*Universitetet i Bergen*

*Institutt for lingvistiske, litterære og estetiske studier*

Ling350

Linguistics - Master's Thesis

# *Complexity-based Ordering of Suffixes in Norwegian*

Helge Sverre Nåvik Ulvolden

# Acknowledgements

Writing this thesis has truly been a fun and interesting process. I would first like to thank my supervisor Koenraad de Smedt for taking so much time to have regular meetings and being extremely helpful. This has made the process much easier and less stressful.

I would also like to thank my friends and family for staying in touch with me and showing interest in my project, whether it be by visiting or online. Our phone calls and gaming nights have definitely helped to keep my motivation going for the past year. And thanks to my dear Frida for keeping me company at home, showing interest and always being there for me.

I would also like to thank the dictionary projects at the University of Bergen for giving me a scholarship to write this thesis. This has enabled me to fully focus on my project without worrying about paying my electricity bills.

Bergen, May 2023

Helge Sverre Nåvik Ulvolden

# Abstract

This study examines the use of 18 derivational suffixes in Norwegian Bokmål, and aims to replicate the study done by Hay and Plag (2004). Their study aimed to explain how restrictions on affix combinations are determined by what is known as *complexity-based ordering*. They found that a suffix is more likely to be able to attach after another suffix if it is more *productive* (able to produce new words) and more *parsable* (easier for a speaker of the language to separate from its base).

In the same way, I used corpus data to find which two-suffix combinations are possible in Norwegian. I then used corpora to gather information on productivity and parsability. Productivity is determined by the proportion of hapaxes (words only occurring once in a corpus) compared to the total amount of tokens with the given suffix. Parsability is determined by comparing the frequency of derivates to the frequency of their base words, e.g., *ærlighet* ('honesty') and *ærlig* ('honest'), thus showing how likely a speaker is to interpret them as single words rather than a base and an affix.

The results of the suffix combinations were organized into a hierarchy based on which suffixes can appear after others. The results showed that unlike English, some combinations are possible in reverse order, and affix ordering therefore does not operate in such a strict hierarchy as in English. The results also showed a correlation between productivity and parsability, i.e., a productive suffix is also more easily separated from its base. A certain connection between these two factors and the suffix hierarchy can also be seen. Although it is hard to determine how connected they are, it is evident that many of the flexible suffixes that attach after others are also more productive and parsable.

# Sammendrag

Denne studien undersøker bruken av 18 avledningssuffikser i norsk bokmål, og har som mål å replikere studien gjort av Hay and Plag (2004). Studien deres forsøkte å forklare hvordan restriksjoner på affikskombinasjoner bestemmes av såkalt *kompleksitetsbasert sortering*. Den viser at et suffiks har større sannsynlighet for å plasseres etter et annet suffiks hvis det er mer *produktivt* (i stand til å produsere nye ord) og mer *parserbart* (lettere for en språkbruker å separere fra baseordet sitt).

På samme måte har jeg brukt korpusmateriale til å finne ut hvilke kombinasjoner av to suffikser som finnes i norsk. Deretter har jeg brukt korpus til å samle informasjon om suffiksenes produktivitet og parserbarhet. Produktivitet måles ved å sammenligne proporsjonen av *hapaxer* (ord som kun dukker opp én gang i et korpus) med det totale antallet ting som inneholder nevnt suffiks. Parserbarhet måles ved å sammenligne frekvensen av avledninger med frekvensen av baseord (f.eks. *ærlighet* og *ærlig*), som viser hvor sannsynlig det er at en språkbruker oppfatter dem som enkelte ord istedenfor et baseord med et suffiks.

Suffikskombinasjoner ble organisert i et hierarki basert på hvilke suffikser som kan opptre etter andre. Resultatene viser at i motsetning til i engelsk kan noen kombinasjoner også opptre i motsatt rekkefølge, og affikser kan derfor ikke organiseres i et like strengt hierarki som i engelsk. Resultatene viser også en korrelasjon mellom produktivitet og parserbarhet, dvs. at et produktivt suffiks også har lettere for å separeres fra baseordet sitt. Vi ser også en viss sammenheng mellom disse to faktorene og suffikshierarkiet. Selv om det er vanskelig å si hvor mye dette henger sammen, er det tydelig at mange av de fleksible suffiksene som kan plasseres etter andre også er mer produktive og parserbare.

# Contents

1 Introduction ........ 8

1.1 Purpose of the study ........ 8

1.2 Motivation ........ 9

1.3 Definitions ........ 9

1.3.1 Derivation ........ 9

1.3.2 Compounding ........ 10

1.3.3 Words and affixes ........ 11

1.3.4 Productivity ........ 11

1.3.5 Lexicalization ........ 12

1.4 Previous research ........ 12

1.4.1 Level ordering ........ 13

1.4.2 Selectional restrictions ........ 14

1.4.3 Complexity-based ordering ........ 14

1.4.4 Affix ordering in English ........ 15

1.4.5 Affix ordering in Norwegian ........ 17

1.4.6 My pilot study ........ 17

1.5 Hypothesis ........ 20

1.6 Conclusion ........ 21

2 Data and method ........ 23

2.1 Choosing suffixes ........ 23

2.2 Using corpus data ........ 29

2.3 Sorting suffixes in a hierarchy ........ 33

2.4 Measuring productivity ........ 34

2.5 Measuring parsability ... 36
2.6 Correlating productivity and parsability ... 39
2.7 Conclusion ... 39
3 Results ... 40
3.1 Attested combinations ... 40
3.2 Productivity ... 42
3.3 Parsability ... 44
3.4 Correlation ... 47
3.5 Conclusion ... 50
4 Conclusion ... 51
4.1 Relevance to lexicography ... 52
4.2 Further research ... 53
5 Literature ... 55
6 Appendices ... 58

## Figures

Figure 1. Example of a partial hierarchy.

Figure 2. Productivity.

Figure 3. Type parsing ratio including hapaxes.

Figure 4. Type parsing ratio excluding hapaxes.

Figure 5. Token parsing ratio including hapaxes.

Figure 6. Token parsing ratio excluding hapaxes.

## Tables

Table 1. Suffix combinations in English.

Table 2. Suffix combinations in Norwegian (pilot study).

Table 3. Suffixes and productivity (pilot study).

Table 4. Suffixes investigated.

Table 5. Attested suffix combinations.

# 1 Introduction

## 1.1 Purpose of the study

The aim of this study is to examine suffix combinations in Norwegian and what determines which combinations are possible. Some suffixes can attach to words that already contain a suffix, but this system has extensive restrictions that only allow some suffixes to combine with each other. An example can be seen in (1) and (2), showing how a suffix first attaches to a base word, forming a new word with a different meaning, which is then combined with another suffix, giving it yet a new meaning.

(1) *kjær* ('dear') + *-lig* → *kjærlig* ('affectionate')

(2) *kjærlig* + *-het* → *kjærlighet* ('love')

The study consists of three different parts. The first is to find out which combinations are possible within a list of suffixes that I have selected. This also includes an attempt to organize them into a hierarchy, based on which ones are the most *flexible* in terms of attaching after others. The next step is to examine their *productivity*, i.e., how often they are used to form new words. While some suffixes mostly appear in established derivations with lexicalized meanings, others are able to attach to many words, forming new words with transparent semantics. The final feature I am interested in is *parsability*, which is how easy it is to separate a suffix from its base word. What this means is that some words can contain suffixes that seem obscure, and the words are more likely to be interpreted as single units rather than a word with a suffix attached; meanwhile, other words are more clearly composed of a base and a suffix, and they can more easily be separated from each other.

The general idea is that these three features, *productivity*, *parsability* and *flexibility*, are all interconnected. This means that if a suffix can be found after other, already affixed words, it is also likely both to be used to form many new words, and to be separable from its bases. This study therefore aims to examine these three features and find out if there is a connection between them.

The study is based on ideas and findings by Hay and Plag (2004), who examined a list of English derivational suffixes. They found that suffixes can be organized into a hierarchy based on which suffix can attach after another, and that none of these combinations could be reversed. They also found a correlation between a suffix's position in the hierarchy and its productivity and parsability. That is, if a suffix is more flexible or appears after other flexible suffixes, it is also more likely to be productive and parsable.

## 1.2 Motivation

Understanding the underlying rules for affixation and affix ordering can teach us a lot about the structure of a language, and in particular its vocabulary. It is easy to overlook the role that morphology plays, and simply regard a language as a collection of words and not words that consist of several meaningful units. The questions examined in this paper give us an insight into how the use of affixes is not determined just by some basic rules, but rather a combination of variables like productivity and phonological and semantical transparency.

Affix ordering in Norwegian is a topic that has not been thoroughly researched before, apart from Indridason (2022), showing which combinations are possible among some Norwegian derivational suffixes (see chapter 1.4.5 for more information). It is therefore interesting to examine whether it can be defined by the same rules as in English. They are both Germanic languages with suffixes that share the same origin, and they have similar structures that contain relatively little inflection and often prefers compounding rather than derivation. Finding out if there is a correlation between productivity, parsability and the ability to attach outside other suffixes could strengthen the theory from Hay and Plag (2004) and show that this is a system that can be applied at least to other similar languages. It could also turn out that this system is unique for English, and that despite the similarity of the two languages, they operate in partly or completely different ways.

## 1.3 Definitions

Before going into detail about the previous research on the topic of affix ordering and presenting my hypothesis, it is important to define the basic terms necessary for understanding this study. In this section, I will describe a selection of fundamental terms related to word-formation, including *derivation*, *compounding*, *word*, *affix*, *productivity* and *lexicalization*.

### 1.3.1 Derivation

Derivation is a word formation process where a word is combined with an affix to create a new word. An affix can, in the case of Norwegian, be a prefix or a suffix. A prefix attaches before the base word, as in (3) whereas a suffix attaches after the base word, as in (4). Derivation differs from inflection, which is a morphological process that does not create new words (lexemes), but rather inflected forms of the same lexeme, like the verb *snakke* (‘speak’, infinitive) and its conjugated forms *snakker* (present) and *snakket* (past/perfect).

(3) *u-* + *hyggelig* (‘nice, pleasant’) → *uhyggelig* (‘unpleasant’)

(4) *kjærlig* (‘affectionate’) + *-het* → *kjærlighet* (‘love’)

There is a certain similarity between derivation and inflection, but they can be distinguished in some ways. First, derivation often creates words in a different word class from the base word, as in (5), while an inflected form of a word always belongs to the same class. In Norwegian, prefixes are always derivational. Inflection is also characterized by being much more consistent and applicable to almost every word of a class. The plural form of nouns, for example, can be applied to all nouns except mass nouns. Furthermore, inflectional suffixes are placed after derivational suffixes, as in *verdiløs* + *e* (definite singular/definite and indefinite plural marker), but not vice versa.

(5) *verdi* (‘worth’, noun) → *verdiløs* (‘worthless’, adjective)

These criteria are examples of ways to distinguish between the two categories, but they apply mostly for Norwegian and cannot be generalized to other languages. Another language would require its own description to explain the difference between derivation and inflection. In chapter 3.3 we will see examples of suffixes that operate as derivational suffixes but also share some features with inflectional suffixes. This makes it hard to determine whether they can fully be classified as derivational.

### 1.3.2 Compounding

Derivation also differs from compounding, which combines two words to form a new one, as in (6). Unlike affixes, both entities in a compound can be segmented and exist as separate words, while affixes need to be a part of a derivation. There are, however, cases where it is not quite intuitively clear if an entity is a word or an affix. Examples in Norwegian are the suffixes *-løs* and *-full*, which correspond to the English suffixes *-less* and *-ful*, respectively. These also exist as independent words, *løs* (‘loose’) and *full* (‘full’), and words containing these could therefore be interpreted as compounds instead of derivations.

(6) *brann* (‘fire’) + *mann* (‘man’) → *brannmann* (‘firefighter’)

### 1.3.3 Words and affixes

Faarlund et al. (1997, p. 59) considers the distinction between words and affixes a gradual one. Since affixes originate in words, they are more similar to their words of origin to begin with and grow apart from these words over time. An example is the prefix *hoved-* (‘main’), which originates from the noun *hode* (‘head’) and has been altered phonetically and obtained a separate meaning. *Hode* is also used in compounds, but has the literal meaning of ‘head’, as in *hodeplagg* (‘headdress’), while *hoved-* can only exist together with another word, like *hovedperson* (‘main character, main person’).

Kenesei (2007) does not regard this as a gradual process, but rather one that can be separated into four steps: word → semi-word → affixoid → affix. A morpheme’s status as one of these four steps is determined by its ability to stand on its own; a word would be able to exist without being attached to anything, while an affix is dependent on its base word. The intermediate stages have a varying degree of flexibility.

### 1.3.4 Productivity

A word formation process can be considered productive when it is used to form new words. In Norwegian, adding the suffix *-ing* after verbs to create nouns, as in example (7) is productive, just like the process of adding *-het* to form nouns from adjectives, such as (8) (Faarlund et al., 1997, p. 55). Other word formation processes can only be found in a limited set of words, like using the suffix *-de* to form nouns from adjectives, seen in (9). This cannot be considered productive, because it is not found in any new words. It is also an example of a less transparent word, which means that it is not easy to segment the different units of the word and understand them. This is because of the altered vowel of the base, and the rarely occurring suffix. *Kjøring*, on the other hand, is a transparent word because it can easily be divided into the base form *kjør-*, of the verb *kjøre*, and *-ing* is a recognizable suffix that exists in many other words that follow the same pattern.

(7) *kjøre* (‘drive’) + *-ing* → *kjøring* (‘driving’, noun)

(8) *god* (‘good’) + *-het* → *godhet* (‘goodness’)

(9) *lang* (‘long’) + *-de* → *lengde* (‘length’)

One challenging question about productivity is what we can attribute the productivity to. Is it the word formation processes, the affixes themselves, the rules or the words? Bauer (2001, p. 12) argues that productivity cannot be solely attributed to affixes, because other word formation processes without affixes such as reduplication also exist. Saussure (1969) has a different view on morphology, and considers the word itself productive, because new words are formed by analogy. That is, a word that is easily decomposable more likely contains an affix that can form new words, because this is a prerequisite for it to be productive. Bauer (2001) argues that this would mean that analogy makes the pattern productive but not the word itself, and that words can therefore not be called productive.

