International Journal Of Lexicography, 2023, **XX**, 1–19
https://doi.org/10.1093/ijl/ecad013
Advance access publication 30 May 2023
Article

Wheat or Chaff? A Compound Selection Model Based on Look-Up Data

Mikkel Ekeland Paulsen

University of Bergen, Norway Norway
(Mikkel.Paulsen@uib.no)

Abstract

Which compounds should be included in general-purpose dictionaries is often an open question that is answered with a case-by-case consideration of all compounds above a certain corpus frequency threshold. Another way to determine which compounds should be listed, is to examine which compounds, or rather which compound properties, are in demand by the users. This study uses look-up data from the two officially sanctioned, general-purpose dictionaries of Norwegian (Bokmålsordboka and Nynorskordboka) to derive an explicit compound selection model that performs with comparable sensitivity and specificity as the traditional procedure. These findings demonstrate that it is indeed possible to arrive at a fully operational and explicit compound selection model that meets the needs of users. With such a tool at their disposal, lexicographers would be able to separate the wheat from the chaff in the boundless field that is the compound lexicon of North Germanic Languages.

Keywords: compound selection, Norwegian, look-up data, corpus data, conditional inference trees

1. Background

Selecting compounds for general-purpose dictionaries of any North Germanic language resembles the act of separating the wheat from the chaff, aside from the fact that in the compound selection scenario there are no physical properties that disclose which compounds correspond to the respective roles. Instead, lexicographers must base their decisions on linguistic and distributional properties of each compound. One way to determine which properties warrant inclusion is to observe which properties are in demand. This study aims to identify the variables that govern the look-up interest into compound lemmas in online dictionaries.

A particular challenge that pertains to compound selection in these languages is the ubiquity and productivity of compounds in language use. As an example, in the medium-sized corpus *Leksikografisk bokmålskorpus* (henceforth LBK), (Fjeld et al. 2020) the string that makes up the semi-frequent word *maskin* 'machine' enters into approximately 2,000 different compound types. Obviously, not all of these compounds can or should go into a general-purpose dictionary. Lexicographers therefore need rigorous methods and criteria for compound selection to be able to extract a selection of compounds that serves the needs of dictionary users. Although there can never be a gold standard for what makes the ideal lemmalist of compounds across dictionaries with different scopes and purposes, look-up behaviour in an online dictionary may give insight into the interests and needs of the users of that particular dictionary. This user interest should at least be a part of the equation in the assessment of the lemmalist of a given dictionary. We may therefore view look-up statistics as an important empirical foundation for both assessment and enhancement of an existing lemmalist.

Corpus frequency is normally understood as a numerical value that reflects how many occurrences there are of a corpus item relative to the size of the corpus. In an online Swahili-English dictionary, De Schryver et al. (2006) found a positive correlation between look-up and corpus frequency among the 5000 most frequent words in the corpus. Trap-Jensen et al. (2014) suggest that corpus frequency is an important predictor of look-up frequency in a monolingual Danish dictionary for about the 20 000 most frequent lemmas in a balanced corpus. The findings of Wolfer et al. (2014) and Müller-Spitzer et al. (2015) also indicate that corpus frequency is an important predictor of look-up frequency in monolingual German dictionaries. This finding is confirmed by De Schryver et al. (2019) in a further study on the same online Swahili-English dictionary. Although one might assume that there are important differences between the look-up behaviour in monolingual and bilingual dictionaries, corpus frequency seems to be an important predictor of look-up frequency in both dictionary types.

A weakness of the above-mentioned studies is that their quantitative focus is limited to corpus frequency, which is an unreliable and somewhat invalid measure of frequency of occurrence, especially if frequency of occurrence is perceived as a token of overall word importance or commonness (Gries 2008: 404, Paulsen 2022). For one thing, it suffices for a corpus item to be very frequent in one limited part of the corpus in order to seem frequent in the corpus as a whole. For example, if the corpus contains a book on goatfish, then the word *goatfish* would occur surprisingly frequently in the corpus as a whole. Goatfish-effects make frequency unreliable in that different language samples will generate very different frequencies for such words. If the frequency of a word is completely different for every language sample, one should be cautious about inferring from frequency estimates to properties of the language variety that the corpus is sampled from.

A way to escape arbitrary goatfish-effects is to include a measure of dispersion that evaluates the degree to which the occurrences of a corpus item are evenly spread throughout the corpus. A word which is both well-dispersed and frequent in the corpus sample is more likely to be commonly occurring in the language variety that the corpus is sampled from than a frequent word whose occurrences are clustered together. Put briefly, dispersion influences what inferences we can make from the frequency score (Paulsen 2022).

Frequency and dispersion, then, are quantitative variables that with all likelihood have a bearing on the look-up frequency of compounds. Since dictionaries are consulted both in productive and receptive contexts, e.g., both when writing and reading, it is reasonable to assume that words which are frequently read or written will typically be looked up more often than words which are seldom read or written.

The present study will also survey a number of **qualitative** variables that may affect which Norwegian compounds users look for, namely semantic transparency, part of speech (POS), whether or not the compound has an interfix, the number of spelling variants of a compound, and one that is particular to parallel dictionaries: whether or not there is equivalence between the two Norwegian written languages *Bokmål* and *Nynorsk*.[1],[2]

A satisfactory model for compound inclusion must integrate considerations pertaining to both linguistic properties and usage. In addition, the criteria and variables on which the model operates must be explicit and reproducible. The major aim of this study is to propose such a model for compound selection in the officially sanctioned general-purpose dictionaries of Norwegian, *Bokmålsordboka* (henceforth *BOB*) and *Nynorskordboka* (henceforth *NOB*) (and, by extension, of other languages that are morphologically similar in terms of compounding). In order to develop such a model, the following steps are taken:

- Based on a sample of compounds, the implicit standard model (henceforth *StandMod*) that has served to create the current lemmalist of compounds in the BOB is examined with respect to its association with the variables in the study.

- Through a mixed set of methods, the variables in the study are investigated with regard to their association with look-up statistics.
- Based on the results of the investigation in the previous step, an alternative model comprised of explicit criteria is developed (henceforth *The Look-Up Predictor Model*, abbreviated *LookMod*).
- Using the look-up data, the performance of LookMod is evaluated and compared with the performance of StandMod.

Historically, it is unknown which set of criteria has prompted the current compound list in the BOB. In other words, it is not explicitly stated what reasoning lies behind the inclusion of a given compound and it is likely that different editors have partly based their decisions on their individual and idiosyncratic intuition.

The findings of the above-mentioned procedure in part validate the StandMod by demonstrating that its output, the current lemmalist of the BOB, is to a large extent in harmony with the look-up interests of the dictionary users. Using variable levels deducted from look-up statistics, the LookMod performs at nearly the same level of specificity and sensitivity as StandMod. Although these are promising results for LookMod, it nevertheless shows that performance alone does not motivate a shift in lexicographical practice. However, the LookMod has some other advantages. Firstly, it is more transparent than the StandMod. Secondly, it is more objective and less reliant on the intuition of a given lexicographer. And thirdly, it could potentially be less time-consuming than the StandMod overall, if one could find a way to automatically annotate all compounds in a corpus according to (most of) the variables in the model. It should be noted that this is somewhat far fetched currently, since we still lack a procedure for automatic detection of compounds.

