LiLT volume 11, issue 2 December 2014

Kolmogorov complexity of morphs and constructions in English

Katharina Ehret[1]

This chapter demonstrates how compression algorithms can be used to address morphological and syntactic complexity in detail by analysing the contribution of specific linguistic features to English texts. The point of departure is the ongoing complexity debate and quest for complexity metrics. After decades of adhering to the equal complexity axiom, recent research seeks to define and measure linguistic complexity (Dahl 2004; Kortmann and Szmrecsanyi 2012; Miestamo et al. 2008).

Against this backdrop, I present a new flavour of the Juola-style compression technique (Juola 1998), targeted manipulation. Essentially, compression algorithms are used to measure linguistic complexity via the relative informativeness in text samples. Thus, I assess the contribution of morphs such as *−ing* or *−ed*, and functional constructions such as progressive (*be* + verb-*ing*) or perfect (*have* + verb past participle) to the syntactic and morphological complexity in a mixed-genre corpus of *Alice's Adventures in Wonderland*, the Gospel of Mark and newspaper texts. I find that a higher number of marker types leads to higher amounts of morphological complexity in the corpus. Syntactic complexity is reduced because the presence of morphological markers enhances the algorithmic prediction of linguistic patterns. To conclude, I show that information-theoretic methods yield linguistically meaningful results and can be used to measure the complexity of specific linguistic features in naturalistic copora.

1 Introduction

Linguistic complexity has been in the limelight of typological research for some time and is still one of the most hotly debated notions in

[1] University of Freiburg

LiLT Volume 11, Issue 2, December 2014.
Theoretical and Computational Morphology: New Trends and Synergies.

linguistics in general (Dahl 2004; Kortmann and Szmrecsanyi 2012; Miestamo et al. 2008; Sampson 2009). At the core of the linguistic complexity debate stands the question whether all languages are, overall, equally complex—or not. For much of the twentieth century it was generally assumed that all languages are of equal complexity (Bickerton 1995; Crystal 1987; Hockett 1958; O'Grady et al. 1997). This alleged truism has recently been challenged and evidence was produced that indeed some languages seem to be less complex than others (Kusters 2003; McWhorter 2001b). The major topics of the current complexity debate are how exactly linguistic complexity can be defined and, then, how this complexity can be measured and quantified.

Despite vastly differing terminology and definitions of complexity in the literature, a distinction between *absolute* and *relative complexity* is generally made (Miestamo 2006; Miestamo et al. 2008; Miestamo 2009). Absolute complexity is a theory-oriented notion of complexity and refers to the amount of complexity inherent in a linguistic system. As such it is independent of and unrelated to a language user. Examples of absolute metrics are *quantitative complexity* and *irregularity-based complexity*. Quantitative complexity is interested in the number of grammatical contrasts, markers or rules in a linguistic system, where more constrasts / markers / rules equate to greater complexity. (Dahl 2004; McWhorter 2001a; Shosted 2006). According to the irregularity-based metric, irregular grammatical markers are more complex than regular markers. For instance, irregular past tense forms (*brought*, *went*) are regarded as more complex than regular forms (*walked*, *used*). Therefore, the more irregular markers a linguistic system exhibits, the more complex it is (Kusters 2003; McWhorter 2001a; Trudgill 2004). Relative complexity notions, on the other hand, define complexity in relation to a language user, in terms of cost, processing or acquisition difficulty (Miestamo 2006; Miestamo et al. 2008). Relative complexity is often equated to *second language acquisition difficulty* which is, in fact, the most popular notion of relative complexity. Linguistic features which are difficult to process or acquire for adult language learners are considered complex (Kusters 2003; Szmrecsanyi and Kortmann 2009; Trudgill 2001).

This paper takes an absolute approach to complexity and builds on proposals to use an unsupervised, algorithmic, information-theoretic measure (Bane 2008; Ehret and Szmrecsanyi to appear; Juola 1998, 2008; Moscoso del Prado Martin et al. 2004; Sadeniemi et al. 2008) to assess linguistic complexity in texts. Essentially, this measure boils down to the notion of Kolmogorov complexity, which can be conveniently approximated by using ready-to-go file compression programmes. Kolmogorov complexity measures the information content of a (text) string by the length of the algorithm which is required to (re)construct the exact string (Juola 2008: 92; see also Sadeniemi et al. 2008; Li et al. 2004). Linguistically speaking, Kolmogorov complexity is a quantitative, (ir)regularity-based type of complexity and does not encompass agent-related, subjective complexity (Kusters 2003, 2008; Szmrecsanyi and Kortmann 2009; Trudgill 2001).

In this spirit, I set out to measure the contribution of morphs and constructions to the syntactic and morphological complexity in English texts. In order to explore the possibilities of targeted file manipulation with compression algorithms, I draw up a set of $N = 10$ features comprising

(i) morphs: *–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.*

The database is a corpus which comprises three genres of written English—literary writing, religious writing and newspaper texts—and samples texts from Carroll's *Alice's Adventures in Wonderland*, the Gospel of Mark in the English Standard Version, and a custom-made newspaper corpus. The newspaper corpus consists of articles on the "Euro-Crisis" and the political situation in Congo which were retrieved between December 2011 and February 2012. Each subcorpus counts roughtly the same number of words so that each genre is equally represented. Table 1 shows the composition of the mixed-genre corpus.

Text / Corpus	Genre	Number of words
Alice's Adventures in Wonderland	literary	14,010
The Gospel of Mark	religious	14,009
Euro-Congo corpus	newspaper	14,007
Total		42,026

TABLE 1 Number of words per corpus component of the mixed-genre corpus.

Foregrounding methodological aspects, this paper seeks to demonstrate that Kolmogorov complexity measurements yield linguistically meaningful results and can be used to assess the contribution of specific linguistic features to the syntactic and morphological complexity of a text. Moreover, I show how the complexity of these features in a given text can be inferred from their complexity contribution to the text.

This paper is structured as follows. Section 2 provides some background on information theory. In Sections 3 and 4 the methodology will be outlined. Sections 5 and 6 analyse the complexity contribution of morphs and constructions, respectively. I conclude with a brief summary sketching the advantages and drawbacks of the methodology.

2 Information-theoretic background

Information theory is "the science which deals with the concept 'information', its measurement and its applications"' (van der Lubbe 1997: 1). In this context, the term 'information' refers to the unpredictability or unexpectedness of a proposition, an event or, in terms of communication, a message (Shannon 1948). In his landmark paper "A Mathematical Theory of Communication", Claude Shannon derives the first ever quantitative measure of information, *Shannon entropy*, thereby laying the cornerstone for modern information theory. Shannon entropy measures the information content of a message by quantifying the amount of uncertainty or choice which is involved in the selection of a message from a possible set of messages. As such, Shannon entropy is always measured in relation to a set of possible messages and their probabilities (Li and Vitanyi 1997: 65). This implies that Shannon entropy is not suitable for measuring the information content of an individual message or, say, a linguistic object independently of the probabilities in the set. In order to measure the information content of an individual object, an absolute measure of information is needed which refers to the information inherent in the object alone (Li and Vitanyi 1997: 93).