### 1.3.5 Lexicalization

Derivations and compounds may over time lose their original transparent meaning while diverging from the words they were made from. The suffix *-aktig* is used to form words meaning ‘similar to x’ and is mostly transparent, but the combination of *fabel* (‘fable’) and *-aktig* has given us the word *fabelaktig* (‘excellent’), with a meaning that originates from the older, more transparent ‘fable-like, relating to fables’. This means that the meaning has become lexicalized. Another example of a lexicalized word is *fordufte* (‘disappear’), created with the prefix *for*- and *dufte* (‘smell’). This bears very little relation to the original verb, and therefore has a new, lexicalized meaning.

The definitions given in this section are the most fundamental ones needed for understanding the topic of this thesis. Because my study examines word-formation processes, knowing the basic principles such as the difference between words and affixes or derivation and inflection is vital for understanding this paper.

## 1.4 Previous research

In this section, I will present the most relevant previous research that has been done about affix ordering. This includes the theory of level ordering and the later theory of selectional restrictions, which were proposed before complexity-based ordering, and are now considered insufficient to describe the complex patterns of how affixes interact with each other. I will then present the principles of complexity-based ordering, and studies examining productivity, parsability and suffix combinations, which they rely on. These studies have used both psycholinguistics and corpus-based methods to obtain a better understanding of our ability to separate affixes from their bases, and how this interacts with productivity and the affixes’ flexibility. I will also present the research that has been done about affix ordering in Norwegian. Finally, I will summarize the pilot study I did before I began writing this thesis.

### 1.4.1 Level ordering

One theory proposed in order to explain the ordering of affixes in English is called level ordering, first introduced by Siegel (1974). Here, affixes are divided into two levels for suffixes and two levels for prefixes, where an affix of class 2 will not appear inside (i.e., in the case of suffixes, before) an affix of class 1. For example, **atomlessity* is ruled out, because -*less* is a level 2 suffix, while -*ity* is a level 1 suffix. Meanwhile, the existence of words like *rationalize* can be explained by this principle, as -*al* is a level 1 suffix and -*ize* is level 2. These levels can be seen in example (10). Most of level 1 affixes are less transparent, and they are usually Latinate (i.e., of Latin or Greek origin). Level 2 affixes are more often native to English, more flexible and more transparent. The problem with level ordering as an explanation is that it does not explain combinations within a level, like *helplessness*, or exceptions where a level I affix appears outside a level II affix. It also does not explain why some combinations of suffixes from the same level are not possible, like **darknessless*.

(10)

Level I suffixes: +ion, +ity, +y, +al, +ic, +ate, +ous, +ive, +able, +ize

Level I prefixes: re+, con+, de+, sub+, pre+, in+, en+, be+

Level II suffixes: #ness, #less, #hood, #ful, #ly, #y, #like, #ist, #able, #ize

Level II prefixes: re#, sub#, un#, non#, de#, semi#, anti#

(from Spencer, 1991, p. 79)

Another problem is the lack of an explanation for why these levels exist in the first place. One explanation could be the difference in etymology, as most of the level I affixes are borrowings and most of the level II affixes are native, but this does not explain how a native speaker without etymological knowledge could distinguish between them. A possible explanation for this could be the difference in phonology, as non-native affixes can be stress-shifting and alter the base word in other ways, but even these phonological properties do not have a consistent pattern.

### 1.4.2 Selectional restrictions

Fabb (1988) suggests a different explanation, namely that affixes operate based on restrictions such as phonological, morphological and syntactical ones. Instead of sorting the affixes into levels, he divides them into the following four classes:

(11)

1. Suffixes that do not attach to already affixed words
2. Suffixes that attach outside one other suffix
3. Suffixes that attach freely
4. Problematic suffixes

This explanation has been criticized by Plag (1996); (1999), because these rules have many exceptions. There is also no clear reason why a suffix should be placed in a certain category, and they therefore seem to have been arbitrarily placed. The third argument against this is that it does not explain the restrictions on which suffixes can be combined.

### 1.4.3 Complexity-based ordering

One attempt to explain this further is known as complexity-based ordering, proposed by Hay (2000) and Hay (2002). Complex words in this context mean words that are more decomposable and can be seen as a word with a suffix attached rather than a single entity. This theory proposes that rather than organizing affixes into two levels, they can be organized into a hierarchy where the most transparent, productive and parsable affixes can be attached outside the less transparent ones. As words containing less transparent affixes tend to be analyzed as a single unit, it is logical to believe that we will find more transparent affixes attaching outside of these words.

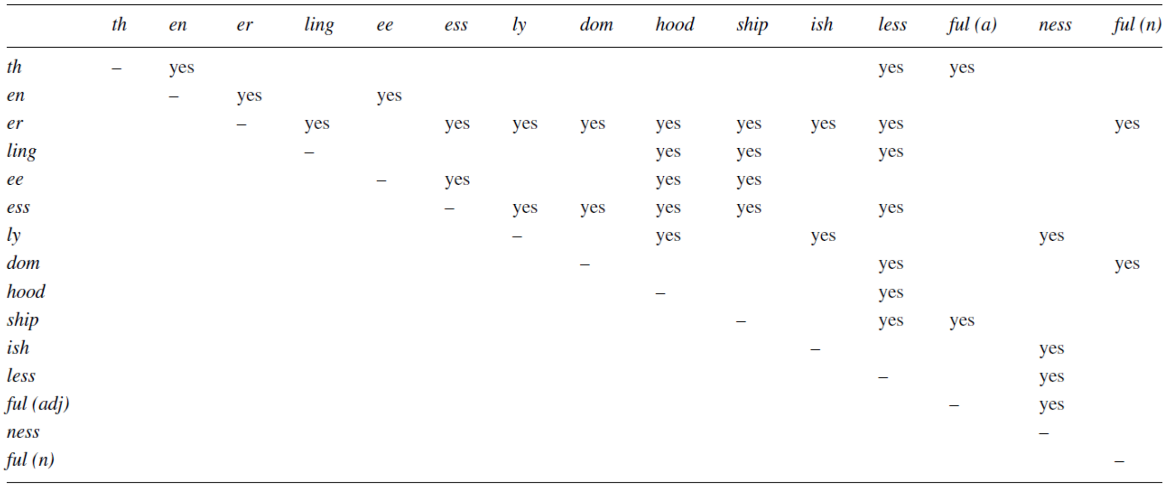
Hay (2002) found that the word *government*, for example, is more frequent than its base *govern*, and that people therefore are more likely to analyze it as a single entity than the word *discernment*, which is less frequent than its base *discern*. This explains why a construction such as *governmental* is acceptable, while **discernmental* is not. It explains why the suffix *-al* can be attached outside *-ment* in some words but not in others, which cannot be explained by level ordering or selectional restrictions.

Hay and Baayen (2001) looked at the relationship between parsing and productivity and determined how parsable a suffix was based on the proportion of derivates that were less common than their bases. That is, if more words containing *-ment* are more frequent than their bases, *-ment* is a less parsable suffix. This was then compared to the suffixes' productivity, which was measured by dividing the number of hapaxes (words that only appear once in a corpus) with the desired suffix by the total number of tokens containing it. This means that if a suffix appears in many hapaxes, it is an indicator of it regularly forming new words. What they found was a strong correlation between productivity and parsability. This has a logical explanation, as parsability should be a prerequisite for a suffix to be productive. If it is transparent and easily separable from its bases, this would also cause it to be more accessible in the mental lexicon, thus allowing us to form more new words with it.

How we perceive affixed words on a psycholinguistic level has been researched by Hay (2001), who focused on how relative frequency of a word compared to the parts it is derived from can determine its decomposability, rather than absolute frequency. This study included a psycholinguistic experiment, which contained 17 pairs of affixed words. Each pair contained the same suffix and had the same syllable count and stress pattern, but one was more frequent than the base it was derived from, and the other was less frequent. The participants were asked to decide which word from each pair was more complex. Here, a complex word is a word that can be broken into several meaningful units, such as English *writer*, composed of *write* and *-er*, while a simplex word is word that is usually analyzed as one single entity. The results showed that the words that were more frequent than their bases were rated as the less complex ones. Additionally, they examined the frequencies of derivates compared to their bases and found that derivates are generally less frequent than their bases. This also confirms what has been suggested in previous research about base and derived frequency (see Harwood & Wright, 1956).

### 1.4.4 Affix ordering in English

Hay and Plag (2004) wanted to figure out how the findings presented so far connect to suffix combinations. They investigated a selection of English suffixes, which ones of them can be combined and how this relates to productivity and parsability. They chose 15 suffixes that were either native or behaving like native suffixes, i.e., stress-neutral, and used corpus data to find which combinations were possible. They found that 36 out of 210 combinations were attested (17%), and these were rearranged to represent a partial hierarchy to test their hypothesis about complexity-based ordering, which can be seen in Table 1. The table shows that no attested combination can be found below the diagonal line, which means that no combination works in reverse and no inner suffix can be found after an outer suffix, e.g., -*ess* cannot be found after -*ly*, because -*ly* is further to the right in the hierarchy.



| | *th* | *en* | *er* | *ling* | *ee* | *ess* | *ly* | *dom* | *hood* | *ship* | *ish* | *less* | *ful (a)* | *ness* | *ful (n)* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *th* | – | yes | | | | | | | | | | yes | yes | | |
| *en* | | – | yes | | yes | | | | | | | | | | |
| *er* | | | – | yes | | yes | yes | yes | yes | yes | yes | yes | | | yes |
| *ling* | | | | – | | | | | yes | yes | | yes | | | |
| *ee* | | | | | – | yes | | | yes | yes | | | | | |
| *ess* | | | | | | – | yes | yes | yes | yes | | yes | | | |
| *ly* | | | | | | | – | | yes | | yes | | | yes | |
| *dom* | | | | | | | | – | | | | yes | | | yes |
| *hood* | | | | | | | | | – | | | yes | | | |
| *ship* | | | | | | | | | | – | | yes | yes | | |
| *ish* | | | | | | | | | | | – | | | yes | |
| *less* | | | | | | | | | | | | – | | yes | |
| *ful (adj)* | | | | | | | | | | | | | – | yes | |
| *ness* | | | | | | | | | | | | | | – | |
| *ful (n)* | | | | | | | | | | | | | | | – |

Table 1: Attested suffix combinations organized into a hierarchy, from Hay and Plag (2004)

After this was done, they also compared the results from this table to data taken from Hay and Baayen (2001) on productivity and parsability. They found a correlation between the suffixes' rank in the hierarchy and how parsable and productive they were. This confirms their hypothesis and shows that affix ordering, rather than being explained by simply organizing them into two levels, can be explained as a process where productivity and parsability determine the likelihood of an affix attaching outside other affixes.

Plag and Baayen (2009) found that this model could also be applied to a larger set of suffixes. By looking at a list of 31 suffixes and which of them can be combined, they too arranged them in a hierarchy and found a strong correlation between their rank and productivity. They also criticized Hay and Plag (2004) for limiting their set to mostly level 2 suffixes, and therefore included suffixes from level 1 in this dataset. Unlike Hay and Plag (2004), they did find several combinations below the diagonal line, showing that their model is not completely solid, although these were very rare. These rare examples were also not taken from the corpus, but from the Oxford English Dictionary, and many of them seem obsolete.

### 1.4.5 Affix ordering in Norwegian

Suffix combinations have also been researched in Norwegian. Indridason (2022) looked at adjective forming suffixes and studied which combinations of them were possible by using corpus data; like Hay and Plag (2004), he excluded non-native suffixes. The study found that out of 72 combinations, 15 of them were attested in the corpus (20.8%), which is similar to the number in English. He also discusses the reasons for some of the ordering restrictions. One reason is that some suffixes cannot attach to words that are already suffixed. They can, however, attach to other morphologically complex words, such as prefixed bases. Some suffixes can attach after others when the connector -*s*- is used, like the adjective forming suffixes -*messig* and -*aktig*, and this seems to slightly compensate for the low number of other suffix combinations.

This study did not organize the suffixes into a hierarchy, neither did it tell anything about productivity and parsability. It did, however, mention that one reason for the ordering restrictions was that combinations rarely worked in reverse, i. e. the construction *kjær-lig-het* ('love'), consisting of a base word and two suffixes, could not be rearranged to **kjær-het-lig*, even though there is no grammatical reason for this, since we do find e.g. *hel-het-lig* ('as a whole'). This shows that unlike in Hay and Plag (2004), Norwegian affix ordering is not completely acyclic. Another explanation he mentions is that in Norwegian and Nordic languages in general affix use is not very flexible, affix combinations are rarely used, and instead these languages tend to use compounds and syntactical constructions, while some languages like Bulgarian and Serbian can have up to five suffixes in one derivation (Körtvélyessy et al., 2020).

### 1.4.6 My pilot study

Before I began working on this thesis, I did a small pilot study of suffix ordering in Norwegian. This study attempted to replicate Hay and Plag (2004) for Norwegian. I chose 13 derivational suffixes that I wanted to examine, and used Norwegian Newspaper Corpus Bokmål (2020) to look up all possible combinations of these, and Corpuscle (Meurer, 2022) for collecting data to calculate their productivity. Because of the limited time, this study did not look at parsability, which would have involved finding the base word for every word containing a suffix and searching for them in a corpus.

The list of suffixes was inspired by Aronoff and Fuhrhop (2002), which was the list that Hay and Plag (2004) based their study on. This was a different though related study on suffix combinations and their restrictions. The suffixes I chose were often cognates of these English suffixes, and in both cases, they were all either native or native-like, which means that they do not change the stress of the base word.