Although the advantages of the LookMod are not trivial, its primary usefulness is the information that it conveys about the variables it contains (and ignore). First and foremost, corpus dispersion stands out as an important predictor of look-up interest. In other words, dispersion is a variable that ought to be included in whatever model lexicographers apply.

2. Norwegian compounds

For the purpose of this study, a compound is defined as a lexeme whose stem is comprised of two individual stems that are both found as stems of separately occurring lexemes. [3] These stems may be either compounds, derivatives or root words. Borrowed compounds (e.g., *airbag*) are excluded from this definition unless both constituents are stems of separately occurring lexemes, in this study operationalised as listed in the BOB, while borderline cases between derivation and compounding are generally accepted as compounds.

Norwegian compounds are generally right-headed, which means that the second constituent is the semantic and grammatical head, whereas the first constituent functions as a modifier. Norwegian compounds may also contain an interfix that is added as a suffix on the first constituent, as in *fortauskant* 'pavement' + $[\text{s}]_{\text{interfix}}$ + 'edge'. Generally speaking, interfix inclusion is a property of the modifier stem. For instance, *fortau* 'pavement' is consistently suffixed, whereas e.g., *vind* 'wind' is never suffixed. There are however many examples of stems that are variably suffixed (see Kulbrandstad & Kinn 2016 for examples).

Compounds are ubiquitous in Norwegian. This fact compels lexicographers to pick and choose from an unbounded list of candidates for dictionary inclusion. The productivity of Norwegian compounds also extends to grammatical variability: Nominals, adjectives, verbs, prepositions and adverbs are all fairly productive as compound constituents. Still there is no doubt that noun-noun compounds are the most productive type (see also Section 3.1).

Fjeld & Vikør (2008) argue that 'semantic transparency' is an important factor in a compound selecting scheme. There is no consensus on the exact meaning of 'semantic transparency', but it roughly refers to whether the meaning of the compound can be inferred from the meaning of its parts (Schäfer (2018) gives an overview over vastly different definitions of semantic transparency). A simple operationalisation of semantic transparency that is utilised in this study is *degree of motivation* as it is defined by Svanlund (2002).

3. Data and method

This section outlines the dataset and the variables of the study.

3.1. Compound sample

The analyses of the current study will be performed on a sample of compounds. The sample is collected with the following steps:

1. Each compound in the sample belongs to one of the five alphabetical stretches (henceforth segments) afrikaans – -aktig, bryllup – bukt, dverg – dørk, einstøing – eksterritorialrett and forstokka – forårsake. These are segments which have recently received a full revision in the BOB, each by a different editor. The BOB contains 1560 lemmas from these segments, 802 of which are compounds.
2. From these segments, any given compound is included in the sample if it fulfils at least one of the following criteria:
   a. It has an entry in the BOB.
   b. It has a minimum of 20 occurrences in the corpus LBK, which equals approx. 0.2 occurrences per million words (pmw).
   c. It occurs in the look-up data (see Section 3.2).

The criteria a, b and c yield 802, 570 and 919 compounds, respectively. There is substantial overlap between the groups and the unique contribution from each category amount to 112, 153 and 196 respectively. The complete sample amounts to 1206 compounds and contains both frequent and infrequent items that show a presumed typical spread across different parts of speech for both constituents. See Supplementary Material Online for tables 1 and 2 that display examples of compounds from different frequency bands in the LBK and the distribution of different compound types.

In the sample, nouns are dominant in both the modifier and head position. Noun-noun compounds account for approximately 75% of the sample, which probably reflects the linguistic reality that Norwegian lexicographers deal with, i.e., that the noun-noun pattern is far more productive than any other compound pattern. As a reference, 84% of the lexeme entries[4] in the BOB are nouns, and about 78% of all entries in the *Norwegian Academy Dictionary* are nouns. This nominal predominance indicates that part of speech needs to be controlled for in statistical analyses.

3.2. Look-up data

Every look-up in the BOB and the NOB is saved in a search log. Both dictionaries are freely accessible from the same interface, currently at ordbokene.no[5] and the look-up data have been obtained through analysis of log files generated from both dictionary

Table 1: Counts and proportions of BOB status at levels of look-up regularity.

look-up regularity	0		1-10		10-100		>100		total
In BOB	count	prop %	count	prop %	count	prop %	count	prop %	
0	153	53	151	46	83	23	16	7	403
1	134	47	178	54	285	77	206	93	803
total	287		329		368		222		1206

look-up and access through general-purpose search engines. With this approach, an inventory of search queries and their number of occurrences within a certain time frame has been generated.[6]

The look-up data are accessible through a webpage[7] that contains a selection of files with lists of all queries that have been performed more than a certain number of times in a given year. For this study, every query that has been performed at least 10 times inside one of the calendar years 2016 – 2020 is included in the look-up data. This means that the data contain information about the accumulated number of look-ups of each compound in the sample for each of the five years in which there are 10 or more look-ups.[8]

The look-up frequency data are not lemmatised and there is no way of knowing for certain which dictionary entry a user is seeking when they use for example the query expression *fortelt*. One could assume that they are after the compound *fortelt* lit. 'front tent', but it is not at all unlikely that they have misspelled the perfect participle form *fortalt* of the frequent verb *fortelle* 'tell' (not a compound). Not knowing the users' intentions, we cannot know what they are seeking in all instances. This fact makes the following operationalisations inescapable: 1) We assume that users do not misspell their queries, and 2) we count certain queries as a look-up for more than one entry. This way, queries which are homographical with the compounds in the sample and inflectional variants thereof count in the summation of the overall number of look-ups for those compounds. Punctuation and frontal or final white space in the queries are ignored. This means that the above example will count as a look-up of *fortelt* and not *fortelle*.

Some dictionary entries have multiple spelling variants, for instance are *dypsnø* and *djupsnø* 'deep snow' parallel headwords of the same entry. Such spelling variants are treated as members of the same lexeme. A further technicality relates to the fact that most queries at *ordbok.uib.no* return values from both BOB and NOB. Since querying into both dictionaries simultaneously is the default when people first enter the website either directly or via a search engine, the parallel look-ups are by far the most frequent, and the look-up data are therefore based on these parallel look-ups. For this reason, query expressions that match compounds in *Nynorsk* that are equivalent to the sampled compound lemmas also enter into the overall number of look-ups for their equivalents. This means that for example both the query *forståelsesfull* (which matches a compound in Bokmål meaning 'understanding' (adjective)) and *forståingsfull* (which matches the Nynorsk equivalent of the compound *forståelsesfull*) add to the look-ups of the same compound lemma.

3.3. Configuration of look-up variables

Four look-up variables are employed in this study.

- **Number of look-ups** is the accumulated number of look-up events that match a compound lexeme.

