Coronastalgia and covidience:

A corpus linguistic study of the productivity of *corona* and *covid* as constituents in word-formation

Daniel Loen

Master's Thesis in English Linguistics

Department of Foreign Languages

University of Bergen

Fall 2022

Abstract in Norwegian

Denne masteroppgaven utforsker hvordan orddanning har skjedd under *covid-19* pandemien, og hvordan konstituentene *corona* og *covid* blir brukt i dannelsen av nye ord samt produktiviteten av prosessene som skaper de. I tillegg ble disse sammenlignet med hvordan ordet *aske* ble brukt i dannelsen av nye ord i 2010 etter et vulkanutbrudd på Island skapte problemer for flytrafikken i Norden. Dette er gjort for å kunne se om orddannelsen som har skjedd under pandemien bærer likhet til den som foregikk i 2010.

For å kunne gjøre denne sammenligningen ble data samlet inn fra *The Coronavirus Corpus* (2019-). Dette korpuset er et som samler inn data fra diverse magasiner, nyhetsartikler samt forum som inneholder ord relatert til pandemien, *covid-19* og korona. I tillegg ble et sammenligningskorpus laget ut ifra artikler fra NRK angående utbruddet på Island i 2010 for å skape et såkalt *aske*-korpus. Dataen fra *corona*-korpuset er samlet inn ved hjelp av søkeord som baserer seg på orddannelses prosessene relevante for oppgaven som da er sammensetning, derivasjon, konvertering, klipping og blanding. Dataene fra *corona*-korpuset er separert i to tidsperioder, 01-06 2020 og 01-06 2021. Denne splittelsen gjør det mulig å sammenligne hvordan produktiviteten av prosessene og konstituentene endret seg i løpet av et år.

I oppgaven ble det funnet at *corona* er generelt mer produktiv enn *covid* som konstituent i orddanning og at den mest produktive orddannelsesprosessen er blandinger. Ingen annen orddannelsesprosess enn sammensetning ble funnet i *aske*-korpuset, som indikerer at dette er den mest produktive prosessen hos NRK. I tillegg ble det funnet at av sammensetninger så er *aske* den mest produktive konstituenten hvor *corona* i 2021 dataen er den nest mest produktive konstituenten. Noen sammensetninger dannet under pandemien har lignende struktur til de dannet etter vulkanutbruddet, noe som viser at ord som refererer til store hendelser som påvirker folk fører til dannelsen av ord med lignende referanser.

Acknowledgements

Firstly, I would like to express my gratitude to my first supervisor Jacopo Romoli for supporting me in a subject outside his primary field and for doing his utmost to motivate me in a trying time. I would like to express my immense gratitude to my second supervisor Dagmar Haumann for stepping in for Jacopo, and for her exceptional assistance, encouragement, patience, guidance and advice throughout the long writing process despite various setbacks. I would like to thank my fellow students for feedback and encouragement during the university seminars. Furthermore, I am thankful to my family and friends for supporting me through this thesis. Last but not least I would like to thank my fiancé Aleksandra for her support, encouragement, care and love during this thesis. You have all helped make this thesis possible, and for that I am immensely thankful.

Daniel Loen

Bergen, November 2022

Table of contents

List of Tables:

List of Figures:

List of abbreviations

Aff – Affixation

Bld – Blending

Clp – Clipping

Cnv – Conversion

HyC – Hyphenated compound

SoC – Solid compound

SpC – Spaced compound

N-N – Noun-Noun

N-P – Noun-Preposition

A-N – Adjective-Noun

N-A – Noun-adjective

WHO – World health organization

FHI – Folkehelseinstituttet (Norwegian institute of public health)

NRK – Norsk rikskringkasting (Norwegian broadcasting corporation)

AU – Australia

CA – Canada

GB – Great Britain

HK – Hong Kong

IE – Ireland

IN – India

LK – Sri Lanka

NZ – New Zealand

US – United States

ZA – South Africa

1 Introduction

The effect of the *coronavirus* pandemic on the creation of new words through word-formation processes is considerable as many new words are created with *corona*, *covid* or the pandemic in general as reference. Other past events such as the volcanic eruption in Iceland in 2010 had a similar effect on Norwegian word formation. Comparing two such events could shed light on the similarities and differences that exist between them.

Subjects in English grammar as well as morphology are my favourite subjects at the university. In addition, exploring the link between societal changes and language shown in sociolinguistics was also interesting (Tagliamonte, 2011). Knowing about all this and living through a worldwide pandemic made me interested in writing this thesis. The eruption of the volcano Eyjafjallajökull in Iceland in 2010 affected many individuals in the northern countries of Europe. This event, much like the current pandemic, resulted in the creation of neologisms that describe various sides of this event. The focus of this thesis is on the processes that creates the neologisms, in addition to the productivity of these processes with the constituents *corona* and *covid*.

This thesis is a corpus study that uses *The Coronavirus Corpus* (Davies, 2019-) as its primary source of data as well as a self-compiled "ash-corpus" based on data collected from the NRK event section about the eruption in Iceland in 2010. The thesis collects data from *The Coronavirus Corpus* (2019-) based on search strings that are introduced in section 3.3 and collects every tenth entry from the output of these search strings.