Kolmogorov complexity is closely related to Shannon entropy but, in constrast, refers to the information content of an individual object or string, not a set of messages (Li and Vitanyi 1997: 521–525). It measures the information content or complexity of a string by the length of the binary programme which is required to (re)construct the exact string (Juola 2008: 92; Sadeniemi et al. 2008: 191; Li et al. 2004). In plain English, Kolmogorov complexity measures the complexity of a string of symbols as the length of the shortest possible description of this string. For illustration, let us assume the two strings of symbols in Example (1) are the objects whose complexity we want to measure. Both strings count ten symbols, yet the length of the shortest possible description of string (1-a) is 5×ab counting four symbols whereas the length of the shortest possible description of (1-b) is the string itself. Measuring the complexity of the two example strings in terms of the length of their shortest possible description necessary to reconstruct them, string (1-b) is more complex than string (1-a).

(1) a. ababababab (10 symbols) → 5×ab (4 symbols)
b. ab?x58gjy9 (10 symbols) → ab?x58gjy9 (10 symbols)

For mathematically non-trivial reasons, Kolmogorov complexity is not computable (Kolmogorov 1965; Li and Vitanyi 1997). Yet, adaptive entropy estimation methods can be used to calculate and approximate its upper bounds. In fact, file compression programmes of the Lempel-Ziv family such as `gzip` use a variant of adaptive entropy estimation that approximates Kolmogorov complexity (Li et al. 2004; Ziv and Lempel 1977). These programmes largely work on the assumption that text strings contain—more or less—structural regularities and redundancies which can be reduced. Compression algorithms like `gzip` compress new text strings on the basis of previously encountered and "memorised" strings taking advantage of the structural redundancy in the text (Juola 2008: 93; Ziv and Lempel 1977: 337). Technically speaking, in a first step compression is achieved by back-referencing redundant (sub)strings with the length of the copied sequence and the distance in the buffer to the previous identical sequence (Ziv and Lempel 1977: 337). In a second step, these length-distance pairs as well as literal unmatched strings are further reduced using *Huffman coding*, a statistical compression method (Salomon 2007: 320–332). In simplified terms, the algorithm "loads" a certain amount of text and "stores" it in a temporary lexicon. While "looking" at further text segments, the programme can "recognise" newly encountered text (sub)strings on the basis of the strings in the lexicon and compress them by eliminating redundancy. Thus, the amount of information measured in a given text string is essentially a measure of the structural surface redundancy in this string. The idea is to measure complexity via the information content in naturalistic text samples. A higher amount of information can be equated with a higher amount of complexity—all other things being equal—in a given text sample. Better compression rates of a given text sample indicate lower information content and thus lower Kolmogorov complexity (Juola 1998, 2008).

In linguistic terms, Kolmogorov complexity is a measure of structural surface redundancy. This is another way of saying that, even though the algorithm to some extent picks up on recurrent linguistic structures, it is absolutely agnostic about their communicative necessity and functions or about form-meaning relationships. In a nutshell, algorithmically measured linguistic complexity is based on the form of structures, not on their function and meaning. It is a quantitative, irregularity-based and text-based metric of absolute linguistic complexity.

3 Measuring Kolmogorov complexity

On a methodological plane, I use `gzip`[2], an open source compression programme of the Lempel-Ziv family, to approximate Kolmogorov complexity and thus measure linguistic complexity on the syntactic and morphological plane. Morphological and syntactic complexity are addressed by distorting the respective information in text samples prior to compression (Juola 2008: 98).

Largely following Juola (2008), morphological distortion is achieved by randomly deleting 10% of the orthographic characters in a text sample. Through this procedure new word forms are created while at the same time the morphological regularity of surface structures is compromised. In other words, the morphological information and hence the complexity is increased. Morphologically complex languages which exhibit overall a large number of word forms anyway, should not be greatly affected. Therefore, morphological distortion should not hurt them as badly as morphologically simple languages, in which distortion creates comparatively more random noise, i.e. complexity. Subsequently, the distorted samples are compressed in order to determine how well or badly the compression programme deals with the distortion. 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 a sample. This procedure is assumed to have little impact on languages with simple syntax—which is essentially defined here as maximum flexibility, i.e. free word order (see Bakker (1998))—as they lack between-word interdependencies. Syntactically complex languages, on the other hand, should be greatly affected as word order regularities and interdependencies are compromised. In short, comparatively bad compression ratios after syntactic distortion indicate high syntactic complexity.

On a more technical note, I take two measures for each text file: the compressed file size in bytes of the distorted file and the compressed file size in bytes of the original undistorted file. Based on these measures

[2] Gzip Version 1.2.4. Published under the GNU General Public License. Written by Jean-Loup Gailly and Mark Adler. URL http://www.gzip.org/

I calculate the *morphological complexity score* defined as $-\frac{m}{c}$, where m is the compressed file size after morphological distortion and c the original compressed file size[3]; and the *syntactic complexity score* $\frac{s}{c}$, where s is the compressed file size after syntactic distortion and c the file size before distortion.

In order to obtain a statistically robust measure, each text file is multiply distorted and compressed with $N = 1000$ iterations.[4] Consider example (2), which illustrates random syntactic distortion. Sentence (2-a) is the original undistorted sentence. In (2-b) and (2-c) three words each were randomly deleted. While (2-b) is still syntactically intact after the distortion, (2-c) is badly compromised as word-interdependencies have been destroyed. Thus, neither the compression of (2-b) nor of (2-c) in isolation would adequately reflect the sentence's syntactic complexity. This issue is easily solved by applying multiple distortion and compression.

(2) a. It was the White Rabbit returning, splendidly dressed, with a pair of white kid gloves in one hand and a large fan in the other: [. . .]
b. It was the ___ Rabbit returning, splendidly dressed, with a pair of ___ kid gloves in one hand and a ___ fan in the other: [. . .]
c. It ___ the White Rabbit returning, splendidly dressed, with a pair ___ white kid gloves in one hand and a large fan in the ___: [. . .] [Alice]

For every iteration of the distortion and compression script, the file size of the compressed original and the compressed distorted sample are returned and the morphological and syntactic complexity scores calculated. The *average morphological complexity score* and *average syntactic complexity score* are subsequently obtained by taking the mean of the total number of measuring points ($N = 1,000$), respectively.

4 Targeted manipulation

In this section a special flavour of the compression technique, *targeted manipulation*, will be presented and used to measure the contribution of morphs and functional constructions to the morphological and syntactic complexity in the mixed-genre corpus thereby demonstrating how compression algorithms can measure detailed morphological and syn-

[3] The morphological complexity score is defined as negative so that higher scores indicate higher complexity.

[4] Unless otherwise indicated, all statistics are implemented in R Version 3.2. R: A language and environment for statistical computing. Developing Core Team 2008. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

tactic complexity.

Targeted manipulation is, essentially, the systematic removal of specific target structures from a text. The idea is to measure the contribution of specific linguistic structures to the morphological and syntactic complexity in text samples by comparing the complexity of manipulated texts and unmanipulated texts. To this end, I combine targeted manipulation, i.e. systematic removal, with the Juola-style compression technique, i.e. random distortion and subsequent compression. Specifically, one feature at a time is removed from the mixed-genre corpus to obtain a set of feature-manipulated text samples. These manipulated texts and the original, intact version of the corpus are then subjected to random distortion and compression. For each text sample I thus obtain an average morphological and an average syntactic complexity score as described above. On the basis of these scores, the morphological and syntactic complexity of the feature-manipulated texts—the texts without the respective feature—and the complexity of the original text can be compared. The difference in complexity between the manipulated texts and the original text is taken as indicator for the amount of morphological / syntactic complexity that an individual feature contributes to the text.[5] On an interpretational plane, the morphological and syntactic complexity of each feature is inferred from the amount of complexity it contributes to the text. This complexity—being to some extent text-dependent—is dubbed *textual complexity*. Generally, a feature that increases the complexity of the original text should be complex, while a feature that decreases the complexity of the original text should be less complex (simple).