Table 2: Performance of both models on the original dataset up regularity

look-up regularity		0		1-10		10-100		>100		
		count	prop %	count	prop %	count	prop %	count	prop %	Sum
StandMod	0	153	53	151	46	83	23	16	7	403
	1	134	47	178	54	285	77	206	93	803
LookMod	0	126	44	156	47	102	28	6	3	390
	1	161	56	173	53	266	72	216	97	816

- **Look-up frequency** is the number of look-up events for a given compound lexeme divided by the total number of look-ups.
- **Look-up dispersion** reflects the distribution of look-up events matching a compound lexeme over look-up year and is operationalised as the average of a *Deviation of Proportions* estimation (as described by Gries 2008, henceforth *DP*) and a *Juilland's D* estimation (as described by Juilland et. al. 1971).[9] In other words, the look-up dispersion reflects the degree to which the look-up frequency for a compound lexeme is stable over the five years that the look-up data represent. In short, dispersion is high when the look-up frequency remains stable over time.
- **Look-up regularity** is the product of number of look-ups and look-up dispersion. This variable reflects number of look-ups while controlling for look-up dispersion. Since none of the compound lexemes in the dataset exhibits optimal dispersion, the look-up regularity scores are slightly lower than number of look-ups for each compound. Figure 1 shows the distribution of log10-scaled look-up regularity. The purpose of the scaling is to make the x-scale of the histogram easier to read, but note that 287 compounds with zero look-ups are left out of the figure. The histogram shows that many compounds lie in the 1-10 and 10-100 ranges, and that the number of items decreases gradually toward a look-up regularity of 1000.

The look-up variables serve as response variables in Sections 5 and 6. There are a number of predictors included in the study which are all listed and briefly explained below. For a detailed account of the distribution and configuration of the predictors, please see Supplementary Material Online.

- **Number of occurrences in corpus (NO):** The absolute number of occurrences of a compound lexeme in the LBK.
- **Corpus dispersion (disp**): The average dispersion of a compound lexeme in the LBK based on a DP- and Juilland's D-estimation of the domain- and yearwise distribution of lemma.
- **Degree of motivation (DoM**): The degree to which the conventional denotations of the constituents are active in the denotation of the compound as a whole, 0 = No motivation, 1 = Motivated modifier, 2 = Motivated head, 3 = Fully motivated.
- POS of modifier and head (**POS_m** and **POS_h**).
- **Interfix**: Whether the compound has an interfix.
- **Parallelism (paral):** The degree to which a compound in Bokmål has an equivalent in Nynorsk, *No* = No obvious equivalent, *Partial* = Semantic equivalence without homography, *Full* = Semantic and homographic equivalence

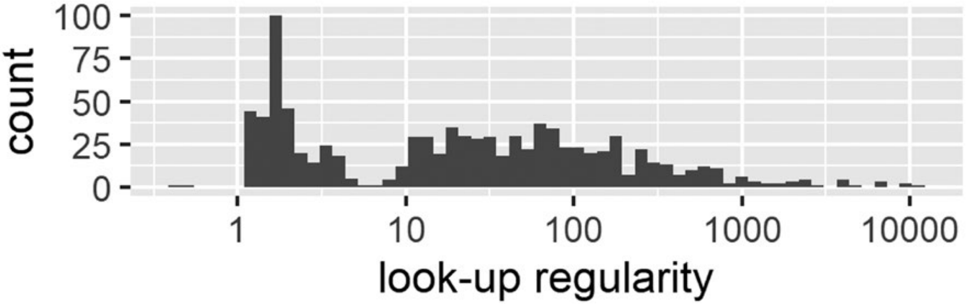

Figure 1: Histogram of logged look-up regularity in sample.

- **Number of spelling variants (Nvar):** Whether the compound has multiple spelling variants.
- **Salience:** The degree to which a compound deviates from the statistically most common properties with respect to the variables POS_m, POS_h, DoM, Nvar and paral.

Now that we are familiar with the data and variables of the study, we may proceed with an investigation of StandMod.

4. StandMod

In this chapter I present the background and broad tendencies of the current compound lemmalist in the BOB within the five segments that are investigated in this study. This list has emerged through a variety of methods and may be seen as the returned value of the StandMod. After an exploration of some of the characteristics of this model, I will evaluate its performance based on the degree to which it reflects the look-up interests of the dictionary users.

4.1. Background and broad tendencies

The set of compounds that are currently listed (*In BOB* = 1) in the BOB has come about through 12 years of compilation before the first publication, and 35 subsequent years of sporadic revisions and updates, the most recent in 2019-2020. During this period, the lexicographical practice has become increasingly corpus-based, and one might assume that the empirical foundation for lemma selection has been gradually strengthened. Another change is that the dictionary has moved from a printed to a digital format. Intuitively, one would assume that this change facilitates the listing of a greater number of compound lemmas since space is much less sparse in digital dictionaries compared to printed ones. The StandMod is nevertheless not explicitly formulated, and it is not known exactly which criteria the various lexicographers that have edited the dictionary have employed. All that is known is that many considerations underlie the selection of compound entries, and a given entry may be justified by for example corpus frequency, grammatical or semantic properties, or internal systematicity within the BOB or between the BOB and the NOB.

Figure 2 displays the relationship between log10-scaled NO (in the LBK) and BOB status. The inter-quartal ranges and the mean values (indicated by the dots) indicate that listed lemmas are associated with higher NO. However, there is considerable variation in the NO-values of both listed and unlisted compounds. In fact, unlisted compounds have a higher median NO than listed ones. This demonstrates that NO is not the only variable that governs dictionary inclusion.

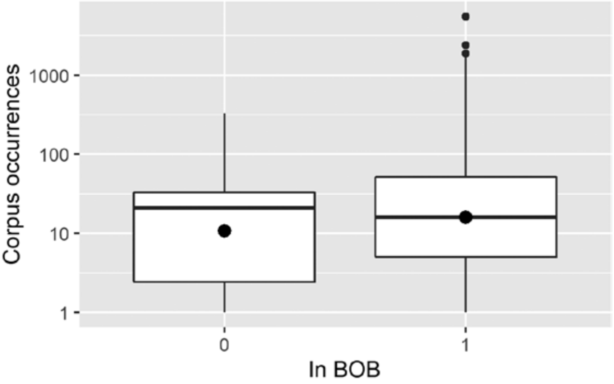

Figure 2: Distribution of NO over BOB status.

Furthermore, unlisted compounds are overrepresented among compounds with NO = 0 and 10–50, whereas listed compounds are overrepresented in the other categories (including 1–10) in Figure 3. In other words, NO seems to have a strong influence on dictionary inclusion when it exceeds 50, but there is substantial and seemingly random variation below this point. This indicates that high frequency may be an important qualifying factor for dictionary inclusion, especially beyond 400, but low frequency is evidently not systematically employed as a disqualifying criterion in StandMod.

There is substantial internal variation with respect to the dispersion score among listed and unlisted compounds, see Figure 4. The median disp value is higher for listed than unlisted items, but many of the listed items have a very low dispersion score.

More enlightening is perhaps Figure 5, which shows the proportions and absolute numbers of listed and unlisted compounds in different dispersion bands. The horizontal line indicates

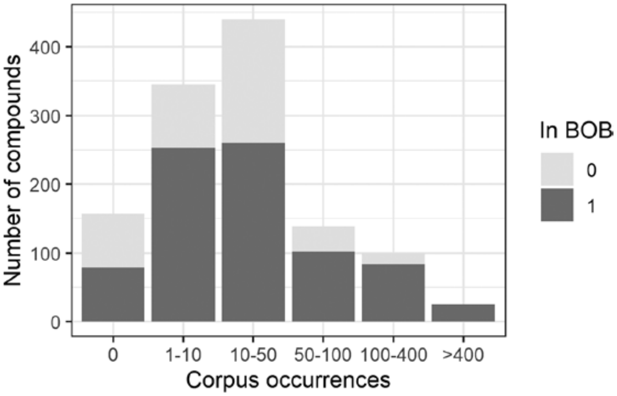

Figure 3: BOB status according to NO bands.