Productivity was, like the study by Hay and Plag (2004), calculated by dividing the number of hapaxes by the total number of tokens containing the suffix. The data collected to calculate productivity had to be manually cleaned due to a lot of items that were not relevant for the study. Irrelevant tokens included words ending in a string that was identical to the suffix but actually unrelated to it, and words that were not direct derivations created with it (e.g., *menneskerettighet* ('human right') is a compound of *menneske* ('human') and *rettighet* ('right'), and not a derivation made using the suffix -*het*). Again, due to lack of time, I did not do a thorough clean-up, and the removal of samples from Nynorsk was not thorough. This might have especially affected the productivity of certain suffixes that are more predominant and maybe even highly productive in Nynorsk but very rare in Bokmål, such as -*nad* and -*dom*.

Out of all possible combinations, 28 were attested in the corpus (17.94%, a proportion which is almost the same as Hay and Plag (2004), who mention 17%). Unlike their study, this one found some combinations below the diagonal line, as seen in Table 2. These were combinations that were all possible in reverse order, a finding which shows that Norwegian has a more cyclic system of affix ordering than English. This was also indicated by Indridason (2022).

| | *-nad* | *-isk* | *-ling* | *-sel* | *-skap* | *-dom* | *-bar* | *-ing* | *-lig* | *-else* | *-het* | *-løs* | *-full* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *-nad* | - | | | | | | | | | | | | |
| *-isk* | | - | | | | | | | | | Yes | | |
| *-ling* | | | - | | | | | | | | | Yes | |
| *-sel* | | | | - | | | | | | | | Yes | Yes |
| *-skap* | | | | | - | | | | Yes | | | Yes | Yes |
| *-dom* | | | | | | - | | | Yes | | | Yes | |
| *-bar* | | | | | | | - | Yes | Yes | Yes | Yes | | |
| *-ing* | | | | | | | | - | | Yes | | Yes | Yes |
| *-lig* | | | | | | Yes | | | - | Yes | Yes | Yes | |
| *-else* | | | | | | | | | | - | | Yes | Yes |
| *-het* | | | | | | | | | Yes | | - | Yes | Yes |
| *-løs* | | | | | | | | | | | Yes | - | |
| *-full* | | | | | | | | | | | Yes | | - |

Table 2: Attested suffix combinations, organized into a hierarchy.

For productivity, I organized the results into a table where the suffixes were ranked by their position in the hierarchy. I did not do any statistical analysis of these results. Still, we can see that some of the most productive ones, *-het*, *-løs* and *-full*, were also found furthest to the right in this hierarchy, which suggests a certain connection between productivity and hierarchical position.

| Suffix | Productivity | Position in hierarchy |
|---|---|---|
| *-nad* | 0.00599455 | 1 |
| *-isk* | 0.01667748 | 2 |
| *-ling* | 0 | 3 |
| *-sel* | 0.004161712 | 4 |
| *-skap* | 0.005248885 | 5 |
| *-dom* | 0.001753019 | 6 |
| *-bar* | 0.02544407 | 7 |
| *-ing* | 0.01722017 | 8 |
| *-lig* | 0.00581091 | 9 |
| *-else* | 0.01278701 | 10 |
| *-het* | 0.02333402 | 11 |
| *-løs* | 0,04858757 | 12 |
| *-full* | 0,03928325 | 13 |

Table 3: Suffixes and their productivity, ranked by their position in the hierarchy.

Although this study lacked several vital parts of what Hay and Plag (2004) did for their study, such as an analysis of parsability and a statistical analysis of the results, we can see similar results to those in English. The percentage of possible combinations is the same, and some of the most productive suffixes are also the ones found to the right in the hierarchy. The conclusion of this study was that further work was needed to be able to more fully replicate the original study, adapted to Norwegian.

## 1.5 Hypothesis

The expected outcome of the present study is that the results will resemble those of Hay and Plag (2004), where suffixes can be organized into a hierarchy and a suffix is unlikely to appear before the ones that are to the left of it in this hierarchical table. Proving a correlation between this hierarchy and productivity and parsability is not easy but I expect that there will be a certain connection, where more flexible suffixes will also be more productive and parsable. For productivity, I expect Norwegian suffixes like *-het*, *-løs* and *-full* to be among the most productive, because they intuitively seem productive, and because that is what my pilot study showed. I also expect that suffixes that are described as unproductive in Faarlund et al. (1997), like *-else* and *-nad*, to have low productivity here as well. The same applies for suffixes that they describe as productive, like *-ing*, having a high productivity here as well.

For parsability, I expect the type and token parsing ratio to be lower the higher the productivity is, because derivates containing productive suffixes are expected to be less frequent than their bases. That is, if *-het* is highly productive and is represented in many non-established words that have been coined spontaneously, these words are also likely to be far less frequent than their bases. The same applies for unproductive suffixes; If *-nad* is unproductive and only appears in lexicalized words, these are more likely to have diverged from their bases and are more likely to have bases that have become obsolete or archaic and rare.

The predictions about productivity and hierarchical order are what my pilot study indicated, where the suffixes to the right in the hierarchy were also the most productive ones. That study did, however, find that some combinations work in reverse order, and that there is not a complete cyclicity in Norwegian affix ordering, unlike what we have seen in Hay and Plag (2004). The two suffixes *-het* and *-løs,* for example, were found in both orders, as in *kjærlighetsløs* ('love-less') and *barnløshet* ('childlessness'). A similar acyclicity was also shown in Plag and Baayen (2009), where more English suffixes were included in the study. I therefore expect to find at least the same combinations in reverse order here as in my pilot study.

The prediction that the more parsable suffixes will also be more productive is intuitive, because a certain degree of transparency should be a prerequisite for an affix to be productive. If a speaker does not analyze a word as a base and an affix, but rather one single entity, they should not be able to use this affix further to produce new words from it. This also applies to affix ordering: If an affix is not likely to be parsed and is unproductive, it should be less likely to be able to attach outside other, more productive affixes.

## 1.6 Conclusion

In this chapter, I have presented the theoretical framework that my study is based on. This includes the purpose of my study and the previous research that exists on this topic. The previous research includes the fundamental question about how affix ordering works, its first suggested explanations, level ordering and selectional restrictions, and why they are insufficient when describing why certain affixes can attach outside others in English. I have then explained the alternative complexity-based ordering hypothesis and the research in both psycholinguistics and corpus linguistics that supports it. This information as well as the research on affix combinations in Norwegian and the findings of my own pilot study give us a foundation for understanding the purpose and interest of this study.

# 2 Data and method

In this chapter, I will describe the methodology of my study in detail. This includes my use and processing of corpus data to measure productivity and parsability and to find the suffix combinations, which constitutes a very large part of this project. It also includes the statistical analysis and my reasoning behind the choice of suffixes and corpora as a tool for measuring productivity, parsability and suffix combinations.

## 2.1 Choosing suffixes

For the present study, I chose to only examine suffixes. Although prefixes are also interesting, including them would have practically meant conducting two separate studies, because prefix combinations would have to be examined separately from suffixes. Because this would have been too much work, I limited the study to suffixes and tried to examine as many of them as possible instead. In this section, I will present the 18 suffixes I have chosen (cf. Table 4) and explain why these in particular are interesting. I will also mention why some suffixes were ruled out, either if they were too hard to analyze due to morphological obscurity or due to the limitations of the corpus I used. Additionally, I will explain the uses of some of them, which ones are likely to be productive or unproductive, and reasons for this such as transparency and word-likeness.

The list of suffixes used is an extension of that used in my pilot study, inspired by Hay and Plag (2004). They chose these suffixes from Aronoff and Fuhrhop (2002) in particular because they wanted to exclude any suffixes that did not behave as native suffixes, that is, suffixes that alter the stress of the base word and only attach to non-native bases. In the same way, I chose to include only stress-neutral suffixes that attach to native bases, some of which being native to Norwegian and others being borrowings from other Germanic languages. The idea is that Norwegian has a similar system of suffixes to that of English, as they are both Germanic languages with similar grammar and many suffixes being cognates, as well as many of the same borrowings of Latin and Greek origin.

One reason why suffixes that do not behave as native should be excluded, is that analyzing them is often difficult. Although they might seem like they attach to a variety of bases and are therefore in use in Norwegian, these are often word formation processes that have already taken place in the languages they were borrowed from. The originally Latin suffix -*ment*, for example, and its French equivalent (usually pronounced /mɑŋ/ in Norwegian), are found in many words like *argument* (‘argument’), *abonnement* (‘subscription’) and *arrangement* (‘event’). These words are all, however, borrowed from Latin and French, where they have been formed from bases that either do not exist in Norwegian or have been borrowed separately. Latin *arguere* does not exist in Norwegian, while French *abonner* and *arranger* have in fact been borrowed into the language: *abonnere* and *arrangere* are Norwegian words but are not the bases for the derivates *abonnement* and *arrangement*.

The selection of suffixes was therefore restricted to native-behaving suffixes, and I attempted to include a range from both intuitively productive ones such as -*het*, -*løs* and -*full*, which can be found after many different words, to unproductive ones like -*ling*, -*sel* and -*nad*, which rarely appear. But there are certain exceptions where a suffix can be stress shifting; when *feilbar*, pronounced /ˈfæɪlbɑːɾ/, with the stress on the first syllable, is suffixed with -*lig*, it moves the stress to the second syllable: / fæɪlˈbɑːɭɪ /. This is very rare, however, and the suffix should still be considered stress neutral. The suffixes can be seen in Table 4.

| **Suffix** | **Base word class** | **Derivate word class** | **Example** | **Translation** |
|---|---|---|---|---|
| **-dom** | A, N, V | N | *syk* → *sykdom*<br>*alder* → *alderdom*<br>*spå* → *spådom* | ‘sick’, ‘disease’<br>‘age’, ‘old age’<br>‘predict’, ‘prediction’ |
| **-nad** | A, V | N | *koste* → *kostnad* | ‘cost’ (V), ‘cost’ (N) |
| **-else** | A, V | N | *spøke* → *spøkelse*<br>*stiv* → *stivelse* | ‘haunt’, ‘ghost’<br>‘stiff’, ‘amylum’ |
| **-full** | N | A | *fordom* → *fordomsfull* | ‘bias’, ‘biased’ |
| **-het** | A, N | N | *fri* → *frihet*<br>*menneske* → *menneskehet* | ‘free’, ‘freedom’<br>‘human’, ‘humanity’ |
| **-løs** | N | A | *verdi* → *verdiløs* | ‘worth’, ‘worthless’ |
| **-ling** | A, N, V | N | *lære* → *lærling*<br>*mann* → *mannsling*<br>*ussel* → *usling* | ‘learn’, ‘trainee’<br>‘man’, ‘small man’<br>‘miserable’, ‘wretch’ |
| **-lig** | A, N, V | A | *blå* → *blålig*<br>*menneske* → *menneskelig*<br>*tro* → *trolig* | ‘blue’, ‘blue-like’<br>‘human’, ‘humane’<br>‘believe’, ‘likely, believable’ |
| **-sel** | A, V | N | *føde* → *fødsel*<br>*redd* → *redsel* | ‘give birth’, ‘birth’ (N)<br>‘afraid’, ‘fear’ |
| **-skap** | A, N | N | *dum* → *dumskap*<br>*bror* → *brorskap*<br>*kjenne* → *kjennskap* | ‘stupid’, ‘stupidity’<br>‘brother’, ‘brotherhood’<br>‘know’, ‘knowledge, familiarity’ |
| **-ing** | A, N, V | N | *handle* → *handling*<br>*rar* → *raring*<br>*Voss* → *vossing* | ‘act’, ‘action’<br>‘weird’, ‘weirdo’<br>‘Voss’, ‘person from Voss’ |
| **-bar** | A, N, V | A | *brenne* → *brennbar* | ‘burn’, ‘ignitable’ |

| | | | | |
|---|---|---|---|---|
| | | | *åpen* → *åpenbar*<br>*frukt* → *fruktbar* | ‘open’, ‘obvious’<br>‘fruit’, ‘fertile’ |
| **-ert** | A, N, V | N | *kikke* → *kikkert* | ‘look’, ‘binoculars’ |
| **-aktig** | A, N, V | A | *feil* → *feilaktig*<br>*diamant* → *diamantaktig*<br>*skape* (*seg*) → *skapaktig* | ‘wrong, amiss’<br>‘diamond’, ‘diamond-like’<br>‘pose’, ‘pretentious’ |
| **-som** | A, N, V | A | *prat* → *pratsom*<br>*lang* → *langsom*<br>*hjelpe* → *hjelpsom* | ‘chat’, ‘talkative’<br>‘long’, ‘slow’<br>‘help’, ‘helpful’ |
| **-messig** | N | A | *by* → *bymessig* | ‘city, town’, ‘townlike’ |
| **-ete** | N, V | A | *rot* → *rotate*<br>*mumle* → *mumlete* | ‘mess’, ‘messy’<br>‘mumble’, ‘mumbling’ |
| **-is** | A, N, V | N | *tygge* → *tyggis* | ‘chew’, ‘chewing gum’ |

Table 4: The eighteen derivational suffixes selected for this study, with their base and derivate word classes.

Attempts to include suffixes that are more opaque and harder to decompose can be challenging. This is visible in the previously mentioned unproductive *-de*, often found together with vowel-altered bases as in *lengde* (‘length’) and *tyngde* (‘weight, gravity’) from *lang* (‘long’) and *tung* (‘heavy’), respectively. This is a suffix that is not so easily recognizable, and its derivates are likely to be interpreted as simplex monomorphemic words. Although this does not mean that it is irrelevant to this study, it would be challenging to analyze such an obscure suffix on a large scale. -*de* is not listed among the derivational suffixes in Faarlund et al. (1997), and for these reasons I chose not to include it.

An example of a suffix that is not native to Norwegian but behaves like it is *-het*, which originally did not exist in Norwegian and was borrowed from Low German *-heit, -hēt,* related to the Norwegian word *heder* (‘glory’). This suffix is found in a lot of words, and forms nouns from adjectives, usually describing abstract words like in (12), meaning “being sensitive”. The suffix is found in some established words with bases that do not occur independently, like *leilighet* (‘apartment’), where **leilig* is not a word. This might lead us to think that *-het*, being a borrowing, does not behave like a native suffix, as *leilighet* is borrowed as a whole from Low German *legelicheit*. But *-het* also occurs in many native Norwegian words that are rare and not established, like *planløshet* ('lack of plans'), which is clearly derived from *planløs* ('plan-less'), and it is therefore productive in Norwegian.

