Variation in word frequency distributions: Definitions, measures and implications for a corpus-based language typology*

Christian Bentz\textsuperscript{a,b}, Dimitrios Alikaniotis\textsuperscript{a}, Tanja Samardžić\textsuperscript{c} and Paula Buttery\textsuperscript{a}
\textsuperscript{a}Department of Theoretical and Applied Linguistics, University of Cambridge, Cambridge, United Kingdom; \textsuperscript{b}Department of General Linguistics, University of Tübingen, Tübingen, Germany; \textsuperscript{c}URPP on Language and Space, University of Zürich, Zürich, Switzerland

* Address correspondence to: Christian Bentz, University of Tübingen, Nauklerstraße 35, 72074 Tübingen, Germany. E-mail: chris@christianbentz.de
Abstract

Word frequencies are central to linguistic studies investigating processing difficulty, learnability, age of acquisition, diachronic transmission and the relative weight given to a concept in society. However, there are few cross-linguistic studies on entire distributions of word frequencies, and even less on systematic changes within them. Here, we first define and test an exact measure for the relative difference between distributions – the \textit{Normalized Frequency Difference (NFD)}. We then apply this measure to parallel corpora, explaining systematic variation in the frequency distributions within the same language and across different languages. We further establish the NFD between lemmatized and unlemmatized corpora as a frequency-based measure of inflectional productivity of a language. Finally, we argue that quantitative measures like the NFD can advance language typology beyond abstract, theory-driven expert judgments, towards more corpus-based, empirical and reproducible analyses.

Keywords: Language typology, corpus linguistics, quantitative linguistics, inflectional productivity, word frequency distributions

*We made an \textit{R} package available for NFD calculation and plotting via https://github.com/dimalik/nfd/
1. Introduction


In psycholinguistics, the effect of word frequency on lexical decision and word naming is one of the most robust and well-known findings since the 1950s (Solomon & Howes 1951, Freedman & Loftus 1971). It was argued early on that any psycholinguistic task involving lexical stimuli needs to control for frequency of occurrence (Loftus & Suppes 1972, Whaley 1978). Importantly, this effect is not limited to surface frequencies of base forms, but extents to complex morphological forms (see Moscoso del Prado Martín, Kostić & Baayen 2004 for an overview). In consequence, the link between word frequency and processing difficulty has repercussions on first and second language acquisition (Ellis 2002, Ellis & Collins 2009, Goldschneider & DeKeyser 2001, Larsen-Freeman 1975, 1976). A recent large-scale longitudinal study on first language acquisition suggests that frequencies in the caregivers input also predict the age of acquisition of words, notably for both content words (e.g. nouns) and closed class function words (Roy, Frank & Roy 2009).

In historical linguistics, frequencies are taken as indicators of synchronic and diachronic transmission processes and change. For example, Bybee (2007: 28) points out that Old English strong verbs had a higher probability of becoming weak verbs if they had low frequencies. This directly relates to the synchronic fact that the most frequent verbs in Modern English tend to be irregulars (i.e. strong verbs). This observation was taken up in a large-scale quantitative study by Lieberman et al. (2007) showing that the “rate of decay” of irregulars can be estimated based on their frequencies (see also Cuskley et al. 2014 and Colaoiri et al. 2015 for quantitative
analyses and agent-based modelling). In a similar vein, dialectological studies report significant effects of frequencies on standardization and change in Dutch and Tuscan dialects (Wieling et al. 2011, Wieling et al. 2014).

Recently, massive diachronic corpora such as the google ngram corpus (Michel et al. 2011) have become available and allow researchers to track word frequency changes since ca. the 18th century in a considerable proportion of the books printed across 7 languages - though see Koplenig (2015a) for several issues concerning a meaningful interpretation of this data.

Besides an abundance of literature on frequency changes in specific words (or groups of words), there is also research on word frequency distributions as a whole. This is one of the core subjects of quantitative linguistics in the spirit of Zipf (1932, 1935, 1949), Yule (1944), and Köhler, Altmann & Piotrowski (2005). The quantitative models available are most exhaustively discussed by Baayen (2001), as well as Popescu, Altmann, Grzybek, Jayaram, Köhler, Krupa, Macutek, Pustet, Uhlirova and Vidya (2009). More recently, several studies have attempted to quantify linguistically meaningful variation in word frequency distributions over time (Bentz, Kiela, Hill & Buttery 2014, Bochkarev, Solovyev & Wichmann 2014, Koplenig 2015b), and across many languages (Bentz, Verkerk, Kiela, Hill & Buttery 2015, Corral, Boleda & Ferreri-Cancho 2014).

However, it is still not well understood exactly which factors influence the shape of word frequency distributions to what extent. Especially in the context of studying potential causes of changes it is important to know which proportion of variance can be attributed to factors such as lexical change (i.e. changes in the base vocabulary due to neologisms or loanwords) and morphological marking (i.e. inflection, derivation, compounding and contractions or clitics). In the following we set out to start disentangling these factors.

We first define the Normalized Frequency Difference (NFD) as a measure of the relative difference in two frequency distributions (Section 3). This measure is then applied in Analysis 1 (Section 4) to assess the distributional differences in English and German parallel corpora before and after removing inflectional markers, derivational morphology, compounds and contractions/clitics. Analysis 2 (Section 5) focuses on inflectional morphology and measures the NFD difference for lemmatized and
unlemmatized parallel corpora across 19 languages. It is shown that the NFD can be used as a frequency-based, cross-linguistic inflection index. The sensitivity of this inflection index to corpus size is tested in Analysis 3 (Section 6). Our final Analysis (Section 7) then adds another level of detail by comparing the impact of lemmatization on different parts of speech for English and Estonian.

Besides an overall discussion of our results in Section 8, we further point towards other lexical diversity measure that could be used in parallel to the NFD, and why we think the NFD has some advantages over these (Section 8.1). Finally, we argue that quantitative measures like the NFD in combination with state-of-the-art computational tools and corpora enables an empirical and reproducible language typology that does not longer have to rely on expert judgements only (Section 8.2).

2. Definition of word types and word tokens

Any measure of variation in word frequency distributions has to be based on the distinction between *word types* and *word tokens*. Since we work with written language, we assume a technical definition. A *word type* is here defined as a unique string of unicode characters (lower case) delimited by non-alphanumeric characters (e.g. white spaces and punctuation marks). A *word token* is then defined as any recurring instance of a specific word type.

Though these or similar definitions of wordhood are taken as a given in most corpus and computational linguistic studies, they are not necessarily uncontroversial from a linguistically more informed point of view. Haspelmath (2011) and Wray (2014) point out that there is a whole range of orthographic, phonetic and distributional definitions of wordhood, which can yield different results for specific cases. For example, writing compounds with or without white spaces is an orthographic convention that does not necessarily reflect a difference in pronunciation. Arguably, there is no more of a pause between the English *car park* than the German *Parkplatz*.

In theory, such orthographic conventions change word types and hence the corresponding token frequencies. However, in practice the important question is *how much* of a difference we actually find.

In the following we propose an exact method to measure the variance in word frequency distributions. This method allows us to measure the difference between any
two distributions in general, and the actual impact that changes in word types will have on their token frequencies in language corpora more specifically.

3. The Normalized Frequency Difference (NFD)
An example of two differing frequency distributions, namely a *uniform* distribution of equal frequencies and a *non-uniform* distribution of varying frequencies, can be seen in the lower left panel of Figure 1. In linguistic examples the ranks (x-axis) of these ordered frequency distributions correspond to word types, and the frequencies on the y-axis to the number of tokens per word type in a given corpus.

Now, let $T = \{t_1, t_2, \ldots, t_V\}$ be the set of word types of size $V$ in a corpus, i.e. its *vocabulary*, and $F = (f_1, f_2, \ldots, f_V)$ be the distribution of values corresponding to the frequencies of occurrences of each word type in the corpus such that $f_i = \text{freq}(t_i)$. The overall number of tokens $N$ in the corpus $C$ is therefore:

$$N^C = \sum_{i=1}^{V} f_i$$  \hspace{1cm} (1)

To get a better overview of the rank/frequency profile we follow Zipf (1932, 1935, 1949) and rank the distribution of token-counts (i.e. the distribution $F$) from highest to lowest. Let $F^A, F^B$ be the two ranked word frequency distributions with vocabulary sizes $V^A$ and $V^B$ taken from two different corpora. We can proceed to calculate the absolute difference in token frequencies for any given rank $i$ as:

$$\Delta \text{Freq}_{A,B,i} = \begin{cases} 
|f_i^A - f_i^B| & \text{if } i \leq V^A \land i \leq V^B \\
 f_i^A & \text{if } i \leq V^A \land i > V^B \\
 f_i^B & \text{otherwise}
\end{cases}$$ \hspace{1cm} (2)

Note that the number of ranks in two distributions might differ due to differing numbers of types, i.e. differing vocabulary sizes $V^A$ and $V^B$. For every rank we take the absolute difference in frequencies if frequencies for both $F^A$ and $F^B$ are available. If there are no frequencies available in either $F^A$ or $F^B$, i.e. they are effectively 0, then we take the frequency of the other vector as absolute difference.
These absolute frequency differences per rank are indicated in the upper left panel of Figure 1 as $\Delta\text{Freq}$. It is important to note here that in many cases the token frequencies compared per rank belong to different word types. For example, if we compare the word frequency distribution of an English and a German text, then the word types in rank 1 are likely to be *the* and *und 'and'* . So there is no direct correspondence between them in terms of a translational equivalent or the like. What brings them together in rank 1 is solely the fact that both have the highest token frequencies in the respective texts.

