close(tc)
        pagetree <- htmlTreeParse(webpage,
           useInternalNodes = TRUE, encoding='
           UTF-8')
        body <- xpathSApply(pagetree, "//div[
           @id='article-body-blocks']/p",
           xmlValue)

        return(body)
}

#get follow-up links from a given webpage
getlinks <- function(url){
        webpage <- getURL(url)
        webpage <- readLines(tc <-
           textConnection(webpage)); close(tc)
        pagetree <- htmlTreeParse(webpage,
           useInternalNodes = TRUE, encoding='
           UTF-8')

  relcon <- xpathSApply(pagetree, "//ul[@id='
     auto-trail-block']")

        links <- unlist(sapply(relcon,
           function(x) as.character(xpathSApply
           (x, "li/div/div/h3/a/@href")),
           simplify = F))

        return(links)
}

#scraper function which takes a startseed (url
   ) and the number of pages (pagemax) to be
   visited
saveData <- function (startseed, pagemax = 2)
   {
done = c()
oldlinks = c()
newlinks <- c()
currentseed <- startseed
i = 0

stopCond=FALSE
```

```
while (! stopCond){
  oldlinks = c(oldlinks, newlinks)
  newlinks = getlinks(paste(currentseed, "?
     page=", i, sep=""))
  if(length(newlinks) < 1) stopCond = TRUE
  newlinks <- newlinks[! newlinks %in%
     oldlinks]
  contents <- sapply(newlinks, getcontent)
  names(contents) <- NULL
  writeout(contents, filename="english.txt")
  i <- i + 1
  if (i >= pagemax){
    stopCond=T
    }
  }
}
```

B.2. Multiple distortion and compression script

```
#compress with GZIP
compress <- function(original.dir, target.dir)
    {

  lapply(list.files(original.dir), function(x)
      {

    filename <- paste(original.dir, x, sep="")

    new.filename <- paste(target.dir, x, sep="
       ")

    file.copy(filename, new.filename)

    system(paste("gzip", new.filename, sep="␣"
       ))    #"gzip"

  })

}
```

```
#morphological distortion
#delete characters function
drop.chars <- function(wordlist, proportion.
   keep=0.9) {       #proportion.keep=0.9
   deletes 10%

  splitvec <- unlist(mapply(function(x, y) rep
     (x, y),

                    1:length(wordlist), nchar
                       (wordlist)))

  characters <- unlist(sapply(wordlist,
     strsplit, ""))

  drop <- sample.int(length(characters), floor
     (length(characters) *

(1-proportion.keep)))

  characters[drop] <- "|"

  new.words <- sapply(split(characters,
     splitvec), paste, collapse="")

  result <- sapply(new.words, gsub, pattern="|
     ", replace="", fixed=TRUE)

  names(result) <-  NULL

  return(result)

}


#function to apply drop.chars to whole
   directory and print to target directory
morphdistort <- function(original.dir, target.
   dir) {

  lapply(list.files(original.dir, full.names =
      TRUE), function(x){
```

```
      old.filename <- paste(x, sep="")

      from <- original.dir

      new.filename <- paste(target.dir, gsub(
         pattern = from, "", old.filename), sep
         ="")  #D:/test/dutch.txt"

      corpusfile <- readLines(x, n = -1L,
         encoding = "UTF-8")

      wordcorpfile <- strsplit(corpusfile, "␣"
         )

      distcorpfile <- paste(drop.chars(unlist(
         wordcorpfile)), collapse = "␣")

      write <- function (distcorp, filename =
         "text.txt") {

                output = paste(sapply(
                   distcorp, paste, collapse="
                   \n"), collapse="\n\n")

                writeLines(output, con =
                   filename, sep = "\n",
                   useBytes = TRUE)

      }
      write(distcorpfile, new.filename)
  })
}


#syntactic distortion function
drop.words = function (cont) {

  cont2 <- unlist(strsplit(cont, "␣"))     #
     split into words

  N <-  length(cont2)
```

```
  sample.vec <- sample.int(N, floor(0.9*N),
     replace=FALSE)     #sample.int(N, floor
     (0.9*N) deletes 10%

  sample.vec <- sample.vec[order(sample.vec)]

  paste(cont2[sample.vec], collapse="␣")

}


#function to apply drop.words to whole
   directory and print to target directory
syndistort <- function(original.dir, target.
   dir) {

  lapply(list.files(original.dir, full.names =
     TRUE), function(x){

     old.filename <- paste(x, sep="")

     from <- original.dir

     new.filename <- paste(target.dir, gsub(
        pattern = from, "", old.filename), sep
        ="")

     corpusfile <- readLines(x, n = -1L,
        encoding = "UTF-8")

     distcorpfile <- paste(drop.words(unlist(
        corpusfile)), collapse = "␣")

    write <- function (distcorp, filename = "
       text.txt") {

            output = paste(sapply(distcorp,
               paste, collapse="\n"),
               collapse="\n\n")

            writeLines(output, con =
               filename, sep = "\n", useBytes
                = TRUE)
```

```
    }
      write(distcorpfile, new.filename)
  })
}


#writeout
writeout <- function (cont3, filename = "C:/
   Users/Kat/Desktop/news.words/text.txt") {
  output <- paste(sapply(cont3, paste,
     collapse="\n"), collapse="\n\n")

  writeLines(output, con = filename, sep = "\n
     ", useBytes = TRUE)