(12) *følsom* ('sensitive') + *-het* → *følsomhet* ('sensitivity')

Other productive suffixes include *-full*, *-løs*, *-messig* and *aktig*, which all have in common that their meaning is almost always completely transparent: *-full* and *-løs* come from the words *full* ('full') and *løs* ('loose') and are equivalents of the English *-ful*, meaning "having a lot of x" and *-less*, meaning "lacking x", respectively; *-aktig* is a Low German borrowing and usually means "similar to x", as in *drømmeaktig* ('dream-like'), but it can also be less transparent in words that are actually Low German borrowings, as in *delaktig* ('part-taking'), not derived directly from *del* ('part'). Furthermore, *-messig* is also a German loan and usually means "related to x" or "regarding x", for example in *værmessig* ('weather-related, regarding the weather'), but it has also produced more lexicalized derivations, such as *regelmessig* ('regular, regularly') from *regel* ('rule').

Other suffixes can be native but seemingly unproductive, like *-ling*, which forms nouns from other nouns and from adjectives and verbs, sometimes functioning as a diminutive suffix or at least forming words with a diminutive-like character, as in (13). It may also designate other properties or roles, such as (14). It is, however, not found in many words, and does not seem to form any new ones. Another example is *-sel*, which is often deverbal, as in (15), which is perhaps more productive but is mainly formed by compounding rather than derivation with this suffix, like *drapstrussel* ('death threat'), which might sound like it is derived from *drapstrue* ('threaten with death') but is more likely a compound of *drap* ('murder') + *trussel* ('threat'). Other unproductive suffixes include *-nad*, which only appears in nouns borrowed from Nynorsk, like *kostnad* ('cost'(N)).

(13) *svak* ('weak') + *-ling* → *svekling* ('weak person')

(14) *rømme* ('escape') + *-ling* → *rømling* ('escapee')

(15) *høre* ('hear') + *-sel* → *hørsel* ('hearing')

Other suffixes are not clearly productive or unproductive: *-ert* is not found in many words in the dictionary, and often has the specific meaning of a tool or a vehicle derived from a verb or a noun, as in (16) and (17). Due to the fact that it occurs rarely and is borrowed from Low German, it is not a fully productive suffix, as many of the words containing it are borrowings themselves. It is, however, sometimes used in analogical word formations, like (18), and also forms other nouns from verbs, like (19).

(16) *rope* (‘yell’) + *-ert* → *ropert* (‘megaphone’)

(17) *knall* (‘bang’) + *-ert* → *knallert* (‘moped’)

(18) *trommel* (‘drum’) + *-ert* → *trommert* (a cylinder-shaped metal box)

(19) *dukke* (‘bathe, bend’) + *-ert* → *dukkert* (‘dip, bath’)

Another suffix that might appear unproductive is *-is*, which is a Swedish borrowing and appears in some words borrowed into Norwegian, like *kjendis* (‘celebrity’). Most of the words it appears in have a colloquial and even pejorative character and are therefore unlikely to appear in a corpus based on more formal texts, although a newspaper corpus also contains interviews and citations that represent informal speech. It is also unique because it sometimes shortens the base it is derived from, as in (20).

(20) *pakistaner* (‘Pakistani’) + *-is* → *pakkis* (‘Paki’ (offensive))

As mentioned before, this study is limited to only native or native-behaving suffixes. An example of problems that appear when trying to analyze words composed with a foreign suffix is the suffix *-isk*, which is a different variant of the older Norwegian *-sk*. While *-sk* is more often used to form adjectives from Norwegian words, *-isk* is borrowed from German *-isch* and sometimes also from Latin *-icus* and Greek *-ikos* and is mostly used on loanwords (Faarlund et al., 1997, pp. 115-116; NAOB, 2022). Words containing *-isk* are often different from their supposed bases, which might not even be their bases at all. For example, we might consider *kritisk* (‘critical’) a derivate of *kritikk* (‘criticism’), but this would not work in the same way as word formations with other productive suffixes in Norwegian, as *kritikk* + *isk* would then form **kritikkisk*. *Kritisk* is in fact borrowed from German *kritisch*, which originally comes from Greek *kritikos*, while *kritikk* is a separate borrowing (NAOB, 2022).

Other words’ bases, where the word formation has in fact happened in Norwegian, might be hard to determine. For example, one might assume that *biologisk* (‘biological’) is derived by combining *biologi* (‘biology’) + *-isk*, or that they are two separate borrowings like *kritikk* and *kritisk*, but *biologisk* is actually derived from *biolog* (‘biologist’) + *isk*. However, *-isk* also occurs in combination with other suffixes that would not otherwise exist on their own, i. e. *abrahamittisk* (‘Abrahamic’), composed of *Abraham* + *-itt* + *-isk*. These examples show that massive borrowing of words containing *-isk* creates an illusion that it is in use in Norwegian or has productively been used to form new words. Given all the problems around this suffix, I therefore chose to exclude it from my study.

## 2.2 Using corpus data

This section describes how I have used corpora to create the dataset for this study. First, I will argue why corpora are better for documenting language use on a large scale, and why it is preferred over dictionaries or other methods. I will then explain how I have collected the data and processed it, which included a lot of manual cleaning to remove all the irrelevant results.

For a study that researches the use of affixes to form new words, using corpora is a quantitative method of attesting how the language is used by a large number of people. Corpora have the advantage that they contain large text collections that produce many examples, which can be generalizable to how a language is used. In comparison to introspection, which only looks at individual speakers and tests whether they can produce a construction or not, a corpus contains empirical data that introspection cannot compete with.

Using a corpus based on newspaper articles raises the question of how representative it is for the language. As newspapers are written in a formal variant of the language, they are unlikely to contain colloquial words and slang, which would also be interesting for this study, particularly for suffixes that are currently productive and form new words that are not yet well documented. A good example of this is *-is*, which creates nouns with a colloquial character. It is unlikely that a formal text in a newspaper would use informal and offensive words such as *rompis* (‘faggot’). Still, newspapers also contain citations and interviews as accounts of spoken language and can therefore be sufficiently representative for the present purposes.

For collecting data to analyze the suffixes’ productivity, I used *Aviskorpus ann.*, a corpus system developed by the Clarino Bergen Centre (Andersen, 2012; Meurer, 2022). This system provides access to a newspaper corpus with 35 692 210 grammatically annotated tokens. It does not provide morphological structure but enables searching for specific word classes and other grammatical features in combination with searching for words ending in a specific string of letters. This is useful because some derivational suffixes might be spelled the same way as other suffixes or parts of words but result in particular word classes. A search for the noun forming suffix -*ert*, for example, would also match every verb with the verbal ending -*ere* that is conjugated in the perfect tense, as in (21). By choosing to only match nouns, this removes a lot of the irrelevant results, although some words still occur that are incorrectly annotated.

(21) *redigere* (‘edit’) → *redigert* (‘edited’)

The Corpuscle search tool also allows us to search for every inflected form by just typing the lemma form, i.e., by searching for every word ending in -*ing*, the search will also match inflected forms like -*ingen*, -*inger* and -*ingene*. The corpus contains articles in both Nynorsk and Bokmål, but since I am currently researching Bokmål, I excluded every result in Nynorsk. A search in all articles written in Bokmål for every noun ending in either -*ing* or any of its inflected forms, is exemplified in (22). Unfortunately, the result list for this suffix specifically was too large to all be downloaded at once and was therefore split into three separate lists, one containing words from *a-i*, the second one containing *j-s* and the third containing *t-z* as well as *æ, ø* and *å*.

(22) `[language="nob"&lemma=".+ing" & pos="subst"]`

The use of the *lemma* function also makes this study different from my pilot study, which only looked at one form of each word. Because this function also gives us every inflected form of the words ending in a particular suffix, the sample size is obviously larger and requires more manual clean-up. A search that includes every form of a word is also more representative, since some nouns might be more frequent in the plural or definite form. For documenting the frequency of every base word, which is discussed in Section 2.5, modal verbs such as *kunne* (‘can’) and *ville* (‘want’) rarely appear as infinitives.

Although every word is annotated with its grammatical features, this corpus does not include any morphological analysis of the words. This means that a word containing a suffix will not have any information about its composition, which would have been extremely useful for this study. As some suffixes are overlapping in form, like -*ing* and -*ling*, a search for -*ling* also includes many words with bases ending in *l*, like (23). These had to be manually removed. Another suffix that could have been part of the study but had to be excluded, was the agentive *-er*, which usually forms nouns from verbs, describing a profession, as in (24), equivalent to *-er* in English. Its exclusion was due to it being identical to the indefinite plural marker, as in (25), so that a search for *-er* in the corpus includes every word with this plural ending. The same problem occurs when a suffix is identical to another word that can also be found as the second part of compounds, like *-dom*. This suffix forms nouns from verbs and other nouns, but the identical word *dom* ('judgement, sentence') can be found in many compounds, like *dødsdom* ('death sentence'), which also had to be manually removed. Clearly, a corpus with morphological annotation of words could have prevented such issues.

(23) *anbefale* ('recommend') + *-ing* → *anbefaling* ('recommendation')

(24) *lære* ('teach, learn') + *-er* → *lærer* ('teacher')

(25) *kvinne* → *kvinner* ('woman', 'women')

In order to use the data for calculating the suffixes' productivity, it had to be manually cleaned. This means removing every word that is not derived using the given suffix. For example, *frihet* ('freedom') is derived directly from a noun, as shown in (26), while *ytringsfrihet* ('freedom of speech') is a compound, seen in (27), and therefore not a derivate formed using this suffix. Again, a corpus with morphological structure annotation would have allowed automated selection and prevented this manual cleanup.

(26) *fri* ('free') + *-het* → *frihet* ('freedom')

(27) *ytring* ('utterance') + *s* + *frihet* → *ytringsfrihet* ('freedom of speech')

Furthermore, some words have endings that are identical to certain suffixes but are in fact not suffixes at all. In a corpus search for any word ending with *ing* for example, this would match words like *ting* ('thing'), which does not contain this suffix. I therefore annotated every word that was not a derivate of the relevant suffix with "0", and the derivates with "1", in a spreadsheet. Words that were spelled incorrectly were annotated with "typo", and words in other languages were annotated with the respective language, such as "Danish", "Swedish" etc. It was especially important to exclude foreign words from the data, as many of them are similar but spelled differently, and many occur only once, which would have interfered with the productivity measure (e.g., Norwegian *rettighet* vs Swedish *rättighet*).

A useful method of checking if the word is derived using a relevant suffix is to try to remove the suffix to see if the rest of the word could still function as an independent word. Removing *-het* from *frihet* gives us the word *fri*, which is a word and therefore shows that this is a derivate of that word, while removing it from *ytringsfrihet* gives us *ytringsfri* ('utterance free'), which is not a word we find in the dictionary. Although it could work as an independent word, it has another meaning and would not be the base for the derived form *ytringsfrihet*.

This method does not always work, however, because whether a word is derived from a suffix or derived from a word already containing the suffix is sometimes hard to decide. For example, the word *nytenkning* ('thinking new, originally') is clearly not derived from the base verb **nytenke* and would therefore be ruled out on this basis. Rather, it could be analyzed either as a compound in (28) or as combined compounding and derivation in

(29), with the adjective *ny* and the noun forming suffix *–(n)ing* attaching simultaneously. NAOB (2022) states that the word is analyzed as in

(29), and therefore not composed of a single base and affix, but rather a phrase and an affix.

(28) *ny* + *tenkning*

new + thinking

(29) *ny* + *tenke* + *-(n)ing*

new + think (V) + SUFF

Attesting the suffix combinations was done by looking up every combination in the corpus, in every inflected form. This too led to some matches that were irrelevant and had to be gone through manually. In many of these searches, I also had to look for a possible connector between the two suffixes. Norwegian compounds and derivations often use connectors such as *-s-* and *-e-*; for compounds this is illustrated in (30) and (31). The use of connectors is often arbitrary and does not follow a specific pattern, although some rules exist (Faarlund et al., 1997, pp. 70-74). The connector *-s-*, for example, rarely occurs after vowels and is most common after native and other Nordic suffixes. It is also common when the first part of a compound is a compound itself; *-e-* is more common when the first word is monosyllabic and is not common after compounds. When looking up the combinations of each suffix, the use of a connector needed to be considered, as some combinations require a connector. For example, when *-else* is followed by another suffix, the connector *-s-* is required, as *-else* is a Nordic suffix borrowed from Danish (NAOB, 2022). For instance, a combination of *-else* and *-aktig* would result in *-elsesaktig*, like in (32).

