DE GRUYTER MOUTON
DOI 10.1515/cllt-2014-0009 — Corpus Linguistics and Ling. Theory 2014; 10(2): 175–211

Christian Bentz*, Douwe Kiela, Felix Hill and Paula Buttery

# Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts

**Abstract:** This paper reports a quantitative analysis of the relationship between word frequency distributions and morphological features in languages. We analyze a commonly-observed process of historical language change: The loss of inflected forms in favour of 'analytic' periphrastic constructions. These tendencies are observed in parallel translations of the Book of Genesis in Old English and Modern English. We show that there are significant differences in the frequency distributions of the two texts, and that parts of these differences are independent of total number of words, style of translation, orthography or contents. We argue that they derive instead from the trade-off between synthetic inflectional marking in Old English and analytic constructions in Modern English. By exploiting the earliest ideas of Zipf, we show that the syntheticity of the language in these texts can be captured mathematically, a property we tentatively call their *grammatical fingerprint*. Our findings suggest implications for both the specific historical process of inflection loss and more generally for the characterization of languages based on statistical properties.

**Keywords:** Zipf's law, vocabulary growth curves, diachronic corpus linguistics, syntheticity, analyticity, parallel texts, historical linguistics, Old English

***Corresponding author: Christian Bentz:** Department of Theoretical and Applied Linguistics, University of Cambridge, UK. E-mail: chris@christianbentz.de
**Douwe Kiela:** Computer Laboratory, University of Cambridge, UK.
**Felix Hill:** Computer Laboratory, University of Cambridge, UK.
**Paula Buttery:** Department of Theoretical and Applied Linguistics, University of Cambridge, UK.

## 1 Introduction

The usage, frequencies and distributions of words change throughout time. Neologisms and foreign words are integrated into the lexicons of languages, derivation and inflection are productively applied to build new grammatical forms and combinations of these forms become more or less acceptable. Historical linguistics aims to elicit the pathways along which these lexical, morphological and

syntactic processes develop. For example, when and why do new words like *to google* get introduced into the lexicon? When and why do irregular past tense forms like *dove* get traded off for regular *–ed* past tense markers as in *dived*? When and why does canonical word order change?

Traditionally, historical linguists have looked for exemplar-based evidence to answer these questions. However, there are drawbacks to this method. Using single examples to illustrate points about diachronic tendencies can lead to 'cherry picking' of the instances that best fit a particular theory.

The development of diachronic corpora and improved computational tools facilitate an alternative approach, involving the statistical analysis of whole corpora. This approach has proved effective at demonstrating subtle tendencies in the usage of morphological patterns. For example, using permutation testing methods for the BNC (British National Corpus) Säily (2011) has shown that women tend to use the derivational suffix –ity less productively than men, and that this could have interesting implications for gender-based sociolinguistics. In a similar vein, Lieberman et al. (2007) constructed a quantitative model for the trade-off between irregular and regular morphology in English verbs, and how the productivity of these morphological markers changed over time. And Baayen (2008: 118–126) uses texts of different registers in combination with a clustering approach to point out how stylistic factors impact the productivity of derivational affixes. It is unlikely that lower level example-based approaches would have noticed the subtle differences demonstrated in these studies. The data-driven approach can be particularly advantageous in the context of diachronic studies, because minor differences in usage patterns might accumulate over several generations and form a successive process of gaining or loosing new inflectional paradigms.

Against this backdrop, this paper focuses on the loss of inflectional morphology from Old English towards Modern English. We propose that quantifying the differences in the frequency distributions of lexical items in parallel texts can help to capture the 'synthetic state' of a language, i.e. whether inflectional morphology and clitics (as in *John's book*) or periphrastic constructions (as in *the book of John*) are predominant. We analyze parsed and POS tagged digital versions of the Book of Genesis in Old English and Modern English, employing a range of quantitative methods to measure grammatical differences, which we show to be significant even if confounding factors are controlled for. We propose that this represents an important first step in modeling and understanding how morphological encoding strategies change over time.[1]

**1** We do not, however, claim that the particular text we are using is representative of Old English on the whole, for any claim along these lines would require a larger sample of different texts.

To this end, in Section 2 we illustrate the process of inflection loss with examples. Section 3 analyzes the statistical distributions of the parallel texts, and explains how the process of inflection loss is reflected in these distributions. In Section 4 we outline how these statistical distributions can be modeled mathematically, enabling the syntheticity of a text to be represented by mathematical parameters, or a 'grammatical fingerprint'. We conclude by considering the implications of our findings with respect to the process of inflection loss, and discuss the potential value of the grammatical fingerprint construct as a means to characterize or classify languages.

# 2 Historical language change and Zipf distributions

## 2.1 The loss of inflection from Old English to Modern English

It is generally assumed that morphological markers were used abundantly in Old English in the noun, verb and adjective paradigms (see for example Campbell [1959: 222–352]; Lass [1994: 123–178]; and Hogg and Fulk [2011: 69–323] for a detailed discussion; but also Thomason and Kaufman 1991, for a critical review). On the other hand, Modern English tends to circumscribe the same information using periphrastic constructions. This effect can be shown for case marking in Example (1), which is taken from the Old English Bible (see Appendix 1 for the texts used).

Gen. 1:2

| | | | | | |
|---|---|---|---|---|---|
| (1) | *God-es* | *gast* | *wæs* | *geferod* | *ofer* |
| | god-GEN.SG | spirit.NOM.SG | be.PRF.3SG | move.PST.PTCP | over |
| | *water-u* | | | | |
| | water-ACC.PL | | | | |
| | 'The spirit of god had moved over the waters.' | | | | |

In the Modern English parallel sentence there is an interesting difference to be noted: The inflectional genitive *godes gast* is replaced by the periphrastic construction *spirit of god*. These patterns can be found across a multitude of sentences. Quite generally, inflectional morphemes indicating genitive, dative and accusative case in Old English are traded off in Modern English for prepositional phrases headed by *of*, *for*, *from*, *to* etc. (see examples in 2–4).

Gen. 2:12

(2) [. . .] *Þæs* *land-es* *gold*
That.GEN.SG land-GEN.SG gold.NOM.SG
'[. . .] the gold of that land'

Gen. 35:12

(3) *Þæt* *land* *Þæt* *ic* *sealde*
That.ACC.SG land.ACC.SG that.ACC.SG I give.PRF
*Abraham-e* *ond* *Isaac-e* [. . .]
Abraham-DAT.SG and Isaac-DAT.SG
'The land (that) I gave to Abraham and Isaac [. . .]'

Gen. 3:21

(4) *God* *worhte* *eac* *Adam-e* *ond* *his* *wif-e*
God.NOM.SG make.PERF also Adam-DAT.SG and his wife-DAT.SG
*fellen-e* *reaf* [. . .]
made of skins-ACC.PL garment.ACC.PL
'The Lord God made garments from skin for Adam and his wife [. . .]'

Such trade-offs are not restricted to the noun paradigm, and extend to verbal inflections, adjectives, pronouns and other word classes. For example, in OE we can find synthetic encoding of future tenses[2] and subjunctives as in (5) and (6).

Gen. 3:5

(5) *eowre* *eag-an* *beo-ð* *geopenode*
your.GEN.PL eye-NOM.PL be-PRS.IND.PL open. PST.PTCP
'Your eyes shall be opened.' (KJV)

Gen. 3:1

(6) *ge* *ne* *æt-en*
you.NOM.PL not eat.PRF-SBJV.PL
'you must not eat.' (NET)

Comparing the Old English sentences to the Modern English parallel translations, we note the different strategies for encoding future and subjunctive meanings as in OE *beoð* 'shall be' and OE *æten* 'should eat'. While OE uses inflectional

**2** Note, that strictly speaking OE does not have an inflectional future tense, since the forms used to encode future meaning are not separable from present forms. However, in some contexts it is clear that a future meaning is encoded, and in these cases the KJV and NET translations will use the MnE periphrastic constructions to encode the future meaning.

marking, MnE employs auxiliary verbs like *should, shall, would, will, must, etc.* The richness of grammatical marking in OE is also reflected in the multitude of verb forms. For example, the strong verb *sing* could correspond to OE *singan, singe, singest, sing, singaþ, singen, singenne*. A full-blown description of the inflectional paradigms can be found in Campbell (1959: 222–352), Lass (1994: 123–178) and Hogg and Fulk (2011: 69– 323). These also include a discussion of adjective classes and pronoun paradigms which have likewise declined in terms of the richness of forms from OE towards MnE.

## 2.2 Zipf's law and the degree of inflection

The idea that such grammatical differences can be reflected in quantitative analyses goes back to the earliest writings of Zipf (1932, 1965 [1935]). He considered the number and distributions of unique word forms in different languages to be linguistically interesting. Analyzing the patterns of word frequencies in Latin, Chinese and English (Zipf 1932) as well as Old English, French, Hebrew, Plains Cree and others (Zipf 1949: 95, Zipf 1949: 129, 1965 [1935]: 252) he suggests that it is possible to measure the "degree of inflection" in what he calls "positional" (i.e. analytic) and "inflected" (i.e. synthetic) languages.

Zipf focused his analyses on subtle differences in the frequency distributions of languages (*Zipf distributions*), which he considered to be indicative of the syntheticity of texts. Recently, the availability of digital corpora has made it possible to evaluate this intuition more precisely. Baayen (2001: 39–133, 2008: 223–236, 2009), for example, examined and criticized measures for lexical richness, such as the type/token ratio and the ratio of hapax legomena versus total number of words. Baroni (2009) showed that the shapes of the Zipf distributions of lemmatized and non-lemmatized versions of the BNC (British National Corpus), the Japanese Web Corpus and the Italian *la Repubblica* Newspaper Corpus differ, a finding that suggests an interesting connection between morphology and frequency distributions (Baroni 2009: 811–812). Further, Ha et al. (2006) noted explicitly that there are systematic differences between the frequency distributions of strongly inflected languages (Latin and Irish) and largely analytic languages (Spanish and English). Another, more wide-ranging study of the variance in word frequency distributions can be found in Popescu et al. (2009). They use texts of different genres from 20 languages, ranging from highly analytic (Maori) to highly synthetic (Hungarian), and consider a variety of measures that could be used to assess the relative syntheticity of a language (Popescu et al. 2009: 18–71). Popescu et al. (2010) elaborate this approach by developing quantitative models to

compare systematic differences between low- and high-frequency regions of Zipf distributions.