}

#get filesizes
filesizes <- function(directory) {

  sapply(list.files(directory, full.names=T),

         function(x) file.info(x)$size)

}


#set up directories

getDataDir <- function(basedir){

file.path(basedir, "data")

}


getTempDir <- function(basedir){

file.path(basedir, "temp")

}


setupTempDir <- function(basedir){
```

```
dirname <- getTempDir(basedir)

dir.create(file.path(dirname, "corpus/"))

dir.create(file.path(dirname, "corpus.
   compressed/"))

dir.create(file.path(dirname, "morphdistort/")
   )

dir.create(file.path(dirname, "morphdistort.
   compressed/"))

dir.create(file.path(dirname, "syndistort/"))

dir.create(file.path(dirname, "syndistort.
   compressed/"))

}


cleanTempDir <- function(basedir) {

  lapply(list.files(getTempDir(basedir), full.
     names=TRUE), unlink, recursive=TRUE)

}


#distortion and compression loop

measure.complexity = function(basedir,
   repetitions) {

  lapply(1:repetitions, function(x) {

   #randomSubset <-  getRandomSample(
      getDataDir(basedir)) #only active for
      resampled datasets

   #get input data
   randomSubset = sapply(list.files("working/
      data/", full.names=TRUE), function (x)
```

```
  readLines(x, n = -1L, encoding = "UTF-8")
  , simplify = FALSE)

##name files and writeout whole corpus

setupTempDir(basedir)


lapply(names(randomSubset), function(x) {

  cleanname = gsub(".txt", "", basename(x)
     )

  newname = file.path(getTempDir(basedir),
      "corpus",

    paste(cleanname, "_random.txt", sep=""
       ))

  writeout(paste(randomSubset[[x]],
     collapse = ""), newname)

})


##distort

morphdistort(file.path(getTempDir(basedir)
   , "corpus/"),

             file.path(getTempDir(basedir)
                , "morphdistort/"))


syndistort(file.path(getTempDir(basedir),
   "corpus/"),

             file.path(getTempDir(basedir)
                , "syndistort/"))

##compress

compress(file.path(getTempDir(basedir), "
   corpus/"),
```

```
          file.path(getTempDir(basedir), "
             corpus.compressed/"))


compress(file.path(getTempDir(basedir), "
   morphdistort/"),

          file.path(getTempDir(basedir), "
             morphdistort.compressed/"))


compress(file.path(getTempDir(basedir), "
   syndistort/"),

          file.path(getTempDir(basedir), "
             syndistort.compressed/"))


##create dataframe

df = data.frame(orig.uncomp = filesizes(

  file.path(getTempDir(basedir), "corpus/"
     )))

##take file sizes and stick in df

df$orig.comp = filesizes(

  file.path(getTempDir(basedir), "corpus.
     compressed/"))

df$morphdist.uncomp =  filesizes(

  file.path(getTempDir(basedir), "
     morphdistort/"))

df$morphdist.comp = filesizes(

  file.path(getTempDir(basedir), "
     morphdistort.compressed/"))

df$syndist.uncomp = filesizes(
```

```
      file.path(getTempDir(basedir), "
         syndistort/"))

    df$syndist.comp = filesizes(

      file.path(getTempDir(basedir), "
         syndistort.compressed/"))


    rownames(df) <- gsub("_.*", "", basename(
       rownames(df)))


    ##add ratios to df

    df$morphratio <- df$morphdist.comp / df$
       orig.comp * -1

    df$synratio <- df$syndist.comp / df$orig.
       comp

    cleanTempDir(basedir)

    return(df)

  })}


#result <- measure.complexity("working/", 1)
```

B.3. Random sampling function

```
#sample random number of sentences
getRandomSample <- function(datadir) {

  randomSubset = sapply(list.files(datadir,
     full.names=TRUE), function (x) readLines(x
     , n = -1L, encoding = "UTF-8"), simplify =
      FALSE)
```

```
  ##split into sentences

  sentences <- sapply(randomSubset, strsplit,
     "\\.")

  ##unlist files, get smallest number of
     sentences

   min.sent <- min(sapply(sentences, function(
      x) length(unlist(x))))

   sample.size = min.sent*0.1 #keep percentage
       of data

  ##list of 10 files containing the same
     number of random sentences

  randomSubset <- lapply(sentences, function(x
     ) {

    corp <- unlist(x)

    sample.corp <- sample.int(length(corp),
       size = sample.size, replace=F)

    paste(corp[sample.corp], collapse="␣")

  })

  return(randomSubset)
}
```

C. Python script

```
#!/usr/bin/env python3

#POS-tag data and subsitute morphs with word
   lemmas for morph complexity measurement;
   possible POS-tags NNS VBZ POS VBG VBD/VBN

import sys

if len(sys.argv) != 2:
    print("Usage: cat stanford-lemmatized-data
       | readtags.py <POSTAG-to-be-lemmatized>
       ", file=sys.stderr)
    sys.exit(1)

postag = sys.argv[1]
num_pos_replaced = 0
num_sentences = 0

# Eliminate irregular VBD and NNS
def hacky_conditions(part_of_speech, word):
    if part_of_speech in ['VBD','VBN'] and not
        word.endswith('ed'):
        return False
    elif part_of_speech == 'NNS' and not word.
       endswith('s'):
        return False
    return True

def convert(bracketline):
    global num_pos_replaced
    bracketline = bracketline.strip()
    assert(bracketline[0] == '[' and
       bracketline[-1] == ']')
```