(30) *land* + *-s-* + *by* → *landsby*

'country, countryside' + CONN + 'town, city' → 'village'

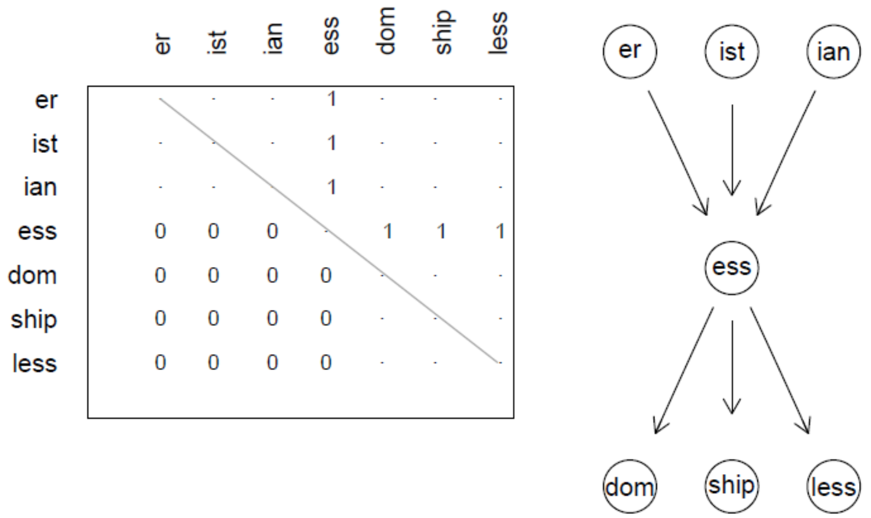
(31) *barn* + *-e-* + *hage* → *barnehage*

'child' + CONN + 'garden' → 'kindergarden'

(32) *spøkelse* ('ghost') + *-s-* + *-aktig* → *spøkelsesaktig* ('ghost-like')

## 2.3 Sorting suffixes in a hierarchy

After finding all combinations, I arranged them into a partial hierarchy. This hierarchy is not based on how many combinations a suffix participates in, but rather the possible suffix sequences. If *-het* appears after *-lig*, for example, then *-het* is hierarchically placed after *-lig*. Then, if *-løs* appears after *-het*, it is placed after *-het*, and so forth. A visual representation of such a hierarchy for English is shown in Figure 1. Some suffixes were not found in any combinations, and therefore had to be ordered arbitrarily, just like the *-er*, *-ist* and *-ian* in this figure. The same applies to suffixes that were also found in reverse combinations, such as Norwegian *-het* and *-full*. The resulting hierarchy for Norwegian will be presented in Section 3.



|  | er | ist | ian | ess | dom | ship | less |
|---|---|---|---|---|---|---|---|
| er |  | · | · | 1 | · | · | · |
| ist | · | · | · | 1 | · | · | · |
| ian | · | · |  | 1 | · | · | · |
| ess | 0 | 0 | 0 | · | 1 | 1 | 1 |
| dom | 0 | 0 | 0 | 0 | · | · | · |
| ship | 0 | 0 | 0 | 0 | · | · | · |
| less | 0 | 0 | 0 | 0 | · | · | · |

Figure 1: The panel to the left shows an example of a partial hierarchy (from Plag & Baayen, 2009). The panel to the right illustrates how this hierarchy can be visualized, where the English suffix *-ess* can appear after *-er, -ist,* and *-ian,* and before *-dom, -ship,* and *-less*. Since none of the combinations occur in reverse, they are ruled out and appear as 0s below the diagonal line. A suffix breaking the principle of the hierarchical order would appear as a 1 below the line in the left panel, and an arrow pointing upwards in the right panel.

## 2.4 Measuring productivity

Productivity is measured, following Hay and Baayen (2001), by dividing the number of hapaxes containing the given affix with the number of tokens containing it. This means that if a suffix is productive, it is likely to appear in many unique words that only appear once in a corpus, as opposed to unproductive suffixes that only appear in frequent, already established words.

There are alternative ways of measuring productivity. One could, for example, argue that high frequency of derived words with a suffix is evidence of productivity itself. The problem with this method is that it does not tell us anything about diachrony. A suffix that has been very productive in the past can appear in many well-established words and still be completely unproductive now. In the same way, a very productive suffix can be represented in few words in a corpus because few of the words it appears in are established and frequent. Aronoff (1983) argues that there is a link between lexical complexity and token frequency, which is exemplified in words ending in *-iveness* and *-ivity*. Although *-ivity* has a higher token frequency, speakers use *-iveness* to produce new words. Because words containing *-ivity* more often tend to have lexicalized meanings, this suffix becomes less accessible for the speaker to produce a new word.

Another way of measuring productivity is using data from dictionaries. This is also problematic, because dictionaries are usually limited to the established words of a language, and therefore not a good account of how a language is currently changing and forming new words. An example is the arguably productive Norwegian suffix *-aktig*, which has a very transparent meaning corresponding to English *-like* or *-ish*. This suffix can attach to a lot of words, with the meaning still being consistent. Some words are not so clear in their meaning, like *delaktig* (composed of *del* ('part') and *-aktig*) not meaning 'part-like', but 'part-taking'. These words are often the ones listed in the dictionary because of the need to define them. Other suffixes, like deverbal *-else*, are likely less productive, but still have more occurrences in the dictionary. A search for *-else* in NAOB (2022) gives us 3375 results, while a search for *-aktig* only gives us 340 results.

I therefore chose to use hapax-conditioned productivity for this study. This is not to say that this is a method without its problems. If a word only appears once in a corpus, this is no guarantee that it is a newly coined word. The unproductive suffix *-ling*, for example, is represented by very few words (17 types in this corpus), and the hapaxes in this case include *brisling* ('sprat') and *krekling* ('crowberry'), which are not newly formed words and not transparent at all. With such a low number of types and with these hapaxes, measuring the productivity using this method might cause this suffix to appear much more productive than it is. Other hapaxes include *trilling* ('triplet') and *seksling* ('sextuplet'), which might indicate that there is still some productivity, as you can take any number and add the suffix and have a transparent meaning. This could also be due to analogy of the original *tvilling* ('twin'), however, and is perhaps only applied to the numbers 1-6 (as those are the only ones that are listed in the dictionary).

To calculate the productivity, I used R, a statistical software tool (R Development Core Team, 2022). Here, I imported every TXT file with the previously cleaned data from the corpus searches and arranged them into a table sorted by frequency, showing each type and their frequency, which was useful in order to look at which words were hapaxes. This was also useful in order to find spelling errors that had been overlooked, and to remove these. These tables were also saved separately to have the full dataset available. I then divided the sum of all types that only occurred once by the total frequency of all items in the table.

## 2.5 Measuring parsability

For measuring parsability, Hay and Baayen (2001) created a dataset with all the words containing a suffix, and compared their frequencies to their base words. To evaluate how parsable a word was, they used a psycholinguistic model called Matcheck (Baayen & Schreuder, 2000; Baayen et al., 2000). This model was used to locate a *parsing line* which determines if a word is parsable or not. The *parsing ratio* was measured based on the proportion of words falling above this line, i.e., a suffix with a lot of words above this line would be considered more parsable. I did not attempt to use a model like this to calculate parsability. Instead, I focused on gathering the base frequencies and compare these to the derivates. Even without an evaluation from a model like Matcheck, I believe that we can tell a lot about a suffix's parsability just by comparing the base frequencies to derived frequencies.

To achieve this, I used the data collected for measuring the productivity, i.e., the frequency tables of each suffix. I then had to search for the base form of each word from these lists. To partly automate this process, I wrote a short Python script that created a new txt file with the suffix removed and added the rest of the search expression, including language, the *lemma* function and word class for each word. Example (33) shows the Python script for creating the list of every base for words ending in *-ing*, with the verb ending *-e* added. An example of a search for the word *anmelde* ('report') can be seen in (34). For suffixes that derive new words from verbs, I added an *e* as a replacement for the suffix, to represent the infinitive ending of the base verb. For verbs with irregular infinitive endings, I had to change this manually in the file. For suffixes that can form words from bases from several word classes, I chose the most common word class for the Python script and changed it manually for the words that belonged to another class.

(33)

```
import csv

import re

filename = "ing"

with open(filename + '.csv', 'r') as csv_file:
    csv_reader = csv.reader(csv_file)

    with open(filename + '_bases.txt', 'w') as new_file:
        csv_writer = csv.writer(new_file)
        for line in csv_reader:
            new_file.write('[language="nob" & lemma="' +
re.sub('ing','', line[0]) + 'e" & pos = "verb"]|\n')
```

(34)

```
[language="nob" & lemma="anmelde" & pos = "verb"]
```

Corpuscle was not able to search for all the base words at once, and I therefore divided the searches into intervals of 20 words at a time. Corpuscle can also show the results as a frequency list. These lists were put together into an xlsx document and sorted alphabetically, so they could then be aligned with their respective derivates.

The use of this corpus causes some problems. The *lemma* function, which is supposed to include only the base and inflected forms of a word, sometimes includes other, similar words. For example, a search for the verb *eie* ('own') will include all inflected forms such as *eier* (present), *eide* (past) and *eid* (perfect), but also includes the noun *eier* ('owner'). The search even includes three occurrences of *åtte* ('eight'), because *åtte* can also be an irregular perfect participle of the verb found in Nynorsk, even though only results in Bokmål were selected. When looking at the context of these words, they were all instances of the number, and not the verb. This means that some results probably include words that are incorrectly annotated and should be removed. There is, however, usually a small amount of them. A search for *eie*, for example, shows a total of 1210 results, 140 of which are the string *eier*, where approximately half of these are nouns. The three instances of *åtte* are not large enough to affect the results. Therefore, even though this is not an ideal way of finding the base frequencies, the lemma function was used without removing every word that is falsely annotated, since such cleanup would be unfeasible and not worth the effort.

Another question when deciding on the base form of a word is at what point the base of the derived word has diverged too much from the word it was derived from. A derivate is likely to have caused some kind of phonetic alteration to the base, even when this is not represented in spelling. An example of a more regular phonological process is when suffixes attach to verbs, and the final /ə/ is deleted, as in (35). Sometimes, the final sound of the verb can also be affected, as the merging of /ɾl/ into /ɭ/ in (36). These are changes that can be regarded as regular, and the involved verbs should therefore be regarded as the bases of the derivates.

(35) *blunke* ('blink') + *-ing* → *blunking* ('blinking')

(36) *avgjøre* (‘decide’, pronounced /ˈɑːʋjøːɾə/) + *-else* → *avgjørelse* (‘decision’, pronounced /ˈɑːʋjøːɭsə/)

Other, more difficult examples are cases where the base verb itself has undergone phonetical or phonological changes and the base in the derivate has remained frozen, like (37). Here, the verb was originally *antage*, which comes from Danish (NAOB, 2022). Other derivates have bases where the vowel has been altered, as in *krekling* (‘crowberry’), which is likely derived from *kråke* (‘crow’). Here, the original base has been altered enough for it to be unrecognizable and should therefore not be considered the base it is to be compared with. Other examples of this include derivates that are loanwords from other languages, like *støpsel* (‘plug’), which is a German loan related to the verb *stoppe* (‘stop’), but not a derivation of it. In all the cases when a word did not have a base that could function as an independent word, its field in the document was left empty so it could be counted as zero occurrences of the base word.

(37) *anta* (‘assume’) + *-else* → *antagelse* (‘assumption’)

I processed these tables created in Excel to obtain the results. First, I calculated the sum of all base frequencies and the sum of all derived frequencies of a suffix, and then divided the latter by the former. This is what is called *type parsing ratio*. Second, I divided all derived frequencies by their base frequencies, and calculated the mean of these. This is called *token parsing ratio*. It is worth noting that type and token parsing ratio in this case do not mean the same as when used by Hay and Baayen (2001). In their study, parsing ratio was determined using Matcheck, and a high parsing ratio therefore means that a high proportion of derivates are parsed. In the present study, a high parsing ratio means a high frequency of derived words compared to their bases, making the given suffix less parsable.

In the frequency tables containing base and derived frequencies there are a lot of hapaxes among the derived frequencies, and words not occurring at all among the bases. The suffix *-aktig* in particular includes a lot of hapaxes that are derived extremely *ad hoc* from names and even movie titles and English phrases. Although these are relevant for measuring hapax-conditioned productivity, they are problematic when comparing them to their bases. Examples are *«hack'n'slash»-aktig*, *Batman-aktig*, and *Disney-aktig*. Words that are derived from phrases, which have been spelled with whitespaces instead of hyphens, only include the last word of the phrase in the search result, i.e., *«Fight Club»-aktig* only includes *Club»-aktig* as the result. Therefore, searching for these bases often give no results, which might lead to an artificially high parsing ratio for -*aktig*, even though such bases are likely more frequent than the derivates.

Base words occurring 0 times includes base words that do not exist, such as **krek-* for *krekling* ('crowberry'), but existing but perhaps rare or archaic base words, such as *snerpe* ('contract, draw together') for *snerpete* ('prissy'). Because it was interesting to see how the results would change when hapaxes and bases that do not occur were removed, I created a separate table without these. I then included both the parsing ratio with and without hapaxes in the table.

## 2.6 Correlating productivity and parsability

In order to test for a correlation between productivity the four different parsing ratios, I used an R script, written with the help of Koenraad De Smedt. I first ran a Shapiro test to find out if the data were normally distributed. I then ran both a Kendall's rank correlation and a Spearman's rank correlation in order to test for a negative correlation between type parsing ratio, token parsing ratio, both with and without hapaxes, and productivity.

## 2.7 Conclusion

This chapter has discussed the data I have gathered and the method for the study. First, I discussed why I specifically chose to work with only native and native-behaving suffixes. I then explained my choice of suffixes, where I attempted to include a range of both productive and unproductive suffixes, as well as some that do not quite clearly categorize as one or the other. I have then argued why the analysis of corpus data is a better method for documenting language use than dictionaries or other methods, and how I used it to obtain the data I wanted.

Manual cleaning of these data was necessary, as they included a lot of noise and irrelevant words like compounds including suffixed words, or derivations with the suffixed word. I also used corpora to attest which suffix combinations existed, and the results were organized into a table with a hierarchical structure. After this, I have explained the method for calculating productivity and why I have chosen this rather than for example raw frequency. Using the processed data, I have then created the second dataset for calculating parsability, and manually cleaned these results so that base and derived frequencies could be compared. Finally, I used R to test for a correlation between productivity and parsability.

# 3 Results

In this chapter, I will present the results of the data I have collected and processed in chapter 2. I will first present the attested combinations and discuss possible reasons why certain combinations appear, and how these results can be organized into a hierarchy. I will then present and discuss the results of the suffixes' productivity and parsability, and the results of the correlation test. Finally, I will discuss how the suffix hierarchy relates to productivity and parsability.