However, despite the quantitative elaboration of the accounts by Baroni (2009), Ha et al. (2006) and Popescu et al. (2009, 2010) each of these studies lacks a thorough linguistic analysis of the causes of differences in frequency distributions. Indeed, Popescu et al. (2010, Conclusion) state that a deep "linguistic or textological" analysis is still necessary to interpret the quantitative models they represent. In light of these observations, in this study we aim to bridge the gap between quantitative models of differences in frequency distributions and the grammatical causes underlying these differences. We start by outlining how morphological features of a language are reflected in the frequency distributions for texts of this language.

# 3 The impact of analytic and synthetic constructions on frequency distributions

In this section, we demonstrate how the syntheticity of a language, represented by its morphological coding strategies, comes to be reflected in frequency distributions. This enables us, in the following section, to show that such frequency-based analyses can be a useful means of classifying languages, reflecting interesting grammatical differences. Our demonstration of the connection between morphological coding and frequency distributions involves two stages. First, we show how morphological patterns relate to frequency counts, thus making them a likely candidate for the cause of the observed cross-lingual statistical variation. Second, we conduct analyses to assess the effect of other possible confounds.

## 3.1 The grammatical phenomena in focus

As pointed out in Section 2, the development from synthetic structures in OE to analytic structures in MnE is mainly reflected in two changes:

a. The trade-off between case marked *content words* in OE and periphrastic constructions with prepositions in MnE
b. The trade-off between verbs marked for subjunctive in OE and periphrastic constructions with modal auxiliaries in MnE

In the context of (a), *content words* are open class items such as nouns, proper nouns, adjectives, adverbs, verbs in the infinitive and past participles. In Old

English most of these content words can, in principle, be *distinctively* marked for case. Distinctively means that there are at least two separate inflectional forms which are used without mutual overlap, i.e. without *case syncretism*. For example, the OE noun *land* (MnE. *land, country*) displays case syncretism for the nominative and accusative, but can be distinctively marked for the dative: *land<u>e</u>*, and genitive: *land<u>es</u>*. This is not the case in MnE. The frequencies of the different word types for the same lemma *land* in OE as well as the frequency of the direct translation in MnE can be seen in Table 1. Note that in the Old English Genesis the total number of occurrences is spread over three distinct types, resulting in relatively low frequencies for each, but that together they add up to roughly the same number as in MnE (139 versus 151). Crucially, for every sentence in which we find the case marked forms *lande* or *landes* in Old English there must be a parallel sentence in Modern English in which either periphrastic constructions (e.g. *to the land*, *of the land*) or fixed word orders are used to encode the same information. Both cases result in higher frequencies for prepositions and/or the nouns involved.

**Table 1:** OE and MnE frequencies for a single noun

| | Word | Freq. | POS* |
|---|---|---|---|
| **OE Genesis** | *lande* | 79 | [79 N^D] |
| | *land* | 49 | [43 N^A] [6 N^N] |
| | *landes* | 11 | [11 N^G] |
| | *Total* | **139** | |
| **MnE Genesis** | *land* | **151** | [151 NH] |

* part-of-speech: N^N (nominative noun), N^D (dative noun), N^A (accusative noun), N^G (genitive noun)

The effect described in (b) refers to other inflectional paradigms beyond nouns, such as the synthetic subjunctive, as seen in (6), where the synthetic form *æten* is replaced by 'must eat' using a modal auxiliary verb in Modern English (see also Lass 1994: 123 for a detailed discussion).

The effects (a) and (b) both suggest that the OE grammar using bound morphemes will in general give rise to low frequency items, whereas analytical structures as in MnE favor even higher frequencies for already common items such as *function words* (prepositions, determiners, quantifiers, etc). This hypothesis is tested for the Old English and Modern English Genesis in the next sections.

## 3.2 Methods and Materials

### 3.2.1 Frequency distributions

The first step towards systematically analyzing the differences in frequency distributions is to generate so-called *rank/frequency profiles* (Popescu 2009; Baayen 2001; Baroni 2009) for the words in the relevant texts. In such a profile every type is assigned a rank according to its frequency. The most frequent type is assigned rank one, the second most frequent type is assigned rank two, and so on. The first ten ranks of the profiles for the Modern English as well as the Old English translations of the Genesis can be seen in Table 2.

Unsurprisingly, function words such as conjunctions, prepositions and personal pronouns are strongly represented in both corpora. Moreover, the frequencies seem roughly to follow Zipf's power law distribution (or a modified version of it). Zipf's law states that there is a systematic relationship between the frequencies of words in a corpus and their rank in a list which sorts the words by number of occurrences (Zipf 1932, 1965 [1935], 1949).

Generally, there are only few highly frequent words found in almost all texts and a long tail of so-called *hapax legomena*, i.e. words which occur only once. In consequence, the frequency of a word can be predicted by its rank in a rank-frequency profile like the one seen in Table 2. However, to assess the differences in these distributions more clearly, we need to introduce another concept: type accumulation curves.

**Table 2:** Old English and Modern English ranked frequencies

| Word | Freq. | Rank | Word | Freq. | Rank |
|---|---|---|---|---|---|
| **OE Genesis** | | | **MnE Genesis** | | |
| *and* | 1731 | 1 | *the* | 1775 | 1 |
| *he* | 535 | 2 | *and* | 1024 | 2 |
| *to* | 497 | 3 | *of* | 821 | 3 |
| *ða* | 468 | 4 | *to* | 808 | 4 |
| *on* | 439 | 5 | *you* | 521 | 5 |
| *þæt* | 352 | 6 | *he* | 479 | 6 |
| *ic* | 342 | 7 | *his* | 420 | 7 |
| *þa* | 326 | 8 | *in* | 366 | 8 |
| *ðe* | 282 | 9 | *i* | 359 | 9 |
| *hi* | 278 | 10 | *will* | 300 | 10 |

### 3.2.2 Type accumulation curves

The same information about type/token ratios that was displayed by means of Zipf-distributions can also be captured in so-called *type accumulation curves*. In this case, instead of plotting frequencies of types according to their ranks, the number of types is counted for a subset of the running tokens (in our example a chunk of 1000 tokens). Hence, the type accumulation curves start off with a single point denoting the number of types for the first 1000 tokens in the text. Moving along the x-axis, the numbers of types for the next chunks of tokens are added to the overall type count, i.e. they are accumulated over the whole text. Type accumulation curves in general have positive but declining slopes, since the likelihood of encountering new types decreases as a function of the types already seen.

## 3.3 Results: Distributions for the Old English and Modern English Genesis

Using frequency distributions and type accumulation curves we can now compare the similarities and differences in the quantitative patterns of the Old English and Modern English Genesis. Figure 1a illustrates the overall picture.

There are a high number of distinct types in Old English. This is reflected in the long black tail of hapax legomena in Figure 1a, whereas Modern English has higher frequencies for the first ~100 ranks. This is illustrated in Figure 1b, which is a zoom into the 100 highest ranks of both distributions. The empirical type accumulation curves are given in 1c by the dotted lines. To test whether the differences in type accumulation curves are significantly different, we apply the method of Säily (2011) and Säily and Suomela (2009), using their open source software for computing confidence intervals for type and hapax accumulation curves (Suomela 2007) (see Appendix 4 for a description of the method).

As can be seen in 1c, although for the first 2 chunks of text the numbers of types in MnE would still fall within the 95% confidence interval of OE, from the third chunk onwards the distributions clearly diverge. Again, this illustrates that overall the OE Genesis employs more different words or types to encode essentially the same information, i.e. it has a higher type richness.[3] In the following

**3** It is important to note that we counted abbreviations and clitics as separate types in MnE. The following instances of types derived this way can be found in the MnE Genesis, with their respective frequencies in parentheses: *s* (199, both genitive *'s* and abbreviated *is*), *t* (31), *ll* (14), *ve* (3), *re* (2), and *d* (1). These instances sum up to 6 extra types and 250 extra tokens. Now, the overall

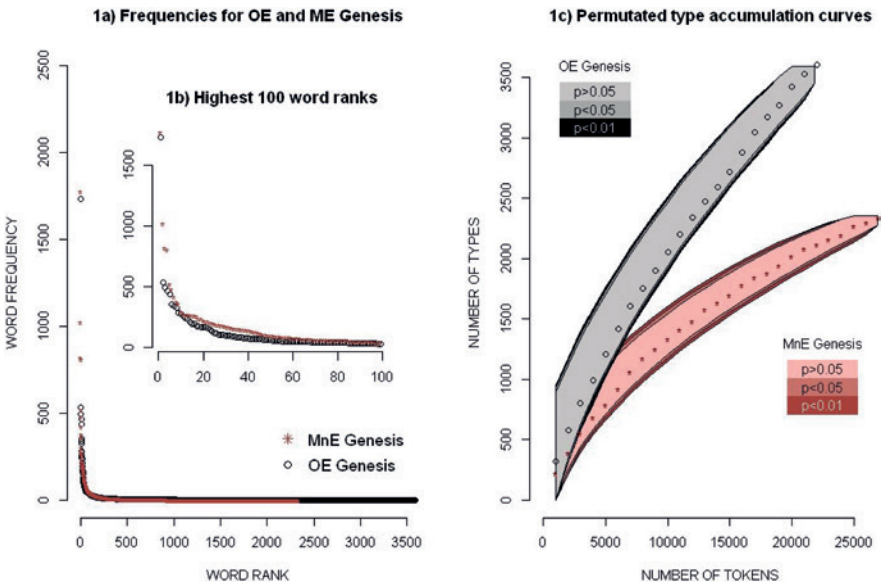


**Fig. 1:** Complete frequency distributions for Old English and Modern English Genesis (1a), and for first one hundred ranks (1b). The plot in 1b is just a zoom into the highest 100 ranks. 1a illustrates that the tail of hapax legomena in MnE (red stars) is shorter than in OE (black dots). The difference in frequencies for higher ranks can be seen in 1b. The empirical type accumulation curves for OE (black dots) and MnE (red stars) with confidence intervals are then given in 1c

sections we show how these significant differences in frequency distributions for OE and MnE stem from the different morphological marking strategies.

