An information-theoretic approach to assess linguistic complexity

Katharina Ehret and Benedikt Szmrecsanyi (FRIAS)

Acknowledgments

We would like to thank Annemarie Verkerk, of the Max Planck Institute for Psycholinguistics (Nijmegen), for making her Alice database available to us. We are also grateful to the audience of the 2011 FRIAS workshop on "Complexity and Isolation" for constructive feedback, and to Christoph Wolk for methodical advice. The usual disclaimers apply.

1. Introduction

Linguistic complexity is one of the currently most hotly debated notions in linguistics. The long-standing assumption that all languages are of equal complexity (Bickerton 1995; Crystal 1987; Edwards 1994; Hockett 1958; O'Grady, Dobrovolsky, and Aronoff 1997) had remained unchallenged for much of the twentieth century. Recently, however, this dogma has been questioned and scrutinized (Kusters 2003; McWhorter 2001), and the notion of linguistic complexity has received a considerable amount of interest (Dahl 2004; Kortmann and Szmrecsanyi 2012; Miestamo, Sinnemäki, and Karlsson 2008; Sampson, Gil, and Trudgill 2009). Two central issues in the linguistic complexity debate are, firstly, the problem of finding a generally applicable definition of what exactly complexity is and secondly, how to measure this complexity.

In this paper, we explore proposals to use an unsupervised, algorithmic, information-theoretic measure (Bane 2008; Juola 1998, 2008; Moscoso del Prado Martin, Kostic, and Baayen 2004; Sadeniemi et al. 2008) for assessing linguistic complexity in various languages, based on more or less naturalistic text corpora. We specifically draw on Kolmogorov complexity, which defines the complexity of a string/text as the length of the shortest possible description of that string/text. Kolmogorov complexity is a quantitative, (ir)regularity-based type of complexity, which is completely agnostic about subjective agent-related complexity such as second language acquisition difficulty (Kusters 2003, 2008; Szmrecsanyi and Kortmann 2009; Trudgill 2001).

Kolmogorov complexity can be conveniently approximated by using modern, off-the-shelf file compression programmes. Obtaining numerical estimates of the relative informativeness of text samples via file compression, we will assess linguistic complexity on the (i) overall, (ii) syntactic, and (iii) morphological level. To furnish a case study, we tap into three datasets:

1. a parallel text database sampling the Gospel of Mark in six languages (Esperanto, Finnish, French, German, and Hungarian) as well as some ten historical varieties of English;
2. 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):
3. and a non-parallel sample of newspaper texts covering nine European languages (Dutch, English, Finnish, French, German, Hungarian, Italian, Romanian, and Spanish).

In this paper, which is methodological in nature, we aim to demonstrate that the compression technique yields linguistically rather meaningful results, because it provides complexity rankings that are in line with what more orthodox complexity notions would lead one to expect. Second, we show that the measurements work on parallel as well as non-parallel corpus data.

This paper is structured as follows. In Section 2, we discuss information theory. Section 3 explains how to measure Kolmogorov complexity. In Section 4, we present our empirical analyses. Section 5 concludes by considering the advantages and drawbacks of the method, and by sketching directions for future research.

2. Information theory

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) analysed the information content along a channel between a message source and a listener, establishing the maximum bounds for the efficiency with which messages can be transmitted. In the framework of this theory, the term 'information' refers to the unpredictability or unexpectedness of a proposition, event or, in terms of communication, a message. Thus, the information content of a message is directly related to its unpredictability, i.e. a message is informative if it is not predictable or expected and conveys something surprising and new. The information content of a message is measured in *Shannon entropy*, a measure of unpredictability or disorder, which calculates the information contained in a given message in relation to the predictable part of the message (which is not informative as it is already given).

A related measure of information is *Kolmogorov complexity* which, in contrast to Shannon entropy, refers to the information content of a given string (not message source). Shannon entropy "is an upper bound on (and asymptotically equal to) Kolmogorov complexity" (Juola 2008: 92), and measures the information content or complexity of a string of symbols as the shortest possible description of it. Mathematically speaking, the complexity of a string is measured by the length of the algorithm which is required to (re)generate the exact string (Juola 2008: 92; Sadeniemi et al. 2008: 191; see also Li et al. 2004). Let us take a look at the two example strings of symbols in (1). Both strings consist of the same number of characters, yet string (1a) can be compressed to the expression *5×cd*, counting four characters, whereas the shortest description of string (1b) is the string itself. Measuring the complexity of string (1a) and string (1b) according to the length of their shortest possible description, string (1a) is obviously less complex than string (1b).

(1) a. cdcdcdcdcd (10 characters) ⟹ 5×cd (4 characters)
b. cdgh39aby7 (10 characters) ⟹ cdgh39aby7 (10 characters)

Adaptive entropy estimation methods can be used to compute and approximate the upper bounds for Kolmogorov complexity (Juola 1998; Ziv and Lempel 1977). File compression programmes (such as `gzip`) actually use a variant of adaptive entropy estimation that approximates Kolmogorov complexity. More specifically, file compression programmes compress text strings by describing new strings on the basis of previously seen and memorised (sub-)strings so that the amount of information and redundancy in a given string can be measured (Juola 2008: 93). The idea, then, is to measure complexity by measuring the information content – and hence, unpredictability – in text samples. In this endeavour, a higher amount of information (unpredictability) is taken to indicate increased linguistic complexity of the linguistic sample under analysis (Juola 1998).

Even though Kolmogorov complexity is not fully compatible with traditional notions of linguistic complexity because compression tools are agnostic of, say, form-meaning relationships, they do capture recurrent (linguistic) patterns and (ir)regularities. Kolmogorov complexity conflates to some extent structural complexity and system complexity (Dahl 2004: 42–44) and also adds a substantial amount of frequency weighting. In sum, Kolmogorov complexity is a quantitative, frequency-based, and corpus-based measure of absolute linguistic complexity; it measures linguistic surface complexity by describing new structures on the basis of previously encountered structures.

3. How to measure Kolmogorov complexity

Methodologically, we utilise an open source compression programme, namely `gzip` (version 1.2.4), to approximate Kolmogorov complexity and thus to assess linguistic complexity on the overall, syntactic and morphological plane. The overall complexity of our text samples is measured by obtaining two measurements for each text file analysed: the file size (in bytes) before compression, which roughly corresponds to Dahl's notion of verbosity (Dahl 2004), and the file size (in bytes) after compression. Subsequently, the two values are subjected to regression analysis in order to eliminate any trivial correlation between the two measurements. The resulting *adjusted overall complexity scores* (regression residuals, in bytes), which measure left-over variance, are taken as indicators of the overall complexity of a given language 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 tier can be addressed by manipulating the text files prior to compression. Largely following Juola (2008), morphological distortion randomly deletes 10% of the orthographic characters in each file. Through this procedure new word forms are created while at the same time morphological regularity is compromised. Subsequently, the distorted samples are compressed in order to determine how well or badly the compression programme deals with the distortion. As morphologically complex languages exhibit overall a relatively large amount of word forms in any case, distortion should not hurt 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 simple syntax (which is defined here as, essentially, free word order) as they lack between-word interdependencies which could be compromised. Syntactically complex languages, however, should be greatly affected as word order regularities are compromised. The auxiliary sequence *would have been* (2a), for instance, which occurs twice in one of our text samples (the Bible in Basic English), could be altered to *would* ___ *been* (2b) through distortion. In this case the compression algorithm would encounter two hapax legomenon patterns – instead of encountering one pattern twice – which leads to uncompressible entropy. This will hurt compression efficiency. To make a long story short, comparatively bad compression ratios after syntactic distortion indicate high syntactic complexity.