Based on the frequency difference per rank given in Equation (2) we then define the *Normalized Frequency Difference* (NFD) between two distributions as:

$$
\text{NFD}(A, B) = \frac{\max(V^A, V^B)}{\sum_{i=1}^{V^A} f_i^A + \sum_{i=1}^{V^B} f_i^B} \sum_{i=1}^{\max(V^A, V^B)} \Delta\text{Freq}(A, B, i)\
$$

Figure 1. An example of visually comparing a uniform (grey) to a non-uniform (black) frequency distribution. The left panel illustrates the frequencies of the two distributions ranked from highest to lowest. The frequency differences are the differences in height of the black and grey bars. These differences are projected onto the upper panel. The right panel illustrates the log frequencies and log ranks for the uniform (grey triangles) and the non-uniform distribution (black dots). Note that the frequency differences are *not* logged in the upper panel but the ranks are logged in order to align them with the plot in the lower panel.
or, by substituting the denominator with (1):

$$NFD(A, B) = \frac{\sum_{i=1}^{\max\{V^A, V^B\}} \DeltaFreq(A, B, i)}{N^A + N^B}$$  \hspace{1cm} (4)

In the numerator we have the sum of frequency differences per rank, i.e. the sum of all \(\DeltaFreq\) in the left upper panel of Figure 1. The denominator corresponds to the sum of the overall token frequencies for both distributions, i.e. the sum of the grey and black bars in the lower left panel. An intuitive interpretation of the NFD is that it is the percentage of token frequency differences per overall number of tokens.

The actual non-uniform and uniform token frequency distributions chosen for illustration in Figure 1 are:

\[
\begin{align*}
F^A &= (45, 20, 15, 10, 5, 1, 1, 1, 1, 1) \\
F^B &= (10, 10, 10, 10, 10, 10, 10, 10, 10, 10)
\end{align*}
\]

The NFD for these is:

$$NFD(A, B) = \frac{100}{100 + 100} = 0.5$$

This means that the sum of token frequency differences amounts to 50% of the overall number of tokens in the uniform and non-uniform distribution together. Generally, we have that \(0 \leq NFD \leq 1\). The normalized frequency difference is 0 when both vectors are exactly the same \((F^A = F^B)\). The NFD is 1 if and only if one of the vectors consists of zeros and the other of at least one non-zero element. Hence, the NFD for real word frequency distributions will range in between 0 and 1, with values closer to 1 indicating bigger frequency differences.

Note, also, that the ranking of frequencies of two distributions from highest to lowest before calculating the frequency differences will yield the minimum NFD, whereas ranking one of the distributions from highest to lowest frequency and the other
from lowest to highest would render the *maximum* NFD.

In theory it does not make any difference which of these we choose to measure the difference between two distributions, as long as our choice is consistent. However, conceptually it makes sense to rank frequencies from highest to lowest, since this way we get a better overview of the frequency profile.

Finally, the lower right panel of Figure 1 illustrates another convention in word frequency research. We transform the ranks and frequencies by applying the natural logarithm and we use a scatterplot of dots instead of a barplot. This is a convention for plotting – not for calculating the NFD – which helps to better see the shapes of the whole distributions even if they have very long tails. Note that we do *not* logarithmically transform the $\Delta$Freqs in the upper panel, since we do not want to reduce the visual salience of frequency differences. However, we *do* logarithmically transform the values of the ranks in order to align them with the lower plot.

4. Analysis 1: Inflection, derivation, compounds and contractions/clitics in English and German

In our first analysis we use the NFD to measure differences between word frequency distributions before and after changing word types by 1) removing inflectional morphology, 2) removing derivational morphology, 3) splitting compounds and 4) removing contractions and clitics. We want to know the impact each of these transformations has on frequency distributions independent of the others. Hence, we proceed by applying them separately. The corpora used, methods and results are outlined in the following.

4.1 Materials

We compiled parallel corpora for English and German using parts of the *Open Subtitles Corpus* (2013, http://opus.lingfil.uu.se/OpenSubtitles2013.php), the *Europarl Corpus* (Koehn 2005, http://www.statmt.org/europarl/), the *Universal Declaration of Human Rights* (http://www.unicode.org/udhr/index_by_name.html) and the *Book of Genesis* (http://homepages.inf.ed.ac.uk/s0787820/bible/). More detailed information about the composition of the parallel corpus sample can be found in Table 1. The advantage of this sample is that the text passages are sentence aligned and hence exact translational
equivalents. This parallel structure provides a natural means of controlling for constant content, which is a confound in non-parallel texts. Moreover, the sample is balanced between spoken and written language as well as different registers (colloquial, political, legal, religious). The disadvantage is that the sample is small (9211 tokens in English, 8304 in German). However, for this analysis keeping the sample small was necessary to enable maximally informed, manual transformations in the morphology.

4.2 Methods
For the English and German parallel corpora outlined in Table 1 we first set all letters to lower case. Consistent with our definition in Section 2 we take non-alphanumeric characters as word type delimiters. However, in this analysis we do not by default split word types on hyphens and apostrophes, since they are relevant for the formation of compounds as well as contractions and clitics. The English and German original corpora are then manually transformed according to the following principles (see Appendix 1 for more detailed explanations).

4.2.1 Inflections
Inflections are neutralized in regular and irregular verbs (e.g. decides/decide/decided $\rightarrow$ decide, sings/sang/sung $\rightarrow$ sing) and nouns (e.g. noses $\rightarrow$ nose, children $\rightarrow$ child, Johanne's $\rightarrow$ Johanne) in English. Note that we include the 's genitive as nominal inflection here, but we do not include -ing forms that are used as adjectives (flaming sword) or nouns (the teaching of). These are categorized as derivational suffixes. In German, inflection is more extensive in the sense that there are suffixes that also have to be removed from adjectives, articles and demonstratives (e.g. verdammte $\rightarrow$ verdammt, dem/den/des $\rightarrow$ der, dieser/diese/dieses $\rightarrow$ dies).

4.2.2 Derivation
There is a whole range of Germanic and Latin prefixes and suffixes that are considered derivational in English (e.g. in-alien-able $\rightarrow$ alien, hope-ful-ly $\rightarrow$ hope, childhood $\rightarrow$ child). We include nominalizing -ing as in mewling thing $\rightarrow$ mewl thing, and teaching
Note that derivation and inflection can overlap so that removing
derivational suffixes renders non-existent words (e.g. realized → reald). For German the
picture is again a bit more complicated since multiple derivational affixes are commonly
attached to the same root (e.g. Anerkennung → kennen, Errungenschaften → ringen)
and can overlap with compounding (e.g. Dringlichkeitsdebatte → Dringensdebatte) and
inflectional morphology (e.g. abgeändert → ändert).

4.2.3 Compounds
Different parts of speech can be compounded (e.g. noun-noun, adjective-noun,
preposition-noun, among others). We split these back into two separate word types (e.g.
daytime → day time, downstairs → down stairs, gentlemen → gentle men). However, we
do not “de-compound” proper names such as Hellfish. Similar principles apply to
German (e.g. Arbeitsschutzregelungen → Arbeit schutz regelungen, kräuterstinkender
→ kräuter stinkender).

4.2.4 Contractions and clitics
Since with the OSC we include spoken language, there are a range of contractions and
clitics to be found in both the English (e.g. you've → you have, you're → you are, I'll →
I will, won't → will not, parliament's → parliament') and German (geht's → geht es,
rührt's → rührt es, dir's → dir es, beim → bei dem, ins → in das).

For each language we separately neutralize inflections, derivations, compounds and
contractions/clitics as outlined above and compare the resulting word frequency
distributions with the original ones.