### 3.3.1 Frequencies of inflected nouns and verbs in Old English

In a first step, we select all the distinctively case marked content words and all the distinctively subjunctive marked verbs in the OE Book of Genesis and mark them by different colors in the overall rank frequency profiles. The results are plotted in Figures 2a–b. The two plots should be considered in parallel to the plots in Figure 1a and 1b, the only difference being that in figure 1a–b the ranked types with their

difference between OE and MnE types is 1245, and the overall difference in tokens is 5096. Hence, the decision to count clitics and abbreviations as separate types accounts for a mere 0.5% and 5% of the variance in the type and token counts.

token frequencies were represented as points, whereas in 2a–b they are represented as thin black columns in a histogram.

A clear trend can be observed: The case marked content words and the subjunctive verb forms are represented more strongly towards the lower frequencies in the tails of hapax legomena, whereas other word classes like function words will display higher frequencies.

Quantifying these trends, we find that 1016 of the 2031 hapax legomena in OE (~50%) are either case marked content words or verbs distinctively marked for subjunctive. Most interestingly, the diachronic comparison shows that the MnE distribution has 1046 hapax legomena less than OE. Since we know that case marking as well as distinctive subjunctive forms are lost in Modern English, we can conclude that the loss of the 1016 morphologically marked types (with frequency = 1) from Old English to Middle English and Modern English should make up for ~97% of the length difference, given that every inflected form occurs in parallel with a non-inflected form and/or another inflected form of the same lemma. In this case the originally inflected forms will add up to a common type with frequency > 1 when the distinct marker is lost.

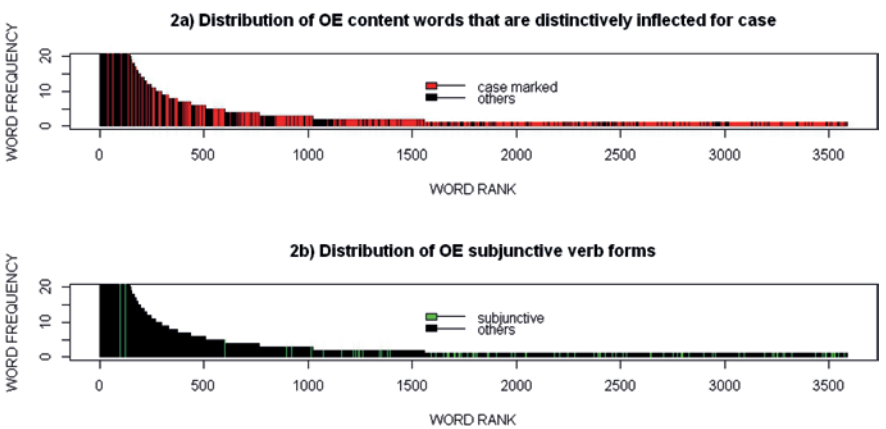


**Fig. 2:** Histograms for frequency distributions of the Old English Genesis. Every word type in the distribution is indicated by a column (although the upper tails of the highest frequency ranks on the y-axis are cut in order to make the tails of hapax legomena on the x-axis visible). In 2a all content words with distinct case markers are indicated by red lines. In 2b all subjunctive verbs are indicated by green lines

Overall, a length difference of 97% for the tails of hapax legomena is sufficient to create a significant difference in the type accumulation curves, as seen in Figures 1a–c.

Moreover, we can measure the degree of deviation for the empirical type accumulation curves by summing up the differences in numbers of types for every parallel chunk and dividing the result by the sum of the overall numbers of types (up to the last parallel chunk) of the OE and the MnE texts. This results in a 23% difference between the OE and MnE type accumulation curves. Hence, it is justified to conclude that having abundant inflectional marking in OE will enhance the number of hapax legomena and account for a significant part of the discrepancies between MnE and OE type accumulation curves.

### 3.3.2 Higher frequencies of prepositions and modal verbs in Modern English

Besides the effect of a shorter tail of the MnE distribution, any differences in frequencies of the higher ranked types will also have an impact on type accumulation curves. We therefore investigate whether loss of inflections might in turn enhance the frequencies of other word classes such as prepositions (a) and modal verbs (b). Specifically, we aim to explore whether the higher frequencies for the highest ranks in Modern English are at least partly due to higher frequencies in items such as prepositions (*of, to, for, from*, etc.) and modal verbs (*will*, *can*, *should*, etc.).

To test this, the Modern English Genesis was parsed using the *Stanford-Parser* (Klein and Manning 2003) and the Old English POS tags were extracted with the *CorpusSearch* software (see Appendices 2 and 3 for the respective tag sets). Both have distinct tags for the categories 'prepositions and subordinating conjunctions' and 'modal verbs' respectively. Since we are dealing with a limited set of prepositions and modal verbs, we are able to check the differences for every single item, as per Figure 3.

Figure 3a displays prepositions and subordinating conjunctions for the highest 100 ranks only. By adding the labels of word types in the plot we can 'trace' them within the Old English and Modern English distributions and pin down the change of their frequencies. The frequencies of prepositions in MnE are systematically higher. Namely, *of* and *to* both occur more than 800 times in MnE, compared to 149 and 497 times in OE, the preposition *for* occurs 264 times in MnE and 120 times in OE, and *from* occurs 188 times in MnE and its counterpart *fram* in OE occurs 41 times.

In the case of modal verbs, a one-to-one comparison would be less straightforward, since they are more prone to semantic change (e.g. differing meanings for *will* in OE and MnE). However, the overall trend (Figure 3b) is similar to the one observed for complete frequency distributions and prepositions only. Note that modal verbs in OE can themselves be inflected (e.g. *sceolde* and *scealt* from

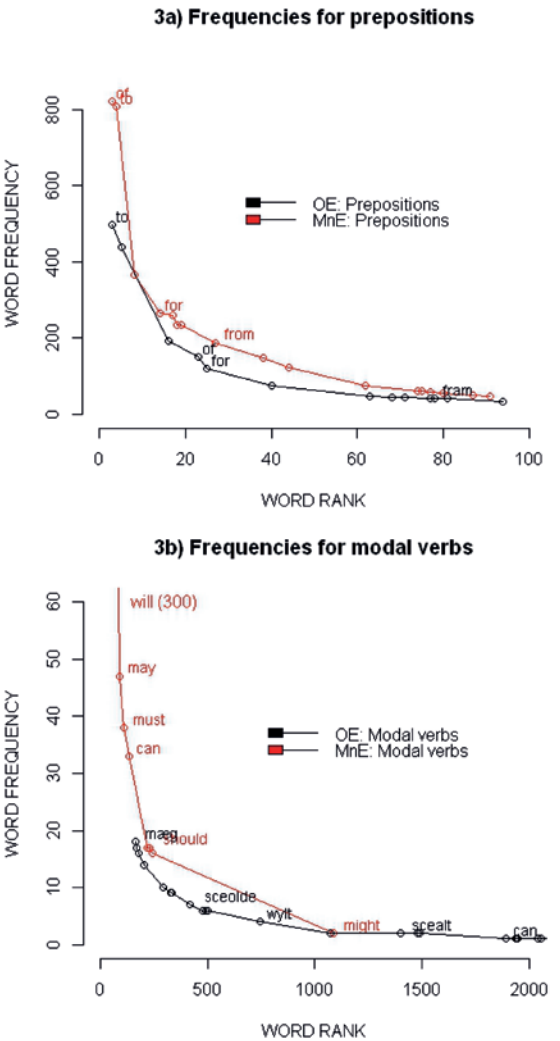


**Fig. 3:** Comparison between frequencies of selected prepositions in OE (black dots) and MnE Genesis (red stars) (3a). A direct comparison of the prepositions corresponding to each other will render higher frequencies for the MnE prepositions: *of* (MnE) > *of* (OE), *to* (MnE) > *to* (OE), *for* (MnE) > *for* (OE), *from* (MnE) > *fram* (OE). Likewise, modal verbs also generally have higher frequencies in MnE (3b)

**sculan* 'shall, must'). This results in more different types of modal verbs for OE (41) than for MnE (8), which again goes hand in hand with generally lower frequencies in OE.

To sum up, by tracing the frequencies of specific types in the MnE and OE Genesis two phenomena were observed: 1) In general, distinctively case marked nouns as well as subjunctive marked verbs in OE have relatively low frequencies (Section 3.3.1); 2) there are differences in the frequency distributions of prepositions and modal verbs (Section 3.3.2). These findings are consistent with the fact that such trade-offs represent a part of the overall tendency from a synthetic to analytic encoding strategy. Further, they are a quantifiable, straightforward explanation for the observed differences in frequency distributions. Nevertheless, in order to be sure of the causation we now investigate whether the effect could in fact be driven by other factors such as style of translation (Section 3.4.1), contents and length of texts (Section 3.4.2) as well as orthography (Section 3.4.3).

## 3.4 Discussion: Checking alternative explanations for differing distributions

### 3.4.1 Style of translation

It is possible that differences in frequency distributions stem from variables that could also occur in synchronic comparisons, such as authorship, style of translation and the particular source text. To test whether this is the case, the King James Version (KJV) as found in the *Oxford Text Archive* was used as another parallel text (an *Early Modern English* version) (see Appendix 1 for more details). Note that the NET translation is mainly based on the Masoretic texts and other sources of Hebrew tradition, whereas the KJV translators relied on a variety of ancient source texts, earlier English translations such as the Bishop's Bible and even contemporary translations in other languages (Carroll and Prickett 2008).

The frequency distributions and type accumulation curves for the NET bible and the KJV can be seen in Figure 4.