Methodologically, each feature is identified and manipulated in a way which damages the texts as little as possible. Systematic manipulation was implemented as follows:

(i) morphs[6]

- *–ing*
- *–ed*
- genitive *’s*
- plural *–s*
- third person singular *–s*

[5] Targeted manipulation is a text-based method. Therefore, the exact quantity of a feature’s contribution to the morphological and syntactic complexity of a given text may vary according to the morphological and syntactic complexity of the original text, respectively.

[6] No distinction is made between inflectional and non-inflectional occurrences of the morphs (*he is singing* vs. *he hates singing*).

(ii) constructions

- progressive aspect *be* + verb–*ing*
- perfect aspect *have* + verb past participle
- passive voice *be* + verb past participle
- future marker *will*
- future marker *going to*

The mixed-genre corpus was annotated with part-of-speech tags using the `Stanford Core NLP tagger and lemmatizer` (Toutanova et al. 2003) to permit the automatic manipulation of the morphs and facilitate the manual coding of the functional constructions. The `NLP tool` was furthermore used to lemmatize the corpus for morph manipulation. Thus, a `python`[7] script identified the morphs on the basis of their part-of-speech tags, e.g. *–ing* is identified by the tag VBG, and replaced the inflected verbs and nouns with their lemma (see examples (3)–(7)).

(3) 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]

(4) 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]

(5) 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]

(6) 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]

(7) 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

[7] Python Software Foundation. Python Language Reference, Version 3. URL http://www.python.org.

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]

The manipulation of functional constructions had to be conducted manually as, for example, not every present participle ending in *–ing* and annotated as VBG is part of a progressive construction. The functional constructions progressive, passive and perfect were therefore manually identified and manipulated by deleting the auxiliary *be/have* and replacing the main verb with its lemma (see examples (8)–(10)). Manipulation is implemented as lemma-substitution because it alters the texts as little as possible and does not introduce new irregularity, for instance, by replacing verbal constructions (*was beginning*) with irregular past tense forms (*began*). Each occurrence of the future markers *going to* and *will*—including its variants *'ll* and *won't*—was deleted as illustrated in examples (11)–(12).

(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]

(11) 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 Ø tell you how the Dodo managed it.
[ALICE]

(12) 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 Ø be delivered into the

hands of men and they kill him.
[Mark]

If negative contractions occurred in any of the constructions, for example, *won't*, *hasn't* or *isn't*, the construction was replaced / deleted while *n't* was replaced with the negative particle *not* (13).

(13) 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]

Finally, all annotation was removed before treating the feature-manipulated texts and the original text with the compression technique described in the previous section. Each text is multiply distorted and compressed with $N = 1000$ iterations and the average morphological and average syntactic complexity score is calculated for each text sample.

Intra-sample variation, i.e. the variation between the different iterations, is accounted for by calculating the *standard deviation*, a measure of dispersion. Table 2 and Table 3 list the standard deviation in the morph and construction manipulated data respectively. In both cases the standard deviation is low. Statistically, this means that the average syntactic complexity score and the average morphological complexity score reflect the actual complexity of the text samples well.

Morph	Standard deviation	
	Morphological complexity	Syntactic complexity
original	0.00123	0.00136
–ing	0.00119	0.00132
–ed	0.00127	0.00132
genitive ’s	0.00125	0.00131
plural –s	0.00119	0.00132
3rd person singular –s	0.0012	0.00132

TABLE 2 Dispersion across individual measuring points of morphological and syntactic complexity scores in the mixed-genre corpus by morph.

On a more statistical note, the measurements presented in the following sections are based on compressed file sizes in bytes. Due to the nature of compression, differences between compressed file sizes are very small to start with. As a consequence, differences in the values of the average morphological and syntactic complexity scores between different texts, and especially between manipulated texts and their orig-

Construction	Standard deviation	
	Morphological complexity	Syntactic complexity
original	0.00123	0.00133
going to	0.00124	0.00125
passive	0.00125	0.00132
perfect	0.00125	0.00128
progressive	0.00123	0.00153
will	0.00121	0.00133

TABLE 3 Dispersion across individual measuring points of morphological and syntactic complexity scores in the mixed-genre corpus by construction.

inals, often seem insignificant. This is not the case. *Tukey's honestly significant difference test* (Tukey's HSD) is calculated for all pairs of the morphological and syntactic complexity scores to establish whether differences between the means of each pair are statistically significant (Baayen 2008: 106–107). The tables with the full statistics are presented in Appendix A.

Suffice it to say, all differences between the morph-manipulated texts and the original corpus are statistically significant. The differences between the average morphological complexity scores across all pairs of the morph-manipulated texts are also statistically significant, yet, the average syntactic complexity scores of the morph-manipulated texts fail to achieve statistical significance. This means that the morph-manipulated texts statistically significantly vary in their morphological complexity but that they are roughly of the same syntactic complexity (for a discussion of morph-complexity see Section 5). In case of the construction-manipulated texts, most differences between the syntactic complexity scores and morphological complexity scores across all pairs are statistically significant. An exception are the texts with the future markers, whose complexity is virtually identical to the complexity of the original corpus, and the constructions passive and perfect, which are of roughly the same morphological and syntactic complexity (for a discussion of construction complexity see Section 6).

5 Morphs

This section surveys the contribution of morphs to the morphological and syntactic complexity in the mixed-genre corpus, and establishes a ranking of the textual complexity of the morphs on the morphological and syntactic level. Furthermore, I show that the findings obtained through compression are in line with previously established complexity metrics.