(2) a. no flesh **would have been** kept from destruction

b. no flesh **would ___ been** kept from destruction

(Mark 13:20 [Basic English])

On a more technical note, we calculate two complexity scores: a *morphological complexity* score which is defined as - m/c, where *m* is the compressed file size after morphological distortion and *c* is the compressed file size before distortion. The *syntactic complexity score* is defined as s/c, where *s* is the compressed file size after syntactic distortion and *c* the file size before distortion.

4. Measuring linguistic complexity in corpora

4.1. The Gospel of Mark – complexity in parallel texts

The use of file compression programmes 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 have become quite popular in typological studies (Cysouw and Wälchli 2007; Dahl 2007) as they facilitate comparability across different languages and language varieties due to the fact that differences in propositional content can be ruled out. The classic database in parallel text studies is the Holy Bible (see, e.g., Juola 2008), and in precisely this spirit we set the stage by applying the compression technique to the Gospel of Mark in a number of historical varieties of English and seven other languages listed below.

Varieties of English:

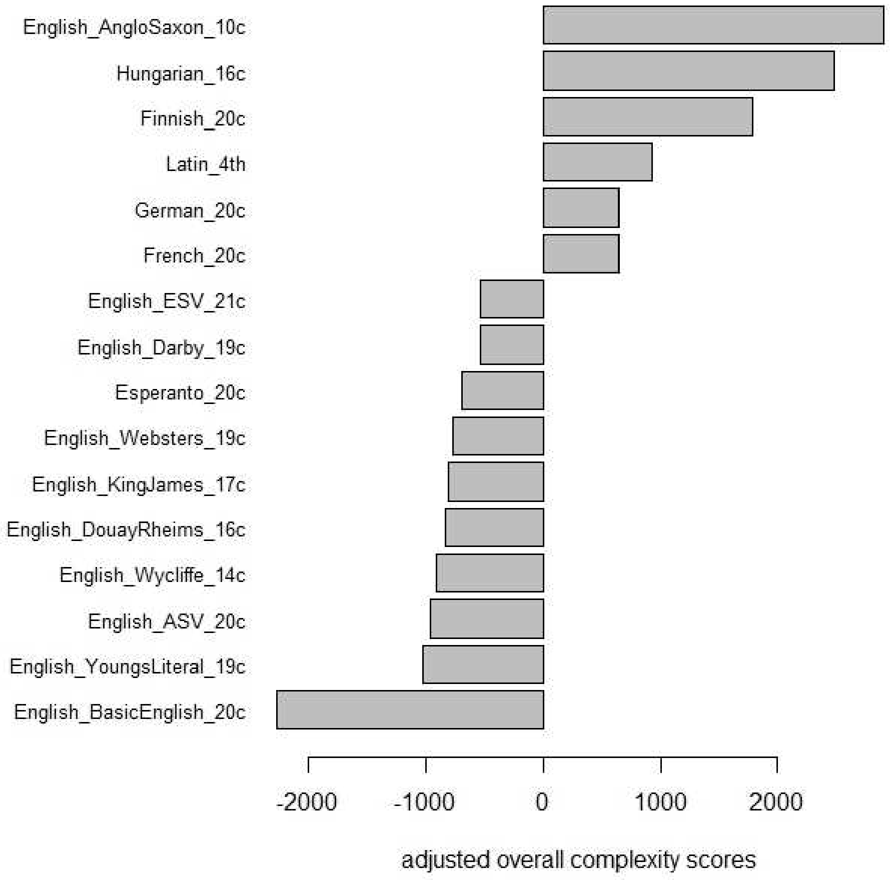
- West Saxon (approx. 10th century [from Bright 1905])
- Wycliffe’s Bible (14th century [1395])
- The Douay-Rheims Bible (16th century [1582])
- The King James Version (17th century [1611])
- Webster’s Revision (19th century [1833])
- Young’s Literal Translation (19th century [1862])
- The Darby Bible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

Other languages:

- Esperanto (Esperanto Londona Biblio, 20th century [1926])
- Finnish (Pyhä Raamattu, 20th century [1992])
- French (Ostervald, 20th century [1996 revision])
- German (Schlachter, "Miniaturbibel", 20th century [1951 revision])
- Hungarian (Vizsoly Bible [a.k.a. Károli Bible], 16th century)
- Latin (Vulgata Clementina, 4thcentury)

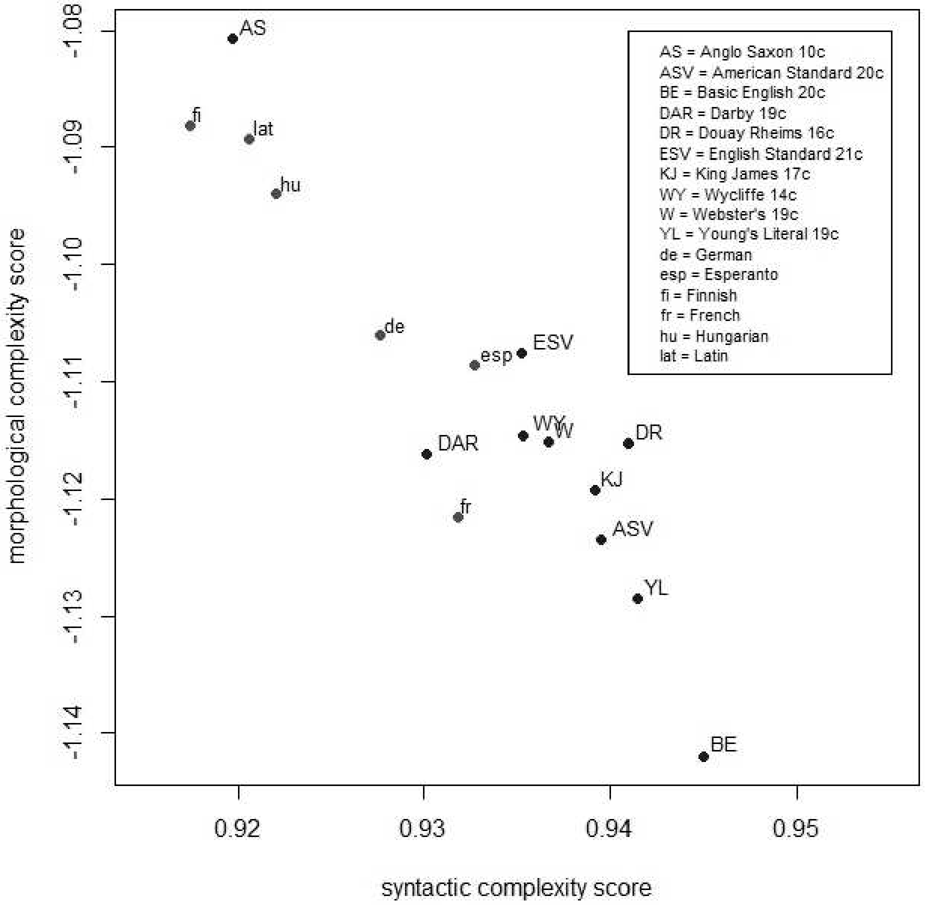
We proceed as described above and establish the file sizes in bytes before and after compression for each file. We then calculate adjusted overall complexity scores for all language samples and obtain a hierarchy of overall complexity (Figure 2). West Saxon, Hungarian, Finnish, Latin, German and French are (in decreasing order) rather complex whereas Esperanto and all English texts after 1066 are rather simple. These findings tie in neatly with previous, more traditional complexity research (Bakker 1998; Nichols 1992).

