Assessing Word Commoness

Adding Dispersion to Frequency

Abstract

The article investigates the two main corpus indicators of word commonness, frequency and dispersion, through a cross-validation analysis of frequency and four dispersion measures (‘Range’, ‘Chi-squared’, ‘Deviation of Proportions’ and ‘Juilland’s D’). The approach provides an estimation of the capacity of the named measures to predict the distribution of corpus items in an extracted language sample. Based on a dataset of 273 Norwegian compounds, the results show that especially Deviation of Proportions is a robust measure of dispersion that can be used in conjunction with frequency to substantiate assertions of word commonness based on corpus data. In addition, dispersion measures do not only reflect what sort of distribution the frequency statistic is generated from, but also how reliable the frequency estimation in the corpus sample is in terms of giving an accurate representation of frequency in the language variety that the corpus is sampled from.

dispersion, frequency, word commonness, cross-validation, lexicography

1. Introduction

Corpus linguistics has become an increasingly popular field over the recent decades. This is particularly the case within the field of lexicography, where data collection traditionally was a time- and space-consuming task with questionable reliability. There are still a number of reasons to think carefully about the use of data in lexicography, but there is no question that corpora offer lexicographers invaluable access to the language variety they seek to describe.

The approach of the present study builds on the premise that corpora reflect corpus-external language use, and consequently that corpus distributions predict language distributions. This is the way corpora are often interpreted in lexicography, where they are, for instance, invoked to make arguments about word commonness. Alongside having sophisticated corpus resources at their disposal, lexicographers can benefit from using sophisticated corpus methods. In this article, I will investigate Gries’ (2008) claim that proficient corpus methods should refrain from using global frequency scores as the sole indicator of commonness. An additional indicator of this is the degree to which the occurrences of a corpus item are evenly distributed throughout the corpus, namely the ‘dispersion’ of the item (see for example Lyne, 1985; Savický & Hlavácová, 2002; Gries, 2008).

In this paper I will apply the term ‘number of occurrences’ (hereafter NO) to refer to the simple count of occurrences in a corpus or a corpus part. The term ‘frequency’ will be reserved for NO/corpus size (also commonly referred to as ‘normalised frequency’). When I state that a corpus item is frequent, it therefore means that it has a high NO relative to corpus size. The term ‘dispersion’ refers to the degree to which a corpus item is proportionally spread out over corpus parts according to their size and the global NO of the corpus item. This is not reflected in the frequency statistic. A perfectly dispersed corpus item has a distribution that mirrors the relative size of the corpus parts. Therefore, a frequent item is not necessarily evenly dispersed, whereas an infrequent one very well can be.

To scrutinise the respective roles of frequency and dispersion in estimating word commonness, I will evaluate the performance of global frequency and the dispersion measures ‘Deviation of Proportions’, ‘Range’, ‘Chi-squared’ and ‘Juilland’s D’ on 273 Norwegian compounds in a Norwegian corpora. Moreover, I will seek to validate the results of these measures by performing a cross-validation. This validation is supplemented with a correlation analysis that encompasses global estimates of frequency and dispersion measures, and the corresponding predictive accuracy of the same measures generated by the cross-validation procedure.

The results of the analyses clearly show the advantages of considering dispersion in addition to frequency when assessing the commonness of corpus items. Moreover, they indicate that dispersion might also play a role in reflecting the reliability of corpus data, in that clumped distributions are more likely than evenly dispersed distributions to be generated by chance by the corpus sample.

This article is a step towards creating both operationalisable and scientifically sound methods for lemma selection in lexicography. Its findings also corroborate the assertion that dispersion needs to be considered as an equally important statistic as frequency, and that frequency scores, when used as a token of commonness, should generally be supported by at least one dispersion estimate. This should ideally be a corpus linguistic convention analogous to the way standard deviation is reported in conjunction with the sample mean.

2. Dictionaries and Distributions

In this chapter, I discuss two main approaches to investigating frequency of use in corpora from a lexicographic point of view. More specifically, I will elaborate on the case of compounds in Norwegian dictionaries, and present five measures that may help to identify which compounds belong to the core vocabulary that a general dictionary seeks to describe.

2.1 Dictionaries and Core Vocabulary

Dictionaries differ in scope and format along with various other parameters (Durkin, 2016). Particularly the question of scope, i.e. what parts of a given vocabulary a dictionary seeks to describe, matters to the quantitative methodology that the lexicographers of a given dictionary project adhere to. In this article I discuss corpus methodology in relation to lexicography that seeks to describe the core vocabulary of a language, that is, the *n* most commonly used lemmas of that language (not the same as the *n* most frequent lemmas). This pertains to mono- or multilingual dictionaries that aim and claim to include lemmas (including affixes, abbreviations, idioms, symbols, etc.) that comprise a representative selection of the synchronic vernacular.

The wordlist of a general dictionary is not a representation of any one language user's mental lexicon. A more precise account would be to say that a dictionary wordlist overlaps with a language user's lexicon. There will be words such as dialectal or sociolectal variants that an ordinary speaker knows, which escape the realms of a general dictionary. Likewise, there is a large number of words that an ordinary speaker, even a speaker with a large vocabulary, does not know, that are in the wordlist. When speaking of a core vocabulary of a language or a representative selection of the synchronic vernacular, one must keep in mind that the reference is a wordlist that intersects and exceeds the individual vocabularies of ordinary speakers of that language.

2.2 The Case of Compounds in Norwegian

Fellbaum (2015) mentions noun compounds as one of the types of multi-word units that can be semantically opaque and therefore clear candidates for inclusion in a dictionary. The treatment of compounds do however vary between dictionaries as they may operate with different thresholds of commonness, frequency of use or what it means to be "semantically opaque". In Norwegian, there are both prosodic and morphological differences that separate compounds from multi-word units. Norwegian compounds are therefore written as single words, and consequently treated as independent lemmas. There are however still varying degrees of lexicality and semantic transparency among Norwegian compounds, as it is both a highly productive and frequent formation type. Bakken (1998) argues that there is a continuum of lexicalisation in compounds, ranging from transparent novel compounds to opaque compounds where one needs a knowledge of the word's etymology in order to classify it as a (historic) compound. A non-lexicalised compound (which I also discuss in chapter 2.3) is *blåtrøye* (lit. "blueshirt"). Although most language speakers of Norwegian would be able to produce some sort of mental representation when encountering this compound, this representation is not necessarily aligned with neither the concept nor the reference that is intended to be evoked when the compound is used. A more lexicalised compound such as *blåbær* ("blueberry") has a more stable and automatic interpretation.

Since compounds may be placed on a continuum with respect to their degree of lexicality, they may also be placed on a continuum with respect to their *compoundhood.* Some words may have originated as compounds, but developed into root words, or they may at least, as already stated, be perceived as root words for someone without knowledge about their etymology. For this study, I will define a compound as a word form consisting of two or more constituents that correspond to individual root words. This working definition will encompass both highly lexicalised and novel compounds. The definition is purposely wide in order to include a wide range of different compound distributions into the study. As further discussed in Section 2.4.6 below, the results of the current study are applicable to all word forms. Since the input of the analysis is distributions, the lexical properties of the word forms behind those distributions are of secondary importance.

Compounds pose a challenge to Norwegian lexicography because of their productivity and frequency. With a medium-sized corpus (*Leksikografisk bokmålskorpus*) of approx. 115 million words, a semi-frequent word such as *maskin* "machine" is a constituent of about 2000 compound types, with a combined number of occurrences of about 12 000.[1] It would be a reckless use of resources to create fully fledged dictionary entries for all of these compound types. Besides, many of these compounds do not have a common or conventional interpretation, so it would be rather pointless to formulate definitions for them. One therefore needs to make a selection, and assessing which compounds make good dictionary candidates can be time-consuming and tedious if one lacks clear-cut quantitative criteria for the selection process.

A central question is therefore which quantitative criteria to use. While the specific threshold will vary according to a project's corpus resources, finances, purpose and so on, the statistic measures to investigate word distribution should be scientifically sound for all such projects, thus not varying a lot between them. In the following sections, I will evaluate different distributional measures and discuss how they might contribute to the lexicographic selection process of Norwegian compounds.

2.3 Disadvantages of Frequency, Advantages of Dispersion

More than a decade ago, Gries (2008) pointed out that frequency of occurrence was still the most frequently used statistic in corpus linguistics. I cannot speak for corpus linguistics in general, but in the Norwegian lexicographic context, this is still the case. One strong indication of this can be found in the two main corpus infrastructures, *Corpuscle* and *Glossa*, which contain the most important Norwegian corpora. None of these infrastructures offer calculation of other distributional statistics than number of occurrences and frequency (number of occurrences/corpus size).[2] One can easily obtain measures of word frequencies relative to various metavariables such as genre or year of production, which gives the user ample opportunity to inspect a word's dispersion by studying where its occurrences are, but dispersion scores based on this information are not available.

Gries (2008) points out that frequencies can often be unreliable and misleading – especially if they are taken to indicate word importance or commonness, or to reflect degree of mental entrenchment. Frequency scores may for instance be highly influenced by the genre of corpus content. In the Norwegian Newspaper Corpus, the compound word form *blåtrøyene* "the blue shirts" is the third most frequent compound form with *blå* "blue" as its first constituent. *Blå* is productive in compounds, both as first and second constituent, and many of these compounds are also lexicalised and well-established in the Norwegian vocabulary (like *blåbær*, see Section 2.2). It is therefore peculiar that a non-lexicalised compound in the definite plural should hold such a prominent position in the corpus. The mystery is however solved by a short inspection of the concordance, which shows that *blåtrøyene* is used in sports news to refer to sports teams with blue jerseys. Since blue isquite a popular jersey colour and most newspapers have a sports section, *blåtrøyene* soars on the frequency list. This is what one could call a classic example of an arbitrary effect,where frequency is not an indication of word commonness nor of mental entrenchment.