4.3 Results
4.3.1 English
The result for removing inflections and derivations in English can be seen in Figure 2.
The NFD between the original corpus and the lemmatized one is 0.072. This means that
removing inflections causes a token frequency change that amounts to 7.2% of the
overall number of tokens in both distributions. The NFD for the original corpus and the
corpus with removed derivational morphology is exactly six times smaller (0.012 or
1.2% change). The upper panels of Figure 2 further illustrate this difference. It is mainly due to the impact that neutralization of inflections has in the high frequency range of the distribution, while removal of derivational morphology does not have this impact. Towards the low frequencies the differences are similar.\footnote{Consider the following example to see why this happens. The high frequency lemma \textit{go} is represented in different word types with respective frequencies in our original text (\textit{go} 10, \textit{going} 6, \textit{went} 3, \textit{gone} 2, \textit{goes} 1, \textit{goeth} 1). If we lemmatize these word types to the lemma \textit{go}, we collapse the whole distribution of different frequencies into a single frequency: 23. In contrast, the words that are modified by derivational morphology are rather in the middle and low frequency range and there is only a fairly limited number of different derivational affixes that apply to the same word (e.g. \textit{hope} 5, \textit{hopefully} 1). Hence, changing \textit{hopefully} to \textit{hope} affects the distribution only minimally.

The token frequency differences in distributions with and without compounds as well as with and without contractions/clitics can be seen in Figure 3. It is remarkable that in our English corpus contractions and clitics change the frequencies more strongly.

![Figure 2. Frequency differences in English illustrated for the removal of inflectional marking (left panel) and the removal of derivational marking (right panel). Original distributions are represented by grey triangles, changed distributions by black dots. Frequency differences per rank (non log-transformed) and NFD values are given in the upper panels.](image)
than compounds or derivational morphology. In fact, for the English corpus derivations have the least impact on the frequency distribution. The order in terms of NFD is: inflections 7.2%, contractions/clitics 2%, compounds 1.3%, and derivations 1.2%. Note, however, that the differences between contractions/clitics, compounds and derivations are minor and might change for different combinations of text types.

4.3.2 German
The results for the German corpus are somewhat different. The NFD order is: inflections 10.9%, derivations 4.8%, compounds 2.1%, and contractions/clitics 1.5% (see Figure 4 and Figure 5). The qualitative asymmetry between inflections and derivations is still given, though in German derivations have a stronger impact on frequency distributions than in English and seem overall more productive than compounds and contractions/clitics.
Figure 4. Frequency differences in German illustrated for the removal of inflectional marking (left panel) and the removal of derivational marking (right panel). Original distributions are represented by grey triangles, changed distributions by black dots. Frequency differences per rank (non log-transformed) and NFD values are given in the upper panels.

Figure 5. Frequency differences in German illustrated for the splitting of compounds (left panel) and the splitting of clitics and contractions (right panel). Original distributions are represented by grey triangles, changed distributions by black dots. Frequency differences per rank (non log-transformed) and NFD values are given in the upper panels.
5. Analysis 2: The NFD as a cross-linguistic inflection index

In Analysis 1 inflectional marking emerged as a predominant factor changing frequency distributions. However, the NFD variation for inflection marking between English and German already suggests that the relative impact on frequency distributions is likely to differ across languages. In fact, the idea that the degree of inflection of a language is reflected in the distribution of word types goes back to Zipf (1932, 1949), and re-appears (among others) in Ha, Stewart, Hanna & Francis (2006), as well as Popescu et al. (2009), and Popescu, Altmann & Köhler (2010). In the following analysis we further quantify these cross-linguistic differences. We use state-of-the-art lemmatization tools to automatically remove inflectional marking in parallel corpora across 19 languages. This allows us to calculate the NFD between unlemmatized (i.e. original) and lemmatized corpora as a frequency based measure of inflectional productivity, i.e. a cross-linguistic “inflection index”. We will henceforth denote this specific kind of normalized frequency difference between lemmatized and unlemmatized texts as NFD_{lem}.

5.1 Materials

We compile parallel corpora by using the full UDHR and the full Parallel Bible Corpus (PBC, Mayer & Cysouw 2014) for each language. The range of texts is limited here by the set of languages for which lemmatization is possible. If we want to exploit this set of languages fully, then we have to restrict the parallel corpora to the UDHR and the PBC. Overall, we arrive at parallel corpora of 12000-17000 tokens for 19 different languages (see Table 2 for details).

5.2 Methods

The splitting of character strings into word types is implemented by the function strsplit() in R (R Core Team 2013, see also Gries 2009). Note that this string splitting method yields clitics and contractions marked by apostrophes as separate word types, i.e. he’ll, it’s and John’s are split to he ll, it s and John s respectively. Likewise, compounds connected by hyphens are split into separate word types.

The word types are then lemmatized using the BTagger (Gesmundo & Samardžić 2012) and TreeTagger (Schmid 1994, 1995). Both the BTagger and the
TreeTagger will first associate the respective word type with a part-of-speech tag (POS tag) and then derive the most likely lemma. For example, for the English word type *rights* the BTagger outputs: *rights, Nc, right*. This is the original word type, the POS tag for *common noun*¹, and the respective lemma.

Automatic processing results in a number of errors, which can influence the observed differences between original and lemmatized texts. The number and the type of errors depend on the lemmatization approach and on the level of difficulty. Both taggers are based on statistical models trained on samples of manually lemmatized text. They are both able to provide high accuracy on words already seen in the training set (close to 100%). The words that are not seen in the training set are harder for both taggers and can be expected to result in more errors. Table 2 shows the percentage of word types unknown for each text and tagger. Using the BTagger with our parallel texts yields more unknown words than for the TreeTagger on average.

Note, however, that this does not necessarily mean that the BTagger will make more errors. Due to its good generalising capacities (Gesmundo and Samardzic 2012), the BTagger obtains a relatively good performance on unknown words as well, whereas the TreeTagger will just output the original word type as lemma for any unknown word. Despite these differences, the overall effect of errors on the $\text{NFD}_{\text{lem}}$ estimation is expected to be similar. Both taggers will transform fewer word types to lemmas than they actually should. In consequence, there will be less of a difference between lemmatized and unlemmatized frequency distributions than there should be, and we will slightly underestimate the $\text{NFD}_{\text{lem}}$.

Also, for both taggers there are generally more unknown words in languages with many inflectional categories. Note, for example, that the percentage of unknown words is higher for Polish than for English for both taggers. We thus expect that our $\text{NFD}_{\text{lem}}$ estimations are somewhat less reliable for the languages with abundant inflection.

In sum, we can expect real $\text{NFD}_{\text{lem}}$ values to be closest to the estimated values in languages such as English, and to be slightly higher than estimated in languages such as...
Polish.

A specific problem with the TreeTagger is that POS tagging and lemmatization for individual languages is based on different treebanks and hence different lemma annotations. For example, for the Estonian word type *inimõiguste* 'of human rights' (GEN.PL) the TreeTagger will output *inim_õiguste* as lemma. The underscore indicates compounding, and the +te indicates the genitive plural marker. However, the actual lemma we want to arrive at is *inimõigus* 'human right' (the BTagger outputs exactly this lemma). In such cases we have to do post-processing of the TreeTagger output to remove the symbols that are not part of the actual lemma.

Note, also, that the TreeTagger can exhibit somewhat unusual behavior with pronouns. For example, in Spanish it lemmatizes all articles (*el, la, lo, los, etc.*) to the masculine form *el*, which as a consequence accumulates a very high frequency (see first rank in the middle panel of Figure 6).

5.3 Results

5.3.1 Cross-linguistic comparison of NFD_{lem} values

In Figure 6 we choose three languages (English, Spanish, Finnish) to represent the range of NFD_{lem} values we find. English has the lowest NFD_{lem} value (5.2%). Note that the value for manual removal of inflection in Analysis 1 was higher (7.2%). We will get back to this difference in Section 6. Spanish is in the middle range (12.2%) and Finnish has the highest value of the 19 languages (19.7%). Interestingly, despite these quantitative differences the patterns of change are similar across languages. Namely, lemmatization universally affects the high frequency ranks and shortens the tail of low frequency word types. Again, this illustrates that inflectional marking systematically creates low frequent word types.

The full range of NFD_{lem} values can be seen in Figure 7. The bulk of languages falls within the middle range from 10-15%, relatively few languages fall below 10% and even fewer above 15%. Hence, languages seem to be approximately normally distributed around an inflectional productivity of 12.5% (measured in NFD_{lem}) with a slight skew towards having rather less than more.