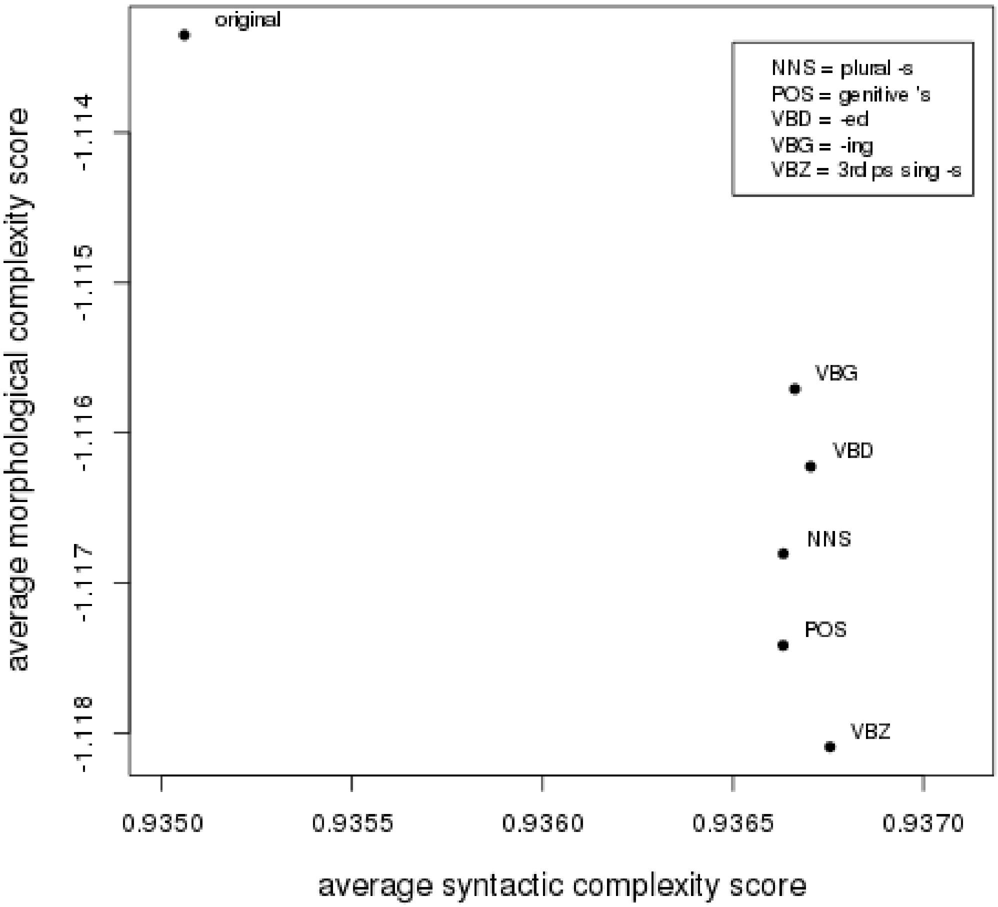

FIGURE 1 Morphological by syntactic complexity of morph-manipulated texts and original text. Abscissa indexes increased syntactic complexity, ordinate indexes increased morphological complexity.

Figure 1 shows the results of the morph analysis. The original text is located in the top left quadrant of the plot and is the morphologically most complex but syntactically most simple text. The morph-manipulated texts, scattered across the right middle to lower part of the plot, all exhibit less morphological but more syntactic complexity than the original. This is another way of saying that the morphs analysed in this section all increase the morphological complexity but at the same time decrease the syntactic complexity of the original text. The former finding is congruent with the assumption of a well-established quantitative metric which holds that “more is more complex” (Arends 2001: 180), i.e. more morphological markers / distinctions generate more morphological complexity (see, for instance, Arends 2001; McWhorter 2001a, 2012; Shosted 2006). While the latter might seem surprising at first, one needs to keep in mind that morphs often form part of morphosyntactic patterns. The morph –*ing*, for instance, often occurs in progressive constructions and is therefore likely to be preceded by a form of the verb *be*. Simply put, one could say that the *ing*-morph facilitates the algorithmic prediction of the progressive pattern *be* + verb-*ing*.

Let us turn to morphological complexity first. In general, all morphs increase morphological complexity, yet, their precise contribution to the morphological complexity in the corpus varies. The degree of variation is assessed by taking the difference between the average morphological complexity scores of the original, unaltered text and each morph-manipulated text. Based on this difference, the textual complexity of each morph at the morphological level is inferred. In this context, morphs which contribute more morphological complexity to the original text are considered information-theoretically more complex than morphs which contribute less morphological complexity to the original text. In this spirit, a ranking of the morphs according to their textual complexity on the morphological level is established (Figure 2). The ranking is in decreasing order of morphological complexity: third person singular –*s*, plural –*s*, genitive –*s*, –*ed* and –*ing*. The endings in –*s*, particularly third person singular –*s*, add comparatively more morphological complexity to the original text than the morphs –*ed* and –*ing* and are therefore more complex.

It is a well-known fact that the English *s*-morph expresses three distinct grammatical meanings, namely third person singular, genitive and plural. For this reason, the *s*-morphs are highly irregular and do not facilitate the algorithmic prediction of patterns, i.e. they are difficult to compress and therefore complex. The –*ed* and –*ing* morphs, on the other hand, are comparatively more regular and encode one grammatical form only, i.e. present and past participle respectively. Thus, the compression technique corroborates the assumption that non-transparency resulting from allomorphy and irregularity of inflectional endings increase morphological complexity (Kusters 2008, 2003; Szmrecsanyi and Kortmann 2009).

Furthermore, the morphological complexity ranking of the five morphs largely coincides with the acquisitional order of morphemes reported in first and second language acquisition studies (e.g. Brown 1973; de Villiers and de Villiers 1973; Bailey et al. 1974; Krashen et al. 1976; Rosansky 1976). Brown (1973) analyses the morpheme acquisition order in a longitudinal study of three children acquiring

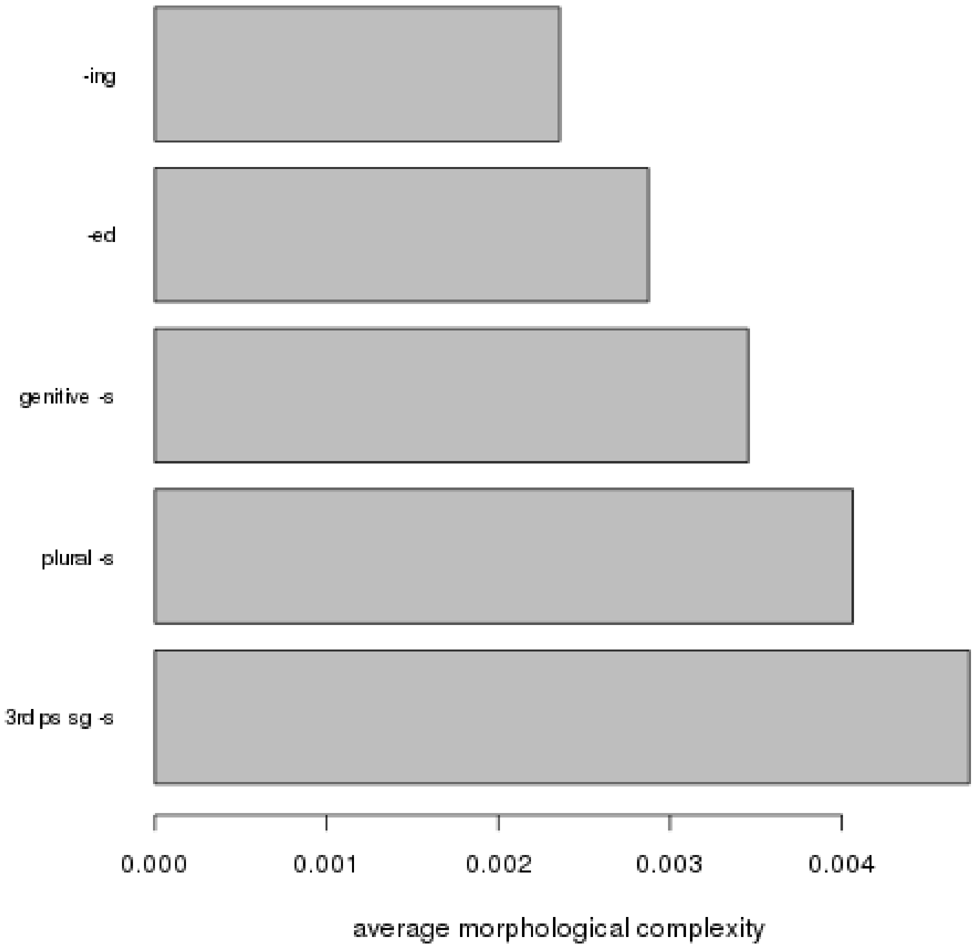

FIGURE 2 Morphological ranking of morphs according to textual morph complexity in the mixed-genre corpus. Abscissa indexes increased morphological complexity.