In the previous paragraph I speak of word 'commonness' and word 'importance'. A word can be important for a specific purpose, e.g. lexicographic description, but it can never be important in itself. Thus, word importance is a contextual phenomenon. A word can furthermore be common in language use, but only in comparison to words that are less common. Word commonness is thus a relative phenomenon. In lexicography, commonness might be an indicator of importance, but not the other way around. In the current study, I will view a common compound as a compound that is more widely used and therefore also more conventional than other compounds. This includes being both frequent (repeatedly occurring in a given text sample) and dispersed (regularly occurring in a given text sample, i.e. across different texts of various types).

The current study has an almost identical outset as Savický & Hlavácová (2002) who acknowledged the weaknesses of frequency scores in reflecting word commonness for lexicographical purposes. They developed three measures of 'corrected frequencies' which in different ways adjust corpus item's frequency depending on their dispersion. Perfectly dispersed items will keep their estimated frequency, while unevenly distributed items are "punished" by having their frequency scores somewhat reduced (which simultaneously reduces the perceived commonness of such items). Different from the parts-based measures that I will investigate in this study, Savický & Hlavácová's (2002) corrected frequencies are distance-based measures. Gries (2008) argues that distance-based measures have a number of weaknesses, e.g. that they "treat a corpus as one homogenous string of words devoid of any structure (in the form of turns, file parts, files, genre/register parts, etc.)" (ibid: 414), with the effect that the order of corpus parts affects the estimates of Savický and Hlavácová's distance-based measures. However, in a different study, Gries (2010) found that two of these distance-based measures (named ALD and AWT) had a strong positive correlation with response time latencies from psycholinguistic experiments[3], suggesting that they are a good indication of word commonness as it is reflected in cognitive entrenchment in individual speakers.[4]

To cope with the weaknesses of frequency scores, Gries (2008) surveys a total number of 17 alternative frequency and dispersion measures. Two of these are his own invention, while the rest are drawn from other sources. Gries (forthcoming) reviews two studies (Biber et al., 2016; Gries, 2010) that evaluate the strengths and weaknesses of the measures on his list. Of particular interest is Gries (2010), since it performs a cluster analysis and a principal component analysis on frequency and dispersion measures. Both of these analyses yielded five distinct groups of measures. In the present study I have selected one measure from each of these groups for validation and evaluation based on a material of Norwegian compounds. One measure was selected because it is commonly used (such as frequency) and one because it is easy to interpret (such as range). Two measures were selected because they have yielded promising results in previous studies

(DP and Juilland's D, see Gries (2008) and Lyne (1985), respectively), and the last one (chi-squared) was selected because it does not cluster with any other measure in Gries' analyses. I outline these measures briefly in the following.

2.4 Comparing Measures

In this chapter I will present five different statistics, four of which are measures of dispersion. A summary of the measures can be found in Table 1.

	Pessimum	Optimum
Frequency	0	1
Deviation of Proportions (DP)	1	0
Relative Range	1/*n*	1
Chi squared	∞	0
Juilland's D for unequal corpus parts (D_uneq)	1	0

Table 1: Summary of measures

2.4.1 *Frequency*

It is quite understandable why frequency is a widely used statistic in corpus linguistics. It is easy to calculate and seemingly easy to interpret. However, it is not obvious what sort of conclusions frequency scores warrant since they are susceptible to random effects, such as over-representation in specific genres or individual texts. This tendency is however not stable across corpora. If an item has a frequency of 0.01 in a corpus of 1 million words, the likelihood of the frequency being inflated or deflated by text type(s) in that particular corpus is greater than if the same item has the same frequency in a corpus which is ten times bigger.

It may seem intuitive that a corpus item *w* which is more frequent than corpus item *y* must also be a more common language item as long as the corpus gives an adequate representation of this language. However, the frequency scores do not reflect what sort of dispersion corpus items have across time, genre and texts, which may be equally important to consider in the assessment of commonness.

2.4.2 *Relative Range*

Range is a simple count of the number of corpus parts that an item occurs in. One may operate with different thresholds of occurrence, but the default is to count every corpus part where an item occurs once or more. In order for range scores to be comparable across corpora, one needs to use ‘relative range’, which is the product of range divided by number of corpus parts. Neither frequency nor differences in the sizes of corpus parts are taken into consideration by range (relative or not). Widely different distributions may therefore result in the same range score. However, if one or more of the corpus parts are very small, this will reduce the likelihood of reaching optimum range. If NO is less than the number of corpus parts, this will obviously eliminate the possibility of reaching optimum range. A low range and a high frequency indicates a highly skewed distribution, whereas a high range will most likely indicate a quite dispersed distribution. The exception is if the high range is a product of over-representation in small corpus parts and under-representation in large corpus parts. The most common items of the language would be expected to have optimum relative range, which is 1. The pessimum is 1 divided by the number of corpus parts, which makes it inversely proportional to the number of corpus parts.

2.4.3 *Chi squared χ^2.*

Chi squared χ^2 (hereafter chisq*)* is also a measure of dispersion. It has the following formula:

$$\sum_{i=1}^{n} \frac{(observed\ v_i - expected\ v_i)^2}{expected\ v_i}$$

where *observed v_i* = NO in corpus parts and *expected v_i* = (*global NO* X *relative size of corpus parts*) and *n* = number of corpus parts.

If one has five equally sized corpus parts *i*, and a perfectly dispersed item *w* that occurs 1000 times in the corpus as a whole, then the chisq score equals 0, which is the optimum. If, however, *w* has a slightly uneven distribution across i_{1-5}, (e.g. 200, 200, 400, 100, 100), the chisq score equals 300. If the distribution were 2, 2, 4, 1, 1, then the chisq score would be 3. The scale of chisq is open at one end since the pessimum is infinitely high.

As seen in the example above, the chisq scores 300 and 3 represent the same dispersion, but fall on different scales according to the magnitude of the distribution they are calculated from. For the purpose of this paper, chisq will be controlled for this difference in magnitude by using the formula chisq/NO.

On average, chisq yields higher values for corpus data consisting of many parts than data with just a few parts. The scale remains the same, but an item with medium dispersion would yield a higher chisq score the more corpus parts there are (provided that the size differences between the corpus parts are more or less equal between the two sets of corpus data).

2.4.4 *DeviationofProportions (DP)*

Deviation of proportions (hereafter DP) is a dispersion measure suggested by Gries (2008) that is calculated from the deviation between the proportion of occurrences of a corpus item in corpus parts and the relative size of those parts. The calculation is made as follows:

$$0.5 \times \sum_{i=1}^{n} \left| \frac{observed\ v_i}{o} - s_i \right|$$

where v_i = NO in corpus parts, o = global NO and s_i = relative sizes of corpus parts.[5]

DP is a measurement of the dispersion of corpus items across corpus parts, and yields values on the scale of 0 to 1, whereby the former is the optimum and the latter is the pessimum. As an example, if one has five equally sized corpus parts *i* and a perfectly dispersed item *w* that occurs twice in each of those parts, then the DP score equals 0. The most frequent items of the language, e.g. function words, would be expected to have a DP value close to 0. Since items like these tend to occur in most texts, there is a strong positive correlation between text size and the NO of such items. This correlation is reflected in a low DP score. The optimum is however only achieved when the distribution of *w* is perfectly even in relation to the relative sizes of the corpus parts. The pessimum, on the other hand, is in practice never 1, but 1 — i_{min}, the relative size of the smallest *i*. The theoretically most uneven distribution possible is if 100% of the occurrences of *w* are attested in i_{min}. In that case, the size of i_{min} will determine what proportion of *w* was already expected in i_{min}, which in turn will decide the DP score's deviation from 1.

DP is similar to chisq in that it is calculated from the deviation between observed and expected occurrences in corpus parts. Chisq is however calculated from absolute numbers, whereas DP is calculated from proportions. DP is therefore less affected by the magnitudes of the NO of a corpus item than chisq.

Moreover, DP is equally sensitive to both positive and negative disproportions. This means that the over-representation of a corpus item in a corpus part has the same effect as its under-representation. This symmetry may be an advantage of DP when it is interpreted in conjunction with frequency, because it indicates whether the distribution of an item is skewed regardless of the direction of the skewness. Moreover, DP benefits from corpora with many parts in that it becomes more stable and less sensitive to outliers.

2.4.5 *Juilland's D (D_uneq)*

Juilland's D for unequal corpus parts (hereafter D_uneq) is calculated with the following formula (taken from Gries, forthcoming: 4), but reversed for the purpose of this study:

$$\frac{sd_{population}\left(\frac{v_i}{s_i}\right)}{mean\left(\frac{v_i}{s_i}\right)} \times \frac{1}{\sqrt{(n-1)}}$$

where v_i = NO in corpus parts, s_i = relative sizes of corpus parts and n = number of corpus parts.