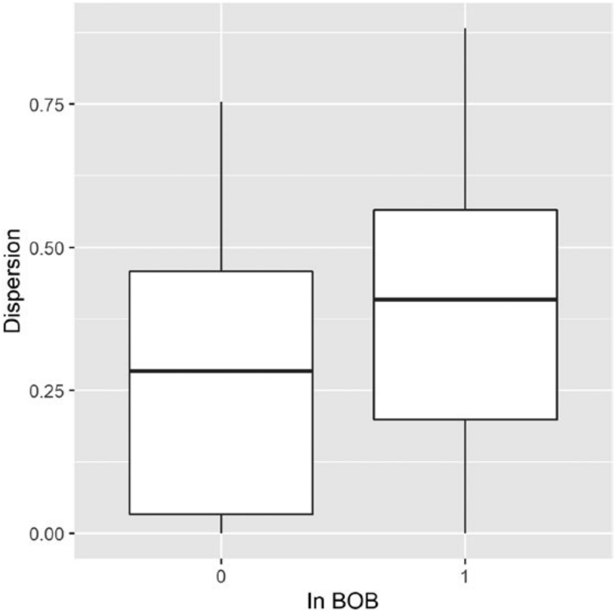

Figure 4: Distribution of disp over BOB status.

the global proportion of listed items. Here, we see a gradual increase in the proportion of listed items as the dispersion value increases. However, it is only in the upper bands 0.5-0.7 and >0.7 that listed lemmas are overrepresented compared to the percentage of listed items in the dataset (66.6%). While higher dispersion increases the likelihood of a compound being listed, low dispersion is evidently not employed as a disqualifying criterion in StandMod.

Deviation from full motivation is associated with an increase in the likelihood that a compound is listed, see Figure 6. Listed compound lemmas are overrepresented for all levels of Degree of Motivation below 3. It should, however, be noted that over 80% of the

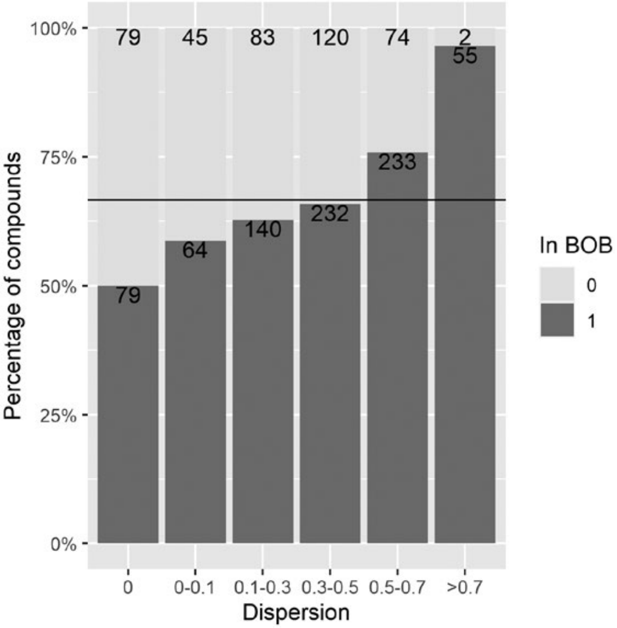

Figure 5: BOB status according to disp bands.

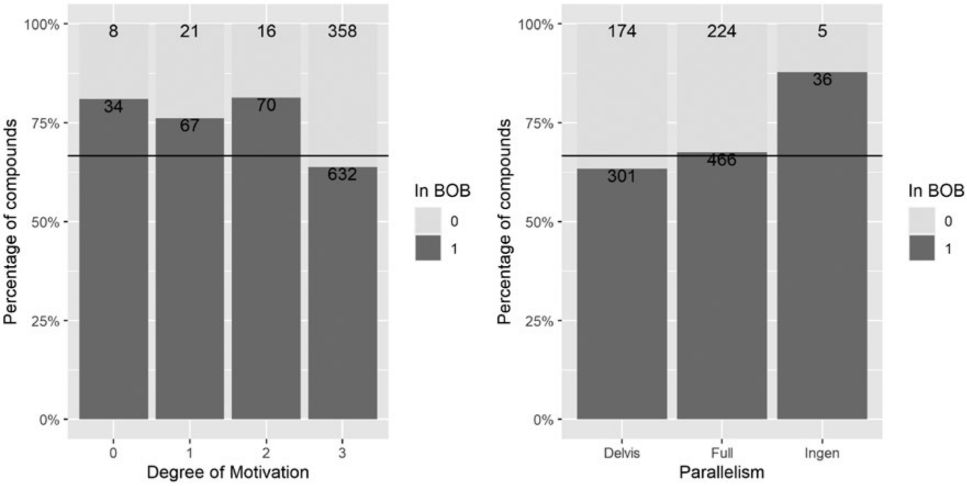

Figure 6: BOB status according to Degree of Motivation and Parallelism.

compounds in the dataset are fully motivated, so the distributions in the levels 0-2 are based on a small proportion of the compounds in the study.

Listed compounds are overrepresented among compounds in Bokmål with no obvious equivalent in Nynorsk, see Figure 6. Only 4% of the compounds in the study fall into this category, which indicates that the role of this variable is at best peripheral in StandMod. Note that the risk of influence from random variation increases in categories with few items. This is especially relevant among compounds with low degree of motivation and no parallelism.

Listed compounds are overrepresented among the adjectival, adverbial, interjectional, prepositional and verbal modifiers, which means that it is only among the compounds with nominal modifiers that listed compounds are underrepresented, see Figure 7. This indicates that StandMod includes a larger proportion of compounds with non-nominal modifiers than with nominal modifiers, presumably because the former are much less productive constructions in Norwegian. Since nominal compounds are much more productive than any other compound construction, one would expect this to be the chief construction among novel compounds also, which in turn would make nominals overrepresented among the unlisted compounds in the sample. The rather small effect of this can be seen on the left side of Figure 7.

A similar pattern can be found with respect to nominal heads. The plot on the right in Figure 7 indicates that listed compounds are overrepresented in all categories except for the nominal one, where they are underrepresented. It should, however, be noted that the distribution of BOB status in compounds with nominal heads is more or less identical with the global distribution in the sample (65%).

Number of spelling variants (Nvar) and interfix status do not demonstrate any particular tendency towards association with BOB status.