```
        lem_sen_words = bracketline[1:-1].split('] [')
        for lw in lem_sen_words:
            lw_els = lw.split()
            d = dict((el.split('=') for el in lw_els))

            assert('Text' in d)
            assert('Lemma' in d)
            assert('PartOfSpeech' in d)

            if (d['PartOfSpeech'] == postag and
                hacky_conditions(d['PartOfSpeech'], d['Text'])):
                num_pos_replaced += 1
                yield d['Lemma']
            else:
                yield d['Text']

lines = sys.stdin.read().splitlines()
for threelines in zip(lines[::3], lines[1::3], lines[2::3]):
    assert(threelines[0].startswith("Sentence "))
    if not threelines[2].startswith('[Text'):
        print(threelines[0] + "\n" + threelines[1] + "\n" + "Strange line: \"" + threelines[2] + "\"", file=sys.stderr)
        assert(False)

    num_sentences += 1

    for w in convert(threelines[2]):
        print(w)

print("Replaced {} POS in {} sentences".format(num_pos_replaced, num_sentences),
        file=sys.stderr)
```

D. Shell scripts

D.1. Fix fullstops

```
#!/bin/sed -f

s/’ / /g;
s/- -/ /g;  # gedankenstriche (emdash / LaTeX
   --)
s/--/ /g;   # andere gedankenstriche
s/- /-/g;   # fix hyphens
s/ -/ /g;
s/’ / /g;
s/\([] "#$%&()*+,/:0123456789<=>@[\\^_‘{|}~\t\
   r]\)\+/␣/g;
s/[?!;]/\./g;
s/\.␣\./\.␣/g;
s/␣\./\.␣/g;
s/\.\.\+/\./g;
s/\.\s\+\.\+/\./g;
s/␣$/␣/;
```

```
#!/bin/sh

echo
fixfullstops.sed | tr -d ’\n’ | tr ’[:upper:]’
    ’[:lower:]’
echo
```

D.2. Remove corpus markup

```
#!/bin/sed -f

/^\s\+<&>side[^<]*<\/&>/d;
/^\s\+<&>[0-9]\+:[0-9]\+<\/&>/d;
s/<[^>]*>//g;
s/\[[^]]*\]//g;
s/{[^}]*}//g;
s/([^)]*)//g;
s/#//g;
s,\/\/,,g;
```

```
#!/bin/sh

echo
iconv -f latin1 -t utf8 | fromdos |
   rmcorpusmarkup.sed
echo
```

D.3. Remove punctuation

```
#!/bin/sed -f

s/’ / /g;
s/- -/ /g;  # gedankenstriche (emdash / LaTeX
   --)
s/--/ /g;   # andere gedankenstriche
s/- /-/g;   # fix hyphens
s/’ / /g;
s/\([] "#$%&()*+,/:;?!0123456789<=>@[\\^_
   ‘{|}~\t\r]\)\+/␣/g;
s/\./␣/g;
s/␣$/␣/;
```

```
#!/bin/sh

echo
rmpunc.sed | tr -d ’\n’ | tr ’[:upper:]’ ’[:
   lower:]’
echo
```

D.4. Remove UTF-8 characters