Overall, the curves look fairly similar, although the type accumulation curves still diverge towards the end of the texts. It is important to note that the aim of the NET was to translate the bible using a literary but not formal style, understandable for every native speaker of present day English, whereas even the revised versions of the KJV are inherently conservative in their style, still strongly influenced by the original text published in 1611. This leads to paratactic sentence structures in KJV. Words and whole phrases are sometimes repeated across sentences. This might be the reason why the KJV has significantly fewer types than the NET Bible. However, in terms of case marking and other inflectional paradigms like the subjunctive and future tense, Modern English and Early Modern English are similar.

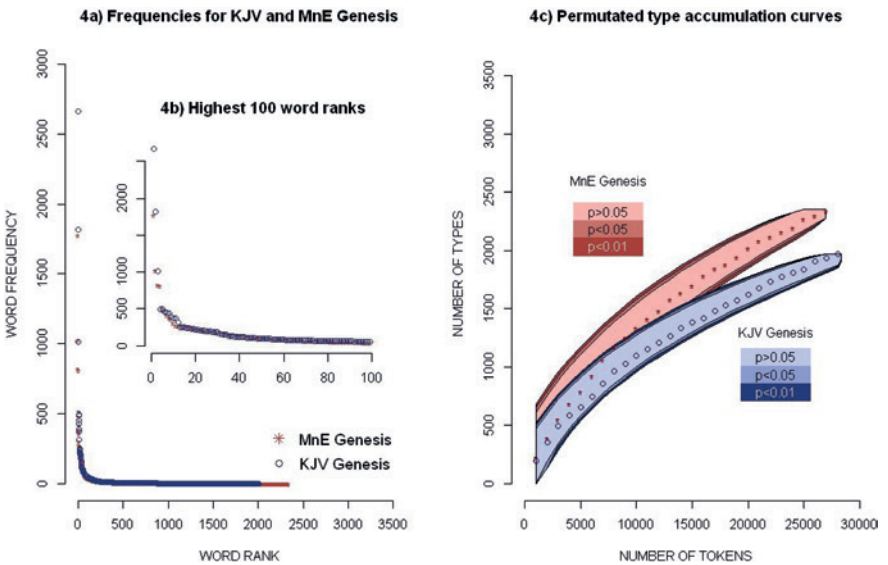


**Fig. 4:** Comparison between the New English Translation (red stars) and the King James Version (blue dotts). 4a and 4b display the empirical Zipf distributions, which are similar in terms of the length of tails (4a) and in terms of the highest frequency items (4b). This is also reflected in the type accumulation curves in 4c, despite a significant deviation towards the ends of the curves

Summing the differences in numbers of types per chunk and dividing them by the overall number of types per chunk, we observe a 10% difference for the KJV and the NET translation. Remember that the difference for the OE and the MnE (NET) texts is 23%. Overall, the higher similarity of the KJV Genesis and the NET Genesis in comparison to the OE Genesis suggests that variation in style and source texts can have an impact on the distributions, but that this impact is not dominant.

### 3.4.2 Content and length

It is conceivable that the differences in the distributions stem from either differences in the information content of the texts or simply from differences in their length. To test this, another text in the NET and the OE translation was analyzed: the Book of Exodus. We made pair-wise comparisons of the MnE Genesis with the MnE Exodus (Figure 5a–c), on the one hand, and the OE Genesis and OE Exodus on the other hand (Figure 6a–c).

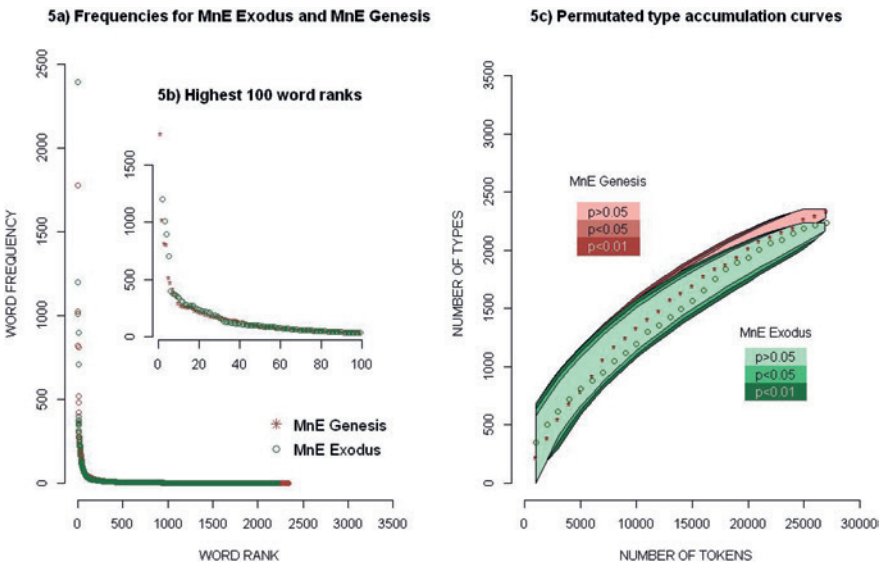


**Fig. 5:** Comparison between MnE Genesis (red stars) and MnE Exodus (green dotts), plots 5a and 5b illustrate the close fit of the two distributions to each other. Both the tails and high frequency items in 5a and 5b are nearly undistinguishable. This results in widely overlapping confidence intervals for type accumulation curves

Although Genesis and Exodus are two different narratives which are encoded with different vocabularies, the type accumulation curves for these texts are strikingly similar. Note, moreover, that while the MnE Genesis and Exodus have almost the same length in terms of tokens (26869 versus 26885 tokens), the OE Genesis and Exodus have vastly differing lengths (21766 versus 14223 tokens) due to a different choice of passages in the text sources we use. Nevertheless, the type accumulation curves for OE Genesis and Exodus are almost indistinguishable, despite the differences in the Zipf curves. This is also reflected in very low percentages of deviation (MnE: 2%; OE 1%) for the summed type differences.

Overall, this suggests that the variance in Zipf curves and type accumulation curves is significantly different for parallel texts with differing inflectional features (Modern English and Old English). In contrast, texts with the same or similar inflectional features exhibit similar distributional patterns even when they encode different information (Genesis versus Exodus) and when they differ with regards to source texts, style of translation and text length (KJV and NET, OE Exodus and Genesis).

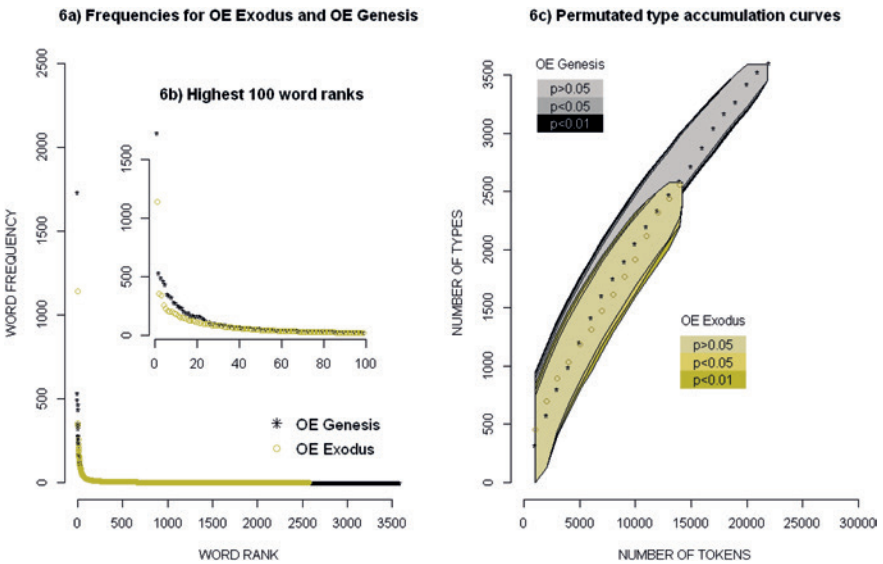


**Fig. 6:** Comparison between OE Genesis (black stars) and OE Exodus (yellow dots). The length difference of the texts renders a shorter tail for the OE Exodus (6a) and lower frequencies for the highest ranks (6b). Despite these differences the type accumulation curves are again very similar, with the OE Genesis points falling into the p > 0.05 confidence interval of the OE Exodus (6c)

### 3.4.3 Orthography

The spelling conventions in OE were generally less rigid than they are in MnE. This is a potential problem for our analysis, since variation in spelling introduces more types into our texts independent of morphological marking. Because it is a laborious task to control the spelling of all words occurring in the Old English text, we restrict our analysis to a subsample, and assess the impact of spelling differences on this sample. For this purpose, we choose as a sample the subset of proper nouns, for three reasons: 1) We do not expect to find many synonyms for the counterparts in Modern English, *Abraham* in OE is *Abraham* in MnE; 2) OE proper nouns can potentially be marked for case just like other content words, e.g. *Abrahame* (Dat.), *Abrahames* (Gen.); 3) we can control for different spellings in OE, e.g. *Effron* vs. *Ephron,* which would be a painstaking task for other lexical categories.

We therefore select all proper nouns which display at least one distinctively case marked form in OE Genesis. Moreover, only the most frequent spelling of an

OE proper noun is left as default spelling in the sample (in order to show that different spellings are not exclusively responsible for the changes). The result is a list of 60 proper nouns in Old English and a list of 22 proper nouns in MnE. The ratio follows from the fact that for almost every proper noun in MnE there are actually three word types in OE (e.g. MnE: *Joseph*, OE: *Iosep*, *Iosepe*, *Iosepes*). Plotting rank/frequency profiles for proper nouns in the MnE and OE texts, we expect the curves to be shaped like the ones seen in Figure 1, with longer tails in OE and higher frequencies for higher ranked types in MnE. As can be seen in Figure 7a and 7b, this is indeed the case.