We may conclude that StandMod has a certain degree of covariation with the distributional variables Number of occurrences (NO) and dispersion (disp), and the semantic variable Degree of Motivation (DoM), and that a vast majority of the unlisted compounds in the dataset have nominal modifiers and/or heads. A generalised linear regression approach has also been attempted, but it has not helped indicate any further characteristics of StandMod that are not visible from the graphs presented above.[10] There is, in other words, substantial variation in BOB status that the variables in this study cannot account for. We

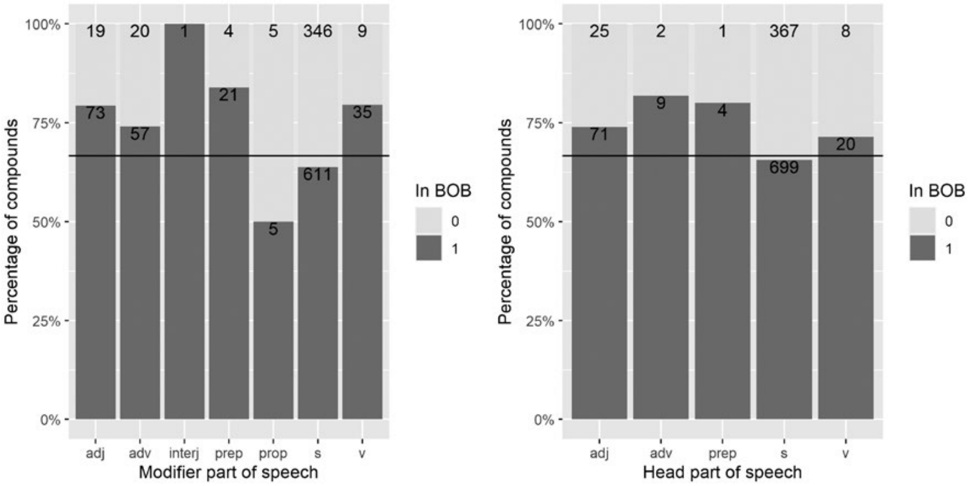

Figure 7: BOB status according to POS_m and POS_h.

may therefore conclude that a defining trait of StandMod is that it is multifaceted and flexible in nature, and that there are grounds for including a compound beyond the variables included here.

4.2. Evaluating StandMod

A way to evaluate the lemmalist of the current version of the BOB, and thereby the model that has produced it, is to investigate the extent to which the lemmalist is in harmony with the interests of the dictionary users. The results of such an investigation are presented in the following, using look-up statistics as an operationalisation of user interest.

A starting point for the evaluation is a simple analysis of the two possible sources of discordance between the dictionary and the look-up statistics, namely *unvisited entries* and *lacunas* (i.e., unlisted lemmas that are looked up fairly regularly) which indicate the specificity and sensitivity, respectively. In practice, unvisited entries are often quite unproblematic. There might be systemic reasons to include certain compounds, regardless of their interest to the dictionary users. However, a large number of unvisited entries coupled with a large number of lacunas might indicate that there is a lack of harmony between the selectional criteria of the dictionary and the needs of the users. The percentage of visited entries relative to unvisited ones is thus worth some scrutiny.

It is not obvious what the appropriate threshold for a lacuna is, but it would be unreasonable to treat every unlisted and looked-up compound as a lacuna. There might for instance be systemic and language-specific reasons for not listing a compound even though it is looked up, e.g., people might look up foreign words.

Table 1 gives an overview of the numbers of listed and unlisted compounds in the sample contingent on look-up regularity. Slightly fewer than 50% of the compounds with a look-up regularity of 0 are listed, while the same number for compounds that are looked up a few times is 54%. Further, the number and proportion of listed compounds increase as the look-up regularity increases beyond 10. Among the most looked up compounds, almost all compounds are listed.

If we turn to the two sources of discordance, there are 134 unvisited entries, which means that approx. 83% of the listed lemmas in the sample are visited often enough to appear in the look-up statistics. Furthermore, 151 + 83 + 16=250 lemmas, i.e., 62% of the unlisted lemmas in the sample, are looked up and are therefore potential lacunas. A majority of these are not obvious lacunas since they reside in the 1–10 look-up regularity range, but 99 of them have a look-up regularity >10, of which 16 have a look-up regularity >100. Based on look-up regularity alone, there appears to be room for improvement of both the specificity and the sensitivity of StandMod. To ascertain whether the unlisted lemmas are in fact lacunas, we can inspect the 16 unlisted compounds with a look-up regularity exceeding 100.

> *ajourholde* 'keep up to date', *bråvåkne* 'wake suddenly', *budrunde* 'bidding round', *boforhold* 'living condition', *dybdeintervju* 'in-depth interview', *dybdelæring* 'in-depth learning', *dyptpløyende* 'that plow deep', *dødslengsel* 'lit. death longing', *dødsspiral* 'death spiral', *dømesvis* 'for example', *fortauskant* 'curb' (lit. 'pavement edge'), *forutberegnelig* 'predictable' (lit. 'precalculable'), *forutenom* 'besides' (preposition), *forutgå* 'precede' (lit. 'pre go'), *forutsaker**

Of these, only *forutsaker* is a questionable candidate since it is most probably a misspelling, although it is not obvious which word it is a misspelling of. Since the remaining 15 lemmas are perfectly acceptable dictionary entries, we may, on the basis of look-up regularity, conclude that there are at least 15 unlisted lemmas that the BOB could benefit from including. If we set the bar at minimum 10 in look-up regularity, StandMod results in between 15 and 99 lacunas, i.e., 2.5 - 17% of the compounds with look-up regularity > 10. Whether or not this is an acceptably low proportion of lacunas is hard to judge without having done the same calculation on other general-purpose dictionaries. But it is clear enough that there is at least some room for improving StandMod. In the following chapter, I will explore predictors of user interest in order to develop an alternative model.

5. An alternative model

Since StandMod results from multiple lexicographers' practices and a certain portion of its variables remains unknown (see Section 1), the best way to achieve further improvements with respect to the level of lacunas in StandMod is to first replace it with a model that can be formulated explicitly. Therefore, this chapter will proceed with an attempt to find predictors of user interest. From this I will derive a model that specifies 1) variables to consider in lexicographic selection, 2) conditions under which to consider each variable, and 3) sensible cut-off points for the selected variables in given circumstances. Finally, the derived model will be evaluated through lacuna analyses of its performance on both the data from which it was derived and an independent set of test data.

5.1. Conditional inference trees and random forests

A starting point for deriving such a model is to use conditional inference trees (henceforth *cits*) with look-up regularity as the response variable to illustrate how the variables included in this study may operate together to identify groupings of compounds that are frequently looked up, see Supplementary Material Online, Strobl et al. (2009), Tagliamonte & Baayen (2012) or Levshina (2015) for a detailed description of this method.

Cits are useful because they have the capacity to capture complex interactions between predictors (Tagliamonte & Baayen, 2012: 164). For example, it might be that DoM is a very useful predictor within a particular NO or disp range but not with others, or that NVar may help discriminate between interesting and uninteresting compounds with adjectival modifiers but not with nominal modifiers. Such interactions may be captured and visualised in a cit.

One should, however, be wary of the fact that cits may also camouflage the contribution of important predictors. The effect of one important variable may for instance be overshadowed by another. A cit-analysis will therefore benefit from a supplementary *random forest analysis*, which is also described in detail in Supplementary Material Online. A random forest analysis computes a conditional importance score for every predictor based on their association with the response variable. Hence, a random forest analysis will detect important variables that may be camouflaged in a cit-analysis.[11]

In the following cit analyses, the response is configured as logarithm to base 10 of look-up regularity + 1 (henceforth *logged look-up regularity*).

In the following, I will perform cit- and random forest analyses to single out the variables and variable interactions that should be included in a look-up based alternative model for compound selection. The results of the analyses will be tested on an independent set of data in Section 6.

5.2. Deriving a new model

5.2.1 Non- and semi-nominal compounds

Since much attention has been devoted to nominal compounds in previous studies (see e.g. Schäfer 2018), I will analyse purely nominal and semi- or non-nominal compounds separately. To begin with the latter, the tree in Figure 8 is generated with a minimum split size of 40, and a significance threshold (henceforth *alpha*) of 0.05 (cf. Supplementary Material Online).