```
#!/bin/sed -f

s/\xe2\x80\xa9//g;  # U+2029
s/\xe2\x80\x90//g;  # U+2010
s/\xef\xbb\xbf//g;  # U+65279
s/\xe2\x80\x98//g; s/\xe2\x80\x99//g;  # weird
    unicode apostrophes
s/\xef\xbb\xbf//g; # BOM or zero width no-
   break space
```

E. Zusammenfassung

Diese Arbeit liegt an der Schnittstelle von quantitativer Korpuslinguistik und Informationstheorie und trägt zur derzeitigen linguistischen Komplexitätsdebatte bei. Es wird eine bisher wenig erforschte Methode untersucht, die Kompressionsalgorithmen zur Messung sprachlicher Komplexität in Textkorpora verwendet. Die Methode hat das Potenzial ein radikal objektives Werkzeug der linguistischen Komplexitätsforschung zu werden, sowohl als komplementäres Diagnostikwerkzeug, als auch als eigenständiges, unabhängiges Analysewerkzeug. Die Hauptziele dieser Arbeit sind erstens die Entwicklung und Anwendbarkeit dieser Methode voranzutreiben und zweitens Einsicht in die Funktionsweise des Kompressionsalgorithmus, der dieser Methode zu Grunde liegt, zu gewinnen, um somit informationstheoretische Komplexität linguistisch definieren zu können.

Der Hintergrund dieser Arbeit ist die derzeitige typologische Komplexitätsdebatte und Suche nach Komplexitätsmetriken, die durch die provokative Behauptung, dass einige Sprachen einfacher als andere seien, in einer Ausgabe von *Linguistic Typology* losgetreten wurde (McWhorter 2001b). Diese Behauptung stellt die Annahme, dass alle Sprachen insgesamt gleich komplex sind (e.g. Crystal 1987; Hockett 1958), in Frage. Seither sind zahlreiche Bücher zum Thema publiziert worden (z.B. Dahl (2004); Kortmann & Szmrecsanyi (2012); Miestamo et al. (2008); Sampson et al. (2009)). Die linguistische Komplexität ist einer der am heftigsten diskutierten Begriffe in der Sprachwissenschaft. Im Zentrum der Komplexitätsforschung stehen die folgenden drei Fragen:

(i) Wie kann linguistische Komplexität definiert werden?

(ii) Wie kann linguistische Komplexität gemessen werden?

(iii) Wie kann linguistische Komplexitätsvariation erklärt werden?

Trotz der intensiven Erforschung von linguistischer Komplexität konnte bisher keine einstimmige Antwort zu diesen Fragen gefunden werden. Ganz im Gegenteil: Es wurde eine große Menge an Definitionen vorgeschlagen, die jeweils innerhalb ihres Forschungskontextes gelten, aber nicht universal anwendbar oder akzeptiert sind. Generell wird jedoch zwischen *absoluter Komplexität* und *relativer Komplexität* unterschieden (Miestamo 2006; Miestamo et al. 2008). Absolute Komplexität ist ein theorieorientierter Begriff

und befasst sich mit der Komplexität, die einem linguistischen System innewohnt. Im Gegensatz dazu definieren relative Komplexitätsbegriffe Komplexität in Bezug auf einen Sprachbenutzer. Die Letzteren tendieren daher dazu, anwendungs- und nutzungsorientierter zu sein als die Ersteren. Auch in Bezug auf Komplexitätsmetriken ist der Status quo ähnlich: Zahlreiche Metriken wurden vorgeschlagen, die jedoch entweder auf empirisch aufwendigen Methoden basieren oder selektiver und subjektiver Natur sind.

Hinsichtlich dieser Forschungslage erforsche und erweitere ich eine unüberwachte (*unsupervised*), algorithmische Metrik—im Folgenden als *Kompressionsmethode* bezeichnet—, deren Wurzeln in der Informationstheorie liegen und erstmals vom finnischen Mathematiker Juola (1998) vorgeschlagen wurde. Dieses Maß basiert auf dem Begriff von *Kolmogorov-Komplexität*, welche die Komplexität eines Textes als die Länge der kürzesten Beschreibung dieses Textes definiert. Um die Methode zu illustrieren, wird angenommen, dass zwei verschiedene Objekte beschrieben werden sollen. Diese Objekte sollen vollständig, aber mit so wenigen Wörtern wie möglich beschrieben werden. Anhand dieser Beschreibungen kann nun die Komplexität der Objekte festgestellt werden: Je mehr Wörter zur vollständigen Beschreibung des Objektes benötigt werden—wobei jedoch so wenige Wörter wie möglich benutzt werden sollen—, desto komplexer ist dieses Objekt. Je länger die kürzeste Beschreibung eines Objektes ist, desto komplexer ist dieses Objekt. In Beispiel (1) liegen zwei Sequenzen mit der gleichen Länge von zehn Zeichen vor. Obwohl die Sequenzen gleich lang sind, sind sie unterschiedlich komplex: Die kürzeste Beschreibung von Sequenz (1-a) ist 5×cd (4 Zeichen). Die kürzeste Beschreibung von Sequenz (1-b) ist die Sequenz selbst und hat eine Länge von 10 Zeichen.

(1) a. cdcdcdcdcd (10 Zeichen) → 5×cd (4 Zeichen)
b. c4gh39aby7 (10 Zeichen) → c4gh39aby7 (10 Zeichen)

Linguistische Komplexität in Texten wird also gemessen, indem ihr Informationsgehalt, d.h. ihre Kolmogorov-Komplexität mit Kompressionsalgorithmen approximiert wird. Die Idee, die dieser Methode zu Grunde liegt, ist, dass Texte, die vergleichsweise besser, d.h. effizienter komprimiert werden können, linguistisch gesehen vergleichsweise weniger komplex sind. Kolmogorov-Komplexität ist ein absolutes Komplexitätskonzept, das auf der Form von Strukturen, und nicht deren Funktion oder Bedeutung basiert. In andere Worte gefasst bedeutet dies, dass Kolmogorov-Komplexität agnostisch ist und kein Wissen über tiefere linguistische Form-Funktionskopplungen besitzt. Informationstheoretische, Kolmogorov-basierte Komplexität ist, linguistisch gesehen, ein Maß *struktureller Oberflächenredundanz*. Vereinfacht ausgedrückt mißt Kolmogorov-Komplexität Wiederholungen orthographischer Buchstabensequenzen (Strukturen) in einem Text. Zu einem gewissen Grad verbindet es die folgenden Komplexitätskonzepte:

- ✔ *Quantitative Komplexität:* die Anzahl grammatischer Kontraste, Marker oder Regeln in einem linguistischen System. Mehr Regeln werden mit mehr Komplexität gleichgesetzt (Dahl 2004; McWhorter 2001b; Shosted 2006).

- ✔ *Irregularitäts-basierende Komplexität:* die Anzahl irregulärer, grammatischer Marker in einem linguistischen System. Irreguläre Marker werden als komplexer angesehen als reguläre Marker (Kusters 2003; McWhorter 2001b; Trudgill 2004).

Es umfasst folglich nicht die folgenden Komplexitätskonzepte:

- ✘ *Redundanz-basierende Komplexität:* linguistische Marker, Formen oder Kategorien ohne grammatische oder kommunikative Funktion gelten als komplex (McWhorter 2001b; Seuren & Wekker 1986; Trudgill 1999).

- ✘ *Zweitspracherwerbsschwierigkeit:* linguistische Merkmale, die für Erwachsene nur schwer zu erwerben sind, sind komplex (Kusters 2003; Szmrecsanyi & Kortmann 2009; Trudgill 2001).

Die zentralen Charakteristiken der Kompressionsmethode, die zugleich auch deren Vorteile sind, lassen sich wie folgt zusammenfassen.

Objektiv.

Eines der Hauptmerkmale und größten Vorteile der Kompressionsmethode ist deren einzigartige Objektivität. Die Kompressionsmethode ist weder auf *a priori* in Komplexitätskategorien eingeteilte linguistische Merkmale, noch auf die subjektive Auswahl von Merkmalen angewiesen, die für viele traditionelle Metriken unabdingbar ist. Im Gegenteil besitzt die Kompressionsmethode keinerlei Wissen über Form-Bedeutungsverhältnisse oder linguistische Kenntnis der Texte, auf die sie angewendet wird. Daher sind ihre Messungen radikal objektiv.

Ökonomisch.

Die Kompressionsmethode ist zudem ein ökonomisches Werkzeug zur Komplexitätsmessung, da sie sich einfach implementieren und prinzipiell auf jede orthographisch transkribierte Textdatenbank anwenden lässt. Darüber hinaus basiert sie nicht auf empirisch aufwendigen Daten, deren Erstellung meistens sehr arbeitsintensiv ist und sich schwer replizieren lässt.

Gebrauchsbasiert.

Die Kompressionsmethode ist eine gebrauchsbasierte (*usage-based*) Methode. Gebrauchsbasierte linguistische Ansätze basieren auf der Annahme, dass sich Grammatik bzw. die kognitive Repräsentation

von grammatischen Strukturen direkt aus dem Sprachgebrauch entwickeln und von diesem beeinflusst werden (e.g. Bybee 2006, 2010; Langacker 1988; Tomasello 2003). Bybee (2006) definiert Grammatik beispielsweise als die kognitive Repräsentation der Erfahrung, die ein Sprachbenutzer mit Sprache gemacht hat (Bybee 2006: 711). Dies bedeutet, dass der Sprachgebrauch das linguistische System eines Sprechers bestimmt, formt und verändert.

Sprachliche Phänomene, deren Muster und Gebrauch können in naturalistischen Textkorpora, welche authentische, geschriebene oder gesprochene Sprache beinhalten, analysiert werden. Kompressionsalgorithmen werden direkt auf orthographisch transkribierte Texte angewendet und basieren nicht auf Paradigmentafeln oder Referenzgrammatiken (wie viele "traditionelle" Methoden). Da die Messungen der Kompressionsmethode auf authentischer Sprache basieren, stellt sie eine inhärent gebrauchsbasierte Methode zur sprachlichen Komplexitätsmessung dar.

Holistisch.

Es wurde bereits erwähnt, dass die Kompressionsmethode nicht auf manuell ausgewählte Merkmale als Input angewiesen ist, sondern direkt auf die Daten, d.h. die Texte angewandt wird. Als solche ist algorithmisch gemessene Komplexität nicht auf spezifische linguistische Merkmale beschränkt, sondern ist eine holistische Methode und Metrik.

Die vorliegende Arbeit geht über die reine Anwendung der Kompressionmethode hinaus, da sie die Funktionsweise des Kompressionsalgorithmus erforscht und damit eine erste linguistische Analyse von Kolmogorov-Komplexität vorgelegt wird. Darüber hinaus wird die Methode weiterentwickelt und vorangetrieben. Die Analyse wird insbesondere durch folgende Forschungsfragen und Ziele geleitet, die im Folgenden diskutiert und zusammen mit den Ergebnissen kurz präsentiert werden.

(i) Können Kompressionsalgorithmen auf nicht-parallele Datensätze angewendet werden?

(ii) Können Kompressionsalgorithmen linguistische Komplexität detailliert messen?

(iii) Was messen Kompressionsalgorithmen aus linguistischer Sicht?

(iv) Wie gut können Kompressionsalgorithmen intra-linguistische (als Gegensatz zu sprachübergreifender) Komplexitätsvariation messen?