This formula produces values between 0 and 1, where the former is the optimum and the latter is the pessimum. Gries (forthcoming) points out that this is the historically most widely used dispersion measure. In theory, a perfectly proportional distribution would yield an optimum of 0, while a maximally disproportionate distribution would give a pessimum of 1. While the first assumption is true (the optimum is attained if and only if the distribution is perfectly proportional to the corpus sizes), the pessimum is also generated whenever a corpus item is attested in a single corpus part, regardless of the number and sizes of those parts. If the corpus is split in two, and one of those parts represents 95% of the corpus, the D_uneq score will still be the pessimum 1 for a corpus item that occurs in only one of those parts, regardless of which.

While D_uneq and DP both consider proportions of occurrences, they behave somewhat differently. While DP is sensitive to negative and positive disproportions only, D_uneq is also sensitive to relative NO variance of corpus items in different corpus parts. Since the D_uneq formula makes use of both standard deviation and sample mean, which are both sensitive to outliers, it is more sensitive to outliers than DP. This effect is however reduced by increasing the number of corpus parts. However, too many corpus parts may render D_uneq quite unreliable: Biber et al. (2016: 452) has found evidence that Juilland's D for equally sized corpus parts "completely fail to discriminate among words with uniform versus skewed distributions" when estimated on a set of 1000 corpus parts. They conclude that it is a property of the formula itself, that "any advantages associated with careful sampling of the corpus are offset by the flaws of the formula, with its inflation of estimates based on a large number of corpus parts" (ibid.: 450). Since I will use 29 unequally sized corpus parts in the current study, it is not obvious how eminent this effect will be here. Biber et al. (2016) shows that the same effect is vividly present with 100 corpus parts, but more or less absent with 10. 29 corpus parts might cause a slight inflation of the D_uneq estimates compared to a smaller number of corpus parts, but it is certainly not canceling out the ability of the formula to discriminate between uniform and skewed distributions, judging from the D_uneq scores obtained in this study. Since the goal of the present study is to evaluate the performance of D_uneq and other measures, and the named effect is expected to be small, I have not made any adjustments to the formula in order to account for the number of corpus parts here.

2.4.6 *Which one(s) to choose?*

There is of course no final answer to the question: "Which dispersion measure should one choose?" The answer would otherwise need to be "it depends on what you want to investigate". My goal is to evaluate the usefulness of these frequency and dispersion measures with respect to one particular task, namely the assessment of commonness of Norwegian compounds. The results of this evaluation have a direct application in lexicographical work, but they are also applicable to any sort of corpus item in any language. In corpus terms, a (Norwegian) compound is just a string that occurs a given number of times in a given number of contexts in the corpus. This string may take on various forms, depending on the inflectional and derivational properties of the compound. In this study, the frequency and dispersion measures will be evaluated based on their performance on different compound distributions. However, it does not really matter what sort of corpus items the distributions are drawn from. What matters is that the measures are tested on a multitude of distributions. The results from this may therefore be applicable to commonness assessments of any sort of corpus items, at least assessments where distributions are used as the main indicator of commonness.

2.5 What are distributions in corpora supposed to tell us?

In this study, I embark upon a discussion of how to best measure word commonness from corpus data. Gries (2008) mentions that frequency of occurrence in corpora is often used to attest degree of mental entrenchment in a given language user. I will not elaborate on the entrenchment, but rather claim that corpora also/instead tell us something different: corpus distributions mirror language distributions (to the extent that the corpus is representative of the language). The collection of material that goes into a given corpus is often designed to comprise materials that resemble what most language users are likely to have been exposed to.

As Stefanowitsch (2020) states, corpus data are necessarily incomplete (as any other sample), but we use them to infer something about what happens outside the data. We do not expect the corpus material to be identical with corpus-external text material, but we expect it to be representative of the given language variety on a distributional level. If something is frequent in the corpus, we expect it to occur often in the language variety as a whole. From a lexicographical point of view, this cross-section of language data is then invaluable, since it represents exactly what general lexicography seeks to describe. With this said, it is then only pertinent to investigate further how corpus data could be utilised and interpreted so that the conclusions we draw from it are valid with respect to the language variety that the corpus data represent.

3. Methodology

The current study is performed on a corpus called *Leksikografisk bokmålskorpus* (hereafter LBK, see Fjeld, Nøklestad & Hagen, 2020), a balanced corpus for Norwegian Bokmål. The LBK has 115 270 577 tokens (102 million of which are words), which makes it a medium-sized corpus (roughly the same size as the British National Corpus).

	Domain	Relative size	No. subdomains	Abbreviation
1	newspapers	5,7%	4	AV
2	non-fictional prose	48,1%	11	SA
3	fictional prose	34,8%	6	SK
4	subtitles from television	5,8%	3	TV
5	leaflets and other short texts	5,6%	5	UN

Table 2: Overview of domains in the LBK

The LBK is divided into five differently sized domains with different numbers of differently sized subdomains, see Table 2. It is typical for “balanced” corpora that they consist of different shares of texts from different genres. In the development of the LBK, the aim of the text selection was to mimic the text types that an average reader is exposed to. This was ensured by basing the selection on a survey of people’s reading habits (Fjeld, Nøklestad & Hagen, 2020). As Stefanowitsch (2020) points out, allowing for a high degree of diversity in the corpus material is perhaps the only way to ensure a certain degree of "balance" or representativity in a corpus. To the degree that the LBK is representative of the language it seeks to represent, the representativity is intended on the text type level, not on the level of the actual texts that are collected in the corpus.[6] Even though there are reasonable proportions of different text types in a corpus, whether or not a particular book about the seafarer Willem Barentsz is in there or not, is an arbitrary choice that is not informed by a statistic saying that this book is particularly popular or representative of its genre. One must therefore expect that the LBK, as most corpora, contains a fair share of arbitrary effects that are captured by the coincidental selection of the particular texts from different genres.

These domains with their corresponding subdomains make up the subcorpora that are used in the present analysis. The frequency and dispersion measures are calculated and tested on a sample of 273 Norwegian compounds with varying NOs, ranging from 1 to 17,018. (see Figure 1 and Table 3 for a short summary of the compounds and their distribution, and appendix A for a complete list).

Maximum	NO	Intermediate	NO	Minimum	NO
omfatte	17018	årsverk	507	svartkopp	2
framtid	16451	arbeidsdeling	503	svart(e)mann	2
omkring	14466	arbeidssøker	494	svartstill	2
bakgrunn	13822	omforme	490	svartsjuke	1
omsorg	10237	omdømme	485	svartsinn	1
omtale	9216	arbeidssituasjon	478	vandrestjerne	1

Table 3: Compound lemmas with maximum, intermediate and minimum number of occurrences

The collection of compounds consist of word forms that are constructed with at least two base words. As an example, in the compound *årstid* (lit. *yeartime* "season") both *år* "year" and *tid* "time" can be used as independent heads of noun phrases (see also Section 2.2 for a working definition of *compound*). The collection is a varied set of compound constructions, encompassing compounds constituted by verbs, adjectives, nouns, prepositions and adverbs, as well as configurations with a combination of these, e.g. adj + verb as in *svartmale* "denigrate", lit. "blackpaint".

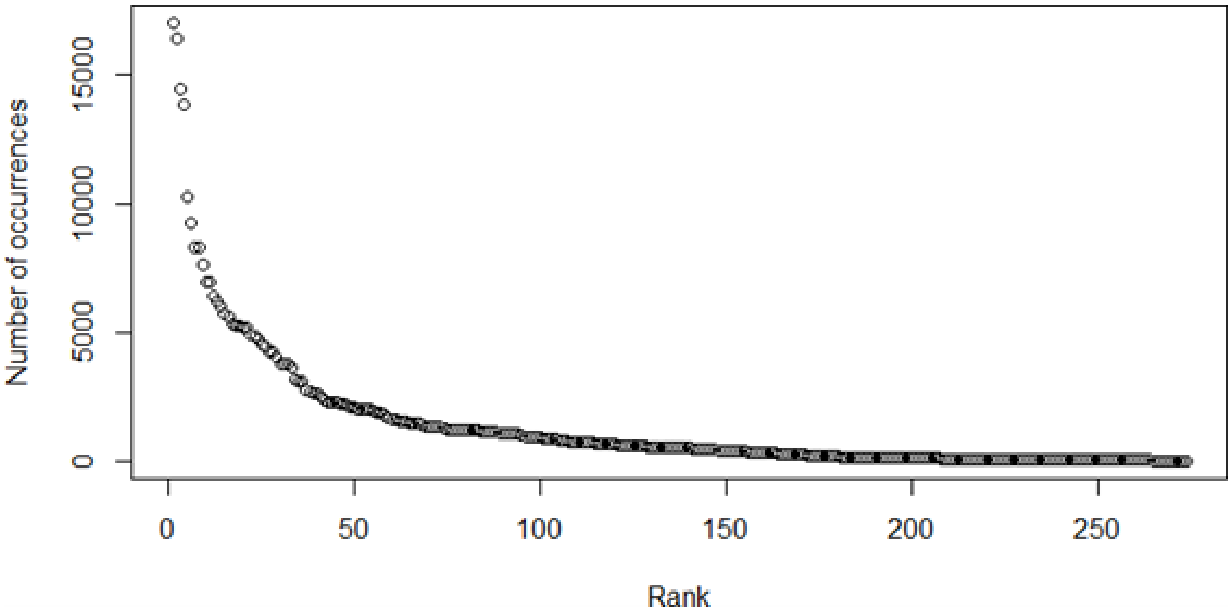

Figure 1. Distribution of compounds in the collection based on NO

The NOs encompass all inflectional forms that belong to the named compound lemmas, as well as any compound word forms with the named compound lemma as their first constituent.[7] Moreover, some words in Norwegian Bokmål have alternate forms, e.g. *framtid* in Table 3 has an alternate form *fremtid*, and *svartemann* has an alternate form *svartmann*. Such alternate forms are also included in NO.