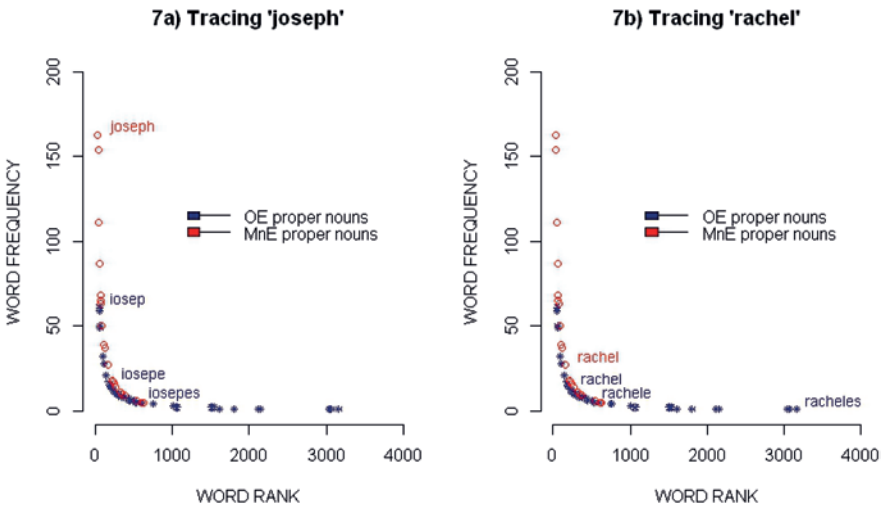


**Fig. 7:** Distributions for proper nouns in OE (blue stars) and MnE (red dots). Individual traces can be seen for the proper noun *jospeh* (7a) and *rachel* (7b). In both cases the highest frequent form of the proper noun in OE is the nominative. The dative and genitive forms 'tail out' towards the ranks with lowest frequencies. In MnE, due to the accumulation of formerly different types to one default type, the frequency for the resulting proper noun is overall higher

The overall distribution of proper nouns for OE is plotted with dark blue stars, the distribution for MnE with red dots. Additionally, the labels for the proper noun with the highest frequency in MnE (*joseph*) and the corresponding labels in OE (*iosep*, *iosepe*, *iosepes*) were added in 7a, and the same was done for a lower frequency proper noun (*rachel*) in 7b. The plots show clearly that having bound morphemes creates low frequency items even for high frequency proper nouns like *joseph* (7a), which might 'tail out' as hapax legomena for already low frequent

items like *rachel* (7b). Since we controlled for orthographic variation, this trend is independent of different writings of the proper nouns.

### 3.5 Conclusion (Section 3)

In this section, we have applied both frequency distributions (Section 3.2.1) and type accumulation curves (Section 3.2.2) to highlight significant statistical differences between the Old English and Modern English Genesis (Section 3.3). In addition, we have explored the extent to which the style of translation (Section 3.4.1), the content and length of texts (Section 3.4.2) as well as orthography (Section 3.4.3) could be responsible for these differences.

In fact, the deviation in type accumulation curves was calculated to be 23% between the Old English and Modern English (New English Translation) Book of Genesis, 10% between the King James Version and the Modern English translation, 2% between the Modern English Genesis and the Modern English Exodus, and 1% between the Old English Genesis and the Old English Exodus (which differ vastly in terms of number of tokens). These percentages in combination with frequency analyses of inflected forms in OE and function words in MnE suggest that (at least for parallel translations) the 'inflectional state' of a language has the biggest impact on the shape of the respective frequency distribution, followed by the style of translation (sentence structure), as well as content and length of texts. Additionally, we have controlled for variation in orthography for proper nouns and found the same pattern of higher frequencies and a shorter tail of hapax legomena in MnE. Of course, this does not rule out orthography as a source of variation, but it shows that the inflectional effects are independent of orthography.

Overall, this suggests that the most important factor shaping frequency distributions in Old English and Modern English parallel texts is the trade-off between synthetic forms and analytic constructions.

## 4 Towards measuring a 'grammatical fingerprint'

We systematically assessed the trend that frequency distributions for rather synthetic languages (e.g. OE) have a longer tail of hapax legomena, whereas the ones for rather analytic languages (e.g. MnE) are associated with higher frequencies towards the first ranks. This suggests that it might be possible to quantify the syntheticity of texts, and potentially even languages at a certain point in time by

using parallel translations. Such an analysis could be applied to explore trends in historical language change and for cross-linguistic comparisons of languages as complex systems (see also Popescu and Altmann [2008] for a related point). In the previous section, we demonstrated that grammatical changes particularly influence the statistical behavior of high and low frequency items in parallel texts. In this section we propose a formal quantitative measure which both captures the trend of longer tails in OE as well as the trend of higher frequencies in MnE. This measure hinges crucially upon the Zipf-Mandelbrot law.

## 4.1 The Zipf-Mandelbrot law

If the rank of a word is defined as $r_i$ and its frequency as $f(r_i)$ then the expected frequency distribution can be captured in the following Equation (i) (derived from Mandelbrot 1953: 491):

$$f(r_i) = \frac{C}{(\beta + r_i)^{\alpha}}, \qquad C > 0, \alpha > 1, \beta > -1, i = 1, 2, \ldots, n \tag{i}$$

In Equation (i) $\alpha$, $\beta$ are parameters and $C$ is a normalizing constant. The parameters and the constant account for modifications of the original law stated by Zipf. More precisely, Mandelbrot (1953, 1966) defined this more general class of functions, of which Zipf's law (ii) is but one specific instance. If we set $\alpha = 1$, and $\beta = 0$ in (i) then we arrive at (ii) (although Zipf [1949: 130] was aware of the fact that $\alpha$ can vary somewhat).

$$f(r_i) = \frac{C}{r_i} \tag{ii}$$

In the following, we will use this default case (Zipf 1949: 24) for the sake of illustration.

In Figure 8 the expected Zipf distributions for the first ten ranks of Old English (8a) and Modern English (8b) rank/frequency profiles are plotted as black dotted lines. The actual distributions as found in the Genesis are plotted as blue lines with crosses and green lines with triangles respectively.

As can be seen in Figure 8a and 8b the expected Zipf distributions are a first approximation of the patterns found for the actual observed frequencies in the Old English and Modern English Genesis.

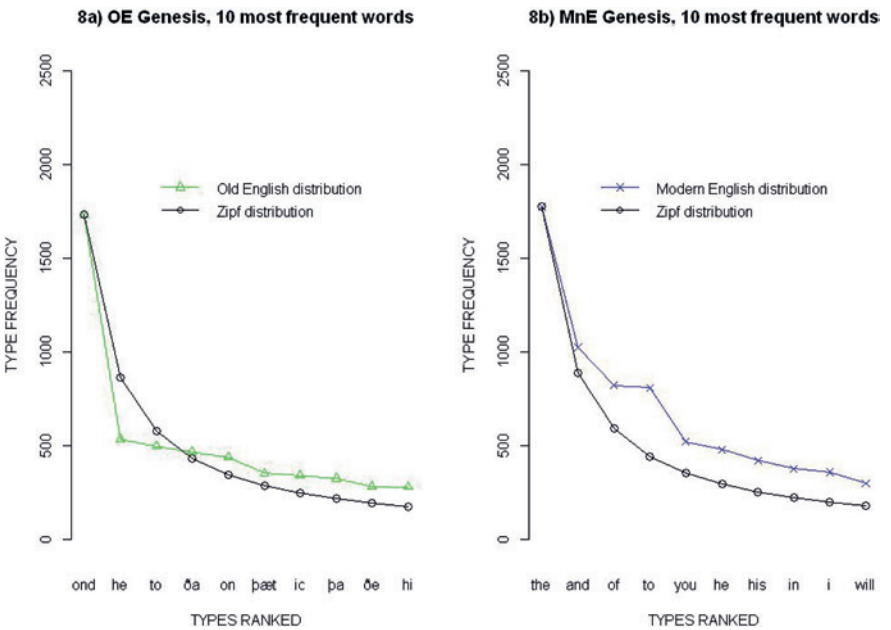


**Fig. 8:** Predicted Zipf distributions for ranked words (black dots, with the actual words printed on the x-axis) compared to actual distributions as found in the Old English (green triangles in 8a) and Modern English Genesis (blue crosses in 8b)

### 4.1.1 The log-log scale

A log-transformation of both the ranks and the frequencies of such Zipf curves produces a straight line as in Figure 9 (black line), representing the original Zipf law in Equation (ii) where $\alpha = 1$ and $\beta = 0$. The addition of a flexible parameter $\beta$ leads to Equation (iii) corresponding to Mandelbrot's (1953) modification of Zipf's original equation.

$$\log f(r) = \log (C) - \alpha^{\star} \log (\beta + r) \qquad \text{(iii)}$$

In this equation, $C$, $\beta$ and $\alpha$ can be adjusted to change the intercept (parameter $C$), the slope (parameter $\alpha$) and the deviance from linearity (parameter $\beta$) of the function. Increasing $C$ (with constant $\alpha$ and $\beta$) shifts the curve upwards (dotted and dashed green line in Figure 9), increasing $\alpha$ (with constant $C$ and $\beta$) produces a steeper slope and results in the dashed blue line in Figure 9, and an increase in $\beta$

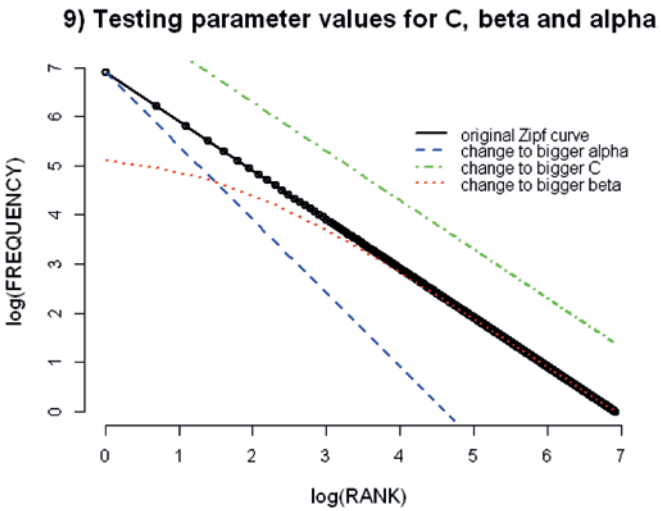


**Fig. 9:** Zipf distributions as log-transformed straight lines with different slopes ($\alpha$) and different $C$s and $\beta$s. The original Zipf's law with $\alpha = 1$ and $\beta = 0$ is represented by the black line with points indicating expected frequencies for declining ranks in a discrete function. Increasing $\alpha$ (with constant $C$ and $\beta$) is equivalent with a steeper slope (blue dashed line), increasing $C$ (with constant $\alpha$ and $\beta$) is shifting the curve upwards (green dashed and dotted line), and increasing $\beta$ (with constant $C$ and $\alpha$) accounts for deviations from linearity (red dotted line)

(with constant $\alpha$ and $C$) bends the higher frequency end of the line downwards (dotted red line in Figure 9).