Die erste Frage beschäftigt sich mit der Anwendbarkeit der Kompressionsmethode auf verschiedene Datentypen. Bisherige Forschung, die Kompressionsalgorithmen zu linguistischer Komplexitätsmessung anwendet, beschränkt sich auf die Analyse von parallelen Textdaten. Parallele Textdaten sind praktisch übersetzungsäquivalente Texte in verschiedenen Sprachen, d.h. es kann ausgeschlossen werden, dass Unterschiede in der Komplexitätsmessung auf dem Inhalt der Texte beruhen. Obwohl parallele Textdaten ideal für die Analyse von sprachübergreifender Komplexitätsvariation sind, stellen sie gleichzeitig eine Limitation für algorithmische Komplexitätsforschung anderer Gebiete (z.B. innersprachlicher Komplexitätsforschung) dar. Aus diesem Grund ist es wichtig, die Anwendung der Kompressionsmethode auf andere Datentypen zu testen. Aus dieser Motivation heraus wird erforscht, inwieweit sich der Inhalt eines Textes auf die Messergebnisse auswirkt, indem die Kompressionsmethode auf parallele, semiparallele und nicht parallele sowie natürliche Textkorpora angewendet wird.

In Kapitel 3 wird die Kompressionsmethode zunächst auf parallele Texte, d.h. das Markusevangelium in sechs verschiedenen Sprachen (Esperanto, Finnisch, Französisch, Deutsch, Latein und Ungarisch) und einigen historischen Varietäten des Englischen angewandt, um zu demonstrieren wie genau linguistische Komplexität mit Kompressionsprogrammen wie `gzip` gemessen werden kann. Es wird gezeigt, dass die Messungen globaler, morphologischer und syntaktischer Komplexität weitgehend mit den Ergebnissen traditionellerer Methoden (Bakker 1998; Nichols 1992) übereinstimmen. Ungarisch und Finnisch sind nach meinen Messungen zum Beispiel global sehr komplexe Sprachen, während alle englischen Bibelversionen—die Westsächsische Version ausgenommen—global vergleichsweise einfach sind. Der Vergleich der morphologischen und syntaktischen Komplexität der englischen Bibelversionen bildet zudem die historische Entwicklung der englischen Sprache von einer morphologisch komplexen zu einer syntaktisch komplexen Sprache ab.

In einem zweiten Schritt wird eine statistisch robustere Version der Kompressionsmethode vorgestellt und auf parallele und—nach der Anwendung von Permutation—semi-parallele Daten von *Alice im Wunderland* in neun europäischen Sprachen (Englisch, Finnisch, Französisch, Deutsch, Italienisch, Niederländisch, Romänisch, Spanisch, Ungarisch) ausgedehnt. Die globale, morphologische und syntaktische Komplexität der neun Sprachen wird in beiden Datensätzen gemessen. Anschließend werden Komplexitätsrankings erstellt. Der Vergleich der beiden Rankings zeigt, dass die Methode sowohl mit parallelen als auch semi-parallen Daten funktioniert, da statistisch eine große Übereinstimmung der Messungen vorliegt.

Im dritten Schritt wird die Kompressionsmethode mit zwei genuin nicht parallelen Zeitungskorpora in denselben neun Sprachen verwendet. Die aus den Messungen resultierenden Rankings werden mit dem Ranking der parallelen Alicedaten, welches als Vergleichsbasis dient, verglichen. Meine Er-

gebnissse zeigen, dass die morphologischen und syntaktischen Rankings der Zeitungsdaten weitgehend kongruent mit den Alicedaten sind. Die globalen Komplexitätsmessungen stimmen jedoch nur moderat überein. Grundsätzlich ist die algorithmische Messung von Komplexität in nicht parallelen Texten möglich, erfordert allerdings ein gewisses Maß an "Inhaltsüberwachung". Dies bedeutet, dass der Inhalt der Komponenten nicht paralleler Korpora ähnlich sein muss, da zufällig ausgewählte Texte verschiedenen Inhaltes nicht zuverlässig mit der Kompressionsmethode gemessen werden können.

Die zweite Frage, ob Kompressionsalgorithmen benutzt werden können, um linguistische Komplexität im Detail zu messen, wird in Kapitel 4 bearbeitet. Die bisherige Forschung im Bereich der algorithmischen Komplexitätsmessung ist vor allem auf das Messen von linguistischer Komplexität aus der Vogelperspektive gerichtet, d.h. morphologische und syntaktische Komplexität wurden insgesamt als Teilbereich einer Sprache gemessen. Die vorliegende Arbeit stellt eine neue, modifizierte Version der klassischen Kompressionsmethode vor, durch die morphologische und syntaktische Komplexität auch im Detail gemessen werden kann (siehe auch Ehret 2014). Die klassische Kompressionsmethode wird mit der systematischen Löschung spezifischer Zielstrukturen kombiniert und wird daher als gezielte Dateimanipulation (*targeted file manipulation*) bezeichnet. Dieses Kapitel stellt eine Ergänzung der bisherigen Forschung mit der gezielten Kompressionsmethode (Ehret 2014) dar und untersucht, inwiefern die bereits erbrachten Messungen textunabhängig sind: In drei Texten, die drei verschiedenen Genres angehören, wird gemessen, wie viel Komplexität morphologische Endungen wie *–ing* und funktionelle Konstruktionen wie *progressive aspect be* + verb*–ing* zur morphologischen und syntaktischen Komplexität eines Textes beitragen.

Aus linguistischer Sicht zeige ich, dass das Vorhandensein morphologischer Endungen in einem Text generell mehr morphologische Komplexität erzeugt, gleichzeitig aber die algorithmische Vorhersage syntaktischer Muster im Text steigert. Die funktionellen Konstruktionen, die untersucht wurden, erhöhen die informationstheoretisch gemessene morphologische Komplexität der Texte, verringern aber deren syntaktische Komplexität. Diese Ergebnisse sind relativ unspektakulär, da sie mit Befunden in der Literatur übereinstimmen und diese lediglich bestätigen (Arends 2001; McWhorter 2001a, 2012; Kusters 2008; Szmrecsanyi & Kortmann 2009; Trudgill 2004). Aus methodischer Sicht belegen diese Ergebnisse jedoch die Effektivität der Kompressionsmethode und demonstrieren, dass Kompressionsalgorithmen durchaus zur detaillierten Komplexitätsmessung eingesetzt werden können. Es wird zudem vorgeführt, dass die Messungen mit gezielter Dateimanipulation größtenteils textunabhängig sind: Obwohl die exakten Messwerte zu einem gewissen Maß von der Komplexität des analysierten Textes abhängen, sind die generellen Komplexitätstrends über verschiedene Texte hinweg sta-