English as first language, setting the benchmark for future research of morpheme acquisition. The fourteen grammatical morphemes he studies include, among others, progressive *–ing*, regular past *–ed*, possessive *s*, plural *–s* and third person singular *–s*, which are listed in Table 4 together with the five morphs in the mixed-genre corpus (a full list and ranking of Brown's morphemes is provided in Appendix B). Brown finds that the semantic and grammatical complexity of the morphemes greatly influence their order of acquisition but that input frequency does have little impact (Brown 1973: 379). In fact, the order of (second language) morpheme acquisition seems to be determined by a cocktail of six factors: perceptual salience, semantic complexity, morpho-phonological regularity, syntactic category and input frequency (Goldschneider and DeKeyser 2005: 47–55). The concept of semantic complexity, i.e. basically the one-form-one-meaning principle, is reflected in the measurements of the compression technique. In essence, the algorithmically measured complexity of the morphs approximates the acquisitional complexity of grammatical morphemes experienced by (second) language learners.

Brown's order of acquisition	Morphological complexity
Morpheme	Morph
Progressive –ing	–ing
Plural –s	–ed
Genitive –s	Genitive –s
Past –ed	Plural –s
3rd ps sg –s regular	3rd ps sg –s

TABLE 4 Mean ranking of grammatical morphemes according to Brown's acquisitional order (Brown 1973 in Goldschneider and DeKeyser 2005: 72) and morphological complexity ranking of the five morphs in the mixed-genre corpus.

On the syntactic level, all morphs decrease the complexity of the original corpus to roughly the same degree. Calculating the difference between the average syntactic complexity score of the original text and each morph-manipulated text, the textual morph complexity at the syntactic level is obtained for each morph. In syntactic terms, this is the amount of complexity a given morph reduces in the original text. The syntactic complexity ranking (Figure 3) of the morphs is in decreasing order of complexity: third person singular *–s*, *–ed*, *–ing*, genitive *–s* and last plural *–s*.

Finally, a two-sided Pearson's correlation test is calculated to test whether textual morph complexity on the morphological and syntactic level is sensitive to frequency effects. In other words, does the token frequency of the morphs in the corpus influence their morphological and syntactic complexity contribution to the corpus? Pearson's correlation coefficient for morphological complexity and token frequency is very low ($r = 0.36$, $p = 0.55$), while the coefficient for syntactic complexity and token frequency is almost zero ($r = -0.04$, $p = 0.95$). This indicates that the complexity contribution of a given morph type to the morphological and syntactic complexity in the corpus does not significantly depend on its token frequency.

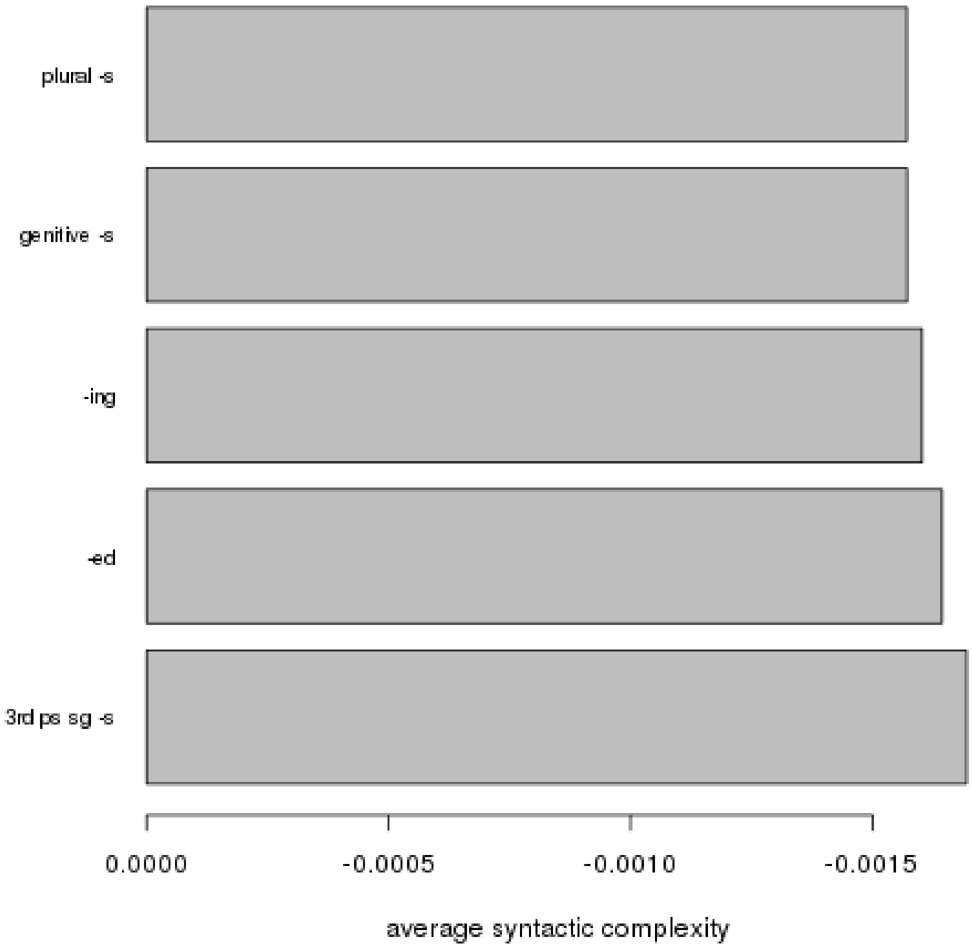

FIGURE 3 Syntactic ranking of morphs according to textual morph complexity in the mixed-genre corpus. Abscissa indexes increased syntactic complexity.

6 Constructions

In this section the functional constructions progressive, passive, perfect and the two future markers *going to* and *will* are analysed, and their quantitative contribution to the morphological and syntactic complexity in the mixed-genre corpus is assessed. Moreover, a ranking of their textual complexity on the syntactic and morphological level is established.

Figure 4 plots the constructions by syntactic and morphological complexity. The original text, situated in the top right quadrant of the plot, is morphologically and syntactically almost the most complex text. The two texts without the future markers are positioned in the top right