As shown in Figure 8, the uppermost split is based on dispersion, where compounds with a disp > 0.468 are grouped on the right, and the rest are grouped on the left. We can inspect the boxplots below each of these splits (namely Nodes 3, 5 and 6 on the left and Nodes 9, 10

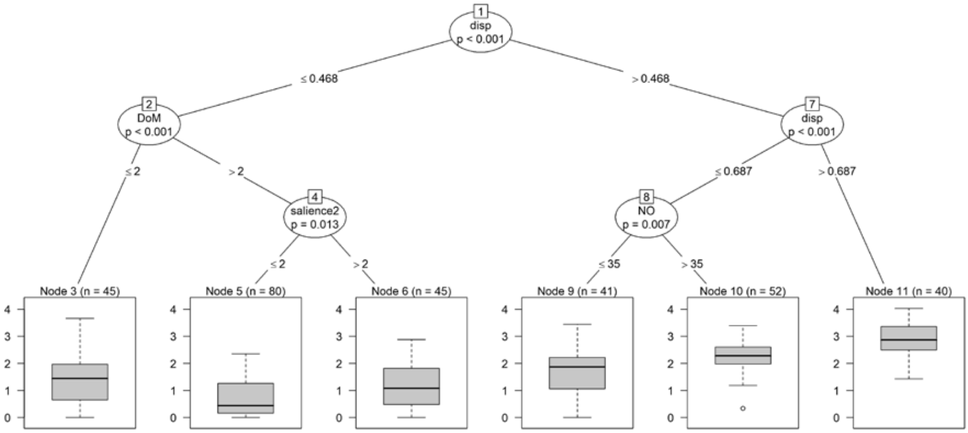

Figure 8: Conditional inference tree predicting the logged look-up regularity of non- and semi-nominal compounds at different levels of the predictors (n = 303).

and 11 on the right) and verify that the ones on the right contain compounds that on average have a higher look-up regularity. In fact, the compounds on the right all have a logged look-up regularity median close to or higher than 2, which corresponds to 99 when we reverse the logarithm. This indicates that dispersion > 0.486 among non- or semi-nominal compounds is associated with medium to high user interest. Since it is not the aim of the new model to differentiate between compounds of various high interest, dispersion > 0.468 is adopted as a qualifying criterion in the new model. See further detailed cit-analysis of non- and semi-nominal compounds with dispersion ≤ 0.468 in Supplementary Material Online. The findings from the multiple cit-analyses over non- and semi-nominal compounds suggest that we may formulate a model based on dispersion, DoM, salience and part of speech that can extract a vast majority of the non- and semi-nominal groupings that at least on a group level appear to be interesting from a user perspective. This will be tested in Section 6.

5.2.2 Nominal compounds

Nominal compounds constitute approximately 75% of the compound data. In the following, I will inspect this section of the data using cits.

Figure 9 contains a cit with a minimum split size of 60 and alpha = 0.05. The cit is complex with many splits. However, if we start by inspecting the boxplots, only two of these, Nodes 3 and 13, present groupings of compounds where the box does not include 0. The groupings in these nodes seem to contain fewer compounds that are not looked up at all than the others, and they have a range that approaches logged look-up regularity of 3 (which corresponds to 999). This latter point is also true for Node 12 and 5. I will nevertheless select the groupings in Node 3 and 13 as compounds that the new model should include, based on the fact that Node 13 has the highest median look-up regularity, and that Node 3 invokes DoM which we already saw was important for the non- and semi-nominal compounds. We may therefore remove compounds with **NO > 61 AND dispersion > 0.613**, and **NO ≤ 61 AND DoM ≤ 2** in order to inspect the remaining compounds further. This inspection involves a random forest analysis and a contingency table and is reported in Supplementary Material Online. What these analyses show is that regardless of which level one chooses, there is no combination of variables that exhaustively predicts the look-up regularity of nominal compounds. However, if one wishes to tune the variables at hand in the new model, the only way to exclude seemingly uninteresting compounds is to also exclude interesting ones, i.e., increasing the specificity of the model comes at the cost of sensitivity, and vice versa.

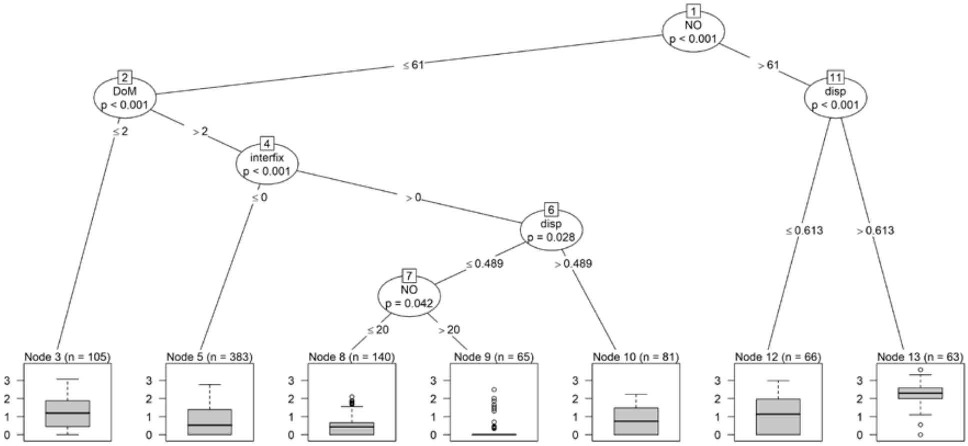

Figure 9: Conditional inference tree predicting logged look-up regularity of nominal compounds at levels of independent variables (n = 903).

The analyses conducted in Sections 5.2.1 and 5.2.2 (and detailed in Supplementary Material Online) single out variables and levels thereof that optimise the proportion of 'interesting' or 'regularly looked up' compounds accepted by the new model. This model is stated and tested in the following.

5.3. The Look-Up Predictor Model

The analyses in Section 5.2. suggest a set of very precise discriminatory levels for the most informative variables. These are accepted without rounding in LookMod that is stated in full below:

Include non- and semi-nominals with the following specifications:

- disp > 0.468
- DoM ≤ 1
- POS-h = adv, n, prep or v AND
    - Salience > 2

  OR

    - POS-m = adv, n or preposition

Include nominals with the following specifications:

- disp > 0.4
- NO > 42 AND disp > 0.507
- NO > 100 AND disp > 0.2
- NO ≤ 61 AND DoM ≤ 2
- disp > 0.2 AND interfix = 0

This model will now be tested on both the data from which it was formulated and on an independent set of compound data. Its performance will be evaluated using lacuna analysis in the same manner that StandMod was evaluated in Section 4.2.

6. Testing the Look-Up Predictor Model

In order to test LookMod, we construct an algorithm that runs through the compounds in the data and assigns the value 1 to those which fulfil one or more of its criteria and the value 0 to those which do not.

Table 2 shows the amount of included and excluded compounds in both models at different levels of look-up regularity. Using the number of unvisited entries and lacunas as performance indicators, we can now compare the performance of the two models.

LookMod produces a slightly higher number of unvisited entries (56%) than StandMod (47%). Further, the numbers for 1–10 are very similar across the two models, while StandMod shows better performance in the 10–100-range. Lastly, LookMod includes a few more of the compounds with look-up regularity > 100.