Figure 1. Overall complexity hierarchy. Negative residuals indicate below-average complexity; positive residuals indicate above-average complexity.



The analysis of morphological and syntactic complexity yields equally intuitive results. Thus in Figure 3, 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 – with the exception of West Saxon – are morphologically simple but syntactically complex and are scattered across the bottom right quadrant. German, Esperanto and French cover the middle ground and seem to be balanced in regard to morphological versus syntactic complexity.

Figure 2. Morphological complexity by syntactic complexity. Abscissa indexes increased syntactic complexity, ordinate indexes increased morphological complexity.



In the Bible sample, morphological complexity trades off against syntactic 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* = -.92, p = 2.374e^{-07} the correlation between the complexity scores indicates a textbook-style trade-off.

We illustrate the workings of the compression technique 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 1), we count nine different segmentable inflected word tokens in the West Saxon version whereas we only count three different tokens and two types (*giv-en, day-s*) in the Basic English version.

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

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

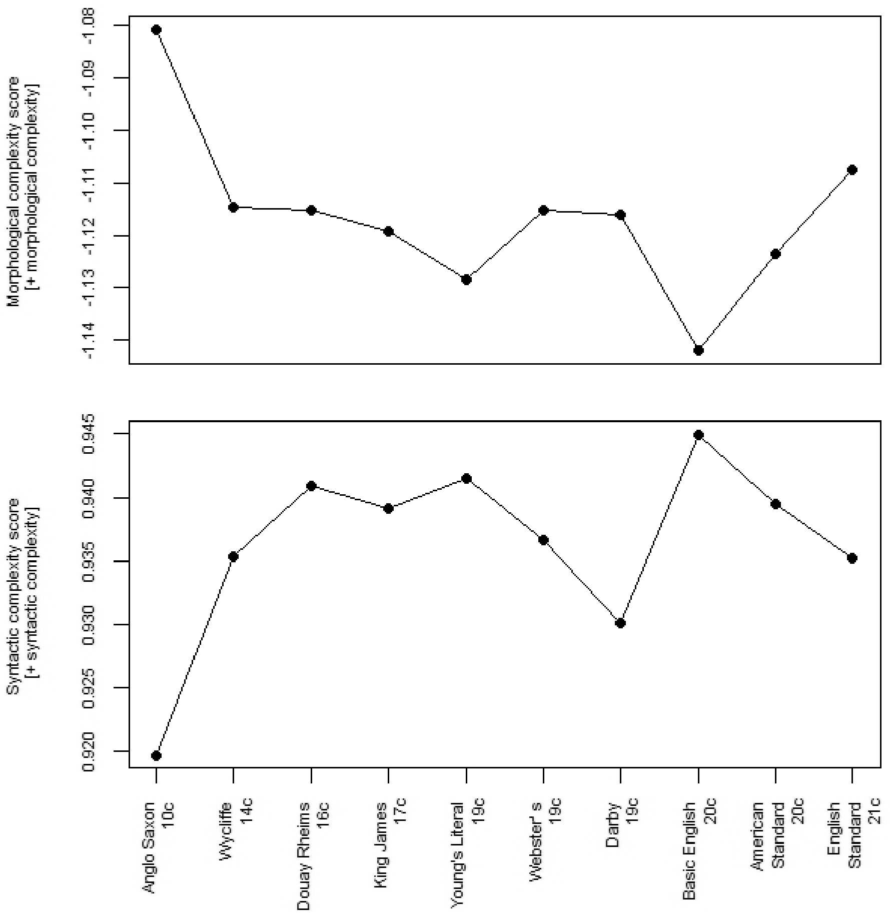
Turning to syntax (Table 2), the West Saxon version features four different word order patterns whereas in the Basic English version word order is relatively rigid (i.e. complex) because the pattern subject-verb dominates throughout the passage. Therefore, Basic English is classified as a syntactically complex language – in contrast to West Saxon, it has many word order rules to break.

Table 2. Word order patterns in Mark 1:8–9.

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

Let us focus now on historical drifts in the English translations of the Bible, as gauged by the compression technique. It is well-known that English has changed from a rather synthetic language – i.e. one that relies heavily on inflections to code grammatical information – in Old English times into a rather analytic language that draws on word order and function words to convey grammatical information. This textbook story is nicely depicted in Figure 3, which plots real time drifts in the history English: our Kolmogorov complexity measurements clearly suggest morphological simplification and syntactic complexification, some outliers notwithstanding.

Figure 3. Real time drifts in English: morphological (upper plot) and syntactic complexity (lower plot). Abscissa arranges Bible translations chronologically.



4.2. Parallel versus non-parallel texts

In this section we aim to demonstrate that the compression technique need not be limited to parallel texts but can also be applied to non-parallel text databases. We draw on a parallel text database and a non-parallel text database in order to explore the reach and limits of the compression technique in two steps. Firstly, we measure and subsequently compare linguistic complexity in a parallel corpus of *Alice's adventures in Wonderland* and a re-sampled semi-parallel version of the same corpus. Secondly, we measure linguistic complexity in non-parallel newspaper texts and compare our results to the complexity hierarchy obtained from the Alice corpus.

4.2.1. *Alice's adventures in Wonderland*

In a first step, we sample *Alice's adventures in Wonderland* (by Lewis Carroll) in nine 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; Kettunen et al. 2006; Sadeniemi et al. 2008) – Dutch, English, Finnish, French, German, Hungarian, Italian, Romanian, Spanish – and measure linguistic complexity on the overall, syntactic, and morphological tier.