3.1 Cross-Validation and Correlation Analysis

The current study assumes the predictive modeling method of cross-validation to evaluate the frequency and dispersion measures. To this end, I calculate the average deviation of the above-mentioned measures with the following procedure:

i. Calculate the value v_1 of a dispersion measure *m1*[8] for compound *w* in corpus domain AV(see table 2);

ii. Calculate the value $v_{2\text{-}5}$ of *m1*for the subdomains (SA01, SA02…) of the remaining corpus domains (here: SA, SV, TV and UN, see table 2);

iii. Calculate the difference *d1* between v_1 and $v_{2\text{-}5}$. This indicates the accuracy of *m1* based on 4 of the 5 domains of the corpus in predicting the remaining 1/5;

iv. Repeat steps i-iii for every corpus domain;

v. Extract the mean $\bar{x}$ of *d1-d5*;[9]

vi. Repeat steps i-iv for each measure *m*;

vii. Repeat steps i-vi for each compound in the collection.

In this procedure, the corpus is divided into 29 parts based on the subdomains. These 29 parts belong to five different domains, where four of them serve as a training set and the remaining genre functions as a test set. The test set represents an extrinsic text collection whose distributions are compared to the distribution in the training set. The ability of the training set to match the distribution of a corpus item in the test set provides an indication of the accuracy with which the frequency and dispersion measures are able to predict the distribution of the corpus item outside of the corpus. This is after all the main thing that corpus distributions are supposed to represent (see Section 2.5).

In addition to the procedure above, the same frequency and dispersion measures are also employed to calculate the global frequency and dispersion of the same items in the corpus as a whole, based on the 29 subdomains that constitute the five corpus domains AV, SA, SK, TV and UN (see table 2 for overview). This approach allows one to analyse how the predictive accuracies of the various measures are affected by different types of distributions.

With the above approach, a certain genre effect is to be expected. The test sets are based on domains corresponding to a particular genre, where the corresponding training sets consist of genres that are different from the one represented in the test set. For instance, distributions in fictional and non-fictional prose subtitles and leaflets are used to predict distributions in newspapers. This design would with all likelihood create a genre effect that lowers prediction estimates across the board. This effect is however stable for all domains and all measures in the study, meaning that any discernible differences between measures will not originate from this genre effect.

The way a corpus is subdivided will always affect dispersion estimates (see Egbert et al. 2020). Using domains as the basis for this subdivision ensures that the dispersion estimates reflect dispersion across text types. One might argue that this kind of dispersion is of particular interest in the pursuit of corpus items belonging to the core vocabulary of a language variety. The LBK can also be subdivided based on the metavariable "year", from which one may measure dispersion across time slots. A weakness of this subdivision could be that text types in the corpus are unevenly distributed across different time slots, so that difference in year in reality reflects difference in genre. One could also state that the same unevenness simultaneously causes problems for the domain-based subdivision, in that difference in "domain" might be a reflection of difference in "year". Against this I would argue that the languages and vocabularies of different genres differ more than the languages and vocabularies of different years, at least when we are talking about year slots inside an interval of 28 years (as is the case with the LBK).

It is also possible to subdivide the corpus by splitting it into *n* random equally-sized parts. This could potentially minimise effects that arise from a structured subdivision. This approach is however impractical since it requires much more preprocessing of the data than basing the subdivision on existing metavariables. More importantly, it obscures the dispersion estimates by masking what sort of dispersion is being estimated. Instead of treating the corpus as a structured body of different texts, which it indeed is, the random subdivision treats the corpus as a kind of bag of words which is devoid of any structure (see also the same argument in Gries' (2008) critique of distance-based measures). By basing both the dispersion estimation and cross-validation on the metavariable "domain", one can expect a certain genre-effect which is undesirable in the cross validation since it affects the predictive accuracy, but desirable in the global dispersion estimation. The advantage in both instances is however that we know what this effect is. A random subdivision would with all certainty also render a certain effect stemming from the qualitative difference among the corpus parts, but this effect would most likely be much less predictable and stable than a universal genre effect.

4. Results and analysis

Figure 2 visualises the relationship between the variables in the study, both the global frequency and dispersion measures and the results of the cross-validation, which is reflected in the -x- values of each of these measures. The relationship is represented by both a plot and Spearman's rank correlation coefficient. In addition, the density curves (shown from the upper left to the bottom right corner) illustrate the distribution of data points for each variable on its own scale. Both the global measure and the predictive accuracy of frequency and chisq have been transformed using logarithm. This does not alter the correlation coefficient (as it is based on Spearman's rank correlation), but it smoothens the plot and the curve, which otherwise would be skewed to the left. The correlations for each variable are discussed in detail below.

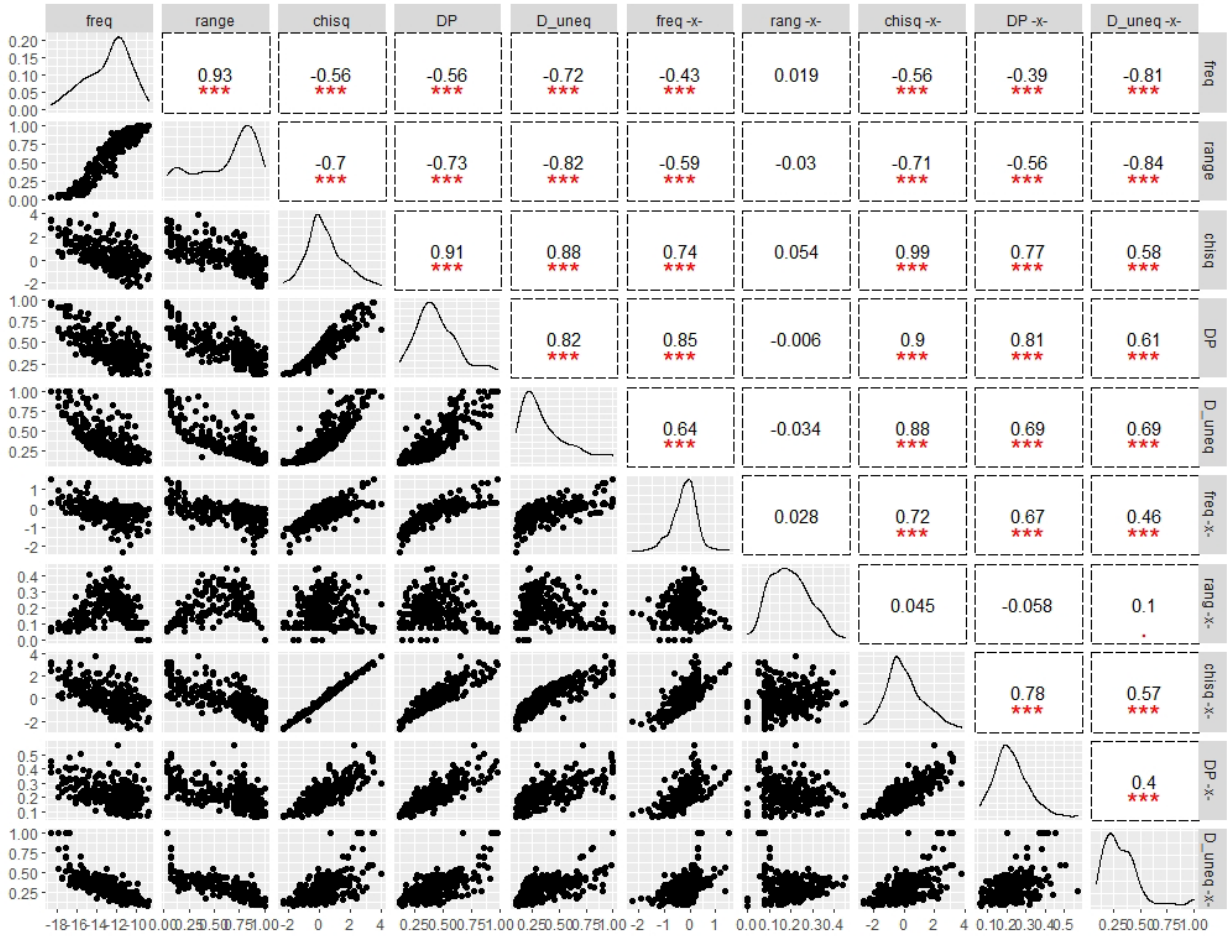

Figure 2. Correlation matrix of correlations between frequency and dispersion measures and the predictive accuracy of the same measures (indicated by -x-)

4.1 Frequency

Frequency is positively correlated (>0.6)[10] with range. This is clearly to be expected, as corpus items that occur often are more likely to occur in many corpus parts than infrequent corpus items.

Furthermore, frequency's negative correlations (<–0.6) with D_uneq and D_uneq -x- indicate that more frequent corpus items in the sample tend to be more evenly dispersed (D_uneq is low) and that the predictive accuracy of D_uneq increases (the deviation decreases) when the frequency of the corpus items increases.

The predictive accuracy of frequency, contained in the variable freq -x-, is positively correlated with the global measures chisq, DP and D_uneq. These correlations are important. They indicate that frequency is a less precise predictor (freq -x- is high) for corpus items with an uneven distribution (chisq, DP and D_uneq are high). The converse is then also true, and this tells us that especially DP (and chisq and D_uneq to a somewhat lesser degree) are important factors to consider when judging the reliability of frequency scores.