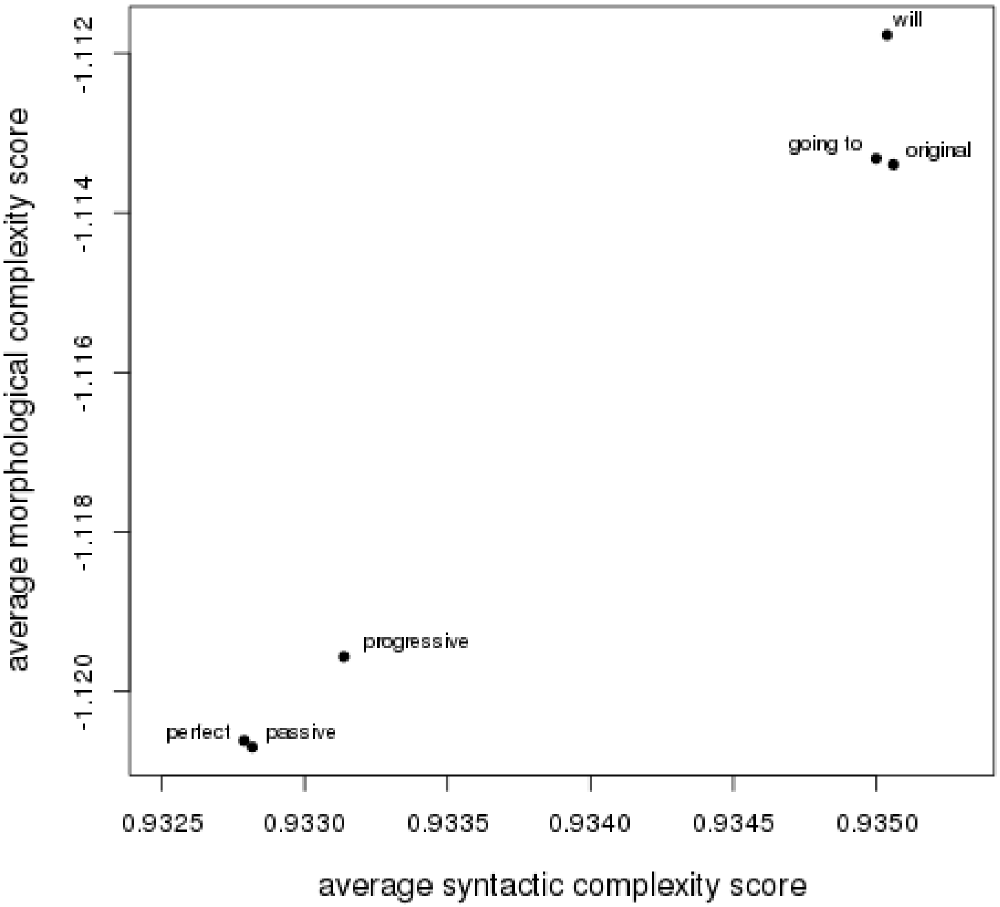

figure 4 Morphological by syntactic complexity of construction-manipulated texts and original text. Abscissa indexes increased syntactic complexity, ordinate indexes increased morphological complexity.

quadrant in close vicinity to the original text: while *going to* is virtually identical to the original and its presence or absence does not affect the complexity of the original corpus, the future marker *will* is the only construction which decreases morphological complexity in the original corpus. The texts without the progressive, perfect and passive constructions, clustering in the bottom left quadrant, are morphologically and syntactically less complex than the original. In other words, their presence increases morphological and syntactic complexity in the text. This finding dovetails with intuitions as the functional constructions should all—apart from the future markers—affect both word order and word form cutting across morphology and syntax. Let me illustrate this point; the passive construction, for example, consists of two discrete components, a form of the auxiliary verb *be* and a verb marked as past participle. It can thus be said that the sequence 'auxiliary *be* + past participle' signals passive. The construction *were arrest-ed* thus cuts across syntax and morphology.

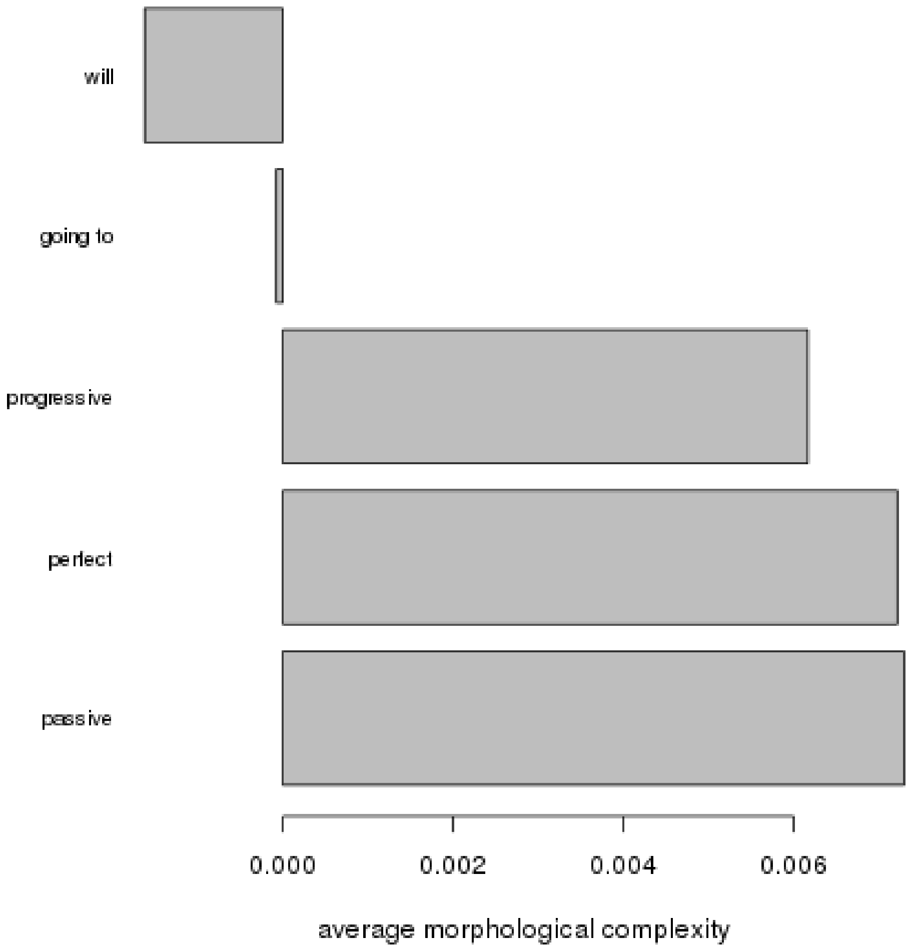

FIGURE 5 Morphological ranking of constructions according to textual construction complexity in the mixed-genre corpus. Abscissa indexes increased morphological complexity.

Although, the constructions passive, perfect and progressive, all increase morphological complexity, the degree of complexity they contribute to the original text varies. The differences between the morphological complexity scores of the construction manipulated texts and the original text are calculated and visualised in Figure 5. The textual complexity of the constructions on the morphological level is inferred from this contribution to the morphological complexity in the corpus. Constructions which contribute more morphological complexity are regarded as comparatively more complex. Passive and perfect are the most morphologically complex constructions and are equally morphologically complex (cf. Section 3). Progressive is slightly less complex. This is probably due to the fact that the *–ing* participle is more regular than the past participles, which can take different forms (e.g. *walk-ed*, *gone*, *eat-en*) and, according to the literature (see e.g. McWhorter 2001b, 2012), irregularity increases complexity.

In contrast, the two future markers affect the complexity of the original text to a much lesser extent than the other constructions—even without the markers the morphological structures of the text remain more or less intact. In fact, the future markers, particularly *will*, decrease morphological complexity. This suggests that analytical, invariant markers are less complex than inflectional markers due to their regularity and transparency (Szmrecsanyi 2009; Szmrecsanyi and Kortmann 2009; Nichols 2009; Trudgill 2004). Despite the fact that frequency is not a factor influencing construction complexity in general (see below), it seems plausible that the complexity of *going to* which occurs only thirteen times in total, might be related to its frequency.

Syntactically, all constructions increase the complexity in the original corpus, thus their textual complexity on the syntactic level is an indicator for the amount of complexity a given construction adds to the original. Figure 6 displays the ranking of the constructions according to their syntactic complexity contributions, i.e. the difference in syntactic complexity between each construction-manipulated text and the original. Perfect and passive are roughly of equal complexity and the most complex constructions, i.e. they add most syntactic complexity to the original text. Progressive also adds a substantial amount of complexity and closely follows perfect and passive in the ranking. The two future markers hardly change the syntactic complexity of the original text as syntactic structures remain largely intact even without the markers.