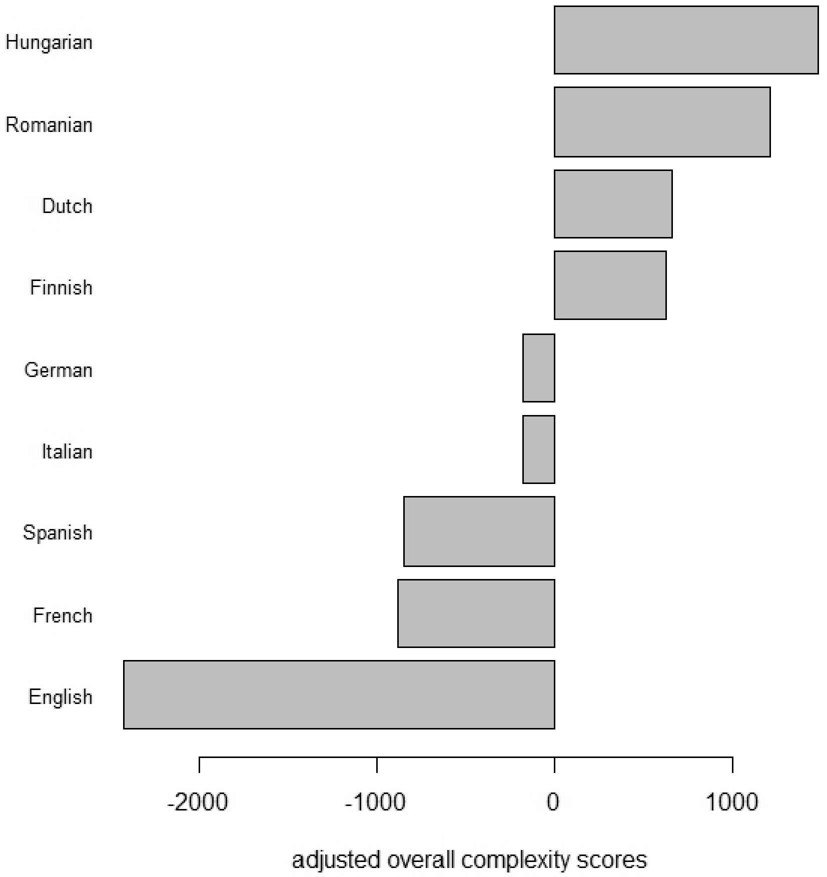
Establishing the file sizes of each text file before and after compression, we calculate the adjusted complexity scores which indicate the overall complexity of each language sample. Subsequently, we address syntactic and morphological complexity by applying multiple distortion and compression – using an R script[1] which implements the methodology as described in Section 3 but allows for multiple iterations – to the complete Alice corpus. By taking multiple measuring points we ensure that our findings are statistically robust. Notice here that in the process of random deletion, any character or word token of a given text file could be modified. However, the impact of the deletion on complexity may, of course, vary according to the precise character/word which was subject to deletion. Consider example (3), which illustrates syntactic distortion. (3a) is the unaltered sentence. In both (3b) and (3c) two words were deleted, but the impact of the deletion differs greatly: while (3b) is still syntactically intact, (3c) has been rendered incomprehensible because syntax is compromised badly.

(3) a. The Rabbit actually took a watch out of its waistcoat-pocket.
b. The Rabbit ___ took a watch out of its waistcoat ___.
c. The ___ actually took a ___ out of its waistcoat-pocket.

In terms of complexity, compression of neither (3b) nor (3c) 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. We therefore apply multiple distortion and compression with *N* = 1,000 iterations to each *Alice* version. Every iteration returns the compressed file sizes for each language sample before and after syntactic/morphological distortion. On the basis of these file sizes we calculate the *average morphological complexity score* and *the average syntactic complexity score*. Intra-sample dispersion turns out to be negligible.[2] The average complexity scores are subsequently obtained by taking the mean of the total number of measuring points (*N* = 1,000) for morphological and syntactic complexity respectively.

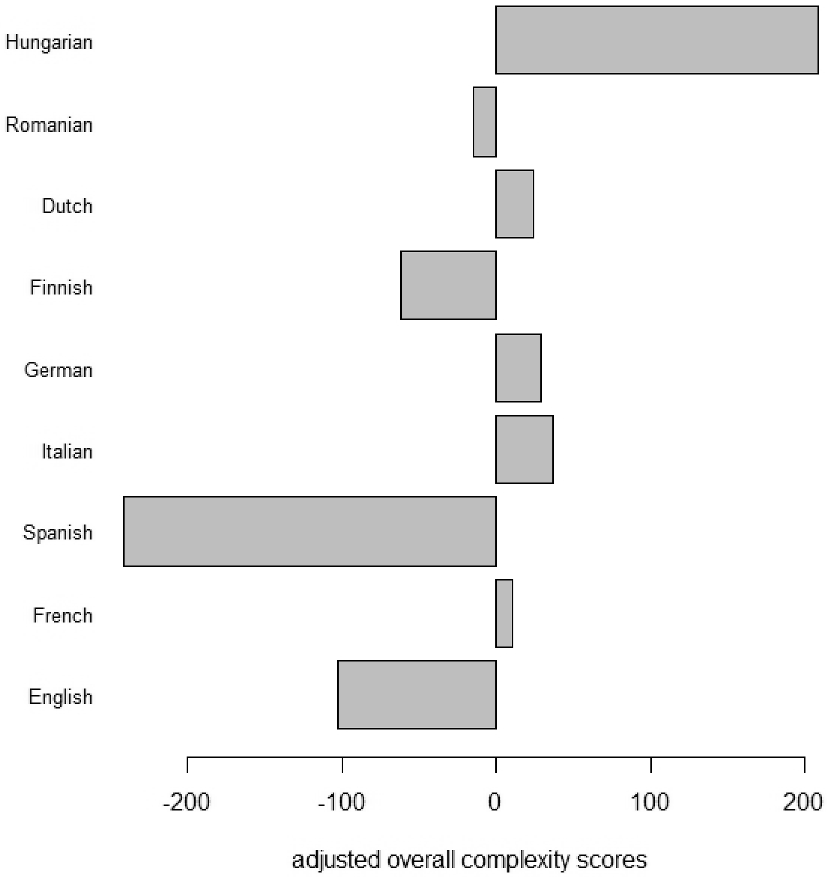
We also create a semi-parallel Alice corpus by means of permutation. Before every iteration of the multiple distortion and compression script, we randomly sample 10% of the total number of sentences[3] from the Alice corpus. Even though the original content of this re-sampled corpus is the same across all languages, the concrete content of each sub-sample varies due to the multiple random permutations. We apply the script with *N* = 1,000 iterations and thus obtain 1,000 measuring points for the compressed and uncompressed file sizes before and after syntactic/morphological distortion of each permutated language sample. Subsequently, we obtain the *average overall complexity score*[4] by calculating regression residuals of the mean compressed file sizes (dependent variable) and the mean uncompressed file sizes (independent variable). The average morphological complexity score and the average syntactic complexity score[5] are subsequently computed as described above. Finally, we compare the overall and morphosyntactic complexity hierarchies of the parallel and semi-parallel Alice corpora.

Figure 4. Overall complexity hierarchy in the parallel Alice corpus.