The variable freq -x- is furthermore positively correlated with the predictive accuracy (-x-) of chisq (0.716) and DP (0.668). This indicates that the predicitve accuracy of chisq and DP are also to a certain extent influenced by the dispersion of the corpus items.

4.2 Range

In order for range to be comparable across training sets and test sets with different amounts of corpus parts, relative range, i.e. range/number of corpus parts, is used in the cross-validation and therefore also in the global measure. I will however denote the global measure as *range* and the predictive accuracy as *range -x-* .

Range is positively correlated with the global measure frequency (see Section 4.1) and negatively correlated with chisq, DP and D_uneq. These negative correlations indicate, not surprisingly, that corpus items that appear in many corpus parts (high range) tend to be more evenly distributed (low chisq, DP and D_uneq).

Range is also negatively correlated with the predictive accuracy (-x-) of chisq, and D_uneq, which suggests that also the predictive accuracy of chisq and D_uneq increases the more corpus parts a corpus item occurs in.

The predictive accuracy of range, the variable range -x-, is neither positively nor negatively correlated with any other variable, suggesting that the predictive accuracy of range is not dependent on any particular kind of distribution.

4.3 Chi squared

As discussed in Section 2.4.3, chisq is heavily influenced by the global frequency of corpus items. In order to control for this, both the global measure and the predictive accuracy of chisq are controlled for global frequency in figure 2.

Chisq's positive correlattion with the global measures DP and D_uneq indicate that these three dispersion measures respond similarily to most distributions, which is not suprising, since they all measure dispersion.

Furthermore, chisq is positively correlated with the predictive accuracy (-x-) of frequency, chisq and DP. This indicates that the dispersion reflected by chisq, when controlled for global frequency, has clear influence on the predictive accuracy of frequency, DP and chisq. The relation between the global measure chisq, with and without frequency control, and its corresponding predictive accuracy is plotted in figure 3.

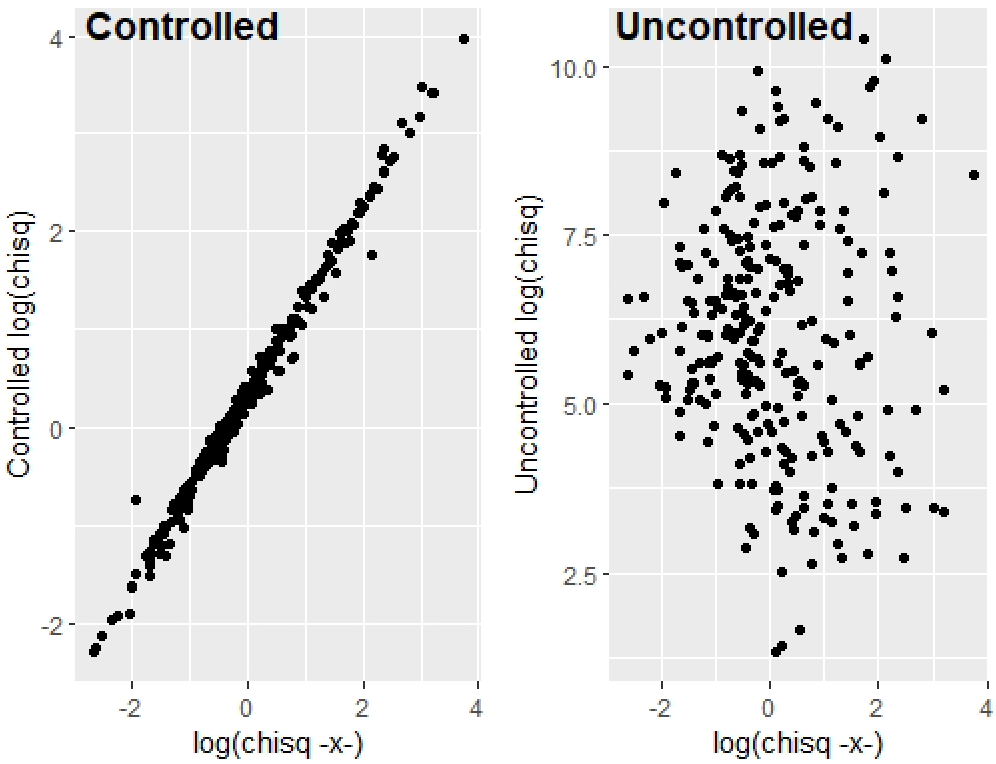

Figure 3: Left plot: Controlled log(chisq) and log(chisq -x-). Right plot: Uncontrolled log(chisq) and log(chisq -x-).

As is reflected by the two plots in figure 3, controlling for frequency in the global measure chisq, makes it an accurate estimator of the sort of dispersion that influences its predictive accuracy. This effect is non-existing when the global measure chisq is used in its original form without controlling for frequency.

Chisq -x- is positively correlated with the global measures DP, chisq and D_uneq, and negatively correlated with range. These positive correlations indicate that predictive accuracy of chisq is dependent on the dispersion of a corpus item (as reflected by DP, chisq and D_uneq). The negative correlation with range is a reflection of the same tendency since items that occur in many corpus parts are generally more evenly dispersed.

Furthermore, chisq -x- is positively correlated with the predictive accuracy (-x-) of frequency (see Section 4.1) and DP. The latter shows that there is a substantial overlap in the predictions made by DP and chisq, most likely due to their tendency to make more accurate predictions for uniform distributions.

4.4 Deviation of Proportions

DP is positively correlated with the global measures chisq (see Section 4.3) and D_uneq and negatively correlated with range (see Section 4.2). The correlation with chisq and

D_uneq shows that DP, chisq and D_uneq yield similar results.

Furthermore, DP is positively correlated with the predictive accuracy (-x-) of all variables except range (see Sections 4.1 and 4.3 for discussions of freq -x- and chisq -x). This indicates that the predictive accuracy of freq, chisq, DP and D_uneq are somewhat dependent on the dispersion of a corpus item, as it is reflected by DP.

The predictive accuracy of DP, as reflected in DP -x-, is positively correlated with the global measures chisq (see Section 4.3), DP (as discussed directly above), D_uneq, and the predictive accuracy of frequency (see Section 4.1) and chisq (see Section 4.3). The correlation with chisq and D_uneq indicates the same tendency as the correlation between DP and DP -x-, namely that the predictive accuracy of DP is greater for corpus items that have a uniform distribution.

4.5 Juilland's D

For the purpose of the cross-validation, the scales of D_uneq and D_uneq -x- have been reversed (compared to the formula in Gries (2008)), so that their scales align with DP. D_uneq is positively correlated with the global measures chisq and DP and negatively correlated with frequency and range. These correlations are accounted for in Sections 4.3, 4.4, 4.1 and 4.2 respectively.

Furthermore, D_uneq is positively correlated with the predictive accuracy (-x-) of frequency, DP, chisq, and D_uneq. All of these predictive accuracies are, as discussed throughout this chapter, to a certain extent dependent on the dispersion of the distribution of corpus items, and it is therefore expected that D_uneq, being a measure of dispersion, correlates positively with these accuracies. This is yet another confirmation that the predictive accuracy is higher (the -x- values are lower) when a corpus item is evenly dispersed (the values of the dispersion measures are low).

D_uneq -x- is positively correlated with the global measures DP and D_uneq, and negatively correlated with frequency and range. These correlations are accounted for in Section 4.4, this section, and Sections 4.1 and 4.2 respectively.

In total, the behaviours of D_uneq and D_uneq -x- resemble the behaviours of DP and DP -x-. There are intercorrelations between all of these variables, except for DP -x- and D_uneq -x- which are not correlated with each other as strongly, although the correlation matrix shows a somewhat positive relation (0.398). The fact that this correlation is not greater may be explained by D_uneq -x- being highly negatively correlated with frequency and range. These negative correlations are stronger (–0.809) and (–0.842) than the positive correlations of D_uneq x¯ with DP (0.609) and D_uneq (0.695), which suggest that the predictive accuracy of D_uneq is more reliant on frequency and range than on dispersion (as reflected by DP and D_uneq). This in turn indicates that D_uneq produces quite inacurrate predictions for infrequent items that occur in just a few corpus parts, whereas DP is more robust in this respect, since DP -x- does not have an equally strong negative correlation with freq (–0.390) and range (–0.556).

5. Summary and concluding discussion

The approach of this study builds on the premise that corpora reflect corpus-external language use, and consequently that corpus distributions predict language distributions. This is the way it is often interpreted in lexicography, where corpora are invoked, among other things, to make arguments about word commonness. With this function in mind, it is of great importance to investigate how one can make predictions and inferences from corpora as accurate as possible.

For this study I have presented the results of a cross-validation of a group of dispersion measures based on 273 Norwegian compounds in the corpus LBK. The cross-validation is supplemented with a Spearman rank-order correlation analysis that sheds light on the question of how the ability of frequency and dispersion measures to predict corpus-external distributions is influenced by the distribution of the corpus item in question. The predicting ability of the frequency and dispersion measures is evaluated through the 'predictive accuracy', which is the mean difference between the distribution estimate of that measure for a corpus item in the training set and in the test set (see Section 3.2 for details on this method).