Interestingly, changing the parameters $C$, $\beta$ and $\alpha$ in the Zipf-Mandlebrot function affects aspects of the curve that correspond to the effects of grammatical differences discussed in the previous section. To illustrate this for our OE and MnE texts, the log-transformed empirical distributions of the OE and MnE Genesis as well as a lemmatized version of the Modern English Genesis are plotted as points in Figures 10a–b.

The lines of best fit are produced by estimating $C$, $\beta$ and $\alpha$ with the maximum likelihood method (using the *likelihood* package in R [Murphy 2012]; see Izsák [2006] for technical details). The parameter estimations of the models as well as $R^2$-values, measuring model fit, can be found in Table 3.

Now, in the typical Zipf-Mandelbrot three parameter models there are several interesting differences between these distributions to be noted.

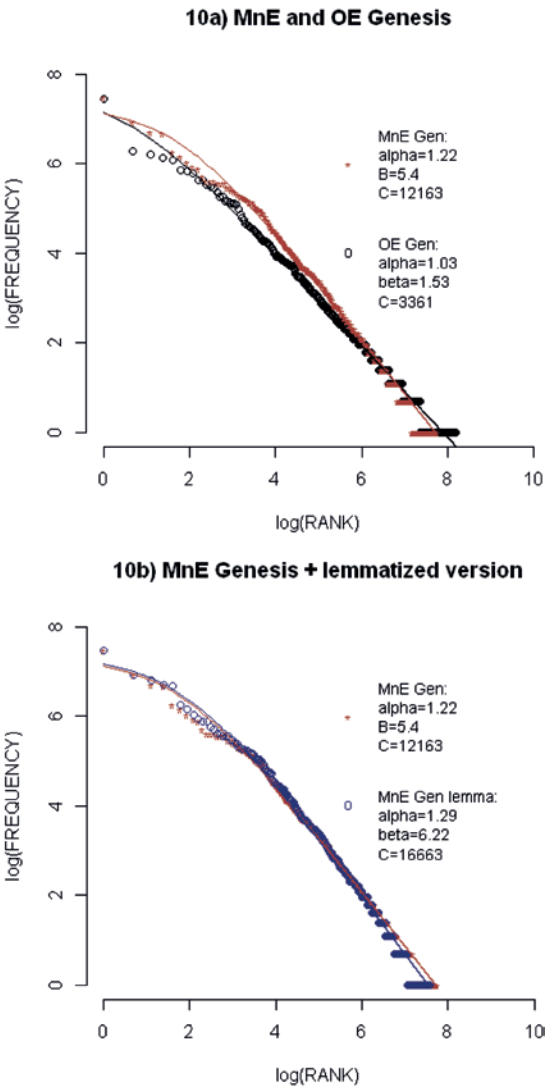


**Fig. 10:** Comparison between the log-transformed MnE (red stars) and OE Genesis (black dots) (10a), as well as the MnE Genesis in its original form and as a lemmatized text (blue dots) (10b). The points in the plots represent the empirical distributions, whereas the lines are lines of best fit attained by maximum likelihood estimation for parameters $\alpha$, $\beta$ and $C$

**Table 3:** Log-likelihood estimated parameter values for Zipf-Mandelbrot models of different MnE and OE text versions. The $R^2$-values are given as a measure of model fit

| Text | $\alpha$ | $\beta$ | $C$ | $R^2$ |
|---|---|---|---|---|
| MnE Gen. | 1.22 | 5.29 | 11951 | 0.95 |
| OE Gen. | 1.03 | 1.49 | 3329 | 0.92 |
| MnE Gen lem. | 1.29 | 6.16 | 16502 | 0.96 |

## 4.2 The parameters of the Zipf-Mandelbrot law

### 4.2.1 The $\alpha$-values

Note that the $\alpha$-values for the Modern and Old English Genesis are different (Figure 10a: 1.22 versus 1.03). In order to understand why this is, consider the lower right corner of Figure 10a: the greater number of hapax legomena in OE results in a longer black line (representing all words with frequency = 1) compared to the red line for MnE. This 'turns' the MnE distribution towards a steeper slope (higher absolute $\alpha$) in the tail. In Section 3.3.1 we showed that morphological marking is one of the most important factors responsible for this length difference of the tails. Hence, the steeper slope $\alpha$ for MnE reflects this diachronic trade-off between morphologically marked, low-frequent forms in OE and the loss of these forms in MnE. As a further illustration, we stripped off the remaining grammatical marking in the MnE text and compared the resulting distribution to the original MnE distribution (Figure 10b). Note that the lemmatized version without grammatical marking has an even steeper slope ($\alpha = 1.29$) than the original ($\alpha = 1.22$), further supporting the view that $\alpha$ actually reflects changes in the morphology of the language.

Finally, we explored whether the overall number of tokens in the texts has an impact on the slope, by cutting the MnE text to the length of the OE text, and re-running the parameter estimation. This resulted in an additive change in $\alpha$ of 0.04 from 1.22 to 1.18, which is small compared to the 0.19 difference between the OE and MnE slopes. Also, note that the length difference between the OE and MnE texts is causally linked to the changes in morphological marking, and therefore part of the phenomenon we are trying to measure.

### 4.2.2 The $\beta$-values

Following the distributions in Figure 10a upwards, for the highest frequencies the slope of the MnE line of best fit (red stars) drops off and even crosses the OE line

(black dots). For the higher ranks the slope of the MnE curve is in fact flatter than the OE curve. This is reflected by the higher $\beta$-value for MnE ($\beta = 5.4$) compared to OE ($\beta = 1.53$). A higher $\beta$-value essentially means that for higher frequencies the MnE Genesis deviates more strongly from the originally predicted slope of $\alpha = 1.22$ than does the OE Genesis for the originally predicted slope of $\alpha = 1.03$. Consistent with the analysis in Section 3.3.2, this difference is due to even higher frequencies for already highly frequent items such as prepositions, conjunctions, articles etc. The $\beta$-value therefore appears to be indicative of the enhancement in higher frequencies for function words.

### 4.2.3 The *C*-values

The *C*-value is higher for MnE (12163) than for OE (3361). This reflects the fact that overall the MnE Genesis is longer in terms of tokens than the OE Genesis. However, the *C*-value is also higher for the lemmatized version of MnE than for the original version (16663 versus 12163), although both versions have exactly the same number of tokens (lemmatized tokens are altered but never deleted). However, note that the lemmatization process artificially lowers the number of types *without* enhancing the number of tokens. This is not a natural trade-off between periphrastic constructions and morphological marking. Therefore, the parameter estimation results in the *C*-value for the lemmatized text adjusting *as if* there were more tokens in the lemmatized version, even though this is not the case.

Overall, the discussion of the trends in Figures 10a–b support the idea that the parameters of the Zipf-Mandelbrot law change systematically according to grammatical encoding strategies, namely whether languages adhere to either analytical or synthetic structures. This is in line with Zipf's original idea of a measure of syntheticity – a 'grammatical fingerprint' of a particular language (Zipf 1949: 95, 1965 [1935]: 252). His proposal appears to work well for the parallel texts we analyzed here. Of course, this does not guarantee that the measure works just as precise for non-parallel texts across languages, a question that has partly been addressed by Popescu et al. (2010, 2009) and Popescu and Altmann (2008), but still needs to be further assessed. However, looking at the behavior of $\alpha$, $\beta$ and $C$ across parallel texts of different time periods and language groups gives a valuable perspective on changing grammatical encoding strategies. Together with other measures proposed earlier (see Popescu et al. 2009) this could be a valuable means to analyze language change, especially when dealing with large amounts of linguistic data.

Finally, it is important to note that the derivation of linguistic and information-theoretic conclusions from Zipf distributions and other statistical patterns is by

no means uncontroversial. While some argue that there is a linguistically 'deep' interpretation for Zipf curves, which can help us understand the similarities and differences in how languages organize and encode information (Baixeries et al. 2012; Ferrer i Cancho and Elvevåg 2010; Ferrer i Cancho 2005), others hold that Zipf curves might be no more than statistical artifacts (Miller 1957; Li 1992). Although we do not enter into the theoretical and philosophical debate here, it is clear that our analyses offer evidence in support of the former position. We do not, of course, claim that frequency distributions reveal everything there is to know about a language, but rather that several of their interesting morpho-syntactic properties are indeed reflected by such statistical analyses.

# 5 Conclusions

Based on the analyses reported in this study, our principal conclusion is that Zipf's idea of a 'grammatical fingerprint', a quantitative measure of the syntheticity of a language based on frequency distributions is a linguistically valuable construct. That is, despite repetitive (and partly justified) criticism from certain corners, Zipf's account can indeed be based on a "serious linguistic theory" – to put it in Mandelbrot's (1953: 492) words.

More specifically, our analyses elicited that the frequency distributions for two diachronic parallel texts of Old English and Modern English are significantly different. This is mainly due to diachronic grammatical changes, namely the trade-off between inflectional marking and periphrastic constructions. Scrutinizing frequency distributions in this way is worthwhile, because it can help us pin down such diachronic tendencies in large-scale and quantitative terms. We showed how these trade-offs can be quantified by a maximum likelihood estimation of the parameters $\alpha$, $\beta$ and $C$, and that these parameters can be interpreted in a linguistically meaningful way. Frequency distributions of corpora have not played and do not yet play a major role in traditional studies of language change and evolution. However, we hope to have illustrated a useful and systematic method for capturing and interpreting morphological trends in language diachrony and synchrony.