Note, also, that for the three languages for which we can use both the BTagger and the TreeTagger the values are fairly similar (English: 5.2% and 5.4%, Estonian:
14.8% and 15.8%, Polish: 11.6% and 12.9%). This is reassuring, since it suggests that differences in NFD_{lem} are not strongly driven by idiosyncrasies of the taggers.

Figure 6. Changes of frequency distributions between unlemmatized (grey triangles) and lemmatized (black dots) texts. English, Spanish and Finnish are chosen to represent the range of the original 19 languages.

Figure 7. NFD_{lem} as an inflection index across 19 languages. The x-axis represents languages with respective ISO 639-3 codes. The y-axis represents the NFD_{lem} between unlemmatized and lemmatized versions of the UDHR and PBC parallel corpora. The colours of bars indicate whether the texts were lemmatized using the BTagger (red) or TreeTagger (blue). Note that for three languages (English, Polish, Estonian) both options are available.
Another, conceptually somewhat different question is how much of the variance in NFD values across different languages is due to differences in inflectional marking. While the NFD$_{lem}$ values reflect how much token frequencies differ between unlemmatized and lemmatized distributions of *the same language*, here we want to know how much of the difference in unlemmatized distributions across different languages can be attributed to differences in inflectional marking, i.e. which proportion of the variation in cross-linguistic NFD values before lemmatization is covered by variance in NFD values after lemmatization.

5.3.2 Calculating the effect of lemmatization

To assess how much NFD values are reduced by lemmatization, we can create two matrices of pairwise comparisons: one matrix with NFD values of languages before lemmatization, and one matrix with NFD values after lemmatization. This way we can calculate the mean NFD for the original (unlemmatized) distributions as 0.14 (SD=0.08), and the mean NFD for the lemmatized distributions as 0.12 (SD=0.05). Finally, the proportion of NFD variance of lemma distributions over variance in the original distributions is 48.8%. In other words, about 50% of the NFD variance that we find across the original word frequency distributions of 19 languages is due to variance introduced by inflectional marking. Hence, the other 50% will be due to derivational morphology, compounds, contractions/clitics and differences in the base vocabulary.

6. Analysis 3: The effect of text size on the NFD$_{lem}$

In any corpus analysis relating to morphological productivity it is important to control for corpus size. This has been pointed out most clearly in quantitative studies on vocabulary growth (Baayen 1992, 1994, 2001: 2), which suggest that using relatively small texts of the kind we used in the preceding analyses might systematically underestimate the actual NFD$_{lem}$. In the following we test the behavior of the NFD$_{lem}$ with growing text size.

6.1 Materials

One of the biggest parallel corpora in terms of number of tokens is currently the *European Parliament Corpus* (EPC, Koehn 2005). It contains several million words of...
European Parliament discussions in 20 European languages. However, due to the fact that lemmatization and the sampling methods we use are relatively time consuming, we use only the first 1 million words per language instead of the full corpus.

6.2 Methods

Matching the languages in the EPC with the languages of the TreeTagger yields a sample of 10 languages for which lemmatization is possible. As in Analysis 2, we split character strings into word types by using the function `strsplit()` in R (R Core Team 2013, see also Gries 2009). Word types are then lemmatized by the TreeTagger.

In order to estimate a) the true NFD\textsubscript{lem} value per language, and b) its relation to sample size, we use three methods of sampling:

1) **Continuous sampling:** We take increasingly larger chunks from the first 100K words of each corpus (i.e. 10 word tokens, 11 word tokens, …, 100K word tokens from the beginning). For these chunks we calculate the NFD\textsubscript{lem} value between frequency distributions of words and lemmas (see Figure 8 continuous). Note that the Europarl corpus is essentially a concatenation of speeches. Hence, sampling from the beginning can bias the early NFD\textsubscript{lem} values in a specific direction and it will take some time to converge on the actual value.

2) **Matched random sampling:** To overcome the potential bias of continuous sampling we use random sampling. Namely, we randomly sample increasingly larger chunks from the original 1M corpora (e.g. randomly sampling 10 word tokens, 11 word tokens, … 100000 word tokens). Also, word types and lemmas are **matched** (i.e. each word type would be paired with its lemma). The results of this method are represented in Figure 8 as random (matched).

3) **Dissociated random sampling:** Using matched word types and lemmas could potentially introduce a further bias. Hence, in our third sampling method we use the same method as in 2), except that word types and lemmas are not matched (see random (**dissociated**) in Figure 8). Since taking two random samples (one for word types and one for lemmas) at each step would quickly become computationally inefficient, we start by shuffling the word types and lemmas in the output file of the TreeTagger so that the word types are not paired with their corresponding lemmas.
6.3 Results

Figure 8 illustrates how the average NFD_{lem} changes with the number of tokens for parallel texts of 10 languages in the EPC, and for the three different sampling conditions. The vertical dashed lines denote 15K tokens, i.e. roughly the average size of texts we had in the analysis of NFD_{lem} values across 19 languages (Analysis 2).

It is clear from Figure 8 that for some strongly inflected languages like Finnish (fin), Polish (pol), Slovak (slk) and Estonian (est) we might slightly underestimate the actual NFD_{lem} values with a text size of 15K and smaller (also depending on the sampling method).

![Figure 8](image)

**Figure 8.** Relationship between number of tokens (x-axis) and NFD_{lem} (y-axis) for 10 languages and three conditions of sampling. The coloured curves are NFD_{lem} values smoothed with a general additive model (gam). ISO 639-3 codes translate as: fin (Finnish), pol (Polish), slk (Slovak), est (Estonian), deu (German), spa (Spanish), ita (Italian), fra (French), eng (English), nld (Dutch). We used 1M word tokens in the original analyses, but we reduced the size to 100K, since the values already converge at around 50K at the most. The vertical dashed lines represent 15K tokens, which corresponds roughly to the average text size we had in Analysis 2.

However, even for these languages the value converges at around 50K. Also, at that text size the sampling method does not play a role anymore. Overall, this is a surprising and encouraging result. It goes to show that the true NFD_{lem} value (of a specific parallel
corpus and language) can be estimated by using even relatively small subsamples of it.

Note that the NFD$_{lem}$ values we end up with for parallel corpora in the analysis on inflectional marking across 19 languages (Analysis 2) and the NFD$_{lem}$ values of the current analyses might be confounded slightly by the specific register of these corpora (see Table 3). For example, if we take the values of Analysis 2 and compare them to the converged values of condition 3 of the current analysis (presumably the least biased sampling method), we find some minor discrepancies for close languages like English and Dutch (their values are swapped). This is also most likely the reason why we end up with a somewhat higher value for English in Analysis 1 (7.2%) compared to Analysis 2 (5.2%).

### INSERT TABLE 3 AROUND HERE

However, the overall Spearman correlation between values of Analyses 2 and 3 (condition 3) is strong ($r_S = 0.94$, $p < 0.0001$), suggesting that neither using different parallel corpora (UDHR and PBC in Analysis 2, and EPC in Analysis 3) nor using different sampling methods are major confounds for the estimation of NFD$_{lem}$ per language.

7. Analysis 4: The impact of lemmatization on different parts of speech

Besides using the NFD to measure differences in word frequency distributions for different morphological markers (Analysis 1), as an inflection index across different languages (Analysis 2) and for different text sizes (Analysis 3), we might also want to look at how changes of morphology interact with different parts of speech (POS), e.g. inflectional marking for nouns, verbs (content words) compared to the distributions of prepositions (function words).

#### 7.1 Materials

We use the same materials here as in Analysis 2. Namely, a combination of the full UDHR and the full PBC as text samples.

#### 7.2 Methods
Again, the tokenization, tagging and lemmatization procedures are the same as for Analysis 2, except that here we use the BTagger only. This is because the TreeTagger uses different sets of POS tags, which makes it harder to meaningfully compare parts of speech across different languages. The BTagger uses a (largely) consistent set of POS tags taken from the *Multext-East morphosyntactic definitions* (MSD) (see footnote 3). Remember from Section 5.2 that for the English word type *rights* the BTagger outputs: *rights, Nc, right*. This is the original word type, the POS tag for common noun, and the respective lemma. Similarly, for the Estonian equivalent *ōiguste* it outputs: *ōiguste, Nc, ōigus*. Using these outputs we can create frequency distributions of words and lemmas per POS tag.

![INSERT TABLE 4 AROUND HERE](image)

For example, Table 4 gives the first ten ranks of distributions for word types and lemmas of English main verbs (Vm) only. Based on this filtering by POS tags we can plot distributions and calculate NFD_{lem} values for common nouns, main verbs and prepositions separately. As a workable example, we take English and Estonian to represent low-inflection and high-inflection languages respectively. For these two we compare differences in the distributions of nouns, verbs and prepositions before and after lemmatization.