The results clearly show that there is an important distinction between frequency and dispersion, and that the latter has a great influence on the predictive accuracy of the former. Therefore, there are (at least) two good reasons for using dispersion measures when assessing the commonness of Norwegian compounds (or any other word for that matter):

i. Common compounds are more evenly dispersed than uncommon ones, which makes dispersion a factor of commonness;

ii. Dispersed distributions are systematic to such a degree that it is unlikely that they are generated by chance in a corpus sample. It is therefore likely that dispersed distributions in corpora arise from dispersed distributions in the language variety that is sampled. Dispersion is therefore a sign of both commonness and reliability in the results.

Of the five measures that I have tested, four had known weaknesses from the very beginning. Frequency has the potential of being skewed or inflated for items occurring frequently in small parts of the corpus (see Gries (2008: 404) for a critique of frequency scores). Range may reflect a certain degree of dispersion, but it does not take expected occurrences into account, which makes it unsuitable for corpora with differently sized corpus parts.[11] Chisq has the weakness of not controlling for different numbers of occurrences, making it a somewhat self-conflicting measure for the purpose of assessing commonness, since a high chisq score may reflect either a high frequency or a skewed distribution. In this study, I have therefore tested a slightly adjusted chisq-measure, where the global frequency is controlled for by dividing the chisq values by the global number of occurrences. This seems to make chisq a more accurate estimator of dispersion, but the scale of chisq is still hard to interpret as it is unpredictable what its pessimum value is for a given corpus. D_uneq has the weakness of being inflated when it is estimated on corpora with many parts (see Section 2.4.5).

The results of the correlation analysis corroborate weaknesses tied to frequency, as its predictive accuracy is tied to the evenness of the distribution of a corpus item (which is reflected by DP, chisq and D_uneq). In other words, there are clear advantages in using frequency measures in conjunction with at least one dispersion measure, since e.g. a low DP score would indicate a higher probability of the frequency score being an accurate prediction of the occurrence of a corpus item outside of the given corpus. Moreover, DP itself, chisq and D_uneq are also more reliable for more uniform distributions (as reflected in a low DP score), which takes us to the core of what these dispersion measures do: namely to indicate whether the distribution of a corpus item is proportional to the degree that the distribution most likely is generated by certain properties or patterns of the language, rather than properties that are particular to the given sample. Furthermore, these findings suggest that neither frequency, DP, chisq nor D_uneq have noteworthy merit when it comes to making accurate predictions of skewed distributions.[12]

Although it makes little sense to argue against the use of certain dispersion measures (since diversified methods are favourable in most instances), I will here discuss what DP on its own can tell us about commonness. Gries (2008: 421) shows that highly frequent corpus items tend to get low DP scores. However, this relation is not one-to-one. The most frequent corpus item on the list that Gries provides, the word form *the*, has a minimal DP score of 0,168, which seems reasonable. However, the word forms *this* and *not*, whose frequencies are one-tenth of that of *the*, have lower DP scores, indicating that these items are more proportionally distributed than *the*. But this does not mean that these two items are more common than *the*. It only means that there are corpus parts where *the* occurs more often or seldom than expected based on the size of the corpus part, and that this tendency is a little bit greater for *the* than for *not* and *this*. This should not alter our conception of the general commonness of these words.

At the opposite end of Gries' list, we find word forms like *mamluks*, *hathor*, *defender* and *diamond*. While the former two intuitively seem like seldom and unevenly distributed corpus items, the latter two seem like more ordinary and commonplace items, although perhaps somewhat domain specific. But according to the DP measures, there is not much difference between them. This may reflect both a weakness of DP and a random effect of the corpus. Firstly, it may be that DP is not able to sufficiently distinguish common domain-specific corpus items from seldom domain-specific corpus items. This distinction is however kept by the frequency score. In the case of the items on Gries' list (2008), the affirmation that *diamond* is indeed a more common word than *mamluks, hathor* and *defender* is given by the fact that *diamond* is one order of magnitude more frequent than the other ones. Secondly, the equal DP score of these corpus elements could reflect a coincidental effect of the text material in the given corpus. Even if "diamond" intuitively feels like a concept one has been exposed to from time to time, it is not necessarily a common word, or it is an uncommon word in the particular corpus text sample. This again leads us to the question of the representativity of corpora. To claim that the word *diamond* is more frequent than what is reflected in the corpus, is to put more faith in one's personal intuition than in the empirical materials, which is not an ideal way to go about research. Another problem with such a claim is, as Stefanowitsch (2020: 29) states, that we simply do not know what the population, from which corpora are sampled, looks like. Although we might have an idea of it, we do not know with any certainty from our own intuition exactly what is common or not in the population, that is, the language, and this is precisely why we use corpora: to inform us about the population. But since we do not know the population, we will not know if our corpus is an adequate representation of it, and this is, again, precisely why we should put some effort into studying validation techniques, as I have done here.

A potential source of error in the current study is the uneven number of subdomains in the domains AV, SA, SK, TV and UN (see Table 2). As discussed in Section 2.4, the measures DP, chisq and D_uneq can be affected by the number of corpus parts *n*, which may increase or decrease their predictive accuracy. With the current method, some test sets may systematically yield greater or smaller deviation from the training sets depending on how many parts the test sets consist of. TV has the smaller number of parts, only 3, while the biggest test set, SA, consists of 11 parts. The median deviation between test and training sets for DP does not seem to be affected by *n*, whereas D_uneq seems to generate much less deviation for the biggest test set SA than for the other test sets. Chisq also seems to generate less deviation when the difference in *n* between the test and training sets is smaller. Both of these tendencies can be explained by sensitivity to number of corpus parts (which could only be eliminated by having an equal *n* in the test and the training sets).

A solution to the above source of error could be to let *n* be based on the number of documents nested into each domain. However, as the domains in the LBK are comprised of very different numbers of documents, ranging from 11776 documents in AV to 548 in SK, the relative difference in *n* would only be greater with this approach. The undesirable influence of *n* on the predictive accuracy would therefore only be amplified by basing *n* on documents rather than domains.

Another potential source of error is the genre-specific nature of the test sets. The corpus LBK is not randomly divided into subdomains, but rather categorised according to genre. One would therefore expect there to be a certain difference between the test sets (consisting of one genre) and the training sets (consisting of four genres). Therefore, distributions might be systematically more uniform in the test sets than in the training sets. A way to avoid this effect is to accumulate test sets that consist of material from different parts of the corpus. This is however not necessarily ideal since it masks what sort of qualitative differences there are between the corpus parts. It is also impractical since it is not the way corpora generally are structured.

There are many dispersion measures that I have not considered in this article (Gries (2008) introduces 17 of them), and these would also benefit from cross-validation approaches similar to mine. There is also some more work to be done on DP, as we still do not know what its level of predictive accuracy is, or how to increase it.

To conclude, this study has underscored the importance of considering dispersion when measuring the commonness of linguistic phenomena based on corpora, specifically when measuring the commonness of compounds. Moreover, my findings indicate that Gries' DP is a particularly useful measure which among other things can be used to validate the degree to which frequency scores reflect word commonness. I would generally recommend corpus creators to implement this statistic so that it is easily accessible from the corpus interface.

Notes

1. This number includes compounds where *maskin* is a constituent of a compound that is a constituent of another compound, e.g. *maskingeværild* "machine gun fire".

2. Corpuscle has other stastical measures implemented in its collocation module, but none of these are applicable in its regular KWIC module.

3. The experiments in question are Balota & Spieler (1998) and Baayen (2008).

4. I will not investigate the distance-based measures of Savický & Hlavácová (2002) any further in this article, although I reckognise that they might be suitable instruments for measuring word commonness. I do however think that parts-based measures have an advantage in being more reliable since they, opposite from distance-based measures, do not disregard the internal text structures of corpora.

5. The multiplication with 0.5 is made to ensure that DP falls on the scale 0-1 since the theoretical maximum of the accumulated difference between observed and expected proportions is approx. 2.

6. The non-fiction category does for instance contain texts about oral care and texts about the wife of Vidkun Quisling.

7. Second constituents are omitted in order to mimic the conventional lexicographic procedure of working alphabetically with compound candidates.

8. The eligible *m*'s are frequency, relative range, chisq/NO, DP and D_uneq (see Sections 4.1-4.5).

9. The mean d of the measures frequency and chisq are controlled for global frequency and global NO. This step is made to ensure that the deviations between training and test sets are not systematically tied to the global frequency of the corpus item in question. The mean deviation for frequency is divided by the global frequency of the item, whereas the mean deviation of chisq is divided by global NO(see Section 2.3 for an explanation of the difference between frequency and NO).

10. The threshold ±0.6 is chosen in order to restrict the presentation and discussions to variables that are strongly correlated with each other. Since all the measures are more or less influenced by the global NO of the various corpus items, the ranks of the variables are expected to have a certain monotic association with each other. A threshold of ±0.6 captures the strongest half of the pairwise associations in the correlation matrix (25 out of 45), and allows me therefore to restrict the discussion to the most important correlations.

11. When the corpus parts are equally sized, the likelihood of a corpus item occurring in any corpus part is the same. When the parts are differently sized, part *i* might be twice as big as part *j*, thus making the likelihood of occurrence in part *i* twice as big. These differences in

likelihood are however not captured by range which treats all corpus parts equally.

12. There are inconclusive results with respect to the ability of range to predict skewed distributions.

References

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

Bakken, K. (1998). *Leksikalisering av sammensetninger: en studie av leksikaliseringsprosessen belyst ved et gammelnorsk diplommateriale fra 1300-tallet*. [Doctoral dissertation, University of Oslo]. Acta Humaniora.