bil.

Die dritte Frage dreht sich um das *black box conundrum*: Trotz der Tatsache, dass die Ergebnisse der Kompressionsmethode linguistisch Sinn machen und interpretierbar sind, ist es essentiell, herauszufinden, *wie* genau der Algorithmus funktioniert und *was* genau die Methode eigentlich macht. Es ist daher eines der Hauptziele dieser Arbeit zu bestimmen, was Kompressionsalgorithmen wie `gzip` aus linguistischer Sicht messen, und die Funktionsweise des Algorithmus zu analysieren. Hierzu wird die Lexikonausgabe von `gzip` , einer Sammlung von Textsequenzen, die der Algorithmus erkannt hat und die der Komprimierung des Inputtextes zugrunde liegen, untersucht. Das Lexikon von *Alice im Wunderland* sowie die Lexika einer morphologisch und syntaktisch manipulierten Version von Alice, wurden manuell nach linguistischen Kategorien (z.B. Nomen, Verben, Phrasen) und nicht linguistischen Kategorien (zufällig komprimierte Sequenzen) annotiert und anschließend analysiert. In (2) wird jeweils eine Beispielsequenz für jede Kategorie aufgeführt. Leerzeichen sind ein Teil komprimierter Sequenzen und werden am Anfang und Ende von Sequenzen durch "␣" repräsentiert.

(2)
a. Lexikalisch ␣opportunity␣
b. Funktionell her␣
c. Anderes ing␣
d. Phrase do cats eat bats␣
e. Gemischt dance t
f. Zufällig omet

Anhand der analysierten Lexika lässt sich feststellen, dass komprimierte Sequenzen je zur Hälfte aus linguistisch sinnvollen und nicht sinnvollen Buchstabensequenzen bestehen. Kompressionsalgorithmen sind also durchaus in der Lage, wieder vorkommende linguistische Strukturen zu erkennen. Allerdings besitzt der Algorithmus keinerlei sprachliche Kenntnis und bevorzugt linguistisch sinnvolle Strukturen nicht gegenüber zufälligen Sequenzen. Als Folge dessen werden auch Sequenzen komprimiert, die oberflächlich, d.h. von ihrer Oberflächenstruktur her linguistischen Strukturen gleichen (z.B. die Endung *–ing* in *beginnING* vs. *nothING*). Der Algorithmus arbeitet auf der Basis der Form von Strukturen und nicht auf der Basis von deren Funktion oder Bedeutung. Algorithmisch gemessene linguistische Komplexität muss aus diesem Grund als ein Maß struktureller Oberflächenredundanz (*structural surface redundancy*) definiert werden.

Des Weiteren wird in diesem Kapitel dargelegt, dass die Manipulation morphologischer und syntaktischer Informationen in einem Text so funktioniert wie beabsichtigt: Morphologische Manipulation, d.h. das zufällige Entfernen von orthographisch transkribierten Buchstaben in einem Text, führt tatsächlich zu einem Anstieg an morphologischen Formen und somit morphologischer Komplexität. Syntaktische Manipulation, d.h. die zufällige

Entfernung von Wörtern in einem Text, wirkt sich wie beabsichtigt auf Abhängigkeiten zwischen Wörtern und Satzteilen aus und erhöht die syntaktische Komplexität in einem Text.

Die vierte und letzte Forschungsfrage bezieht sich auf die Anwendbarkeit der Kompressionsmethode auf natürliche Textkorpora und der Messung innersprachlicher Komplexitätsvariation im Englischen. Dies wird in zwei Fallstudien wie folgt gezeigt. Zum einen wird die Komplexitätsvariation auf globaler, morphologischer und syntaktischer Ebene in verschiedenen, geschriebenen Textgenres des britischen Englisch, repräsentiert durch Daten aus dem *British National Corpus* (BNC), untersucht. Es stellt sich heraus, dass die Komplexität der untersuchten Textgenres in Zusammenhang mit der Formalität der Genres steht. Weniger formale Genres, wie Email oder Briefe sind generell weniger komplex als formalere Genres, wie z.B. Zeitungsartikel. Diese Ergebnisse stimmen mit der in Biber (1988) diskutierten Registervariation überein. Insbesondere die globale Komplexitätshierarchie der verschiedenen Zeitungsgenres (Qualitätszeitung, Regionalzeitung und Bildzeitung) stellt diesen Trend wie im Lehrbuch dar (Abbildung E.1).

Abbildung E.1.: Globale Komplexitätshiercharchie der Zeitungsgenres im BNC.

Zum anderen wird die globale, morphologische und syntaktische Komplexität in Texten aus dem *International Corpus of Learner English* (ICLE) analysiert. Diese Texte stammen von Lernern mit unterschiedlichen Muttersprachen / Nationalitäten, deren Instruktionshintergrund—die Länge des Zeitraumes in dem Englisch an der Schule und/oder Universität erlernt wurde—variiert. Unter der Annahme, dass alle anderen Faktoren (wie beispielsweise Textlänge) gleich sind, wird der Zeitraum des vorangegangenen Englischunterrichts als Maßstab für die Sprachkompetenz der Lernenden betrachtet. Die Resultate deuten darauf hin, dass mehr Instruktion zu mehr globaler und morphologischer Komplexität führen, wohingegen die Texte von weniger fortgeschrittenen Lernern durch erhöhte syntaktische Komplexität gekennzeichnet sind.