## 3.1 Attested combinations

Out of all 306 possible combinations, 40 of them were found in the corpus, or 13.17%, which is less than attested in both Hay and Plag (2004), Indridason (2022) and my own pilot study. Table 5 shows all attested combinations, with the first suffix appearing in the left column, and the second in the top row. The cells on the diagonal containing a dash represent same suffix repetitions, which are ruled out because of semantical reasons. The empty cells represent unattested combinations. Like Hay and Plag (2004), I did not set a threshold on the number of attested combinations, in order to increase the chance of falsifying the hypothesis. Some combinations were only attested in one example each, like *-else* + *-bar,* found in *ansettelsesbar* ('employable'). This is an unusual example, because *-bar* normally attaches to verbs, and *ansettelse* ('employment') is a noun derived from the verb *ansette* ('employ').

| | *ert* | *is* | *ling* | *nad* | *skap* | *som* | *sel* | *ete* | *else* | *bar* | *ing* | *aktig* | *dom* | *messig* | *lig* | *het* | *løs* | *full* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *ert* | - | | | | | | | | | | | | | | | | | |
| *is* | | - | | | | | | | | | | | | | | | | |
| *ling* | | | - | | | | | | | | | | | | | | | |
| *nad* | | | | - | | | | | | | | | | Yes | | | | |
| *skap* | | | | | - | | | | | | | | | Yes | Yes | | Yes | Yes |
| *som* | | | | | | - | | | | | | | | | Yes | Yes | | |
| *sel* | | | | | | | - | Yes | | | Yes | | | Yes | | Yes | | Yes |
| *ete* | | | | | | | | - | | | | | | | | Yes | | |
| *else* | | | | | | | | | - | Yes | | Yes | | Yes | | | Yes | Yes |
| *bar* | | | | | | | | | | - | Yes | | | | Yes | Yes | | |
| *ing* | | | | | | | | | | | - | Yes | | Yes | | | Yes | Yes |
| *aktig* | | | | | | | | | | | | - | | | | Yes | | |
| *dom* | | | | | | | | | | | | | - | Yes | Yes | | Yes | |
| *messig* | | | | | | | | | | | | | | - | | Yes | | |
| *lig* | | | | | | | | | Yes | | | | Yes | | - | Yes | Yes | |
| *het* | | | | | | | | | | | | | | Yes | Yes | - | Yes | Yes |
| *løs* | | | | | | | | | | | | | | | | Yes | - | |
| *full* | | | | | | | | | | | | | | | | Yes | | - |

Table 5: Attested suffix combinations organized in a hierarchical order.

The table was organized into a hierarchy, where a suffix is placed to the right of other suffixes that it appears outside of. As some suffixes do not appear outside any other suffixes, like -*ert*, -*is* and -*ling*, they had to be organized in an arbitrary order. In the same way, -*løs* and -*het* can be found combined in both orders, and therefore also had to be ordered arbitrarily in the table. Unlike Hay and Plag (2004), some combinations are found below the diagonal, meaning that the ordering of suffixes in Norwegian is not completely acyclic. These combinations are mostly combinations that work in reverse. There is also a difference from Plag and Baayen (2009), who found very rare examples of combinations below the diagonal. While their examples were uncommon words that did not occur in corpora but rather in dictionaries, and that were old and most likely obsolete, the examples found in Norwegian were attested in the corpus and are much more frequent. Words consisting of -*lig* and -*het* and words with the same combination reversed were much more established and common. 13810 tokens contained -*lighet*, and 195 tokens contained -*hetlig*.

As expected, all the combinations below the diagonal from my pilot study seen in Table 2 were also found here. The pilot study found four combinations below the line, while this study found six combinations. The two other combinations, -*ligelse* and -*ligdom*, were also in my pilot study, but because there were fewer suffixes included, I was able to arrange the hierarchy differently and place them above the line. *-else* and *-dom* could have been positioned further to the right if they had not conflicted with *-messig*, which can be placed after both suffixes.

The low number of attested combinations can be explained partly by selectional restrictions. If a suffix can only attach to verbs, any adjective or noun-forming suffix is automatically ruled out. Therefore, the noun forming *-het*, *-nad*, *-dom* and *-else* do not appear in any combinations with each other. These restrictions also help explaining why some combinations only occur once in the entire corpus. An example of this is the combination of *-lig* and *-løs*. Since *-lig* is an adjectival suffix, it only appears in front of noun-forming suffixes in these examples. The only exception is in front of *-løs*, which forms adjectives, in the word *boligløs* ('homeless'). In this exceptional case, *bolig* ('accommodation') is a noun and not an adjective; in other words, *-lig* is not used here to form an adjective like usual. In other instances where no example of an attested combination is found, it can be explained by semantics. *-aktig* does not appear after *-lig* and vice versa, even though they both form adjectives and can appear after adjectives. This is likely because they are very close in meaning: both form words that mean "similar to X".

## 3.2 Productivity

Using the hapax-conditioned productivity measure, we can see in Figure 2 that *-aktig*, *-ete* and *-messig* are by far more productive than the other suffixes. This is also intuitive from their consistent and transparent meanings.

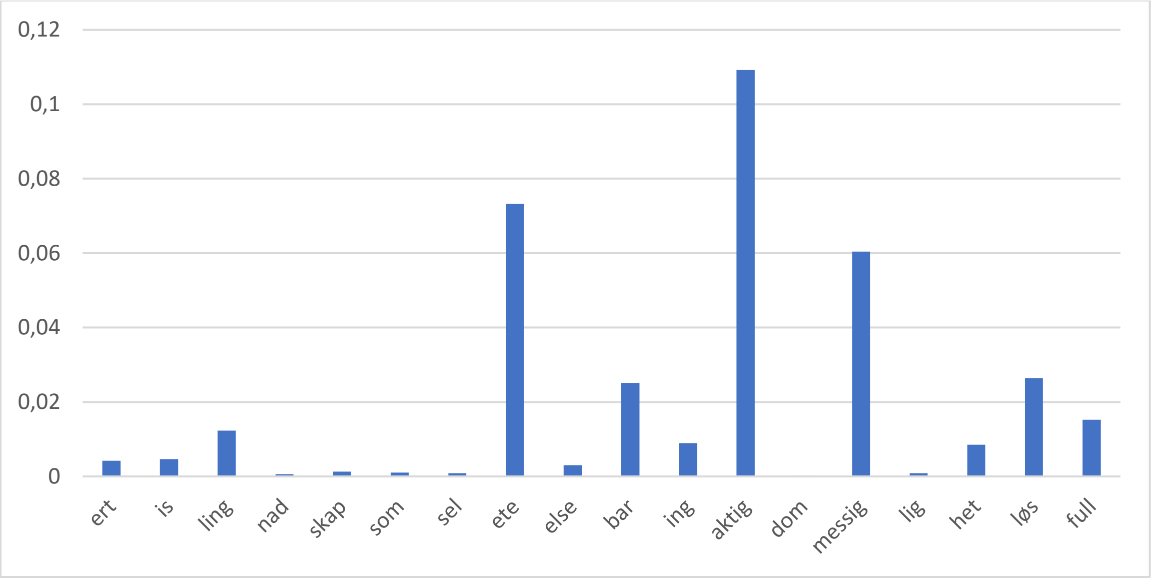


Figure 2: Suffixes ranked by their productivity, arranged in hierarchical order corresponding with Table 5.

The most productive, *-aktig*, is a good example of how a productive suffix is also more transparent and clearer in meaning. Here, most of the types are hapaxes or only appear a few times, and the types that are very frequent are also the less transparent ones such as *nøyaktig* ('precise', 625), *feilaktig* ('amiss', 281), *delaktig* ('part-taking', 93) and *fabelaktig* ('excellent', 57). Jonsbråten (2021) argues that this suffix in particular is on the border between being a word and a suffix, because of how loosely it is attached to words. She also mentions *-messig* having the same characteristics. Still, both of these do not appear as separate words, and are therefore different from suffix-like words, such as *fattig* ('poor') and *vennlig* ('friendly'), which attach to words like suffixes, to form words such as *fantasifattig* ('imagination-poor, unimaginative') and *publikumsvennlig* ('audience-friendly'). She also compares *-aktig* to *-ish*, which has been borrowed from English and has a similar meaning. *-ish* still has a different status, because it is also sometimes used as a separate word, and has a very distinct meaning compared to the English suffix. NAOB (2022) contains two separate entries for the suffix *-ish* and the adverb *ish*, and it is described by Nilssen (2015) to mostly behave as a suffix but sometimes as a word.

Another finding is that *-ling* appears to be one of the more productive suffixes, which is unlikely to be true. As mentioned before, it was represented by few types, and because some of these happened to be hapaxes despite not being newly coined words, it ranks higher than it should. This shows the weakness of measuring productivity based on hapaxes, particularly for a rare suffix with a very small sample size. The same applies for *-ert*, which only had 8 types and one hapax, and could therefore appear more productive than is actually the case. This suffix, however, did not have a high productivity.

In the data of Figure 2 we also see *-lig*, *which* has the lowest productivity of all suffixes, with 0.000886951, despite having a rather high type frequency (495). Some of the most frequent words contain bases that do not exist as independent words, like *mulig* (‘possible’) and *dårlig* (‘bad’). Other words have separable words as bases, but are not always transparent in their meaning, such as *tidlig* (‘early’), composed of *tid* (‘time’) and *-lig*.

## 3.3 Parsability

Figure 3 - Figure 6 show us the type and token parsing ratios, both including and excluding hapaxes. The results from the two different methods of measuring parsability show us that the outcome can be very different depending on which one we use. Although the results for some of the suffixes are quite similar, there are some suffixes with very different results. The type parsing ratio for *-nad* is the highest of all suffixes, while its token parsing ratio is far below average. A high frequency of derivates compared to bases is what was expected from unproductive suffixes like this, and it is therefore surprising to see a low token parsing ratio.

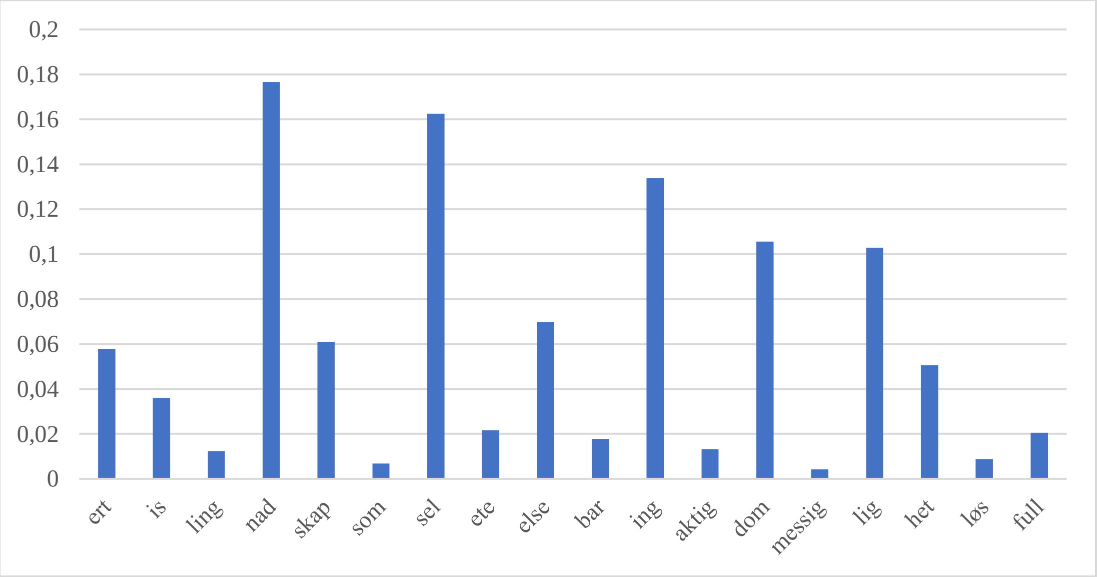


Figure 3: Type parsing ratio including hapaxes, arranged in hierarchical order corresponding with Table 5.

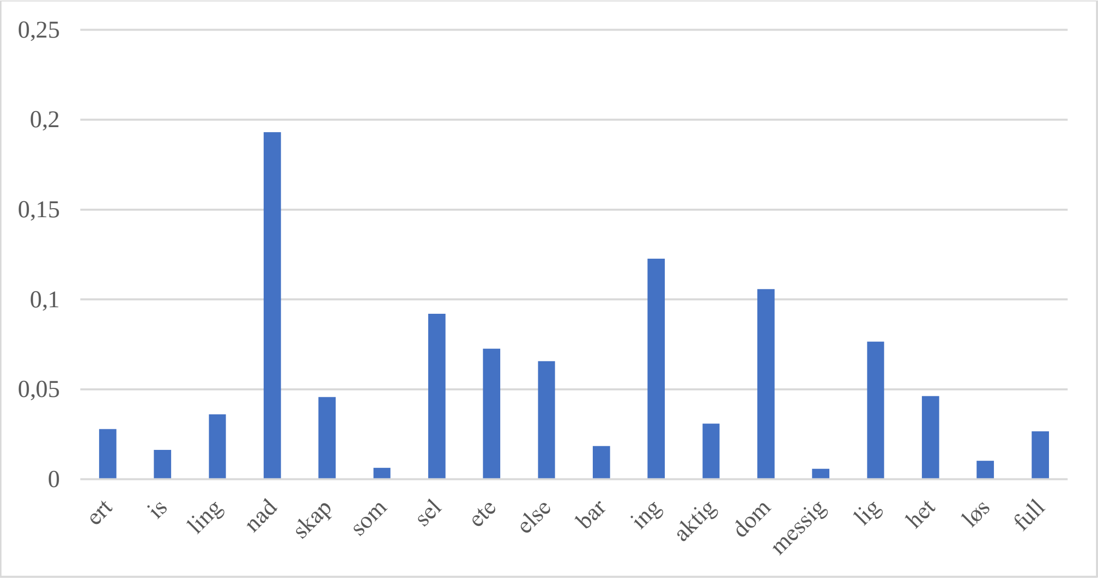


Figure 4: Type parsing ratio excluding hapaxes, arranged in hierarchical order.