The ranking of overall complexity in the parallel corpus (Figure 4) is, in decreasing order of complexity, Hungarian, Romanian, Dutch, Finnish, German, Italian, Spanish, French and English. On the whole, these results are as we would expect them to be. In fact, only Dutch exhibits surprisingly high overall complexity. Comparing this ranking to the results of the semi-parallel corpus (Figure 5) we find that Hungarian still exhibits the highest overall complexity. Spanish and English are still the least complex languages – even though their order is now reversed, and Spanish is less complex than English. However, the ranking of the other languages has changed: Italian, German and Dutch as well as French are (in decreasing order) complex whereas Romanian and Finnish have become less complex. The overall correlation of the two complexity hierarchies is moderate but significant (Pearson's $r = 0.59$, $p = 0.05$)[6].

Figure 5. Overall complexity hierarchy in the semi-parallel Alice corpus.



Turning to morphological and syntactic complexity, the analysis of the parallel Alice corpus (Figure 6) dovetails with intuitions: Finnish, which according to the Kolmogorov metric is the morphologically most complex but syntactically most simple language in the sample, is located in the extreme top left quadrant whereas morphologically simple but syntactically complex languages (Spanish, French and English) cluster in the bottom right quadrant of the plot. The middle ground is covered by more balanced languages; while Romanian and Hungarian as well as Dutch are more on the morphologically complex side, German and Italian seem to be well balanced. In the semi-parallel corpus (Figure 7) we observe a similar distribution. Finnish and Hungarian are clearly the morphologically most complex and syntactically most simple languages while English, in the bottom right quadrant, is the syntactically most complex but morphologically most simple language. Dutch, Italian, Spanish, Romanian and French cluster in the centre whereas German exhibits medium syntactic complexity and low morphological complexity. Both datasets exhibit a very significant trade-off regarding morphology and syntax: with a negative correlation of Pearson’s *r* = -0.93 (*p* = 0.0002) in the parallel and *r* = -0.79 (*p* = 0.006) in the semi-parallel corpus, most languages in our sample trade off morphological for syntactic complexity.

Figure 6. Morphological by syntactic complexity in the parallel Alice corpus.

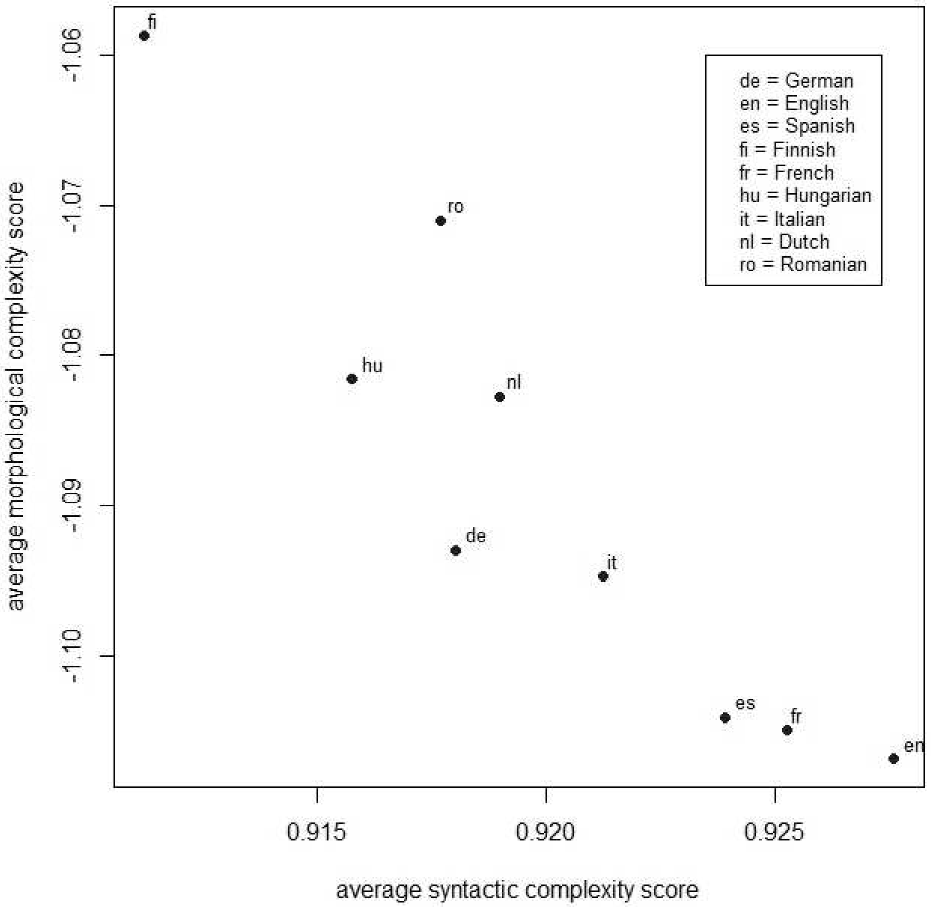
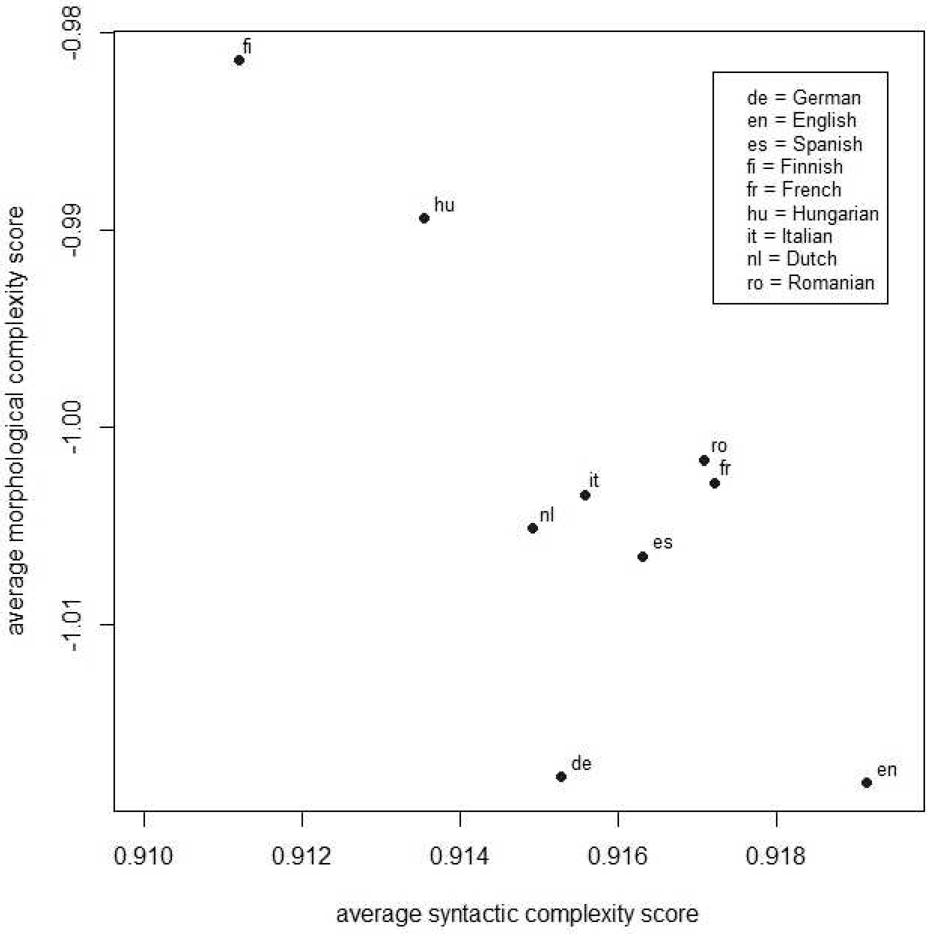


Figure 7. Morphological by syntactic complexity in the semi-parallel Alice corpus.