### 7.3 Results

Figure 9 and Figure 10 illustrate how word frequency distributions differ between different parts of speech in English and Estonian, as well as how much impact inflectional marking has on word types of each part of speech.

Generally speaking, the unlemmatized distributions of nouns and verbs look fairly similar within the same language, whereas prepositions have a “steeper” distribution occupying the high frequency range – as we would expect for function words. Note, however, that there is an interesting asymmetry between distributions of prepositions between the two languages, with English having more and higher frequent prepositions than Estonian. Moreover, lemmatization affects verbs more strongly than nouns in both languages. In fact, this asymmetry in inflectional marking is even stronger for English than for Estonian, as reflected in Table 5.
Namely, the NFD\textsubscript{lem} value for verbs in English is roughly 4 times higher than the value for nouns, whereas in Estonian it is only twice as high.

8. Discussion

In Analysis 1 we calculated NFDs for four different kinds of word type formation.
patterns (inflection, derivation, compounding and contraction/clitics) and two different languages (English and German). Overall, the results of Analysis 1 can be interpreted in two ways: either with a focus on the relative importance of different morphological marking strategies within the same language, or as a comparison of the same marking strategies across the two languages.

With regards to the former it can be said that inflectional marking has the strongest impact on frequency distributions in both English and German. Removing inflectional markers has both an impact on the high frequency ranks and the low frequency ranks, since the frequencies of different word forms add up to the frequency of the respective lemma. In other words, having inflectional marking in a language systematically creates low frequency word types and “pushes” the overall word frequency distribution towards having a longer tail (as predicted for example by Baayen 2001: 155-160). Contractions and clitics have a qualitatively similar effect.

In contrast, the impact is different for derivational morphology in the sense that a) there is almost no change in the high frequency ranks, and b) the overall change in token frequencies is smaller. This seems linked with Moscoso del Prado Martín et al.’s (2004: 5) observation that for predicting lexical decision latencies (i.e. processing difficulty) token frequencies are more important for word types belonging to inflectional paradigms than for word types belonging to derivational paradigms. Take the example of different inflectional variants of go (go 10, going 6, went 3, gone 2, goes 1, goeth 1) and derivational variants of hope (hope 5, hopefully 1) from above. Just taking the token frequency of go as a predictor for reaction times would grossly underestimate the frequency “support” given from other inflectional variants (10 vs. 23), whereas for hope this would barely make a difference (5 vs. 6).

If it holds true that frequencies of words are directly related to their learnability and processing difficulty, then these results suggest that inflectional marking might have a systematically higher “cognitive cost” than derivational morphology, since it systematically creates lower frequency items. Thus, the NFD might emerge as an objective, quantitative way of measuring the learnability and processing difficulty of morphology across languages.

For compounds we find a pattern that is similar to the one for derivation. Splitting compounds affects mainly the middle and low frequency ranks towards the tail
of the distribution, and has overall only a small effect on the distribution.

Comparing the same marking strategies across the two languages we find that inflection and derivation are more productive in German than in English – as we would expect – whereas compounding as well as contraction/cliticization appear to have similar productivity with slight deviations. This is somewhat surprising given that German is often referred to as a language taking compounding to its extremes. This perception might be caused by the fact that in theory almost any number of words can be compounded together in German. Take the example of the *Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz*, which translates into English as "the law for the delegation of monitoring beef labelling". However, despite such extreme examples, we do not find strong evidence in our analysis that compounding is much more productive in German than English looking at the actual frequencies in practice. The actual normalized difference that splitting of compounds causes in token frequencies of our corpus is 1.3% in English and 2.1% in German. If this result is replicated in further studies with bigger and more representative corpora, it might be taken as an example of how some extreme cases can bias our perception of the actual productivity of word formation patterns.

Analysis 2 then focused on a cross-linguistic analysis of inflectional morphology. This is partly motivated by the fact that inflectional morphology turned out to be the predominant factor changing frequency distributions in Analysis 1, and partly by the fact that automated tools to remove derivational affixes, compounds and contractions/clitics across different languages are not available at this point (to our knowledge).

Analysis 2 shows that the NFD can be used as a cross-linguistic, frequency-based inflection index. It further illustrates that inflectional morphology “pushes” word frequency distributions towards the low frequency tail. Importantly, this is not an idiosyncrasy of specific languages, but a general property of inflectional marking. Moreover, there seems to be a “natural” tendency for languages to range around an NFD of 10-15%, with a slight skew towards rather having a lower value (< 10%) than a higher one (> 15%). The “outliers” might be the synchronic outcome of more “extreme” histories of language change and learning pressures, such as language contact versus relative isolation (Bentz, Verkerk, Kiela, Hill & Buttery 2015, Bentz & Winter
The methods and data used in Analysis 2 can also be used to measure how much of the NFD variance we find across different languages is due to differences in the productivity of inflectional marking. This value was estimated to around 50%. Hence, inflectional marking can be said to have a strong cross-linguistic impact on word frequency distributions. This is an important result for studies that try to explain cross-linguistic differences in lexical diversity by learning pressures (e.g. Bentz et al. 2015). The other half of the variance will be divided between other morphological marking strategies and differences in the range of word types in the base vocabulary. To further disentangle these we would need computational tools to automatically remove derivational morphology, compounds and contractions/clitics from a set of languages.

Analysis 3 systematically tested the dependence of the inflection index (NFD_{lem}) on the number of tokens. It turned out that for most languages small text sizes of around 10-15K are enough to get a good approximation, though for strongly inflected languages this number might go up to 50K or more. This analysis also suggested that the register of a text (e.g. European Parliament discussions versus legal texts and Bible translations) might be a slight confound. Of course, in the optimal case we would be able to compile parallel corpora of around 100K across a wide range of registers to closely approximate the actual inflection indices representative for whole languages. Hence, advancing quantitative cross-linguistic comparison is a matter of building larger parallel corpora and elaborating computational tools to process them.

Finally, Analysis 4 added another level of detail by looking at word frequency distributions for different parts of speech in English and Estonian. As we would expect, closed class function words (e.g. prepositions) have much steeper distributions (fewer word types and higher token frequencies) than content words (e.g. nouns and verbs) in both languages. In fact, the idea that overall word frequency distributions are composed of different component distributions according to parts of speech goes back to at least Yule (1944: 19-21). Baayen (2001: 155-160) gives a mathematical account of how to model these components based on mixture models. However, only now the computational tools and corpora are becoming available to empirically estimate the exact values and shapes of component distributions.
Moreover, analyzing parts of speech separately suggests that there might be a measurable trade-off between preposition-heavy encoding strategies (e.g. English), on one hand, and nominal inflectional marking strategies (e.g. Estonian), on the other hand. Based on the NFD$_{sen}$ results we can conjecture that the nominal encoding strategy requires a wide range of low-frequent nominal word types, whereas the preposition heavy strategy relies on a small range of high-frequent word types instead. Again, the occurrence of these differing strategies might be linked to specific histories of language learning and the respective processing pressures. Eventually, measuring such trade-offs can help to disentangle the pathways along which different linguistic features change and interact.

In the following, we want to address two more general points regarding the NFD. Namely, its relation to other lexical diversity measures, and its implications for a language typology based on corpus data rather than expert categorization.

8.1 Comparison to lexical diversity measures

There is a wide range of lexical diversity (LD) measures in quantitative and applied linguistics (see for example Baayen 2001, Jarvis 2002, McCarthy & Jarvis 2007, 2010, Mitchell 2015, Tweedie & Baayen 1998 for an overview, and Michalke 2014 for an R implementation). In fact, Mitchell (2015) reports a total of 50 different models to describe type-token ratios. In principle, any of these LD measures could be used to calculate the lexical diversity difference ($\Delta$LD) between two word frequency distributions instead – or on top of – the NFD. Table 6 gives values for a range of LD measures (mainly the ones represented in the koRpus package by Michalke 2014) applied to the *uniform* and *non-uniform* distributions ($F_A$, $F_B$) used earlier to demonstrate the NFD measure in Section 3.

The advantage of LD measures over the NFD is that they can be applied to a single distribution rather than requiring a pairwise comparison of distributions. This is convenient when comparing lexical diversities across many languages and across different time periods (e.g. Bentz et al. 2015, Bentz et al. 2014, Koplenig 2015b). However, there are also disadvantages.
Now, parametric measures have the disadvantage that they require curve fitting by assuming an underlying model like the Zipf-Mandelbrot model (Mandelbrot 1953). Since it is generally hard to determine the right balance between underfitting by using the most parsimonious model and overfitting by using strongly modified models, the results of such curve fitting procedures are an easy target for criticism.