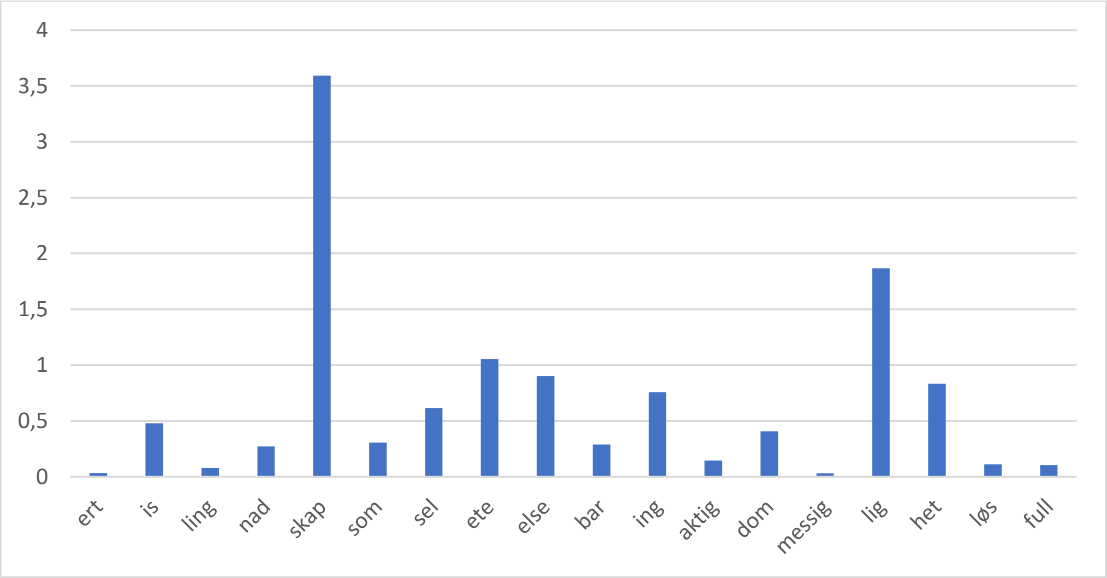


Figure 5: Token parsing ratio including hapaxes, arranged in hierarchical order.

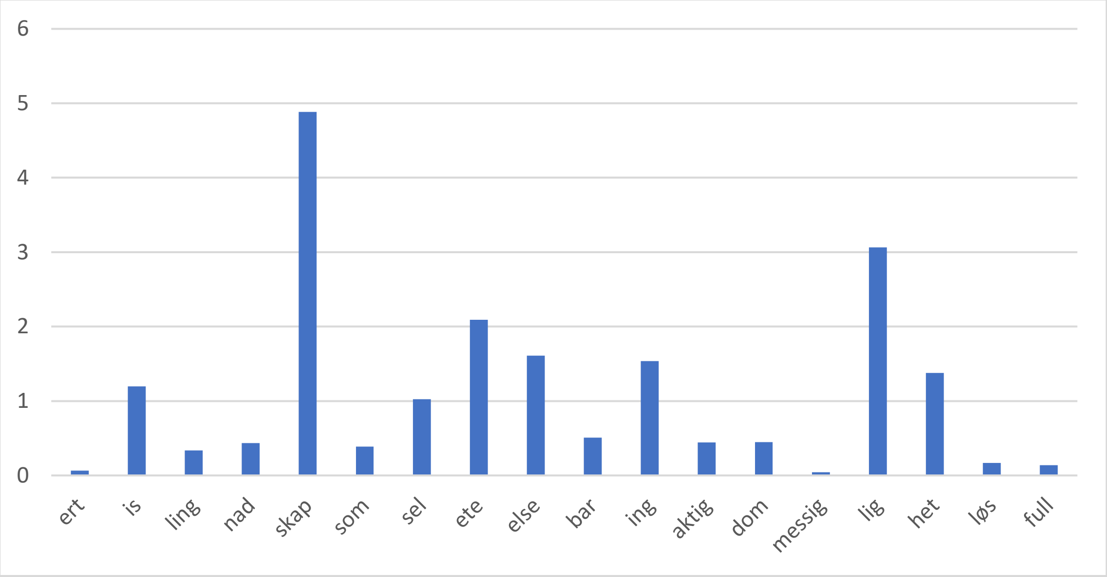


Figure 6: Token parsing ratio excluding hapaxes, arranged in hierarchical order.

Like in the case with productivity, it is likely the small sample size that is the issue. There are only five types containing -*nad*, and the rather lexicalized word *dugnad* (‘volunteer work, community work’) is much more frequent than its base *duge* (‘be useful’), with the former occurring 97 times, and the latter 67. Although one could expect that this would make the token parsing ratio much higher, it does not seem to affect the result, because the token parsing ratio is low while the type parsing ratio is very high. The high type parsing ratio could be explained by the pairs *koste*/*kostnad* (‘cost$_V$, cost$_N$’) and *søke*/*søknad* (‘apply, application’), where both the derivates and the bases have very high frequencies. This difference in token parsing ratio and type parsing ratio is also found in -*dom* and -*sel*.

One interesting observation is that -*ing*, although among the more productive suffixes, still has a very high parsing ratio. This suffix differs from the others in its unusually high number of types (3716), many of these being hapaxes and the derivations being rather transparent in meaning and form. It is therefore reasonable to assume that this is a productive suffix. One could even argue that it is somewhere between a derivational and inflectional suffix because of how the verbs mostly follow a very clear and consistent pattern. NAOB (2022) lists verbs with inflected forms containing verb + -*ing* as a verbal noun, e.g., the entry for *kjøre* shows *kjøring* as one of the inflected forms. Ordbøkene (2022) does not include this in its entries, however, and Faarlund et al. (1997, p. 104) and Leira (1992, p. 30) list it as a derivational suffix. Nouns ending in -*ing* also take inflectional suffixes like the definite singular *kjøringen*/*kjøringa*, and the suffix combinations found in this study show that it appears before several other suffixes, like *-aktig* and *-messig*.

The suffix *-ete* also shows some of these features, being very productive but still having a high parsing ratio. This suffix also attaches to verbs very consistently, with the same meaning usually meaning "doing x a lot". It does, however, also attach to nouns to form words with the meaning "having a lot of x", which makes it different from the more consistent *-ing*.

The reason why this characteristic is important is that a suffix partly behaving as an inflectional suffix can affect the frequencies. While a highly productive derivational suffix is likely to form many sporadic words that are infrequent compared to their bases, the inflected forms of a lexeme would possibly appear frequently enough to compete with its lemma form. For example, the infinitive of the verb *kjøre* ('drive') appears 1583 times, while the present *kjører* appears 863 times and the past *kjørte* appears 2233 times. The derivate *kjøring* is relatively rare compared to these, with a frequency of 190. If we were to measure parsability in the same way for inflected forms as we did for derivates, inflected verbs like this would not appear very parsable. Despite this, a language user is likely able to separate a word from its inflectional suffixes easily, considering their transparency and consistency, and their ability to be used productively. Although this example shows a low relative frequency of *kjøring*, it is somewhat hard to determine how parsable *-ing* is. Even though it clearly operates more as a derivational suffix, its function creating verbal nouns does give it a unique character that renders different results and challenges the method of measuring parsability.

## 3.4 Correlation

The results of correlation testing (cf. Appendix 1) show a negative correlation between productivity and all the other columns, both with and without hapaxes, but only the correlation between productivity and type parsing ratio (with and without hapaxes) is significant. This means that a suffix with relatively high productivity in general has a relatively low type parsing ratio, and vice versa. This is what was expected and has a logical explanation, because a suffix being parsable should be a prerequisite for a speaker to be able to recognize it as a single entity and thus use it to create new words.

As I have previously discussed in section 2.1, some suffixes that were not included in this study have created words so morphologically obscure that they can be hard to recognize as a complex word consisting of several entities rather than just a single word. An example is *-de*, as seen in the words *lengde* (‘length’) and *høyde* (‘height’), where the vowel of the former has also been altered and made the base word less recognizable. Another example from the suffixes included in my study are words such as *bunad* (a traditional Norwegian clothing), derived from the verb *bu* (‘prepare’), which was not found in the corpus. Since *bunad* has a frequency of 25, speakers are more likely to interpret this as one word, and thus less likely to use *-nad* to form new words.

The reason why we see different results in type and token parsing ratio could be explained by how each base-derivate pair can have very different parsing ratios, ultimately affecting the token parsing ratio and making it unrepresentative for that suffix. An example is the aforementioned large difference in type and token parsing ratio for *-nad*, caused by its small sample size. Other examples can be seen in *-sel*, which contains derivates with very high frequencies, such as *oppførsel* (‘behavior’), derived from *oppføre* (*seg*) (‘behave’ + reflexive pronoun). The derivate occurs 618 times, and the base 588 times, giving it a parsing ratio of 1,051020408. This number does not affect the token parsing ratio much, but the high frequency of this derivate, together with other frequent derivates, is likely the reason why we get a higher type parsing ratio.

It is also worth noting that the parsing ratio can be very high for some particular words, like *forespørsel* (‘request’$_N$) occurring 201 times, while its base *forespørre* (‘request’$_V$) occurs 15 times, leaving us with the very high number 13.4. Still, when counted together with all other frequencies, this does not seem to visibly affect the token parsing ratio. Rather, it is the high raw frequency of derived words that causes the high type parsing ratio, thus giving us the difference in type and token parsing ratio.

It is difficult to examine whether there is a correlation between parsability, productivity and a suffix’s rank in the hierarchy. The results from Hay and Plag (2004) showed a correlation between all these three factors, which means that a productive and parsable suffix would also rank higher in the hierarchy based on its ability to attach outside other suffixes. As seen in Table 5, I have arranged the suffixes in the same way, so that a suffix will appear to the right of the suffixes it can be placed after. The problem is that unlike in English, some of these combinations are also found in reverse. In these cases, the suffixes had to be ranked arbitrarily, like *-løs* and *-het*. Suffixes that do not combine with any other suffixes also had to be ordered arbitrarily. Because of this, correlating their ranks cannot be done in the same way as it was done for productivity and parsability. Instead, I will discuss how well these are connected based on the data seen in Figure 2 - Figure 6.

These data are sorted by the suffixes' ranks in the hierarchy in order to give an idea of whether or not there is a connection between the hierarchy and their ranks. In both Figure 3 and Figure 4, which show the type parsing ratio both with and without hapaxes, we can see that suffixes to the right tend to be more parsable. The most productive suffixes in Figure 2 also appear on the right side. This does indicate that there is a connection between all these three factors. Figure 5 and Figure 6, showing token parsing ratio, do not show the same tendencies, as many suffixes to the left appear very parsable. This is not surprising, considering that no significant correlation was found between productivity and token parsing ratio either.

The results in productivity show that the three suffixes *-aktig*, *-messig* and *-ete*, which were unusually productive, do not appear furthest to the right in the hierarchy. *-aktig*, which is the most productive of all the suffixes, only appears after the two suffixes *-else* and *-ing*, and for the latter there is only one occurrence (*festningsaktig*, 'fortress-like'). Its special characteristics as a unique suffix that is very loosely attached to words and even names might explain why it does not fit into this hierarchy, although this high flexibility should leave us to expect it to attach after more suffixes. *-ete*, which is the third most productive, is only found after *-sel*. Judging from the corpus data, this suffix seems to attach mostly to nouns and to some verbs, most of which are monosyllabic or disyllabic with the last syllable replaced with the suffix, as in (38). *-messig* is more liberally attached to suffixed words and is found after seven other suffixes.

(38) *rynke* ('wrinkle') → *rynkete* ('wrinkly')

Another explanation why these three suffixes are not found after more suffixes could be that their total frequency is lower than for some of the suffixes that appear further to the right in the hierarchy. -*het* and -*lig* have a type frequency of 946 and 496, respectively, while for example -the type frequency of -*aktig* is only 190. A suffix that occurs more rarely in a corpus is also likely to be found in less combinations, even if other combinations exist and are considered grammatical by speakers of a language. For example, a speaker of Norwegian would likely consider *kjærlighetsaktig* ('love-like') and *kjærlighetsmessig* ('love-related') acceptable, even though the combinations of -*het* and -*aktig* or -*messig* do not occur in this corpus. However, this does not sufficiently explain their position in the hierarchy, as other suffixes such as -*løs* have roughly the same type frequency (245) as -*aktig* (190) and -*messig* (253). -*full* is even less frequent, with a type frequency of only 92.

The other productive suffixes -*bar, -ing, -het, -løs* and -*full* all appear on the right side of the table, which is what is expected. In the same way, the unproductive suffixes -*ert, -is, -nad, -skap, -som* and -*sel* appear on the right side of the table. -*ling*, which is likely not as productive as it appears, is also on the left side. -*dom* is the least productive of all suffixes, and -*lig* is also very unproductive. Still, they appear on the right side of the hierarchy.

Figure 4, showing type parsing ratio without hapaxes, shows us that some suffixes that are parsable also appear to the right, like -*messig, -løs* and -*full*. However, some suffixes on the left are also very parsable, such as -*ert, -is, -ling*, and -*som*. There are also suffixes on the right side that are some of the least parsable, namely -*ing* and -*dom*. As I have discussed, -*ing* has a special status because its derivates are considered verbal nouns and it may therefore look like an inflectional suffix. Even though it operates more as a derivational than inflectional suffix, it shows similarities with inflectional suffixes. This can explain why it has one of the highest type parsing ratios, as inflected forms of verbs are more common proportionally to their bases than derivates are.

## 3.5 Conclusion

In this chapter I have presented the results from my processed dataset and discussed the connection between combinations, productivity and parsability. The results showed that suffixes can mostly be organized into a hierarchy similar to in English, and that some combinations appear in reverse or break the structure of this hierarchy. The correlation test showed a significant correlation between productivity and type parsing ratio, and both productivity and parsability seem to have a certain connection to the suffix hierarchy.

# 4 Conclusion

In this study I have examined the use of suffix combinations in Norwegian, and how these relate to productivity and parsability. I have selected 18 derivational suffixes and used corpora to collect the information about both combinations, productivity and parsability. The results tell us a lot about the use of suffixes in Norwegian, how a suffix's current status in the language can be shown by these three factors, and how they are interconnected.