If we define desirables as compounds with a sufficient user-interest to be included in the dictionary, then the lacuna rate would be the rate of unlisted lemmas among the desirables. For the purpose of this study, I will operationalise desirables as compounds with look-up regularity > 10. This means that LookMod has 108 unlisted desirables, which gives a lacuna rate of 18.3%. In comparison, the StandMod has only a slightly better coverage with a lacuna rate of 16.8%. The performance of the StandMod is thus only a little bit more accurate than LookMod on this particular dataset, when it comes to both the number of unvisited entries and the lacuna rate. In other words, StandMod has a slightly higher sensitivity and specificity. One would however expect the StandMod to have a certain advantage since it is the product of many rounds of revision with a case-by-case consideration of each compound, while the LookMod is automatically generated using a handful of absolute criteria. On the other hand, the LookMod is based on look-up regularity which in the current procedure is also employed as the evaluation variable. On this background, while LookMod shows promising results in performing at nearly the same level as StandMod, we should perhaps require an even more convincing advantage in performance if LookMod is to become the new standard.

But there is still a potential source of error that we should control for. LookMod has so far been tested on the very same data from which it originated, which means that it might be overfitted to this particular data. To control for this, we should test the models on a different dataset and compare their performances on that data. This will be done in the following.

6.1. Testing the models on an independent dataset

The independent dataset (henceforth *testset*) is harvested using the same criteria as the initial dataset in Section 3.1 and consists of 214 compounds from the segment *gjerdesitting* – *glasur*, which along with the other segments in this study has undergone a recent revision.[12] Approximately 77% of testset is currently listed in the BOB. A lacuna analysis will now be performed to assess the performance of both StandMod and LookMod on the testset.

The performance of StandMod and LookMod with respect to look-up regularity in the testset is summarised in Table 3. The proportions of listed lemmas for each level of look-up regularity resemble the proportions of listed lemmas among the original dataset for both models. The inclusion rates are distinctly higher in the desirable levels of look-up regularity

for both models, although also a majority of the undesirable compounds are listed by both models. On basis of the testset, LookMod produces a slightly lower proportion of unvisited entries (62%) than StandMod (70%), which contrasts with the results from the original dataset. The difference is however too miniscule (4 compounds) to reflect any true difference between the models. When it comes to the other indicator of performance, namely lacunas, there are 79 desirables in the testset. StandMod captures 72 (91%) of these. In other words, it has 7 lacunas (9%), which is a lower lacuna-rate than for the original dataset. In comparison, LookMod captures 67 (85%) of the desirables, leaving 12 lacunas and a lacuna-rate of 15%. This is also lower than on the original data, but it still

Table 3: Performance of both models on testset

look-up regularity		0		1-10		10-100		>100		
		count	prop %	count	prop %	count	prop %	count	prop %	Sum
StandMod	0	14	30	27	31	6	12	1	3	48
	1	33	70	61	69	45	88	27	97	166
LookMod	0	18	38	39	44	10	20	2	7	69
	1	29	62	49	56	41	80	26	93	145

tells us that StandMod, if anything, performs slightly better than LookMod with respect to lacunas, and thereby sensitivity, for both datasets. Although it is somehow expected that StandMod should perform relatively well as it has come about through meticulous choosing and picking, and it therefore is a promising feature of LookMod that it performs at nearly the same level of specificity and sensitivity, we should demand better accuracy for LookMod if it is to become a lexicographic standard.

6.2. Summary of the performance of the models

It can be inferred from the results in the previous paragraphs that LookMod seems to be a viable starting point for collecting compounds. It picks out a vast majority of the compounds that can be viewed as desirables from a look-up perspective, while keeping the number of unvisited entries at a reasonably low level, at least if we compare it to StandMod. Although StandMod seems to be slightly more accurate with respect to both unvisited entries and lacunas, LookMod is advantageous due to the simple fact that its variables and configurations are explicitly formulated. This is not to say that we are completely in the dark with respect to the inner workings of StandMod, but it does not consist of an explicit set of criteria that can be mechanically applied to a set of compounds - which at least some of the variables in LookMod can.

Furthermore, one must expect a certain “dictionary effect” in the datasets that are utilised in this study. Since the dictionary in question is undergoing a revision, it is likely that a certain portion of the look-ups are conducted by the lexicographers working with the revision. This might cause a slight inflation of the look-up frequency of the listed lemmas, which would have a positive effect on the performance of StandMod as it is evaluated here.[13] It should however be noted that this effect is somewhat controlled for by the dispersion estimate that is incorporated in the look-up regularity variable. LookMod is affected by the dictionary effect via the look-up regularity scores that its criteria are derived from.

All in all, the primary utility for LookMod is probably not to be employed as an autonomous machine that creates lemmalists unaided by humans. And its accuracy rate does not at this point warrant a complete refurnishing of lexicographic practice. Rather, its usefulness comes as a working tool for lexicographers and lexicologists. The procedure that has been undertaken in order to generate the LookMod provides valuable information about the usefulness of the variables in question. It also indicates that some variables are not universally useful but have applicability within subsets of compounds. Among these are degree of motivation, which is an important predictor among non- and semi-nominal compounds and infrequent nominal compounds, salience which has applicability within certain parts of speech after controlling for dispersion, and number of occurrences whose primary utility is among compounds in the 0.2-0.4 dispersion range. Dispersion then, is the only variable in this study that seems to be a globally important predictor of look-regularity.

7. Concluding remarks

In this study, several methods have been applied to a material of 1206 Norwegian compounds. The aim has been to evaluate the current lemmalist of compounds in *Bokmålsordboka* and devise an alternative model for lexicographic compound selection. The Standard and Look-Up Predictor models have been evaluated using look-up statistics from the two online dictionaries *Bokmålsordboka* and *Nynorskordboka*.

One might argue that it makes little sense to operate with a set of variables that are only able to a certain extent to predict the look-up distribution of a certain compound, when one can easily consult the look-up counts directly. However, not all lexicographical or lexicological projects can benefit from an existing body of look-up information. Additionally, and perhaps more importantly, the look-up statistics only show what has been looked up in the past, and not what one can expect to be looked up in the future. Although one might expect the user interest of yesterday to resemble the user interest of tomorrow, a set of variables might be able to cover both the commonalities and the disparities between the two, whereas strict adherence to the direct look-up statistics can only cover the former.

Of course, the ideal is not to create an autonomous machine that selects compounds unaided by humans, but rather to supply lexicographers with working tools. No doubt, one can only expect that LookMod would be even more precise, and possibly outperform StandMod, if it were supplemented with the critical assessment of a group of lexicographers. Furthermore, the utility of the LookMod is not necessarily as a meticulous procedure that must be complied with to the letter, but as a sensible list of variables and levels thereof. It also conveys information about the hierarchy of variables, for instance that dispersion is a globally important variability, while DoM's and NO's utilities as variables are chiefly among slightly underdispersed compounds.

There is also something to be said for not sticking too closely to the arbitrary variations and idiosyncrasies of look-up statistics, but rather to conform to rigorous selectional variables that reflect broader patterns of look-up behaviour. Look-ups can for example be catalysed by temporarily socially relevant things such as crossword puzzles, the news cycle, seasons and public holidays (see Bäckerud, Nilsson & Sköldberg (2020) and Wolfer et al. (2014)). The BOB for instance is accessed a lot in connection with nationwide high school exams, where a given compound in the handout materials at such an exam may catalyse thousands of look-ups for that compound. Such arbitrary effects call for a moderate use of look-up frequency as an indicator of word importance and stress the importance of look-up dispersion over time.