On the other hand, most non-parametric models based on TTR will not indicate the difference between the uniform and non-uniform distribution at all, or give it only minor recognition (except for MTLD). This is because they are purely based on the ratio of word types to word tokens, which is actually the same for the uniform and non-uniform distributions in our example. Hence, TTR-based measures tend to be insensitive to the exact distribution of token frequencies. As Analysis 1, 2, and 4 have shown, this is a shortcoming, since changes in grammatical marking can have subtle effects on the exact distributions of word types and tokens.

To capture these differences we are left with Shannon H over word types, ZM parameters, MATTR, MSTTR, Yule's K, HD-D, and MTLD (and potentially others that we have not tested). At least Shannon H and ZM parameters have been applied in earlier cross-linguistic and diachronic studies to measure lexical diversities (e.g. Bentz et al. 2015, Bentz, Kiela, Hill & Buttery 2014, Koplenig 2015b). Generally, difference indices based on these LD measures are expected to strongly correlate with NFDs. For example, for the cross-linguistic inflection index in Analysis 2 Figure 11 illustrates a strong Pearson correlation ($r=0.96$, $p<0.0001$) between the difference in Shannon H and the NFD$_\text{lem}$.

Thus, in practice these LD measures might be just as suitable to measure differences between distributions as the NFD. However, it is not clear if they exhibit the properties of stable convergence even for small text sizes that we have illustrated in Analysis 4 for the NFD$_\text{lem}$. It is beyond the scope of this paper to test the behavior of all of these measures for growing text sizes as we did for the NFD$_\text{lem}$.

In any case, what speaks for the NFD is first that it is a non-parametric measure, which does not require curve fitting by assuming an underlying model. This makes it less theory-dependent and immune to discussions surrounding the correct parametric model to be fitted. Second, it is sensitive to even minimal changes in the exact
distribution of token frequencies over type frequencies, and can hence measure subtle differences at any level of detail (given the right corpora). Third, the NFD has a straightforward, intuitive and frequency-based interpretation. It is the percentage of token frequency differences per overall number of tokens. In contrast, interpreting differences in Shannon entropy (Shannon & Weaver 1949) or the Kullback-Leibler divergence (Kullback & Leibler 1951, see Bochkarev et al. 2014 for an application) requires a thorough understanding of the mathematical underpinnings of information theory.

8.2 Toward a quantitative corpus typology

Our paper mainly focused on defining and testing the NFD as a measure, and applying it to assess what drives differences in word frequency distributions. However, we also would like to make a more general point here about the emerging field of quantitative typology.

In recent years, computational and statistical methods have found wider application in the area of linguistic typology. This is possible mainly through the development of large scale, cross-linguistic databases of language information such as the Ethnologue (Lewis, Simons & Fenning 2013), the World Atlas of Language
Structures (WALS, Dryer & Haspelmath 2013), the AUTOTYPE database (Bickel & Nichols 1999), the Glottolog (Hammarström, Forkel, Haspelmath & Bank 2015), and more recently the development of massive parallel corpora (Mayer & Cysouw 2014, Koehn 2005).

Take the WALS as an example. It contains 151 chapters with expert judgements on how to categorize languages with regards to a range of linguistic features. For example, chapter 49 (Iggesen 2013) gives the “Number of Cases” as a discrete ordering from “no morphological case” to “10 or more cases” for 261 languages. This can be seen as a valuable first impression of cross-linguistic case marking strategies.

However, it is only a very coarse-grained approximation for the actual usage of case markers. Bickel (2015) points out that we would need whole feature matrices of case markers according to the exact properties they have in a specific language. Also – above and beyond description – the productivity of case markers in languages can vary vastly. For example, ranking German and Icelandic as having 4 cases (nominative, accusative, dative and genitive) is a fairly abstract, theory-driven categorization, which conceals the fact that the overall frequencies of usage might be vastly different between the two languages. In fact, even within the same language a specific case marker might have a different productivity for different words. For example, the word land 'country' occurs 5 times in our German corpus used in Analysis 1, the genitive singular marked form land-es occurs 4 times. In contrast, the word gott 'god' occurs 55 times and the genitive singular form gottes still only 4 times. So, if we just take these numbers as our frequency distributions and calculate the NFD for these, we get 0.44 for the productivity of the genitive singular marker with land and only 0.07 for gott. Hence, the genitive singular marker is almost 6 times more productive for land than for gott (in our text sample). Of course, at this level of specificity our results will depend much more on the composition of corpora, and we will generally run into the problem of data sparsity. Again, these problems can only be overcome by compiling bigger, more balanced, and hence more representative parallel corpora.

Ultimately, we want to be able to measure the productivity of morphological markers – or any other linguistic structure – with any depth of specification from corpora directly. The application of POS tagging and lemmatization tools in combination with quantitative measures such as the NFD are a first step in that
direction. Hence, the NFD is a more realistic, empirical, and less theory-driven estimation of morphological productivity in particular, and a measure for the impact that any systematic manipulation of word types might have on frequency distributions more generally.

Of course, applying computational tools in typology does not entirely overcome our reliance on expert judgements and theoretical reasoning, since these are implemented in tools like the BTagger or TreeTagger. However, our analyses become more reducible, and the impact of our theory-driven decisions becomes more directly measureable. We think that a similar reasoning applies to other linguistic features such as phoneme inventories and word/constituent order.

9. Conclusions
We established here the Normalized Frequency Difference as a measure of the deviation between any two frequency distributions. This measure is widely applicable, relatively easy to interpret and – in its application as an inflection index – surprisingly robust to differences in corpus size and register. Our analyses show that it is interesting for two broad lines of linguistic research: a) to estimate the impact that changes in word types have on the word frequency distribution of a specific language, b) to assess the difference in impact of changes across languages. Though we have focused mainly on morphological marking strategies in this paper, the NFD can be used as a measure more generally, namely whenever we define a systematic way of changing word types that is reflected in their written form. Ultimately, we are aiming at making language typology more corpus-based, empirical and hence reproducible.

References


Koplenig, A. (2015a). The impact of lacking metadata for the measurement of cultural and linguistic change using the google ngram data sets – reconstructing the composition of the German corpus in times of WWII. *Digital Scholarship in the Humanities*. http://dx.doi.org/10.1093/llc/fqv037


Verbal Behavior, 17 (2), 143–154.


Notes

1 In the upper panels we log-transform the ranks of the distributions, but not the ΔFreq. This exaggerates the visual differences in frequencies somewhat.

2 Note that we included the ‘s genitive both under inflection and clitics. Theoretically it should be considered a phrasal clitic, since it does not attach directly to nouns, but rather to noun phrases. However, in practice it is found mostly on nouns and might be perceived as noun inflection by learners and speakers.

3 The POS tags used in the BTagger are the first two letters of the Multext-East morphosyntactic definitions (MSD). See a full list here:

Appendixes

Appendix 1

Word types were manually changed according to the principles outlined for English and German below. In difficult cases the Longman Grammar (Biber, Johansson, Leech, Conrad & Finegan 1999) and the Duden Grammar (Fabricius-Hansen, C., Gallmann, P.,
Eisenberg, P., Fiehler, R., Peters, J., Nübling, D., . . . Fritz, T. A. 2009) were consulted for English and German respectively.

English
Inflection
- Verb forms that are marked for third person, past tense, progressive as well as gerunds and present participles (e.g. *is*, *was*, *were* → *be*; *suggests*, *suggested*, *suggesting* → *suggest*). Note that this does not include *-ing* forms that are used as adjectives (flaming sword) or nouns (the teaching of). These are categorized as derivational suffixes.
- Noun inflections such as s-plurals and 's genitives as well as irregular forms (e.g. *Johann's* → *Johann*).
- Personal pronouns that are remnants of the case system (e.g. *him, his* → *he*).
- Omissions resulting in apostrophes that are relevant for inflectional change are converted back to separate word types and lemmatized (e.g. *I'm* → *I be*; *didn't* → *do not*).