Balota, D. A., & Spieler, D. H. (1998). The Utility of Item-Level Analyses in Model Evaluation: A Reply to Seidenberg and Plaut. *Psychological Science*, *9*(3), 238-240. DOI: 10.1111/1467-9280.00047

Biber, D., et al. (2016). On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics*, *21*(4), 439-464. DOI: 10.1075/ijcl.21.4.01bib

Durkin, P. (2016). Introduction. In P. Durkin (Ed.), *The Oxford Handbook of Lexicography*. Oxford University Press.

Egbert, J., et al. (2020). Lexical dispersion and corpus design. *International Journal of Corpus Linguistics, 25*(1), 89-115.

Fellbaum, C. D. (2015). The Treatment of Multi-word Units in Lexicography. In P. Durkin (Ed.), *The Oxford Handbook of Lexicography*. Oxford University Press. DOI: 10.1093/oxfordhb/9780199691630.013.31

Fjeld, R. V., Nøklestad, A., & Hagen, K. (2020). Leksikografisk bokmålskorpus (LBK) – bakgrunn og bruk. In J. B. Johannessen, & K. Hagen (Eds.), *Leksikografi og korpus. En hyllest til Ruth Vatvedt Fjeld*, *Oslo Studies in Language 11*(1) (pp. 47-59). ISSN 1890-9639 / ISBN 978-82-91398-12-9.

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics, 13*(4), 403-437. DOI: 10.1075/ijcl.13.4.02gri

Gries, S. T. (2010). Dispersions and adjusted frequencies in corpora: further explorations. In S. T. Gries, S. Wulff, & M. Davies (Eds.), *Corpus linguistic applications: current studies, new directions* (pp. 197-212). Rodopi.

Gries, S. T. (Forthcoming). Analyzing Dispersion. In M. Paquot & S.T. Gries (Eds.) A *Practical Handbook of Corpus Linguistics*. Springer. DOI: 10.1007/978-3-030-46216-1_5

Leksikografisk bokmålskorpus. Distributed by the CLARINO UiB Portal. Retrieved February 23, 2021, DOI:11495/E1A4-54BE-FCD4-1.

Lyne, A. A. (1985). *The vocabulary of French business correspondence: word frequencies, collocations and problems of lexicometric method*. Slatkine-Champion.

Norwegian Newspaper Corpus Bokmål. Created by Norsk aviskorpus. Distributed by the Clarino UiB Portal. Retrieved February 23, 2021, DOI: 11495/D9B5-0349-4330-0

R Core Team (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

Savický, P., & Hlavácová, J. (2002). Measures of Word Commonness. *Journal of Quantitative Linguistics, 9*(3), 215-231. doi: 10.1076/jqul.9.3.215.14124

Stefanowitsch, A. (2020). *Corpus linguistics. A guide to the methodology*. Language Science press.

Appendix A

List of compounds with corresponding frequency and dispersion estimates

Word	NO	DP	Rel. Range	Chisq/NO	D_uneq
svartand	3	0.91	0.07	9.88	0.69
svartbak	37	0.55	0.14	1.47	0.54
svartebok	21	0.41	0.28	1.55	0.39
svartebørs	188	0.40	0.55	1.68	0.36
svartedauden	72	0.54	0.34	2.33	0.39
svartekunst	7	0.58	0.14	3.95	0.65
svarteliste (verb)	47	0.37	0.48	0.96	0.28
svarteliste (noun)	35	0.43	0.34	1.90	0.38
svartemannen	2	0.91	0.07	15.85	0.80
svartemarje	29	0.57	0.21	1.46	0.41
svarteper	92	0.55	0.41	11.36	0.66
svarthvit	23	0.57	0.21	4.00	0.66
svart-hvitt	370	0.29	0.79	0.43	0.27
svarthyll	12	0.54	0.14	2.38	0.59
svarthåret	51	0.56	0.21	1.45	0.44
svartjord	11	0.52	0.14	2.08	0.59
svartkopp	2	0.64	0.07	8.10	0.92
svartmale	79	0.28	0.55	0.57	0.28
svartrot	6	0.81	0.07	30.68	0.99
svartsinn	1	0.97	0.03	30.28	1.00

svartsjuk	6	0.49	0.14	4.30	0.64
svartsjuke	1	0.97	0.03	32.37	1.00
svartskjorte	10	0.57	0.17	3.39	0.49
svartsladd	3	0.59	0.07	1.76	0.74
svartsmusket	24	0.67	0.07	5.75	0.92
svartspett	18	0.57	0.21	4.13	0.50
svartstill	2	0.67	0.03	2.05	1.00
svartsyn	43	0.41	0.41	1.42	0.35
svarttrost	64	0.39	0.38	1.06	0.49
svartkledd	198	0.44	0.38	0.87	0.34
svartovn	35	0.56	0.10	7.60	0.84
svartbrun	44	0.62	0.17	6.74	0.79
svartbrent	40	0.56	0.24	6.71	0.91
svartlakkere	29	0.57	0.10	1.51	0.62
svartor	23	0.45	0.21	1.35	0.44
vandrefalk	33	0.77	0.14	16.02	0.86
vandrehistorie	76	0.44	0.45	3.42	0.42
vandremotiv	4	0.82	0.07	4.62	0.71
vandrepokal	17	0.81	0.14	5.77	0.54
vandresagn	11	0.68	0.14	6.68	0.63
vandreskjold	6	0.89	0.07	22.52	0.93
vandrestjerne	1	0.94	0.03	15.15	1.00
vandreutstilling	23	0.61	0.21	3.04	0.45
vandreår	18	0.86	0.07	23.75	0.99
vandrearbeider	13	0.59	0.07	3.32	0.84
vandrehall	7	0.61	0.10	1.98	0.61
vandresafari	6	0.67	0.03	2.05	1.00
vandrestav	11	0.48	0.17	2.05	0.70
vandrefugl	6	0.92	0.03	11.67	1.00
vandredue	4	0.86	0.07	6.20	0.70
vandremaur	4	0.61	0.07	3.80	0.85
vandresløyfe	4	0.93	0.03	13.72	1.00
vandretur	14	0.69	0.21	2.70	0.42
tankearbeid	27	0.36	0.24	0.88	0.38
tankebane	63	0.36	0.38	0.72	0.32
tankebygning	12	0.35	0.24	2.19	0.48
tankeeksperiment	90	0.40	0.55	0.98	0.24
tankeflukt	23	0.38	0.28	0.94	0.37
tankegang	974	0.35	0.79	0.62	0.19
tankegods	255	0.42	0.69	1.12	0.36
tankekorn	3	0.55	0.07	1.27	0.70
tankekors	163	0.45	0.69	1.66	0.29
tankeleser	47	0.47	0.38	1.64	0.46
tankelesing	7	0.61	0.07	4.83	0.88
tankemodell	47	0.60	0.38	2.63	0.39
tankeoverføring	19	0.47	0.14	0.92	0.49
tankerekke	199	0.31	0.52	0.55	0.25
tankeretning	39	0.51	0.31	4.13	0.58
tankesprang	30	0.42	0.31	1.48	0.39
tankespredt	24	0.62	0.10	1.76	0.66

tankestrek	22	0.54	0.28	3.83	0.52
tanketom	74	0.46	0.48	1.55	0.39
tankevekkende	267	0.44	0.76	1.09	0.21
tankevekker	37	0.53	0.41	1.99	0.32
tankeverden	77	0.27	0.45	1.28	0.45
tankevirksomhet	110	0.34	0.45	0.87	0.33
tankemønster	88	0.44	0.52	1.12	0.28
tankesett	87	0.48	0.41	1.60	0.33
tankeprosess	75	0.46	0.48	1.50	0.33
tankespinn	75	0.40	0.34	0.83	0.37
tankesmie	75	0.67	0.21	5.47	0.65
årbok	234	0.43	0.55	1.16	0.28
åremål	103	0.55	0.38	3.80	0.56
årgang	355	0.31	0.76	1.06	0.32
årmann	19	0.86	0.07	6.54	0.72
årrekke	567	0.34	0.79	0.69	0.18
årring	122	0.39	0.38	1.99	0.48
årsavgift	178	0.62	0.41	2.86	0.40
årsbasis	81	0.62	0.41	4.52	0.42
årsberetning	307	0.66	0.45	5.43	0.48
årsbest	42	0.84	0.14	17.05	0.90
årsdag	125	0.35	0.69	1.04	0.21
årsgammal	84	0.65	0.34	52.87	0.93
årsinntekt	91	0.51	0.48	2.18	0.34
årsklasse	123	0.50	0.45	1.89	0.34
årskull	212	0.54	0.62	6.51	0.52
årsmelding	160	0.59	0.55	6.51	0.56
årsmøte	611	0.58	0.34	2.55	0.35
årsoppgjør	77	0.62	0.28	2.53	0.44
årsskifte	638	0.49	0.72	1.71	0.30
årsskudd	23	0.76	0.14	4.80	0.60
årsskrift	4	0.86	0.07	8.87	0.81
årstall	349	0.18	0.76	0.27	0.18
årstid	399	0.14	0.90	0.48	0.19
årsunge	11	0.80	0.14	7.24	0.57
årsvekst	12	0.50	0.21	2.68	0.50
årsverk	507	0.57	0.55	5.10	0.45
årti	345	0.27	0.38	0.43	0.17
årtusen	693	0.33	0.90	0.80	0.19
årviss	122	0.40	0.59	1.02	0.29
årslønn	192	0.39	0.69	1.06	0.21
århundre	6434	0.34	0.93	0.80	0.18
framtid	16451	0.22	1.00	0.27	0.13
framover	6926	0.12	0.97	0.10	0.13
framstå	5994	0.38	0.97	0.77	0.16
framfor	5257	0.14	0.97	0.14	0.11
framragende	637	0.33	0.79	0.63	0.20
framstille	4888	0.35	0.93	0.74	0.14
framheve	4064	0.46	0.93	1.29	0.22
framgå	4435	0.61	0.76	7.42	0.55