Data on look-up behaviour in online dictionaries has a wide range of research possibilities. With respect to the question of compound selection, and especially the current unexplained variation among nominal compounds, future research could for example include a wider range of qualitative variables or distributional measurements from more than one corpus. Such adjustments might contribute to make more accurate predictions of look-up behaviour.

Finally, meta-information about the users performing the look-ups might help to uncover what needs different user groups have, and how one can design the lemmalist to meet those needs.

Notes

[1] The search logs that will be inspected for this study are drawn from the website *ordbok.uib.no* which has an interface that enables users to make parallel queries in the official Norwegian dictionaries for the two written standards of Norwegian, *Bokmål* and *Nynorsk*. The interface may therefore be utilised as a bilingual dictionary to check equivalency between the two standards.

[2] Müller-Spitzer et al. (2015) also finds that polysemic words are more frequently looked up than monosemic ones, even when controlling for the fact that the most frequent words are polysemic. I will not

include polysemy as a factor here, because the data consist of many words that are not listed in dictionaries. Therefore, I have no objective way to determine what is mono- or polysemic.

[3] This means that derivations of compounds such as *tospråklighet* "bilingualism" (lit. "twolingualness") do not count as compounds since there is no way to divide it into two individual stems because neither *språklighet* lit. "lingualness" or *-het* "-ness" are stems on their own.

[4] Leaving affixes, symbols, abbreviations and the like aside.

[5] The dictionaries were formerly accesible through an interface located at *ordbok.uib.no* and the look-up stats in question are gathered from this old interface.

[6] It should be noted that the regular expression has some caveats, for instance that people may use URLs as query expressions and thereby obscure the textual surroundings that normally enclose the search query. These shortcomings are however not expected to have any influence on the results of this study.

[7] https://ordbok.uib.no/stats/h/mest.sokt.2.html

[8] The reason why not every query is included is mainly to filter out noise and hapax legomena that only serve to slow down the computational processes involved in obtaining the look-up statistics. Besides, a handful of missed queries here and there does not alter the general tendencies of the look-up frequency variable.

[9] These two measures have opposite scales, but I have reversed the scale of DP so that it aligns with Juilland's D and Number of occurrences.

[10] More specifically, I performed a stepwise regression procedure of a generalised linear binomial model where all variables in the study, including interactions between NO, disp and the qualitative variables, were included. The procedure suggested, based on the Aikake information criterion, a model with disp as its sole predictor of BOB status. Fitting such a model showed that an increase in disp is associated with a higher likelihood of a compound being included in the dictionary. This information can also be easily obtained from Figure 5.

[11] The conditional inference tree and random forest procedure are performed by using the Cforest function of the party package in R (Hothorn et al. 2006, R Core Team 2017).

[12] The lexicographer that has edited the testset has also edited one of the segments in the original dataset.

[13] It is of course hard to pinpoint exactly how often editors visit the dictionary home page and hence how large the dictionary effect is, but from my own experience of editing BOB, the dictionary home page is visited several times a day. It is not unlikely that a single editor is responsible for 10+ look-ups of the same compound lemma.

References

A. Dictionaries

Bokmålsordboka. *Språkrådet and University of Bergen*. (ordbokene.no).

Norwegian Academy Dictionary. *Det Norske Akademi for språk og litteratur*. (naob.no).

Nynorskordboka. *Språkrådet and University of Bergen*. (ordbokene.no).

B. Other literature

Bäckerud, E., Nilsson, P. and E. Sköldberg. 2020. 'Så används Svenska Akademiens ordböcker på nätet. Implicit och explicit feedback från användarna.' *Nordiske Studier i Leksikografi* 15: 91–101.

Fjeld, R. V., Nøklestad, A. and K. Hagen. 2020. *'Leksikografisk bokmålskorpus (LBK) – bakgrunn og bruk.'* In Leksikografi og korpus. En hyllest til Ruth Vatvedt Fjeld. Edited by Johannessen, J.B. and K. Hagen. *Oslo Studies and Language* 11: 47–59.

Fjeld, R. V. and L. Vikør. 2008. *'Ord og ordbøker.'* Høyskoleforlaget.

Gries, S. T. 2008. 'Dispersions and Adjusted Frequencies in Corpora.' *International Journal of Corpus Linguistics* 13: 403–437.

Hothorn, T., Hornik, K. and A. Zeleis. 2006. 'Unbiased Recursive Partitioning: A Conditional Inference Framework.' *Journal of Computational and Graphical Statistics* 15: 651–674.

Juilland, A. G., Brodin, D. R. and C. Davidovitch. 1971. *'Frequency Dictionary of French Words.'* Mouton de Gruyter.

Kulbrandstad, L. and T. Kinn. 2016. 'Språkets mønstre'. 4th edition. Universitetsforlaget.

Leksikografisk bokmålskorpus. Distributed by the CLARINO. UiB portal: https://clarino.uib.no/lex/corpus/concordance

Levshina, N. 2015. *'How to Do Linguistics with R: Data Exploration and Statistical analysis'*. John Benjamins Publishing Company.

Müller-Spitzer, C., Wolfer, S. and A. Koplenig. 2015. 'Observing Online Dictionary Users: Studies Using Wiktionary Log Files'. *International Journal of Lexicography* 28: 1–2.

Paulsen, M. E. 2022. 'Assessing Word Commonness: Adding Dispersion to Frequency'. *International Journal of Corpus Linguistics*. (https://doi.org/10.1075/ijcl.21037.eke).

R Core Team. 2017. *'R: A Language and Environment for Statistical Computing.'* R Foundation for Statistical Computing. Vienna, Austria.

Schryver, G.-M., Joffe, D., Joffe, P. and S. Hillewaert. 2006. 'Do Dictionary Users Really Look Up frequent words? On the Overestimation of the Value of Corpus-Based Lexicography.' *Lexikos* 16: 67–83.

Schryver, G.-M. d., Wolfer, S. and R. Lew. 2019. 'The Relationship between Dictionary Look-Up Frequency and Corpus Frequency Revisited: A Log-File Analysis of a Decade of User Interaction with a Swahili-English Dictionary.' *Gema Online Journal of Language Studies* 19: 1–27.

Schäfer, M. 2018. *'The Semantic Transparency of English Compound Nouns.'* Language Science Press.

Strobl, C., Malley, J. and G. Tutz. 2009. 'An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests.' *Psychological Methods* 14.4: 323–348.

Svanlund, Jan. 2002. 'Lexikalisering.' *Språk och stil* 12. 7–45.

Tagliamonte, S. A. and R. H. Baayen. 2012. 'Models, Forests, and Trees of York English: Was/were Variation as a Case Study for Statistical Practice'. *Language Variation and Change* 24: 135–178.

Trap-Jensen, L., Lorentzen, H. and N.H. Sørensen. 2014. 'An Odd Couple – Corpus Frequency and Look-Up Frequency: What Relationship?' Slovenščina 2.0, 2.2. Edited by: Kosem, I and M. Rundell. Trojina, Institute for Applied Slovene, Slovenia: 94–113.

Wolfer, S., Koplenig, A., Meyer, P. and C. Müller-Spitzer. 2014. 'Dictionary Users Do Look Up Frequent and Socially Relevant Words. Two Log File Analyses.' Proceedings of the 16th EURALEX International Congress. Edited by Abel, A., Vettori, C., and N. Ralli. Eurac research, Bolzano: 281–290.