A two-sided Pearson's correlation test establishes that the complexity a given construction contributes to the syntactic and morphological complexity in the corpus does, generally, not depend on its token frequency in the corpus (syntactic complexity: $r = -0.75$, $p = 0.14$, morphological complexity: $r = -0.67$, $p = 0.21$). However, the quasi "zero effect" of the future marker *going to* on the syntactic and morphological complexity in the corpus suggests that in this particular case, frequency might play a role.

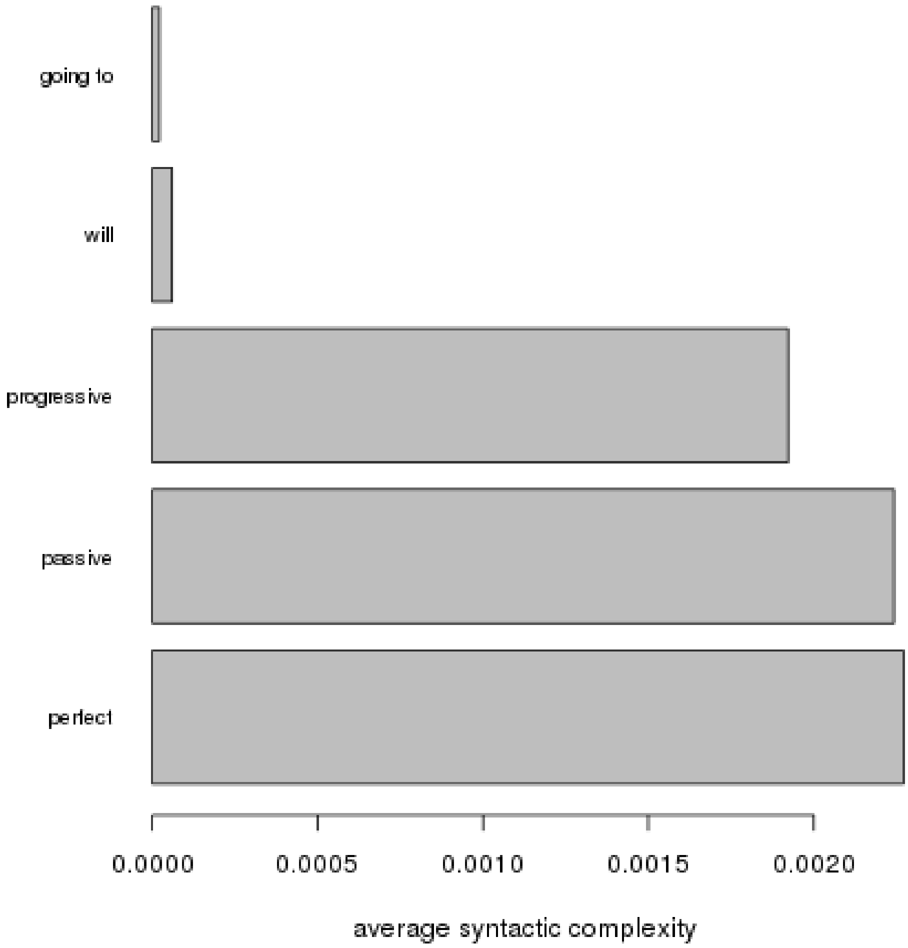

FIGURE 6 Syntactic ranking of constructions according to textual construction complexity in the mixed-genre corpus. Abscissa indexes increased syntactic complexity.

7 Concluding remarks

This chapter ventured into methodologically unexplored territory by experimenting with a new flavour of the Juola-style compression technique, targeted manipulation. Previous studies using compression algorithms have only addressed morphological and syntactic complexity from a bird's eye perspective (e.g. Ehret and Szmrecsanyi to appear; Juola 1998, 2008; Sadeniemi et al. 2008). In filling this gap, I have demonstrated how compression algorithms can be used to address morphology and syntax in detail by measuring the contribution of specific linguistic features to the morphological and syntactic complexity in a mixed-genre corpus. In more general terms, this paper shows that

Kolmogorov measurements yield linguistically meaningful results and provides evidence for the validity of the compression technique.

First, targeted manipulation established that a larger amount of morphological marker types leads to a larger amount of morphological complexity in the corpus. More specifically, more irregular and non-transparent morphs were found to be more complex than regular morphs. On the other hand, the presence of more marker types increases syntactic simplicity by faciliting the prediction of patterns in the text. The functional constructions analysed, with the exception of the future markes *will* and *going to*, increase both morphological and syntactic complexity in the corpus. This finding suggests that analytical, invariant markers are considerably less complex than inflected morphosyntactic patterns. While these results in themselves are nothing to write home about, they underline the effectiveness of targeted manipulation because they are in line with well-established, quantitative complexity metrics (McWhorter 2001b, 2012; Szmrecsanyi 2009; Szmrecsanyi and Kortmann 2009; Trudgill 2004) and connect with findings on morpheme acquisition order (Brown 1973; de Villiers and de Villiers 1973: see also Goldschneider and DeKeyser (2005)).

Second, the textual complexity of the features analysed is generally independent of their token frequency in the corpus. However, the case of the future marker *going to*, which is the feature with the lowest token frequency in the corpus and occurs only 13 times in total, implies that targeted manipulation is not suitable for measuring low-frequency phenomena. Establishing the minimum token frequency which is required for targeted manipulation should therefore be put on the agenda of future research on compression and complexity.

It goes without saying that the compression technique, and thus targeted manipulation, is not without flaws. Compression algorithms are totally agnostic about language intrinsic knowledge such as form-meaning relationships. This ofttimes mentioned blemish is, however, also one of the major assets of the compression technique: it is radically objective. Furthermore, the technique works on and is restricted to written text databases. This is another way of saying that the technique depends to some degree on orthographic conventions and transcription protocols.

In spite of these drawbacks, this chapter provides a hitherto missing part in algorihtmic complexity research by establishing that compression algorithms are sensitive to and capable of capturing the (ir)regularity of specific linguistic patterns. However, much work is still needed to explore the full potential of information-theoretic methods in complexity research. An analysis of intertextual variation, the exact nature of algorithmically measured complexity and the linguistic meaning of compressed strings is outside the scope of this paper but is currently being explored by the author (Ehret in preparation).

Acknowledgements

I am grateful to the audience of the 2013 ICLE workshop on "Theoretical and Computational Morphology: New Trends and Synergies" for constructive feedback, and to Benedikt Szmrecsanyi and Christoph Wolk for methodical advice. In addition, Kimmo Kettunen and two anonymous referees provided me with helpful comments. I would also like to thank Jens Stimpfle for practical support with python programming. The usual disclaimers apply.

Appendix A

Tukey's honestly significant difference test				
	Syntactic complexity		Morphological complexity	
Pair	Difference	p-value	Difference	p-value
orig-nns	-1.571860e-03	0.0000	0.0035	0
pos-nns	-4.814975e-07	1	-0.0006	0
vbd-nns	7.186718e-05	0.8304	0.0006	0
vbg-nns	3.070955e-05	0.9955	0.0012	0
vbz-nns	1.226097e-04	0.3030	-0.0013	0
pos-orig	1.571378e-03	0.0000	-0.0041	0
vbd-orig	1.643727e-03	0.0000	-0.0029	0
vbg-orig	1.602569e-03	0.0000	-0.0024	0
vbz-orig	1.694469e-03	0.0000	-0.0047	0
vbd-pos	7.234867e-05	0.8264	0.0012	0
vbg-pos	3.119105e-05	0.9951	0.0017	0
vbz-pos	1.230912e-04	0.2986	-0.0007	0
vbg-vbd	-4.115763e-05	0.9826	0.0005	0
vbz-vbd	5.074255e-05	0.9565	-0.0019	0
vbz-vbg	9.190018e-05	0.6305	-0.0024	0

TABLE 5 Tukey's HSD for average syntactic and morphological complexity scores of morph-manipulated texts and original in the mixed-genre corpus. The statistics include the difference between the means and the adjusted p-value per pair. Legend: orig = original, pos = genitive $-s$, vbd = $-ed$, vbz = third person singular s, nns = plural $-s$, vbg = *ing*.