Considering wider implications, at least in connection with functionalist and cognitive approaches to language acquisition and change it could be that these quantitative methods will help in understanding the pathways along which lexical items are learned and handed down from generation to generation. Within these paradigms it has been stated that the *frequency* of a type and its *regularity* are systematically linked to render its *learnability* (Bybee 2007; Christiansen and Chater 2008). While low-frequency items need to be regular in order to be

preserved over a couple of generations of learners, high frequency items can afford to be irregular. Liebermann et al. (2007) used this relationship to calculate the 'half-life' of irregular verb forms in English by using their frequencies of occurrence in corpora. This approach could be extended to other grammatical categories by using the information contained in frequency distributions.

In a similar vein, it could be asked how language contact changes the frequency distributions of lexical items in an overall population of speakers and how this affects the language learning of the next generations. If it is true that adult second language speakers by trend reduce the usage of distinct morphological markers (Trudgill 2011; McWhorter 2007; Bentz and Christiansen 2010; Bentz and Winter 2013; Bentz and Winter 2012), then they might skew the distributions of morphologically marked forms and therefore change the input available for the next generation of language learners. This could iteratively, over several generations, lead to the trends observed in this study.

It is undeniable that quantitative methods are a valuable tool for analyzing language variation and change. However, the capacity of this approach to elicit the precise details of such trends is not well understood. We believe, based on the analyses in this study, that the utility of quantitative methods is greater than widely held at present. Of course, it is open to future research to criticize and refine our approach, as well as to broaden the scope of possible applications. In particular, we plan to test comprehensively whether Zipf's 'grammatical fingerprint' is indeed a stable construct, occurring constantly and distinctively in a wider variety of texts and languages.

**Acknowledgments:** We would like to thank Bodo Winter from UC Merced, Richard Samworth, Yining Chen and Arlene Kim from the Statistical Laboratory of Cambridge University, and Ferenc Izsák for discussing appropriate statistical methods with us. Also, thanks go to Stefan Th. Gries and two anonymous reviewers for their insightful comments and suggestions. CB would like to thank the Arts and Humanities Research Council UK, the Cambridge Home and European Scholarship Scheme as well as Cambridge Assessment for financial support.

# Appendix 1

## Sample texts

The texts used for this study are: a) Ælfric's Old English translations of the Book of Genesis and the Book of Exodus, which are available with syntactic annotations in the *York-Toronto-Helsinki Parsed Corpus of Old English Prose* (Taylor et al. 2003); b) the Modern English counterparts of these books as found in the *New English Translation* (NET) (http://bible.org/netbible/) as well as the *King James Version* (KJV) of the *Oxford Text Archive* (http://ota.ahds.ac.uk/desc/3036. Accessed 11/20/2011). These texts were chosen because: 1) they are some of the few parallel texts with syntactic annotations available in Old and Modern English; 2) they are well documented and hence accessible for further philological inquiries; 3) chapters and sentences in the two texts have been manually aligned for meaning. Thus, it is possible to compare the parallel translations precisely, while controlling for potential confounds.

1.1 YCOE

| | |
|---|---|
| ***Text name*** | Heptateuch |
| ***File name*** | cootest.o3 |
| ***Passages*** | Genesis, Exodus (some sentences in the YCOE samples are cut out. Therefore, the parallel sentences NET and KJV had to be found and cut out as well) |
| ***Cameron number*** | B8.1.4, B8.1.7.2 |
| ***Manuscript*** | (1) London, British Museum, Cotton Claudius B.IV (Gen, Exod, Lev, Num, Deut, Josh)<br>(2) Oxford, Bodleian, Laud 509 (Judg.) |
| ***Dialect*** | West Saxon |
| ***Language of source text(s)*** | Latin |
| ***Edition*** | Crawford, Samuel J. 1922. The Old English Version of the Heptateuch. Ælfric's Treatise on the Old and New Testament and His Preface to Genesis. EETS 160. London: OUP. Reprinted with additions by N.R. Ker 1969. |

1.2 New English Translation (NET)

| | |
|---|---|
| *Text name* | New English Translation |
| *Passages* | Genesis, Exodus |
| *Manuscript* | – |
| *Language* | Standard Modern English |
| *Language of source text(s)* | Hebrew |
| *Edition* | – |
| *URL* | http://bible.org/netbible |

1.3 King James Version (KJV)

| | |
|---|---|
| *Text name* | Authorized King James Version |
| *Passages* | Genesis, Exodus |
| *Manuscript* | – |
| *Language* | Standard Modern English |
| *Language of source text(s)* | Ambiguous. English, Spanish, French, Greek, Latin, Hebrew |
| *Edition* | Project Gutenberg version, 10th edition |
| *URL* | http://ota.ox.ac.uk/desc/1691 |

# Appendix 2

YCOE: Part-of-speech labels

Nominals and Pronominals

| | |
|---|---|
| N | Common noun, singular or plural |
| NR | Proper noun, singular or plural |
| MAN | Indefinite "man" |
| PRO | Personal pronoun |
| PRO$ | Possessive pronoun |

Adjectives and Adverbs

| | |
|---|---|
| ADJ | Adjective |
| ADJR | Comparative Adjective |

ADJS Superlative Adjective
ADV Adverb
ADVR Comparative Adverb
ADVS Superlative Adverb

Quantifiers and numerals
Q Quantifier
QR Comparative Quantifier
QS Superlative Quantifier
NUM Numeral

Wh-words
WPRO Wh-pronoun
WADJ Wh-adjective
WADV Wh-adverb
WQ WHETHER

Miscellaneous
CONJ Coordinating conjunction
C Complementizer
D Determiner
P Preposition or subordinating conjunction
NEG Negation (note that NEG can adjoin to verbs, quantifiers, conjunctions, etc.)
RP Adverbial particle (note that RP can adjoin to verbs)
FP Focus particle
FW Foreign word
INTJ Interjection
XX unknown or problematic word

The verb BE
BE infinitive
BEI imperative
BEPH present tense, ambiguous imperative/subjunctive
BEPI present tense, unambiguous indicative
BEPS present tense, unambiguous subjunctive
BEP present tense, ambiguous form
BEDI past tense, unambiguous indicative

BEDS past tense, unambiguous subjunctive
BED past tense, ambiguous form
BAG present participle
BEN past participle

The verb HAVE
HV infinitive
HVI imperative
HVPI present tense, unambiguous indicative
HVPS present tense, unambiguous subjunctive
HVP present tense, ambiguous form
HVDI past tense, unambiguous indicative
HVDS past tense, unambiguous subjunctive
HVD past tense, ambiguous form
HAG present participle
HVN past participle (verbal or adjectival)

Auxiliary verbs
AX infinitive
AXI imperative
AXPI present tense, unambiguous indicative
AXPS present tense, unambiguous subjunctive
AXP present tense, ambiguous form
AXDI past tense, unambiguous indicative
AXDS past tense, unambiguous subjunctive
AXD past tense, ambiguous form
AXG present participle
AXN past participle (verbal or adjectival)

Modal verbs
MD infinitive
MDI imperative
MDPI present tense, unambiguous indicative
MDPS present tense, unambiguous subjunctive
MDP present tense, ambiguous form
MDDI past tense, unambiguous indicative
MDDS past tense, unambiguous subjunctive
MDD past tense, ambiguous form
TO infinitival TO

All other verbs
VB infinitive
VBI imperative
VBPH ambiguous imperative/subjunctive
VBPI present tense, unambiguous indicative
VBPS present tense, unambiguous subjunctive
VBP present tense, ambiguous form
VBDI past tense, unambiguous indicative
VBDS past tense, unambiguous subjunctive
VBD past tense, ambiguous form
VAG present participle
VBN past participle (verbal or adjectival)

Extended POS tags
^N nominative case (case may be marked on N, D, MAN, Q(R/S), NR, NUM,
^A accusative case PRO, WPRO, PRO$, ADJ(R/S), WADJ, participles,
^G genitive case infinitives)
^D dative case
^I instrumental case
^T temporal (marked on ADV, WADV)
^L locative (marked on ADV, WADV)
^D directional (marked on ADV, WADV)

## Appendix 3

Penn Treebank: Part-of-speech labels

CC Coordinating conjunction
CD Cardinal number
DT Determiner
EX Existential there
FW Foreign word
IN Preposition or subordinating conjunction
JJ Adjective
JJR Adjective, comparative
JJS Adjective, superlative
LS List item marker
MD Modal
NN Noun, singular or mass

| | |
|---|---|
| NNS | Noun, plural |
| NNP | Proper noun, singular |
| NNPS | Proper noun, plural |
| PDT | Predeterminer |
| POS | Possessive ending |
| PRP | Personal pronoun |
| PRP$ | Possessive pronoun (prolog version PRP-S) |
| RB | Adverb |
| RBR | Adverb, comparative |
| RBS | Adverb, superlative |
| RP | Particle |
| SYM | Symbol |
| TO | to |
| UH | Interjection |
| VB | Verb, base form |
| VBD | Verb, past tense |
| VBG | Verb, gerund or present participle |
| VBN | Verb, past participle |
| VBP | Verb, non-3rd person singular present |
| VBZ | Verb, 3rd person singular present |
| WDT | Wh-determiner |
| WP | Wh-pronoun |
| WP$ | Possessive wh-pronoun (prolog version WP-S) |
| WRB | Wh-adverb |

# Appendix 4

## 4.1 Methods

Parallel translations were used to plot, analyze and compare the frequency distributions of distinct word types. For example, the case marked forms of OE *god* (Nom.), *godes* (Gen.) and *gode* (Dat.) are treated as distinct types rather than case inflected forms of the same lemma *god*. The same accounts for adjectival and verbal inflections. This is not always trivial, since we need to decide how to deal with borderline cases. For instance, the genitive *'s* in Modern English (as in *Abraham's sons*) can either be analyzed as a distinct type (*Abraham* and *s*) or as a genuine inflection (*Abrahams*). Since the Modern English genitive 's is widely analyzed as a phrasal clitic rather than a genuine case marker (Blevins 2006: 511; Allen 2003;

Anderson 1992: 118; Hudson 1995) we adhere to the first approach. Likewise, abbreviated forms such as *don't*, *I'll* and *we'd* are analyzed as two separate types. Since there are no such abbreviations to be found in OE, this somewhat artificially increases the analyticity of MnE. However, in the actual frequency distributions we find few of these abbreviations.