The hypothesis for the study was that suffixes, similar to the findings of Hay and Plag (2004), could be organized into a hierarchy, where a suffix cannot appear before suffixes that appear to the left of it in this hierarchical table. The results of my pilot study had already shown that some combinations appear in reverse, and I therefore expected to find the same combinations as these. The results from this study showed similar results to those of my pilot study, with the same combinations appearing in reverse and thus breaking the structure of the hierarchy. Most of the combinations, however, do not appear in reverse. Thus, the table of combinations mostly follows the same structure as in English, the main difference being that certain productive suffixes can be combined with each other in both orders.

Furthermore, I expected to find a correlation between productivity and parsability. The results showed that there was a significant correlation between productivity and type parsing ratio, both when hapaxes were included and not included. There was no significant correlation between productivity and token parsing ratio. These results show a logical connection between these two factors, that can be explained by how we process words consisting of a base and a suffix. If a speaker is easily able to separate the suffix from the base, it is also understandable that they are more likely to use this suffix to form new words.

In the same way as I expected productivity and parsability to be correlated, I also expected to find a connection between these two factors and the suffix hierarchy. These results could not simply be correlated with a suffix's rank in the hierarchy, partly because many of them were tied for the same rank and therefore sorted arbitrarily. The results therefore had to be analyzed by looking at what the general tendency was. When comparing the hierarchical order to productivity and parsability, we can see that suffixes on the right side of the table also tend to be more productive and parsable, with some exceptions.

It is difficult to come up with a satisfying conclusion for how connected the hierarchy is with productivity and parsability. As we have seen, the three suffixes *-aktig*, *-messig* and *-ete*, which were by far the most productive, did not appear the furthest to the right in the hierarchy. *-aktig* and *-ete* in particular have a certain unique character that breaks with the principles of complexity-based ordering.

To conclude, I would argue that when applying the complexity-based ordering hypothesis to Norwegian, it is clear that suffix combinations are more cyclic than in English. It is also clear that the principles of this hypothesis are somewhat less consistent, as seen with the aforementioned suffixes. Some of the difficulties also lie in the small sample sizes. This makes it hard to determine how parsable and productive the suffixes are, as we have seen with *-ling*, *-nad*, *-ert* and *-is*. Despite these issues, this paper serves as overall evidence that complexity-based ordering as a fundamental principle exists in Norwegian as well. The link between productivity and parsability is clear, and these features contribute to explaining why a suffix can or cannot attach outside other suffixes.

## 4.1 Relevance to lexicography

This paper is written as a part of the dictionary projects at the University of Bergen, which has given me a scholarship to write about a topic that relates to lexicography. The university owns the two dictionaries *Bokmålsordboka* and *Nynorskordboka* (Ordbøkene, 2022) as well as *Norsk ordbank*, a collection of Norwegian words and their inflected forms, in cooperation with *Språkrådet*, the Norwegian language council. The two dictionaries are currently undergoing a revision, and there is also an attempt to establish a lexicography community at the university. They therefore want more research in this field, which is the background for why I have received this scholarship. In this section, I will explain how the research conducted in this paper relates to lexicography.

Because this is a study of suffixes rather than words, it is worth noting that dictionaries cover suffixes and not only words. As we have seen throughout this paper, it is not always easy to determine the boundary between an affix and a word. Affixes often originate from words, and their transformation can be regarded as either a gradual or a stepwise process. Examples include *-full* and *-løs* with their respective words of origin, and the word-like suffixes *-aktig* and *-messig*. This study helps confirming why interpreting them as words can be reasonable, as they are among both the most productive and most parsable suffixes. The results in this study can therefore help us classifying the current status of a suffix based on these variables.

This is also relevant for identifying which words should be added to the dictionary. Words containing unproductive and unparsable suffixes likely have a lexicalized meaning and a phonetically altered base and should therefore contain their own entry in a dictionary. Other, transparent words created with a parsable suffix like *-aktig* are likely understood easily by a language user as long as they have consistent meanings that can be found in many other words with the same suffix. The most lexicalized words with *-aktig* are, as mentioned in chapter 3.2, also the most frequent ones and serve as exceptions to an otherwise very consistent pattern. They also contain their own entries in Ordbøkene (2022), while words that are clear in meaning usually do not, like *barnebokaktig* ('children's book-like').

Since dictionaries are supposed to represent the modern use of a language, corpus linguistics is a useful resource for documenting language use among a broad sample of speakers. Much like word frequencies can help us decide whether a word is common enough to belong in a dictionary, a suffix's productivity, parsability and ability to participate in suffix combinations can tell us a lot about its usage. This also relates to diachronic linguistics, as a suffix's place in the gradual transformation from word to suffix and from productive to completely unproductive can be documented with the information obtained in this study.

## 4.2 Further research

This study has only researched the use of suffixes in Norwegian Bokmål, which is one of the two written forms of Norwegian. A suggestion for a future study could be to conduct the same study in Norwegian Nynorsk. Because Nynorsk is less influenced by Danish and Low German, its use of suffixes is different from Bokmål. Suffixes such as *-het* and *-else*, for example, are borrowed from Danish, and the former is of Low German origin. They are therefore more prevalent in Bokmål than Nynorsk. Other suffixes, such as *-nad* are unproductive in Bokmål but is used much more in Nynorsk. It would therefore be interesting to see whether a similar hierarchy is found, and which suffixes are the most and least productive and parsable.

Because this study used a corpus without any morphological analysis, some suffixes that have many homophones had to be excluded from the study. This included the agentive suffix *-er*, as in *lærer* (‘teacher’), corresponding to English *-er*. If a corpus with morphological annotations were available, a future study could contain a larger set of suffixes, thus being more representative for the language. Finally, another idea for a similar study is to examine Norwegian prefixes in the same way as in this study.

# 5 Literature

Andersen, G. (2012). Building a large corpus based on newspapers from the web. In *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*. John Benjamins Publishing Company.

Aronoff, M. (1983). Potential words, actual words, productivity and frequency. *Proceedings of the 13th International Congress of Linguists*, 163–171.

Aronoff, M., & Fuhrhop, N. (2002). Restricting Suffix Combinations In German And English: Closing Suffixes And The Monosuffix Constraint. *Natural Language & Linguistic Theory*, *20*, 451–490.

Baayen, R. H., & Schreuder, R. (2000). Towards a Psycholinguistic Computational Model for Morphological Parsing. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, *358*(1769), 1281-1293. http://www.jstor.org/stable/2666818

Baayen, R. H., Schreuder, R., & Sproat, R. (2000). Morphology in the Mental Lexicon: A Computational Model for Visual Word Recognition. In F. Van Eynde & D. Gibbon (Eds.), *Lexicon Development for Speech and Language Processing* (pp. 267-293). Springer Netherlands. https://doi.org/10.1007/978-94-010-9458-0_9

Bauer, L. (2001). *Morphological Productivity*. Cambridge University Press.

Faarlund, J. T., Lie, S., & Vannebo, K. I. (1997). *Norsk referansegrammatikk*. Universitetsforlaget.

Fabb, N. (1988). English Suffixation Is Constrained Only by Selectional Restrictions. *Natural Language & Linguistic Theory*, *6*(4), 527-539. http://www.jstor.org/stable/4047592

Harwood, F. W., & Wright, A. M. (1956). Statistical Study of English Word Formation. *Language*, *32*(2), 260-273.

Hay, J. (2000). *Causes and Consequences of Word Structure* [Dissertation, Northwestern University].

Hay, J. (2001). Lexical frequency in morphology: is everything relative? *Linguistics*, *39*(6), 1041-1070. https://doi.org/doi:10.1515/ling.2001.041

Hay, J. (2002). From Speech Perception to Morphology: Affix Ordering Revisited. *Language*, *78*(3), 527-555.

Hay, J., & Baayen, H. (2001). Parsing and Productivity. In *Yearbook of Morphology 2001* (pp. 203–235).

Hay, J., & Plag, I. (2004). What Constrains Possible Suffix Combinations? On the Interaction of Grammatical and Processing Restrictions in Derivational Morphology. *Natural Language & Linguistic Theory*, *22*, 565–596.

Indridason, T. G. (2022). Suffikskombinasjoner i norsk med adjektivsuffiks på første plass: Hva er tillatt, og hva er ikke tillatt? *Bergen Language and Linguistics Studies (BeLLS)*.

Jonsbråten, S. H. (2021). *Suffiks- eller ordaktig? Grammatisk og semantisk analyse av -aktig i norsk* [Master's thesis, Universitetet i Bergen]. Bergen.

Kenesei, I. (2007). Semiwords and affixoids: The territory between word and affix. *Acta Linguistica Hungarica*, *54*(3), 263-293. https://doi.org/10.1556/aling.54.2007.3.2

Körtvélyessy, L., Bagasheva, A., & Štekauer, P. (2020). *Derivational Networks Across Languages*. De Gruyter Mouton.

Leira, V. (1992). *Ordlaging og ordelement i norsk*. Samlaget.

Meurer, P. (2022). *Aviskorpus ann.* Clarino Bergen Centre. Retrieved 12, 05 from https://clarino.uib.no/korpuskel/page?page-id=korpuskel-main-page

NAOB. (2022). Retrieved 21, 10 from https://naob.no/

Nilssen, S. (2015). *"Ei raudgrøn-ish dame som meg" - Grammatisk og semantisk om det engelske (-)ish brukt i norsk* [Master's thesis, Universitetet i Bergen]. Bergen.

Norwegian Newspaper Corpus Bokmål. (2020). Retrieved 12, 05 from https://www.nb.no/sprakbanken/en/resource-catalogue/oai-clarino-uib-no-avis-plain/

Ordbøkene. (2022). Retrieved 24, 05 from https://ordbokene.no/

Plag, I. (1996). Selectional restrictions in English suffixation revisited: a reply to Fabb (1988).

Plag, I. (1999). *Morphological Productivity. Structural Constraints on English Derivation*. Mouton de Gruyter, 1999. https://doi.org/10.1007/978-94-017-3724-1_13

Plag, I., & Baayen, H. (2009). Suffix ordering and morphological processing. *Language*, *85*, 109-152.

R Development Core Team. (2022). *R: A language and environment for statistical computing*. In R Foundation for Statistical Computing. http://www.R-project.org/

Saussure, F. (1969). *Cours de linguistique générale*. Payot.

Siegel, D. (1974). *Topics in English morphology* [Dissertation, Massachusetts Institute of Technology].

Spencer, A. (1991). *Morphological theory: an introduction to word structure in generative grammar*. Oxford: Blackwell.

# 6 Appendices

## Appendix 1: Statistical tests

```
Correlation between parsing ratio (PR) and productivity
for types and tokens, with and without hapax legomena.

      Spearman's rank correlation rho

data:  Type_PR_hapax and Productivity
S = 1646, p-value = 0.001728
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.6986584


      Kendall's rank correlation tau

data:  Type_PR_hapax and Productivity
T = 33, p-value = 0.0006468
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.5686275


      Spearman's rank correlation rho

data:  Type_PR and Productivity
S = 1462, p-value = 0.03294
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.5087719


      Kendall's rank correlation tau

data:  Type_PR and Productivity
T = 45, p-value = 0.01721
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.4117647


      Spearman's rank correlation rho

data:  Token_PR_hapax and Productivity
S = 1308, p-value = 0.155
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
```

```
-0.3498452 


       Kendall's rank correlation tau

data:  Token_pr_hapax and Productivity
T = 60, p-value = 0.2291
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau 
-0.2156863 


       Spearman's rank correlation rho

data:  Token_PR and Productivity
S = 1206, p-value = 0.3266
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho 
-0.244582 


       Kendall's rank correlation tau

data:  Token_PR and Productivity
T = 63, p-value = 0.3297
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau 
-0.1764706 
```

## Appendix 2: Suffixes investigated

| Suffix | Productivity | Token parsing ratio (with hapaxes) | Type parsing ratio (with hapaxes) | Token parsing ratio | Type parsing ratio |
|---|---|---|---|---|---|
| ert | 0,004201681 | 0,032521329 | 0,057879377 | 0,064770326 | 0,027864746 |
| is | 0,004622496 | 0,478370708 | 0,035969628 | 1,195124305 | 0,016269679 |
| ling | 0,01234568 | 0,079173139 | 0,012365154 | 0,336452376 | 0,036121145 |
| nad | 0,000607903 | 0,272634047 | 0,17657793 | 0,436028774 | 0,193105959 |
| skap | 0,001283148 | 3,593154839 | 0,060936702 | 4,883145942 | 0,045721865 |
| som | 0,001027397 | 0,305945337 | 0,006863127 | 0,387503885 | 0,006246849 |
| sel | 0,00088006 | 0,614209398 | 0,162455832 | 1,023552337 | 0,091972834 |
| ete | 0,07322326 | 1,053863238 | 0,021630719 | 2,092442965 | 0,072585114 |
| else | 0,00304414 | 0,901827577 | 0,069748052 | 1,608780614 | 0,065572406 |
| bar | 0,02514336 | 0,288513144 | 0,017814901 | 0,507530514 | 0,018463224 |
| ing | 0,008989749 | 0,754687778 | 0,133759255 | 1,537596068 | 0,122653332 |
| aktig | 0,1092077 | 0,14553564 | 0,013235796 | 0,44382659 | 0,030932457 |
| dom | 0,000178955 | 0,40669493 | 0,105660019 | 0,446421027 | 0,105671426 |
| messig | 0,06042781 | 0,029567759 | 0,004225844 | 0,044704209 | 0,005762612 |
| lig | 0,000886951 | 1,865618348 | 0,10284511 | 3,062097118 | 0,076462765 |
| het | 0,008552043 | 0,833716115 | 0,050528994 | 1,376590894 | 0,046318127 |
| løs | 0,02643172 | 0,110442672 | 0,008875582 | 0,169334608 | 0,010297702 |
| full | 0,01525424 | 0,103767998 | 0,020454812 | 0,136606632 | 0,026616796 |