Tukey's honestly significant difference test				
	Syntactic complexity		Morphological complexity	
Pair	Difference	p-value	Difference	p-value
orig-going to	6.059570e-05	0.9147	-7.511450e-05	0.7492
passive-going to	-2.182193e-03	0.0000	-7.376026e-03	0.0000
perfect-going to	-2.210936e-03	0.0000	-7.296271e-03	0.0000
progressive-going to	-1.861947e-03	0.0000	-6.244625e-03	0.0000
will-going to	3.833917e-05	0.9881	1.546404e-03	0.0000
passive-orig	-2.242789e-03	0.0000	-7.300911e-03	0.0000
perfect-orig	-2.271532e-03	0.0000	-7.221157e-03	0.0000
progressive-orig	-1.922543e-03	0.0000	-6.169510e-03	0.0000
will-orig	-2.225653e-05	0.9991	1.621519e-03	0.0000
perfect-passive	-2.874334e-05	0.9969	7.975459e-05	0.6979
progressive-passive	3.202462e-04	0.0000	1.131401e-03	0.0000
will-passive	2.220532e-03	0.0000	8.922430e-03	0.0000
progressive-perfect	3.489895e-04	0.0000	1.051646e-03	0.0000
will-perfect	2.249276e-03	0.0000	8.842676e-03	0.0000
will-progressive	1.900286e-03	0.0000	7.791029e-03	0.0000

TABLE 6 Tukey's HSD for average syntactic and morphological complexity scores of construction-manipulated texts and original in the mixed-genre corpus. The statistics include the difference between the means and the adjusted p-value per pair.

Appendix B

Brown's acquisition order of fourteen grammatical morphemes. The mean order across the three children is reported here according to Goldschneider and DeKeyser (2005: 72).

1. progressive *–ing*
2. preposition *in*
3. preposition *on*
4. plural *–s*
5. past irregular (e.g. *went*, *brought*
6. genitive *–s*
7. uncontractible copula
8. articles *a*, *the*
9. past regular *–ed*
10. 3rd ps sg regular (e.g. *sing–s*)
11. 3rd ps sg irregular (e.g. *does*, *has*)
12. uncontractible auxiliary
13. contractible copula
14. uncontractible copula

References

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

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

Bailey, Nathalie, Carolyn Madden, and Stephen D. Krashen. 1974. Is there a "natural sequence" in adult second language learning? *Language Learning* 24(2):235–243.

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

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

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

Brown, Roger. 1973. *A First Language. The Early Stages*. Cambridge, Mass.: Harvard University Press.

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

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

de Villiers, Jill and Peter de Villiers. 1973. A cross-sectional study of the acquisition of grammatical morphemes in child speech. *Journal of Psycholinguistic Research* 2(3):267–278.

Ehret, Katharina. in preparation. *Exploring information-theoretic methods to assess linguistic complexity in English*. Ph.D. thesis, Freiburg.

Ehret, Katharina and Benedikt Szmrecsanyi. to appear. An information-theoretic approach to assess linguistic complexity. In R. Baechler and G. Seiler, eds., *Complexity and Isolation*. Berlin: de Gruyter.

Goldschneider, Jennifer and Robert M. DeKeyser. 2005. Explaining the "natural order of l2 morpheme acquisition" in english: A meta-analysis of multiple determinants. *Language Learning* 55(S1):27–77.

Hockett, Charles Francis. 1958. *A course in modern linguistics*. New York: Macmillan. 58005007 illus. 22 cm.

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

Juola, Patrick. 2008. Assessing linguistic complexity. In M. Miestamo, K. Sinnemäki, and F. Karlsson, eds., *Language Complexity: Typology, Contact, Change*, pages 89–107. Amsterdam, Philadelphia: Benjamins.

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

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

Krashen, Stephen D., Victoria Sferlazza, Lorna Feldman, and Ann Fathman. 1976. Adult performance on the SLOPE test: More evidence for a natural sequence in adult second language acquisition. *Language Learning* 26(1):145–151.

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

Kusters, Wouter. 2008. Complexity in linguistic theory, language learning and language change. In M. Miestamo, K. Sinnemäki, and F. Karlsson, eds., *Language Complexity: Typology, Contact, Change*, pages 3–21. Amsterdam, Philadelphia: Benjamins.

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

Li, Ming and Paul M. B Vitanyi. 1997. *An introduction to Kolmogorov complexity and its applications*. New York: Springer. 96042357 Ming Li, Paul Vitányi. ill. ; 25 cm. Graduate texts in computer science Includes bibliographical references (p. [591]-617) and index.

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

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

McWhorter, John. 2012. Complexity hotspot. In B. Kortmann and B. Szmrecsanyi, eds., *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, Linguae & Litterae, pages 243–246. Berlin/Boston: Walter de Gruyter.

Miestamo, Matti. 2006. On the feasibility of complexity metrics. In K. Kerge and M.-M. Sepper, eds., *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. Tallinn: Tallinna Ülikooli Kirjastus.

Miestamo, Matti. 2009. Implicational hierarchies and grammatical complexity. In G. Sampson, D. Gil, and P. Trudgill, eds., *Language Complexity as an Evolving Variable*, pages 80–97. Oxford: Oxford University Press.

Miestamo, Matti, Kaius Sinnemki, and Fred Karlsson, eds. 2008. *Language complexity: typology, contact, change*. Amsterdam, Philadelphia: Benjamins.

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

Nichols, Johanna. 2009. Linguistic complexity: a comprehensive definition and survey. In G. Sampson, D. Gil, and P. Trudgill, eds., *Language Complexity as an Evolving Variable*, pages 64–79. Oxford: Oxford University Press.

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

Rosansky, Ellen. 1976. Methods and morphemes in second language acquisition research. *Language Learning* 26(2):409–425.

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

Salomon, David. 2007. *Data Compression. The complete Reference.*. London: Springer Verlag, 4th edn.

Sampson, Geoffrey. 2009. A linguistic axiom challenged. In G. Sampson, D. Gil, and P. Trudgill, eds., *Language Complexity as an Evolving Variable*, pages 1–18. Oxford: Oxford University Press.

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

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

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

Szmrecsanyi, Benedikt and Bernd Kortmann. 2009. Between simplification and complexification: non-standard varieties of english around the world. In G. Sampson, D. Gil, and P. Trudgill, eds., *Language Complexity as an Evolving Variable*, pages 64–79. Oxford: Oxford University Press.

Toutanova, Kristina, Dan Klein, Chirstopher Manning, and Yoram Singer. 2003. 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.

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

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

van der Lubbe, J. C. A. 1997. *Information theory*. Cambridge [England] ; New York: Cambridge University Press. 96009421 J.C.A. van der Lubbe ; translated by Hendrik Jan Hoeve and Steve Gee. ill. ; 24 cm. Includes bibliographical references (p. 345-346) and index.

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