framgang	3070	0.30	0.93	0.46	0.13
framskritt	1682	0.31	0.86	0.89	0.22
framføre	1591	0.22	0.90	0.36	0.14
framvekst	916	0.51	0.66	1.19	0.26
framkomme	1987	0.53	0.83	2.71	0.69
framtredende	1554	0.40	0.83	0.91	0.18
framholde	1529	0.47	0.76	1.88	0.28
framsette	1939	0.46	0.76	4.56	0.46
frambringe	1211	0.35	0.76	0.63	0.22
framkalle	1426	0.23	0.93	0.79	0.24
framlegge	2001	0.61	0.79	8.83	0.59
bakgrunn	13822	0.35	0.93	0.82	0.16
bakover	4490	0.39	0.97	0.71	0.18
bakside	1787	0.28	0.90	0.36	0.12
baksete	1585	0.41	0.79	0.74	0.26
bakgård	1053	0.38	0.86	0.71	0.27
bakhode	912	0.32	0.86	0.46	0.15
bakfra	833	0.32	0.90	0.49	0.20
baklengs	601	0.31	0.79	0.45	0.20
bakdør	477	0.46	0.66	0.99	0.25
bakrom	405	0.32	0.69	0.52	0.25
bakpå	340	0.34	0.76	0.51	0.19
bakenfor	615	0.18	0.72	0.30	0.21
bakteppe	312	0.41	0.72	0.88	0.20
medføre	7602	0.45	0.86	1.58	0.27
medarbeider	3137	0.37	0.90	0.64	0.15
medhold	2290	0.66	0.72	10.71	0.69
meddele	1337	0.28	0.86	1.16	0.69
medvirke	1891	0.39	0.86	1.13	0.22
medpasient	125	0.62	0.45	5.36	0.50
medfødt	713	0.26	0.76	0.38	0.17
medfølelse	635	0.28	0.76	0.71	0.35
medgi	554	0.30	0.69	0.73	0.31
medmenneske	536	0.29	0.76	1.33	0.28
medhjelper	454	0.27	0.72	0.95	0.22
medborger	809	0.78	0.69	9.50	0.73
medbringe	285	0.25	0.69	0.92	0.29
medelev	344	0.36	0.69	1.08	0.25
medstudent	337	0.33	0.72	2.71	0.54
medtatt	326	0.33	0.62	0.70	0.26
meddommer	324	0.71	0.48	10.44	0.66
medregne	324	0.40	0.62	0.97	0.23
medbestemmelse	413	0.60	0.66	4.79	0.43
medspiller	263	0.36	0.72	0.82	0.22
medfølende	249	0.45	0.48	0.93	0.29
medeier	220	0.42	0.55	2.17	0.34
medskyldig	219	0.26	0.72	0.39	0.20
medynk	193	0.44	0.55	1.05	0.45
medfange	115	0.40	0.38	1.78	0.44
medsammensvoren	148	0.37	0.55	1.32	0.61

medgå	120	0.71	0.31	11.57	0.74
medfart	132	0.32	0.62	0.79	0.22
medpassasjer	126	0.32	0.59	1.14	0.40
arbeidstaker	5270	0.58	0.79	1.92	0.30
arbeidsgiver	5714	0.50	0.86	1.70	0.27
arbeidsplass	3772	0.37	0.86	0.84	0.17
arbeidsliv	4227	0.57	0.79	3.03	0.42
arbeidskraft	2191	0.44	0.83	1.24	0.26
arbeidstid	2268	0.46	0.86	1.26	0.41
arbeidsoppgave	1341	0.44	0.86	1.50	0.26
arbeidsmarked	2038	0.58	0.79	7.91	0.57
arbeidsforhold	1185	0.50	0.79	2.63	0.54
arbeidsgruppe	1168	0.61	0.72	4.44	0.44
arbeidsdag	1148	0.17	0.90	0.28	0.54
arbeiderklasse	1211	0.44	0.66	1.02	0.23
arbeidsmiljø	2372	0.62	0.76	4.31	0.40
arbeidsledig	894	0.43	0.83	2.89	0.33
arbeiderbevegelse	843	0.50	0.66	1.51	0.29
arbeidsløs	523	0.32	0.72	0.81	0.21
arbeidsdeling	503	0.56	0.45	1.72	0.36
arbeidssøker	494	0.71	0.41	20.33	0.82
arbeidssituasjon	478	0.53	0.72	2.09	0.33
arbeidssted	473	0.48	0.59	1.85	0.34
arbeidsinntekt	428	0.70	0.48	13.36	0.67
arbeidsinnsats	406	0.47	0.59	1.93	0.31
omkring	14466	0.14	1.00	0.20	0.13
omfatte	17018	0.44	1.00	1.19	0.22
omtale	9216	0.35	0.93	0.65	0.15
omsorg	10237	0.47	0.93	1.48	0.26
omgang	5166	0.23	0.97	0.50	0.15
omfang	3735	0.40	0.83	1.41	0.27
omgi	3068	0.21	0.97	0.31	0.15
omhandle	2621	0.50	0.79	2.52	0.33
omsider	2490	0.39	0.90	0.72	0.19
omkomme	1983	0.47	0.83	2.51	0.30
omvendt	2131	0.12	0.93	0.10	0.10
omhyggelig	1175	0.36	0.83	0.67	0.27
omsette	1143	0.36	0.90	0.68	0.17
omdanne	1158	0.48	0.79	2.21	0.34
omfavne	1097	0.34	0.86	0.57	0.20
omstridt	1077	0.40	0.76	0.86	0.20
omverden	1219	0.32	0.83	0.61	0.16
omgjøre	1049	0.29	0.79	0.72	0.18
ombestemme	894	0.42	0.76	0.80	0.23
omringe	826	0.27	0.90	0.91	0.27
omlag	793	0.41	0.76	1.54	0.25
omslag	715	0.21	0.79	0.23	0.16
omgås	1308	0.13	0.90	0.15	0.11
omvei	591	0.24	0.79	0.34	0.19
omslutte	569	0.33	0.86	0.53	0.21

omkranse	518	0.18	0.86	0.31	0.25
omtanke	517	0.17	0.86	0.25	0.24
omforme	490	0.41	0.79	0.86	0.22
omdømme	485	0.40	0.72	0.87	0.29
undersøke	8322	0.31	0.97	1.03	0.19
understreke	6956	0.40	0.86	0.84	0.20
underveis	2769	0.23	0.97	0.43	0.13
undertegne	2272	0.33	0.90	0.73	0.17
underlegge	1875	0.37	0.90	0.94	0.21
undervise	1990	0.23	0.90	0.34	0.18
undertrykke	1062	0.31	0.86	0.75	0.19
underligge	1266	0.45	0.76	1.65	0.27
underbygge	1052	0.50	0.79	2.22	0.28
underholde	1379	0.19	0.90	0.49	0.23
undergrave	949	0.40	0.86	0.78	0.16
underlag	840	0.24	0.86	0.48	0.33
underskrive	783	0.32	0.76	0.97	0.23
underside	746	0.43	0.79	2.82	0.38
undervurdere	740	0.25	0.83	0.34	0.15
underordne	1322	0.36	0.83	0.70	0.18
undergang	742	0.19	0.79	0.27	0.16
undertøy	710	0.38	0.79	0.81	0.20
underarm	614	0.47	0.76	1.01	0.27
understøtte	577	0.46	0.69	1.87	0.32
underkaste	572	0.26	0.72	0.48	0.21
underskrift	641	0.26	0.86	0.81	0.23
underbukse	525	0.45	0.62	0.96	0.26
underliv	535	0.33	0.83	0.53	0.21
underkant	443	0.37	0.76	1.32	0.30
undersått	432	0.34	0.79	1.08	0.26
overta	8325	0.31	1.00	0.67	0.17
overleve	6169	0.17	0.97	0.25	0.11
overføre	5622	0.36	1.00	0.81	0.21
overbevise	5349	0.19	0.93	0.22	0.13
oversikt	5190	0.32	0.93	0.72	0.18
overalt	4229	0.23	0.97	0.28	0.11
overflate	4738	0.26	0.93	0.41	0.18
overgang	4829	0.34	0.93	0.70	0.17
overlate	2688	0.15	0.93	0.12	0.10
overordne	3623	0.43	0.86	1.61	0.29
oversette	3752	0.25	0.93	0.30	0.13
overse	2634	0.14	0.90	0.15	0.12
overgrep	2036	0.37	0.93	0.84	0.28
overtale	1555	0.26	0.90	0.36	0.17
overvåke	1456	0.23	0.86	0.32	0.20
overskrift	1339	0.27	0.83	0.51	0.19
overskride	1211	0.37	0.83	0.69	0.18
overstige	1124	0.48	0.79	2.15	0.50
overdrive	2180	0.15	0.97	0.19	0.12
overgi	1219	0.33	0.90	2.57	0.40

overnatte	1125	0.30	0.90	0.69	0.22
overlege	1471	0.19	0.86	0.46	0.18
overkropp	1034	0.44	0.83	0.90	0.27