## 4.2 Software

Working with the digitalized texts mentioned above, different software tools were used to assess the frequencies of lexical items and compare them diachronically and synchronically. The corpus applications (Gries 2009) of the statistical software *R* (R Development Core Team 2010), the *ZipfR* package (Evert and Baroni 2007), the *stats* package and the *likelihood* package (Murphy 2012) were used to obtain ranked frequency tables for lexical items of texts in different unannotated formats as well as to estimate parameters for Zipf-Mandelbrot models. For the annotated texts of the Penn-Treebank type the freeware *CorpusSearch*[4] was used to build lexicons with frequency of occurrence information as well as part-of-speech tags (POS tags). For the Modern English distributions the texts were parsed separately using the *Stanford-Parser* (Klein and Manning 2003).

Moreover, software developed by Suomela (2007) was used to compute and plot confidence intervals for type accumulation curves. According to this method, we split the OE and MnE Genesis into token chunks of 1000 tokens each. The empirical curves are then derived by counting the types within each of these chunks and successively adding the numbers of types up (in the original order of the chunks). To test whether the resulting type accumulation curves are significantly different and not just due to random variation, the chunks are then randomly re-ordered in 1 million permutations. This way, confidence intervals can be plotted, which indicate how likely it is that the differing shapes of the empirical type accumulation curves are just due to random variation within the chunks.

# References

Allen, Cynthia. 2003. Deflexion and the development of the genitive in English. *English Language and Linguistics* 7 (1). 1–28.

Anderson, Stephen R. 1992. *A-morphus morphology*. Cambridge: Cambridge University Press.

4 http://corpussearch.sourceforge.net/

Baayen, R. Harald. 2009. Corpus linguistics in morphology: morphological productivity. In Anke Lüdeling & Merja Kytö, M. (eds.), *Corpus Linguistics. An international handbook*, 900–919. Berlin & New York: Mouton de Gruyter.

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

Baayen, R. Harald. 2001. *Word frequency distributions*. Dordrecht, Boston & London: Kluwer.

Baixeries, Jaume, Ramon Ferrer i Cancho & Brita Elvevåg. 2012. The exponent of Zipf's law in language ontogeny. In Thomas C. Scott-Phillips, Mónica Tamariz, Erica A. Cartmill, James R. Hurford (eds.), *The evolution of language, proceedings of the 9th international conference (EVOLANG9)*, 409–410. Singapore: World Scientific.

Baroni, Marco. 2009. Distributions in text. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics. An international handbook*, 803–821. Berlin & New York: Mouton de Gruyter.

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

Bentz, Christian & Bodo Winter. 2012. The impact of L2 speakers on the evolution of case marking. In Thomas C. Scott-Phillips, Mónica Tamariz, Erica A. Cartmill, James R. Hurford (eds.), *The evolution of language, proceedings of the 9th international conference (EVOLANG9)*, 409–410. Singapore: World Scientific. 58–63.

Bentz, Christian & Morten H. Christiansen. 2010. Linguistic adaptation at work? The change of word order and case system from Latin to the Romance languages. In Andrew D. Smith, Schouwstra, Bart de Boer and Kenny Smith (eds.), *The evolution of language*, *proceedings of the 8th international conference on the evolution of language*, 26–33. Singapore: World Scientific.

Blevins, James. 2006. English inflection and derivation. In Bas Aarts & April McMahon (eds.), *Handbook of English Linguistics*, 507–536. London: Blackwell.

Bybee, Joan L. 2007. *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Campbell, Alistair. 1959. *Old English grammar*. Oxford: Clarendon Press.

Carroll, Robert & Stephen Prickett (eds.). 2008. *The Bible: Authorized King James Version*. Oxford: Oxford University Press.

Christiansen, Morten H. & Nick Chater. 2008. Language as shaped by the brain. *Behavioral and Brain Sciences* 31(5). 489–509.

Evert, Stefan & Marco Baroni. 2007. zipfR: Word frequency distributions in R. Paper presented at the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions, Prague, Czech Republic.

Ferrer i Cancho, Ramon & Brita Elvevåg. 2010. Random texts do not exhibit the real Zipf's law-like rank distribution. *PloS one* 5 (3), e9411. 1–10.

Ferrer i Cancho. 2005. The variation of Zipf's law in human language. *The European Physical Journal β* 44. 249–257.

Gries, Stefan Th. 2009. *Quantitative Corpus Linguistics with R. A practical Introduction*. New York & London: Routledge.

Ha, Le Quan, Darryl W. Stewart, Philip Hanna & Francis J. Smith. 2006. Zipf and Type-Token rules for the English, Spanish, Irish and Latin languages. *Web Journal of Formal, Computational and Cognitive Linguistics* 8. http://fccl.ksu.ru/issue8/ha_fccl_zipf.pdf. (accessed 22/11/2012).

Hogg, Richard M. & R. D. Fulk. 2011. *A Grammar of Old English. Volume 2: Morphology*. Malden & Oxford: Wiley-Blackwell.

Hudson, Richard. 1995. Does English really have case? *Journal of Linguistics*, 31. 375–392.
Izsák, Ferenc. 2006. Maximum likelihood estimation for constrained parameters of multinomial distributions – Application to Zipf-Mandelbrot models. *Computational Statistics & Data Analysis* 51. 1575–1583.
Klein, Dan & Christoper D. Manning. 2003. Accurate unlexicalized parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423–430.
Lass, Roger. 1994. *Old English*. Cambridge: Cambridge University Press.
Li, Wentian. 1992. Random texts exhibit Zipf's-law-like word frequency distributions. *IEEE Transactions on Information Theory* 38 (6). 1842–1845.
Liebermann, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang & Martin A. Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature* 449. 713–716.
Mandelbrot, Benoît. 1966. Information theory and psycholinguistics: A theory of word frequencies. In Paul F. Lazarsfeld & Neil W. Henry (eds.), *Readings in mathematical social science*, 350–368. Chicago: Science Research Associates Inc.
Mandelbrot, Benoît. 1953. An informational theory of the statistical structure of language. In Willis Jackson (ed.), *Communication Theory*, 468–502. London: Butterworths Scientific Publications.
McWhorter, John. 2007. *Language interrupted: Signs of non-native acquisition in standard language grammars*. Oxford/NewYork: Oxford University Press.
Miller, George A. 1957. Some effects of intermittent silence. *The American Journal of Psychology* 70 (2). 311–314.
Murphy, Lora. 2012. *likelihood: Methods for maximum likelihood estimation*. R package version 1.5. http://CRAN.R-project.org/package=likelihood.
Popescu, Ioan-Iovitz, Gabriel Altmann & Reinhard Köhler. 2010. Zipf's law– another view. *Quality & Quantity* 44 (4). 713–731.
Popescu, Ioan-Iovitz, Gabriel Altmann, Peter Grzybek, Bijapur D Jayaram, Reinhard Köhler, Viktor Krupa, Ján Mačutek, Regina Pustet, Ludmilla Uhlířová, Matummal N Vidya. 2009. *Word frequency studies*. Berlin & New York: Mouton de Gruyter.
Popescu, Ioan-Iovitz and Gabriel Altmann. 2008. Hapax legomena and language typology. *Journal of Quantitative Linguistics* 15 (4). 370–378.
R Development Core Team. 2010. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/.
Säily, Tanja. 2011. Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory* 7 (1). 119–141.
Säily, Tanja and Jukka Suomela. 2009. Comparing type counts: The case of women, men and *-ity* in early English letters. In Antoinette Renouf & Andrew Kehoe (eds.), *Corpus Linguistics: Refinements and Reassessments*, 87–109. Amsterdam: Rodopi.
Suomela, Jukka. 2007. Type and hapax accumulation curves. Computer program. http://www.cs.helsinki.fi/u/josuomel/types/ (accessed 02/11/2012).
Taylor, A., Warner, A., Pintzuk, S. & Beths, F. 2003. *The York-Toronto-Helsinki Parsed Corpus of Old English Prose*. http://www-users.york.ac.uk/~lang22/YCOE/ YcoeHome.htm.
Thomason, Sarah G., and Terrence Kaufman. 1991. *Language Contact, Creolization, and Genetic Linguistics*. Berkeley: University of California Press.
Trudgill, Peter. 2011. *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford: Oxford University Press.
Zipf, George K. 1965 [1935]. *The psycho-biology of language*. Cambridge, MA: The M.I.T. Press.

Zipf, George K.1949. *Human Behavior and the Principle of Least Effort*. Cambridge MA: Addison-Wesley.
Zipf, George K. 1932. *Selected studies of the principle of relative frequency in language*. Cambridge, MA: Harvard University Press.

## Bionotes

*Christian Bentz* is currently a PhD student in Computation, Cognition and Language at the University of Cambridge.

*Douwe Kiela* studied Cognitive Artificial Intelligence and Philosophy at Utrecht University. He obtained his MSc in Logic from the University of Amsterdam's Institute for Logic, Language & Computation, and then complete the MPhil in Advanced Computer Science at the University of Cambridge. He is now a PhD student in Computer Science at the same university's Natural Language & Information Processing group.

*Felix Hill* has a degree in maths from the University of Oxford (MMath) and an MPhil in linguistics from the University of Cambridge. He works in the Cambridge Computer Laboratory on computational semantics with Anna Korhonen.

*Paula Buttery* is a Lecturer in Computational Linguistics at the Department of Theoretical and Applied Linguistics; an associated researcher in the Natural Language and Information Processing Group at the Cambridge Computer Laboratory; and a fellow of Gonville and Caius College, University of Cambridge. Recently she has also worked as an Information Extraction and Data Mining Engineer at the European BioInformatics Institute, part of the European Molecular Biological Laboratory.