The morphosyntactic complexity analysis of the parallel and semi-parallel Alice corpus would, on the whole, seem to yield similar results. This is another way of saying that the compression technique can be effectively used with semi-parallel texts. In order to back up these findings statistically, we correlate the average syntactic and morphological complexity scores of the two datasets. At the syntactic level the two datasets correlate very highly (Pearson's r = 0.89, p = 0.0006)[7]; at the morphological level the correlation is less strong but still high (Pearson's r = 0.73, p = 0.01)[8].

In order to validate our findings, we compare the results obtained via compression to more traditional research. Bakker (1998) investigates syntactic complexity on the basis of word-order patterns and their flexibility. He assigns values between 0 and 1 for syntactic flexibility based on twelve word-order variables. Values close to zero indicate less flexibility and thus increased syntactic complexity (Bakker 1998: 387). Table 3 below shows the syntactic complexity rankings of our sample languages in the parallel and non-parallel Alice corpora as well as according to Bakker's flexibility indices. We compare Bakker's ranking and the rankings in our datasets by correlating our adjusted complexity scores with Bakker's flexibility values: the order of syntactic complexity in the parallel Alice corpus very highly correlates (Pearson's $r = 0.74$, $p = 0.02$)[9] with Bakker's findings. The semi-parallel ranking correlates to a lesser degree but is still significant (Pearson's $r = 0.5$, $p = 0.01$)[10].

Table 3. Syntactic complexity ranking according to Bakker's flexibility indices and our syntactic complexity scores. Languages are listed in decreasing order of syntactic complexity.

			Parallel Alice corpus			Semi-parallel Alice corpus		
rank	language	Bakker's flexibility index	rank	language	syntactic complexity score	rank	language	syntactic complexity score
1.	French	0.1	1.	English	0.928	1.	English	0.919
2.	Spanish	0.3	2.	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

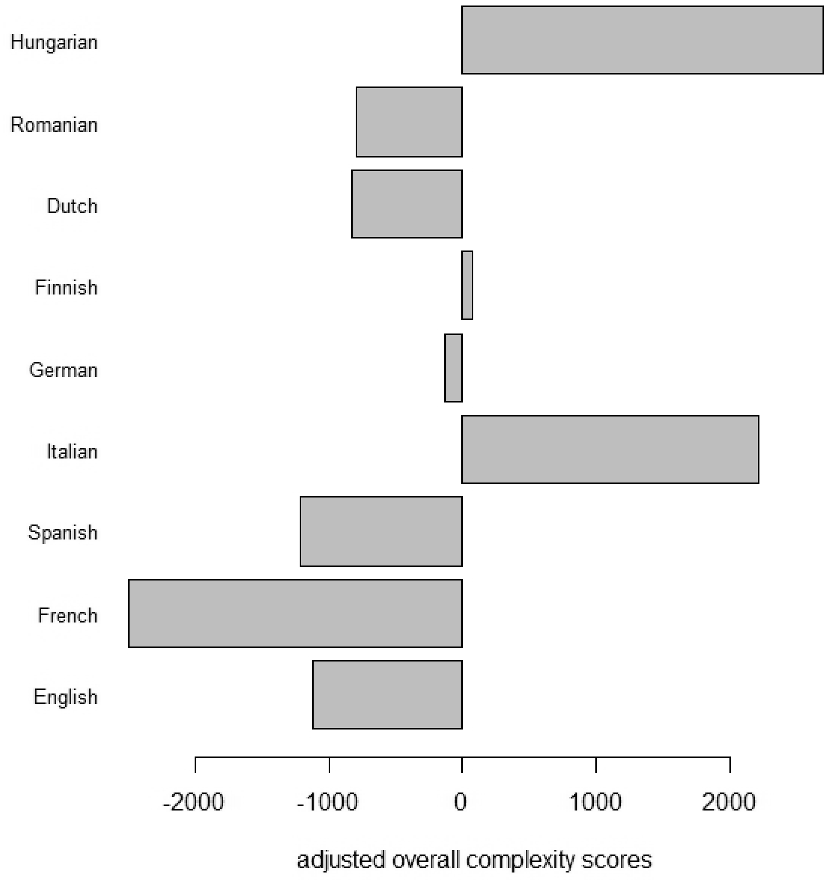
In sum, the results of the parallel and semi-parallel corpus for overall linguistic complexity are fairly similar, and the complexity rankings we obtain correlate with those reported in previous complexity research. For some interesting reason – a detailed discussion of which is reserved for another occasion – dislocations are especially pronounced among the languages that are fairly balanced between morphological and syntactic complexity. But in all, the compression technique achieves good results across both corpora. We have thus demonstrated that the use of the compression technique is not in principle limited to parallel corpora, but can successfully be used with semi-parallel corpora.

4.2.2. Newspaper texts

Next, we compile a non-parallel corpus of newspaper[11] texts on several contemporary topics in Dutch, English, Finnish, French, German, Hungarian, Italian, Romanian and Spanish. The topics were chosen according to their availability across the 9 languages. In this paper we analyse articles which were tagged as dealing with the ‘Euro crisis’ and ‘Congo’[12]. For each topic we sample the same number of sentences in order to keep sample size constant across the different language samples. Methodologically, we proceed as described above and calculate adjusted complexity scores as indicator of overall complexity. For the calculation of morphological and syntactic complexity, the news texts are subjected to multiple distortion and compression with N = 1,000 iterations. The average morphological complexity score[13] and the average syntactic complexity score[14] are subsequently calculated as outlined in the previous section.

Figure 8 shows the overall complexity ranking of the non-parallel newspaper texts. The Kolmogorov metric rates Hungarian as the most complex language in the dataset. It is closely followed by Italian and (in decreasing order of complexity) Finnish, German, Romanian, Dutch, Spanish, English and French. Similarly to the semi-parallel Alice corpus, languages which are balanced between morphological and syntactic complexity – such as Italian, German or Dutch – tend to become increasingly complex with increasing lack of 'content control'. Languages on the extreme end of the complexity scale such as Hungarian and English, on the contrary, seem to be hardly affected.