Derivation
- prefix *en-* (e.g. *en-forced* → *forced*)
- prefix *dis-* (e.g. *dis-content* → *content*)
- prefix *out-* (e.g. *out-come* → *come, outrage* → *rage*)
- prefix *in-* (e.g. *in-alien-able* → *alien*)
- prefix *un-* (e.g. *un-known* → *known, un-animous* → *animous*)
- prefix *inter-* (e.g. *international* → *national*)
- prefix *non-* (e.g. *non-smoking* → *smoking*)
- Latin prefixes as in *re-port, trans-port* are not removed, only in the case where removal of a prefix leads to a word form that exists independently (e.g. *re-consider* → *consider, re-cover* → *cover*)
- suffix *-ly* (e.g. *mere-ly* → *mere, legal-ly* → *legal*)
- suffix *-hood* (e.g. *widowhood* → *widow, childhood* → *child*)
- suffix *-ize* (e.g. *real-ize* → *real*)
- suffix *-ation* (e.g. *conversation* → *converse, information* → *inform*, but not *salvation* since the Latin root is not used as independent word)
- suffix *-al* (e.g. *norm-al, sex-ual, environment-al* but not *annual, general* since the
Latin root is not used as independent word

- suffix -ity (e.g. community → commune, activity → active/act, but not dignity)
- suffix -fy (justify → just, but not verify)
- suffix -ful (e.g. peace-ful → peace, hope-ful-ly → hope)
- suffix -er (e.g. driv-er → drive, killer → kill)
- suffix -ment (e.g. develop-ment → develop)
- suffix -able (e.g. question-able → question)
- nominalizing -ing as in mewling → mewl, teaching → teach, and derived adjectives
  (e.g. living thing → live thing)
- other-wise → other

Note: derivation and inflection can overlap so that removing derivational suffixes
renders non-existent words (e.g. realized → reald). Derivation can take place without
change of surface forms (conversion, e.g. the telephone → to telephone).

Compounds (Longman 1999: 58)
- noun-noun combinations (e.g. daytime → day time, cellphone → cell phone, boyfriend
  → boy friend)
- proper names are not de-compounded (e.g. Hellfish)
- adjective-noun combinations (e.g. gentlemen → gentle men; note that multiannual is an
  exception, because multi serves as a productive prefix)
- preposition-noun combinations (e.g. downstairs → down stairs, anyway → any way;
  note that outcome is an exception, because out- is counted as derivational suffix)
- someone → some one, without → with out
- therefore → there fore
- myself → self, ourselves → our selfs

Contraction and cliticization
- you've → you have , you're → you are
- I'll → I will
- pig’s → pig s and pigs’ → pigs
- won't → will not
isn't → is not

Inflection

- inflected verb forms marked for person and tense, as well as gerunds and participles (e.g. bin, ist, war → sein, gehend → gehen, geschmuggelt → schmuggeln)
- noun morphology, such as plural and case marking (including Umlaut patterns) (e.g. Freiheiten → Freiheit, Gespenstern → Gespenster, Stürme → Sturm)
- the so-called Fugen-S as in Glaubensfreiheit is not removed since it is not considered productive (Duden 2009: 712-713)
- pronouns marked for case: (e.g. mir, dir, ihr, ihm → ich, du, er, sie)
- adjectives marked for case, number and gender (e.g. verdammte → verdammt)
- articles marked for case and number (e.g. dem → der, den, des → der/das, der.PL → die)
- demonstratives marked for gender (e.g. dieser/diese/dieses → dies)
- adjectival gerunds and participles with gender, number and case marking (e.g. kräuterstinkend-er → kräuterstinkend, folgend-er → folgend, gefuerchtete → gefuerchtet)
- adjectives marked for comparison (größer → groß, besser → gut)
- combinations of prepositions with verbs that involve medial zu are considered as the outcome of derivation rather than inflection (e.g. zurückkehren)
- Others: jeder/jede/jedes/jedem → jede, jener/jenem/jene → jene, andere/anderes/anderen → ander, keine/keiner/keines → kein, aller/alles/allem → alle

Derivation/Conversion (Duden 2009: 666pp.)

There are five different kinds of word formation in German that are considered here:

1) derivation: one independent and one (or more) dependent parts (e.g. Un-glück, ver-gehen, Ge-bild-e)

2) change of word class, can include morphological changes as well (e.g. fremd → Fremder, hart → härten)

3) abbreviations (e.g. information → info, not attested in our corpus)

4) particle verbs (ab-aendern, ver-aendern, ein-sehen, auf-stehen)
5) compounds: two independent parts (Vor-bild, Auf-wind, gelb-gruen):

Category 3) is irrelevant here since abbreviations of this kind are not attested in our texts. Category 5) is considered under compounds. This leaves us with categories 1), 2) and 4). For these categories we proceed as follows:

1) remove derivational prefix or suffix if this yields a word in German (e.g. ver-gehen → gehen; leave affixes if removing them yields non-words, i.e. erstatten, abstatten → statten*). Also, inflectional morphology is left if possible (e.g. vergeht → geht)

More examples:
- ver-missen → missen, but not verletzen, vergessen since letzen* and gessen* are non-words
- ent-lang → lang
- er-klären → klären, er-lösen → lösen, er-kranken → kranken
- unter-brechen → brechen, but zurueck-bleiben not, since considered to be compound
- See-un-geheuer → Seegeheuer
- pein-litch → pein, wahrscheinlich → schein, unheimlich → heim, but not möglich, since mög* is not a word

2) reduce to original word form (e.g Ergebnis → geben, erhitzt → Hitze, offensichtlich → offensichtlich, namens → namen)

4) remove particle as well as zu and ge (i.e. abzuaendern → aendern) if this yields an independent word; leave morphology if possible (e.g. abgeändert → ändert, beantworteten → antworteten, but not erschienenen → schienenen*)

More examples:
- zurueck-zu-schicken → schicken

Often, more then one of the above categories can be relevant to the removal of derivational morphology. For example Anerkennung → kennen involves 1), 2) and 4); Errungenschaften → ringen involves 1) and 2).

Composition
- noun-noun compounds (e.g. Fischfabrik → Fisch fabrik)
- preposition-noun compounds (e.g. zurückgehen → zurück gehen)
- noun-adjective compounds (e.g. kräuterstinkender → kräuter stinkender)
- Not proper names like Hellfish
- removal of the Fugen-s (e.g. Geschwindigkeitsmessung → Geschwindigkeit messung)
- multiple elements (e.g. Arbeitsschutzregelungen → Arbeit schutz regelungen)

Contractions and clitics
- neuter pronoun es contracted to ’s (e.g. geht's → geht es, rührt's → rührt es, dir's → dir es)
- omissions and amalgamations ”Verschmelzungen” (e.g. zum → zu dem, zur → zu der, im → in dem, am → an dem, beim → bei dem, vom → von dem, ins → in das, ans → an das, aufs → auf das) (Duden 2009: 616pp.)
<table>
<thead>
<tr>
<th>Corpus</th>
<th>Part</th>
<th>No. tokens English</th>
<th>No. tokens German</th>
<th>Register</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSC*</td>
<td>500 lines</td>
<td>3000</td>
<td>2608</td>
<td>Spoken (subtitles of movies)</td>
</tr>
<tr>
<td>EPC*</td>
<td>100 lines</td>
<td>2333</td>
<td>2134</td>
<td>Speeches (European Parliament)</td>
</tr>
<tr>
<td>UDHR*</td>
<td>30 articles</td>
<td>1753</td>
<td>1644</td>
<td>Written (legal)</td>
</tr>
<tr>
<td>BOG*</td>
<td>Chapters 1:3</td>
<td>2125</td>
<td>1918</td>
<td>Written (religious)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>9211</strong></td>
<td><strong>8304</strong></td>
<td></td>
</tr>
</tbody>
</table>

* OSC: Open Subtitles Corpus; EPC: Europarl Corpus; UDHR: Universal Declaration of Human Rights; BOG: Book of Genesis
Table 2. Information on languages, ISO codes, the tagger used, number of tokens per parallel corpus, number of unknown tokens, and the percentage of unknown tokens for Analysis 2.

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

* Numbers and percentages of word tokens that are unknown to the tagger.
Table 3. *NFD_{lem} values for corpora and languages in the cross-linguistic analysis of 19 languages (Analysis 2) and the current analysis (condition 3).*

<table>
<thead>
<tr>
<th>ISO</th>
<th>Language</th>
<th>Analysis 2</th>
<th>Analysis 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>eng</td>
<td>English</td>
<td>0.052</td>
<td>0.057</td>
</tr>
<tr>
<td>nld</td>
<td>Dutch</td>
<td>0.064</td>
<td>0.045</td>
</tr>
<tr>
<td>fra</td>
<td>French</td>
<td>0.089</td>
<td>0.084</td>
</tr>
<tr>
<td>ita</td>
<td>Italian</td>
<td>0.104</td>
<td>0.091</td>
</tr>
<tr>
<td>spa</td>
<td>Spanish</td>
<td>0.122</td>
<td>0.111</td>
</tr>
<tr>
<td>deu</td>
<td>German</td>
<td>0.123</td>
<td>0.108</td>
</tr>
<tr>
<td>pol</td>
<td>Polish</td>
<td>0.129</td>
<td>0.158</td>
</tr>
<tr>
<td>slk</td>
<td>Slovak</td>
<td>0.145</td>
<td>0.162</td>
</tr>
<tr>
<td>est</td>
<td>Estonian</td>
<td>0.158</td>
<td>0.156</td>
</tr>
<tr>
<td>fin</td>
<td>Finnish</td>
<td>0.197</td>
<td>0.234</td>
</tr>
</tbody>
</table>
Table 4. First 10 rows of original (left) and lemmatized (right) frequency distributions for English verbs.