Die Komplexität der Texte scheint zwar von der jeweiligen Muttersprache / Nationalität der Lerner beeinflusst zu sein, aber die Relation zwischen Instruktion und Komplexität der Texte bleibt hiervon weitgehend unberührt: Längere Instruktionszeiträume führen generell zu mehr Komplexität unabhängig von der Muttersprache der Lerner. Trotzdem ist auf diesem Gebiet

weitere Forschung mit größeren Datensätzen erforderlich, um den Zusammenhang zwischen der Muttersprache der Lerner und der Komplexität ihrer Textproduktion zu klären.

Folglich kann die Kompressionsmethode erfolgreich auf natürliche Korpusdaten angewendet werden. Allerdings müssen Menge und Qualität der Datensätze berücksichtigt werden, da, wie bei allen quantitativen Methoden, der Output vom Dateninput abhängig ist. Die Kompressionsmethode könnte des Weitereren als Werkzeug zur Messung von Sprachkenntnissen im Bereich der Überprüfung von Sprachkompetenz (z.B. TOEFL) eingesetzt werden.

Obwohl die vorliegende Arbeit die Funktionsweise der Kompressionsmethode aus linguistischer Sichtweise eingehend analysiert und deren Implementierung und Anwendungsbereich erweitert, bleiben doch noch viele Fragen offen. Zukünftige Forschung sollte innersprachliche Komplexität in Lernervarietäten unbedingt näher untersuchen, um festzustellen, ob die auf ICLE basierenden Ergebnisse (siehe Kapitel 6.2) in größeren Datensätzen bestätigt werden können. Die Möglichkeit, historische Komplexitätsvariation mit Kompressionsalgorithmen zu messen, wurde in Kapitel 3.1 angeschnitten. Es wurden verschiedene historische Bibeltexte in Englisch verglichen und die Verlagerung von Englisch, von einer morphologisch reichhaltigen Sprache zu einer Sprache, die grammatische Information überwiegend durch Syntax übermittelt, dargestellt. Dieses Forschungsgebiet verdient es jedoch gründlicher bearbeitet zu werden, vor allem, da umfangreiche historische Korpora wie beispielsweise ARCHER (*A Representative Corpus of Historical English Registers*) oder COHA (*Corpus of Historical American English*) bereits vorliegen und sich ausgezeichtnet für die Analyse mit der Kompressionsmethode eignen. So könnte zum Beispiel der Komplexitätswandel in historischen englischen Genres untersucht werden, indem man britisches und amerikanisches Englisch vergleicht; eventuell läßt sich sogar eine Kolloquialisierung, d.h. die Annäherung geschriebener Sprache an gesprochene Sprache (für eine ausführliche Diskussion siehe Hundt & Mair (1999)), anhand von algorithmischen Komplexitätsmessungen belegen. Es ist nämlich bekannt, dass sich gesprochene Sprache eher durch syntaktische Einbettung auszeichnet, wohingegen sich geschriebene Sprache eher durch erhöhte Phrasenkomplexität und der Verwendung vielfältigerer lexiko-grammatischen Kombinationen (wie z.B. *I think that [...]*) auszeichnet (Biber et al. 2011: 29–32). Im Kontext algorithmischer Komplexitätsforschung müsste eine Kolloquialisierung geschriebener Sprache folglich anhand einer Verschiebung von morphologischer Komplexität zu syntaktischer Komplexität bzw. einer Erhöhung syntaktischer Komplexität nachgewiesen werden. Außerdem könnte die Kompressionsmethode Anwendung im Bereich des Erstspracherwerbs und auf digitalisierte gesprochene Sprachdaten finden. Überhaupt ist die algorithmische Komplexitätsforschung anderer Sprachen und Varietäten

als Englisch bisher vernachlässigt worden und sollte unbedingt erforscht werden.

Abgesehen von den eben erwähnten Punkten, kann auch die Methode selbst noch erweitert werden. Diese Arbeit, wie auch bisherige Forschung in diesem Feld, hat sich vor allem auf die Messung von globaler, morphologischer und syntaktischer Komplexität konzentriert—nur Juola (2008) hat zudem auch pragmatische Komplexität analysiert. Andere sprachliche Unterbereiche könnten jedoch auch algorithmisch gemessen werden: Phonologische Komplexität kann man möglicherweise durch die Komprimierung phonologischer Transkripte messen. Interessant wäre es auch `gzip` mit linguistischem Wissen auszustatten, sodass der Algorithmus linguistisch sinnvolle Strukturen erkennen und komprimieren kann. Dies könnte erreicht werden, indem man dem Algorithmus während des Kompressionsvorgangs ein Wörterbuch einspeist, in welchem alle komprimierbaren Sequenzen, d.h. Wörter einer bestimmten Sprache, aufgelistet sind. Eine derartig veränderte Kompressionsmethode wäre selbstverständlich nicht mehr universell anwendbar und würde an Objektivität verlieren, da sie auf eine bestimmte Sprache beschränkt wäre und bereits vor Anwendung die zu komprimierenden Strukturen bekannt wären.

Die Kompressionsmethode ist eine zeitsparende und einfach anwendbare Methode zur Komplexitätsmessung in Korpora, die im Vergleich zu herkömmlichen Methoden nicht nur holistisch, sondern gleichzeitig auch gebrauchsbasiert ist. Trotz dieser Vorteile hat die Methode auch einige Schwächen, da sie beispielsweise nur auf orthographisch transkribierte Texte angewendet werden kann und die Qualität der Ergebnisse von der Qualität der Ausgangsdaten abhängt.