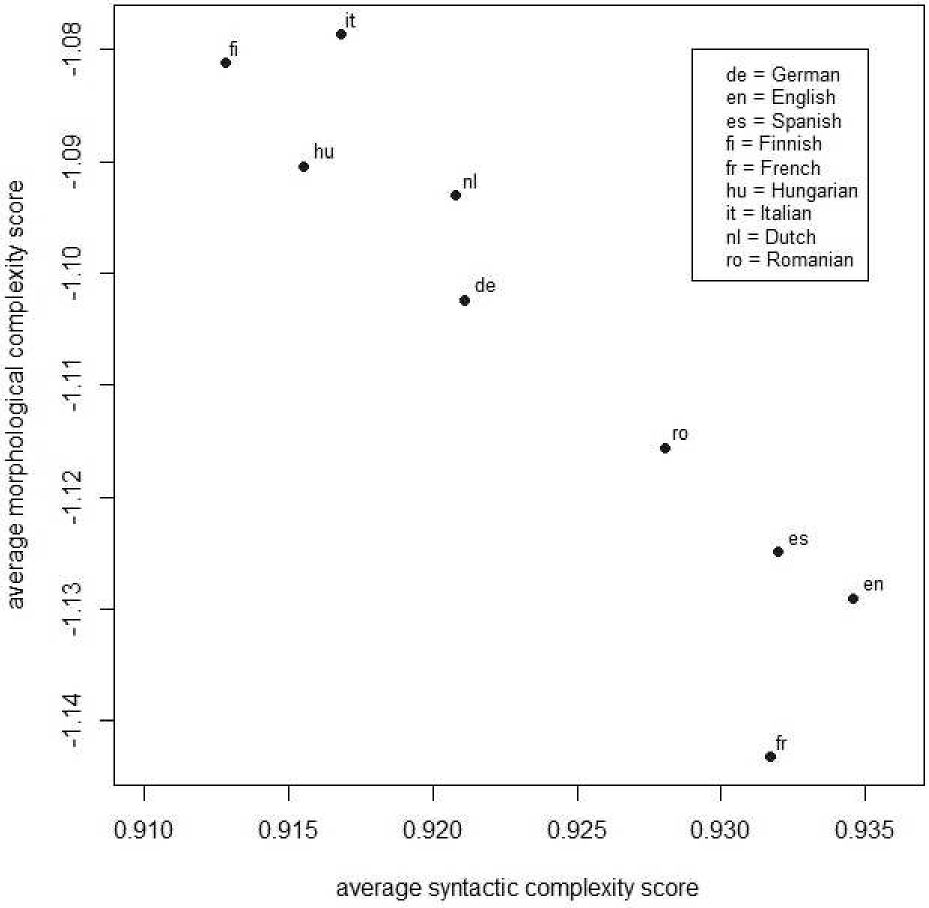
Figure 8. Overall complexity hierarchy in non-parallel newspaper texts.



As far as overall complexity scores are concerned, the correlation between the non-parallel news database and the parallel Alice corpus is moderate (Pearson's $r = 0.49$, $p = 0.09$)[15].

The analysis of morphological and syntactic complexity is shown in Figure 9. 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. Syntactically complex but morphologically simple languages, i.e. Spanish, English and French, occupy the bottom right quadrant.

Figure 9. Morphological by syntactic complexity in non-parallel newspaper texts.



Apart from the Italian data point (which is an as yet inexplicable outlier), these results tie in neatly with our previous findings. In fact, we observe a very high, significant correlation between the parallel and non-parallel complexity scores on the syntactic level (Pearson's *r* = 0.84, *p* = 0.002)[16] and a high, significant correlation on the morphological level (Pearson's *r* = 0.63, *p* = 0.03)[17].

Comparing our findings for syntactic complexity in the non-parallel corpus with Bakker's (1998) flexibility values (Table 4), we observe a moderate correlation (Pearson's *r* = 0.49, *p* = 0.1)[18].

Table 4. Syntactic complexity ranking of according to Bakker's flexibility indices and our syntactic complexity scores in the non-parallel newspaper corpus. Languages are listed in decreasing order of syntactic complexity.

rank	**language**	**Bakker's flexibility index**	**rank**	**language**	**syntactic complexity score**
1.	French	0.1	1.	English	0.935
2.	Spanish	0.3	2.	Spanish	0.932
3.	Italian	0.3	3.	French	0.932
4.	English	0.4	4.	Romanian	0.928
5.	German	0.4	5.	German	0.921
6.	Dutch	0.4	6.	Dutch	0.920
7.	Romanian	0.5	7.	Italian	0.917
8.	Finnish	0.6	8.	Hungarian	0.916
9.	Hungarian	NA	9.	Finnish	0.912

5. Discussion and outlook

In this chapter, we have explored the reach and limits of information theoretic methodologies to measure linguistic complexity on the overall, syntactic and morphological level. We set out by measuring complexity in parallel texts using translational equivalents of the Bible in some ten (historical) varieties of English and six European languages. The analysis of both overall and morphosyntactic complexity yields linguistically meaningful results and corroborates findings from previous, more traditional research (e.g. Bakker 1998; Nichols 1992). Focusing on historical English Bible translations, we can trace the development of English from a morphologically complex and syntactically simple to a morphologically simple and syntactically complex language through time. In fact we find a statistically significant, negative correlation between morphological and syntactic complexity in all our datasets. This indicates a trade-off, à la Hockett (1958: 180–181), between morphological and syntactic complexity.

In a second step, we applied the compression technique in a slightly modified version – using multiple instead of simple distortion and compression – to a parallel and re-sampled semi-parallel corpus of *Alice's adventures in Wonderland* in nine European languages. Furthermore, we tested the compression technique on genuinely non-parallel texts using newspaper articles in the same nine languages. The complexity hierarchies obtained from the semi-parallel and non-parallel corpora were compared and correlated to the complexity ranking of the parallel *Alice* control corpus. In terms of the morphological and syntactic complexity, we achieve high to very high correlations between the parallel, semi-parallel and non-parallel corpora. Overall complexity correlates only moderately. Although control of the topic, if not the precise content, across the different corpus samples seems to be a factor crucial to the successful application of the compression technique, our results show that the compression technique is not limited to parallel texts but can also be successfully applied to semi-parallel and non-parallel texts. In all, using file compression tools to assess linguistic complexity promises an economical and radically objective way to measure linguistic complexity. We also demonstrated that the compression technique reliably works with different text types, i.e. religious, literary and newspaper texts.

The technique has, needless to say, drawbacks. For one thing, the compression technique is completely agnostic about things that are dear to many linguists, such as form-function pairings. The technique, as we use it, is also text based, and so it crucially requires the availability of corpora of text and/or speech (but see Blevins, this volume, for an information-theoretic approach to paradigmatic complexity). Furthermore, even though we seem to obtain linguistically meaningful results by measuring complexity drawing on off-the-shelf file compression programmes, we are not yet able to identify and linguistically interpret the precise patterns and strings which are recognised by these programmes to create compression economy. To address this issue, work is under way by the first-named author to develop a custom-programmed compression algorithm which works transparently to aid linguistic interpretation.

In any event, the full potential of the compression technique has not yet been fully explored. Compression algorithms appear to be a useful and quite reliable tool for measuring cross-linguistic complexity variance. It has yet to be shown how well the algorithms pick up intra-linguistic (for example, dialectal) complexity variance – say, between more and less isolated dialects of the same language – and whether it is possible to determine the precise weight of individual grammatical features, such as for instance genitive markers or tense and aspect markers, via more specific file manipulation (Ehret in preparation).