<table>
<thead>
<tr>
<th>Word</th>
<th>Freq</th>
<th>Rank</th>
<th>POS*</th>
<th>Lemma</th>
<th>Freq</th>
<th>Rank</th>
<th>POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>is</td>
<td>126</td>
<td>1</td>
<td>Vm</td>
<td>be</td>
<td>308</td>
<td>1</td>
<td>Vm</td>
</tr>
<tr>
<td>said</td>
<td>119</td>
<td>2</td>
<td>Vm</td>
<td>say</td>
<td>285</td>
<td>2</td>
<td>Vm</td>
</tr>
<tr>
<td>come</td>
<td>62</td>
<td>3</td>
<td>Vm</td>
<td>come</td>
<td>127</td>
<td>3</td>
<td>Vm</td>
</tr>
<tr>
<td>say</td>
<td>58</td>
<td>4</td>
<td>Vm</td>
<td>go</td>
<td>112</td>
<td>4</td>
<td>Vm</td>
</tr>
<tr>
<td>says</td>
<td>58</td>
<td>5</td>
<td>Vm</td>
<td>take</td>
<td>70</td>
<td>5</td>
<td>Vm</td>
</tr>
<tr>
<td>was</td>
<td>54</td>
<td>6</td>
<td>Vm</td>
<td>see</td>
<td>68</td>
<td>6</td>
<td>Vm</td>
</tr>
<tr>
<td>saying</td>
<td>47</td>
<td>7</td>
<td>Vm</td>
<td>have</td>
<td>67</td>
<td>7</td>
<td>Vm</td>
</tr>
<tr>
<td>be</td>
<td>46</td>
<td>8</td>
<td>Vm</td>
<td>give</td>
<td>44</td>
<td>8</td>
<td>Vm</td>
</tr>
<tr>
<td>went</td>
<td>44</td>
<td>9</td>
<td>Vm</td>
<td>hear</td>
<td>44</td>
<td>9</td>
<td>Vm</td>
</tr>
<tr>
<td>came</td>
<td>35</td>
<td>10</td>
<td>Vm</td>
<td>answer</td>
<td>43</td>
<td>10</td>
<td>Vm</td>
</tr>
</tbody>
</table>

*Vm: main verbs
Table 5. $NFD_{lem}$ values per POS in English and Estonian.

<table>
<thead>
<tr>
<th>Language</th>
<th>Nouns</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>0.046</td>
<td>0.185</td>
</tr>
<tr>
<td>Estonian</td>
<td>0.159</td>
<td>0.304</td>
</tr>
</tbody>
</table>
Table 6. Differences in lexical diversities between uniform and non-uniform distributions.

<table>
<thead>
<tr>
<th>Measure</th>
<th>non-uniform</th>
<th>uniform</th>
<th>ΔLD</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZM $\alpha$</td>
<td>8.67</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>ZM $\beta$</td>
<td>12.45</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>HD-D</td>
<td>7.04</td>
<td>9.97</td>
<td>2.93</td>
<td>parametric</td>
</tr>
<tr>
<td>Shannon H</td>
<td>2.27</td>
<td>3.32</td>
<td>1.05</td>
<td>non-parametric</td>
</tr>
<tr>
<td>Yule's K</td>
<td>2680</td>
<td>900</td>
<td>1780</td>
<td></td>
</tr>
<tr>
<td>TTR</td>
<td>0.10</td>
<td>0.10</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>MSTTR</td>
<td>0.17</td>
<td>0.10</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>MATTR</td>
<td>0.16</td>
<td>0.19</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Herdan's C</td>
<td>0.50</td>
<td>0.50</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Guiraud's R</td>
<td>1.00</td>
<td>1.00</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>CTTR</td>
<td>0.71</td>
<td>0.71</td>
<td>0</td>
<td>(TTR-based)</td>
</tr>
<tr>
<td>Dugast's U</td>
<td>4.00</td>
<td>4.00</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Summer's S 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Maas index</td>
<td>0.50</td>
<td>0.50</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>MTLD</td>
<td>2.20</td>
<td>2.04</td>
<td>0.16</td>
<td></td>
</tr>
</tbody>
</table>

* H: Shannon entropy over word types; ZM: Zipf-Mandelbrot parameters $\alpha$ and $\beta$; TTR: type-token-ratio; MSTTR: Mean Segmental Type-Token Ratio; MATTR: Moving-Average Type-Token Ratio; CTTR: Carroll's Corrected TTR; Dugast's U: Dugast's Uber Index; HD-D: idealized version of vocd-D; MTLD: Measure of Textual Lexical Diversity
Figure captions

Figure 1. An example of visually comparing a uniform (grey) to a non-uniform (black) frequency distribution. The left panel illustrates the frequencies of the two distributions ranked from highest to lowest. The frequency differences are the differences in height of the black and grey bars. These differences are projected onto the upper panel. The right panel illustrates the log frequencies and log ranks for the uniform (grey triangles) and the non-uniform distribution (black dots). Note that the frequency differences are not logged in the upper panel but the ranks are logged in order to align them with the plot in the lower panel.

Figure 2. Frequency differences in English illustrated for the removal of inflectional marking (left panel) and the removal of derivational marking (right panel). Original distributions are represented by grey triangles, changed distributions by black dots. Frequency differences per rank (non log-transformed) and NFD values are given in the upper panels.

Figure 3. Frequency differences in English illustrated for the removal of inflectional marking (left panel) and the removal of derivational marking (right panel). Original distributions are represented by grey triangles, changed distributions by black dots. Frequency differences per rank (non log-transformed) and NFD values are given in the upper panels.

Figure 4. Frequency differences in German illustrated for the removal of inflectional marking (left panel) and the removal of derivational marking (right panel). Original distributions are represented by grey triangles, changed distributions by black dots. Frequency differences per rank (non log-transformed) and NFD values are given in the upper panels.

Figure 5. Frequency differences in German illustrated for the splitting of compounds (left panel) and the splitting of clitics and contractions (right panel). Original distributions are represented by grey triangles, changed distributions by black dots. Frequency differences per rank (non log-transformed) and NFD values are given in the upper panels.

Figure 6. Changes of frequency distributions between unlemmatized (grey triangles) and lemmatized (black dots) texts. English, Spanish and Finnish are chosen to represent the range of the original 19 languages.

Figure 7. NFD$_{lem}$ as an inflection index across 19 languages. The x-axis represents languages with respective ISO 639-3 codes. The y-axis represents the NFD$_{lem}$ between unlemmatized and lemmatized versions of the UDHR and PBC parallel corpora. The colours of bars indicate whether the texts were lemmatized using the BTagger (red) or TreeTagger (blue). Note that for three languages (English, Polish, Estonian) both options are available.

Figure 8. Relationship between number of tokens (x-axis) and NFD$_{lem}$ (y-axis) for 10 languages and three conditions of sampling. The coloured curves are NFD$_{lem}$ values smoothed with a general additive model (gam). ISO 639-3 codes translate as: fin (Finnish), pol (Polish), slk (Slovak), est (Estonian), deu (German), spa (Spanish), ita (Italian), fra (French), eng (English), nld (Dutch). We used 1M word tokens in the original analyses, but we reduced the size to 100K, since the values already converge at around 50K at the most. The vertical dashed lines represent 15K tokens, which corresponds roughly to the average text size we had in Analysis 2.
Figure 9. Distributions of lemmatized (black dots) and unlemmatized (grey triangles) word types by parts of speech (nouns, prepositions and verbs) in English.

Figure 10. Distributions of lemmatized (black dots) and unlemmatized (grey triangles) word types by parts of speech (nouns, prepositions and verbs) in Estonian.

Figure 11. Correlation between NFD_{lem} values (y-axis) and entropy difference (x-axis) for data from Analysis 2. Dots represent languages, different colours indicate lemmatization by either BTagger (red) or TreeTagger (blue). The dashed lines are linear models with confidence intervals.