References

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

Bane, Max 2008 Quantifying and measuring morphological complexity. *Proceedings of the 26th West Coast Conference on Formal Linguistics*, 67–76.

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

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

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

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

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

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

Ehret, Katharina in preparation *A corpus based study of information theoretic complexity in World Englishes*. PhD dissertation, University of Freiburg.

Gries, Stefan Th. 2009 *Quantitative Corpus Linguistics With R. A Practical Introduction*. New York/London: Routledge.

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

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

Juola, Patrick 2008 Assessing linguistic complexity. In: Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson (eds.), *Language Complexity: Typology, Contact, Change*. Amsterdam, Philadelphia: Benjamins.

Kettunen, Kimmo, Markus Sadeniemi, Tiina Lindh-Knuutila, and Timo Honkela 2006 Analysis of EU Languages through Text Compression. In: Tapio Salakoski, Filip Ginter, Sampo Pyysalo, and Tapio Pahikkala (eds.), *Advances in Natural Language Processing*, Lecture Notes in Artificial Intelligence, 99–109. Heidelberg: Springer-Verlag Berlin.

Kortmann, Bernd, and Benedikt Szmrecsanyi (eds.) 2012 *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Berlin, New York: Walter de Gruyter.

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: Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson (eds.), *Language Complexity: Typology, Contact, Change*, 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.

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

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

Miestamo, Matti, Kaius Sinnemäki, 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 1992 *Language in Space and Time*. Chicago, IL: University of Chicago Press.

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

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.

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

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

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

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

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

[1] 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).

[2] We measure the dispersion across the individual data points by calculating the *variation coefficient*, which is defined as the ratio of the standard deviation (sd) to the mean: sd(data1)/mean(data1). This measure is more robust than the standard deviation and handles possible outliers in the data much better as it is independent from the mean size (Gries 2009: 203–204). The smaller the value of the variation coefficient is, the smaller is the dispersion in the sample, i.e. the more reliable is the mean. We report the following variation coefficients for morphological complexity ratios in the parallel corpus: Dutch - 0.0012, English - 0.0014, Finnish - 0.0013, French - 0.0014, German - 0.0012, Hungarian - 0.0013, Italian - 0.0012, Romanian - 0.0014, Spanish - 0.0012. The dispersion measures for syntactic complexity ratios in the parallel corpus are: Dutch - 0.0015, English - 0.0017, Finnish - 0.0016, French - 0.0016, German - 0.0015, Hungarian - 0.0017, Italian - 0.008, Romanian - 0.0016, Spanish - 0.0016.

[3] By taking 10% of the total number of sentences instead, for instance, of words or characters, we ensure that syntactic inter-dependencies remain intact while, at the same time, we keep the sample size across languages constant.

[4] The average overall complexity score is calculated on the basis of compressed and uncompressed file sizes. Variation coefficients for compressed file sizes are as follows: Dutch - 0.076, English - 0.88, Finnish - 0.074, French - 0.07, German - 0.083, Hungarian - 0.063, Italian - 0.08, Romanian - 0.077, Spanish - 0.07, and uncompressed file sizes: Dutch -

0.074, English - 0.084, Finnish - 0.067, French - 0.069, German - 0.08, Hungarian - 0.063, Italian - 0.079, Romanian - 0.079, Spanish - 0.073.

[5] This yields the following variation coefficients in the semi-parallel corpus for

(i) morphological complexity scores: Dutch - 0.0056, English - 0.0059, Finnish - 0.0054, French - 0.006, German - 0.0054, Hungarian - 0.0057, Italian - 0.006, Romanian - 0.006, Spanish - 0.0062,

(ii) syntactic complexity scores: Dutch - 0.0048, English - 0.0047, Finnish - 0.0053, French - 0.0051, German - 0.0045, Hungarian - 0.0059, Italian - 0.0049, Romanian - 0.0053, Spanish - 0.0052.

[6] Spearman's rho of correlation for overall complexity between the parallel and semi-parallel corpus is $r = 0.5$, $p = 0.09$.

[7] Spearman's rho of correlation for syntactic complexity between the parallel and semi-parallel corpus is $r = 0.82$, $p = 0.005$.

[8] Spearman's rho of correlation for morphological complexity between the parallel and semi-parallel corpus is $r = 0.73$, $p = 0.02$.

[9] Spearman's of correlation for the syntactic complexity ranking according to Bakker's flexibility value and syntactic complexity scores in the parallel Alice corpus $r = 0.75$, $p = 0.02$.

[10] Spearman's of correlation for the syntactic complexity ranking according to Bakker's flexibility value and syntactic complexity scores in the semi-parallel Alice corpus $r = 0.43$, $p = 0.1$.

[11] We retrieved articles from the following online newspapers:

Dutch: Volkskrant (http://www.volkskrant.nl/)

English: The Guardian (http://www.guardian.co.uk/)

Finnish: Helsinki Sanomat (http://www.hs.fi/) and Iltasanomat (http://www.iltasanomat.fi)

French: Le Figaro (http://www.lefigaro.fr/)

German: Die Welt (www.welt.de)

Hungarian: HvG (http://hvg.hu/) and Nepszava (http://www.nepszava.hu)

Italian: La repubblica (http://www.repubblica.it/)

Romanian: Adevarul (http://www.adevarul.ro/)

Spanish: ABC (http://www.abc.es)

[12] We also sampled articles on the topics ‘Iran’, ‘Tunisia’, ‘Kim il Jong’ and ‘Putin/Russia’. However, these topics yielded less satisfying results. Due 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, it is likely that – depending also on the political relations/interests among the respective countries – our sources substantially differ in the topics published under the same topic/tag.

[13] This yields the following variation coefficients for the morphological complexity scores in the non-parallel corpus: Dutch - 0.0017, English - 0.0016, Finnish - 0.0017, French - 0.0016, German - 0.0018, Hungarian - 0.0014, Italian - 0.0014, Romanian - 0.0017, Spanish -0.0016.

[14] This yields the following variation coefficients for the syntactic complexity scores in the non-parallel corpus: Dutch - 0.0022, English - 0.002, Finnish - 0.0021, French - 0.0019, German - 0.0023, Hungarian - 0.0019, Italian - 0.0017, Romanian - 0.0021, Spanish - 0.0021.

[15] Spearman’s rho of correlation for overall complexity between the parallel and non-parallel corpus is $r = 0.65$, $p = 0.03$.

[16] Spearman’s rho of correlation for syntactic complexity between the parallel and non-parallel corpus is $r = 0.81$, $p = 0.005$.

[17] Spearman’s rho of correlation for morphological complexity between the parallel and non-parallel corpus is $r = 0.63$, $p = 0.04$.

[18] Spearman’s of correlation for the syntactic complexity ranking according to Bakker’s flexibility value and syntactic complexity scores in the parallel Alice corpus $r = 0.38$, $p = 0.17$.