The *coronavirus* pandemic is a large-scale event that is currently affecting millions of people in the world, the eruption of Eyjafjallajökull was also a large-scale event, but only for the northern countries. Despite this, the two events are similar in that they have affected the daily lives of individuals. This impact is likely what caused the creation of neologisms that relate to these events. In relation to the *coronavirus* pandemic there are two key words, *covid-19* and *coronavirus*. These two words have in recent times been shortened to *covid* and *corona* while they each keep their original semantic meaning. Examining and comparing the analysis based on the data collected from the *ash-corpus* and the data from *The Coronavirus Corpus* (2019-) may show how similar these two events are on various aspects as well as potentially finding consistency in word-formation in both small and large events. It also allows for analysis on which ways these two events differ. The questions "is there more diverse word-formation in the *ash-corpus* as opposed to *The Coronavirus Corpus* (2019-)?" and "which process is most frequent in the two analyses" are examples of this.

The research questions for this thesis are therefore divided into two parts. The questions that only focus on the data from *The Coronavirus Corpus* (2019-) and the questions that focus on the comparison between the analysis of *The Coronavirus Corpus* (2019-) and the *ash-corpus*.

The research questions focusing on *corona* and *covid* are as follows:

1. Is one of the two constituents *corona* and *covid* favoured by word-formation processes?

1a. If one constituent is favoured over the other, which word-formation processes favour which constituent?

1b. Which word-formation process is the most productive one in the data overall?

The research questions that focus on *aske, corona* and *covid* are as follows:

2. What differences and similarities exist between the analysis of the *ash-corpus* and the analysis of *The Coronavirus Corpus* (2019-)?

2a. Which word-formation process, if any, is the most productive in the *ash-corpus*?

2b. Are there any differences in terms of productivity between the two analyses?

The data collected from *The Coronavirus Corpus* (2019-) are limited to only the first half year of the pandemic, 01-06 2020 and one year later, 01-06 2021. The data from the *ash-corpus* are limited to only articles from NRK's event section on this volcanic eruption. These limitations are motivated by the same factor, namely the time constraint on this thesis as it would be too much data to analyse if a longer timeframe were chosen or if multiple news agencies were used.

This thesis has five chapters. The second chapter is the theory chapter and describes the theoretical background for the thesis and introduces some earlier studies that are similar to this thesis. The third chapter introduces the methodology that is used when collecting the data that is displayed in this thesis and also explains the motivation behind the choice of methodology. The fourth chapter presents the results of my study and discusses this data and its implications. In addition, this chapter discusses some data from the same timeframe in 2022 from *The Coronavirus Corpus* (2019-). In the last chapter a conclusion is drawn from the results while the research questions are answered.

1.2 General background

In January 2020 (FHI, 2020) a new virus from the family coronaviridae, known back in 1968, was discovered. The virus, which became known as the novel *coronavirus* SARS-CoV-2 began spreading rapidly all over the world. Many countries enforced restrictions in order to

inhibit the spread of the virus, but the virus still spread rapidly and soon led to a global pandemic. New words were starting to spread as quickly as the virus to refer to the "new normal" the pandemic brought with it. Even countries which were largely unaffected by the *coronavirus* were involved in the creation of new *corona*-related words because of the internet. This connection means that speakers can have a lot more interaction with one another about the *coronavirus* and its consequences, and therefore results in a large variety of new words.

Because most of the information on the internet is in the form of digital writing, it is a simple task to extract this data and produce corpora from it. In addition, most large-scale events tend to have blanket coverage in the beginning that may subside as daily life stabilizes. This is the case with the volcanic eruption in Iceland as well as the *coronavirus* pandemic.

To combat the spread of the virus, there have been set out several different measures. These vary in severity, from the use of facial masks and hand sanitizers to lockdown of countries. Curfews have been implemented in some countries and others have established periodical lockdowns where no one is allowed outside without a facial mask. These measures have one thing in common, namely that they force individuals to accommodate to a new norm. This accommodation has resulted in an increase of terms used to describe this "new normal" (Lawson, 2020). These terms refer to different elements in day-to-day life, some of the terms refer to the lockdowns themselves while other terms refer to the time blurring nature of confinement (i.e. *blursday*). Terms relating directly to the virus also appeared, such as *covidiot* which refers to individuals that ignore these curfews or guidelines. These new words are considered neologisms, which refer to an entirely new word that is increasing in frequency (Bauer et al., 2013, p.30). Seemingly any new word can be classified as a neologism, but generally only those that are used by a community and thus increase in frequency are considered as neologisms rather than nonce words.

Language seems to be affected by large scale events as seen in some studies about these types of events (e.g. Buchstaller & Mearns, 2018 and De Smedt, 2012). The event covered by De Smedt (2012) is the eruption of the Icelandic volcano Eyjafjallajökull (https://www.nrk.no/urix/flere-hundre-evakuert-pa-island-1.7079404) on the 14th of April 2010. This eruption significantly affected flight traffic for many northern countries, disrupting or even cancelling flights (https://www.nrk.no/urix/askeskyen-rammer-europa-1.7081241). During the first days of the eruption there was blanket media coverage of the event and its consequences. During this time journalists in various media tested the waters with new words which took the current predicament into consideration. During the early days of this coverage

the lexicon experienced an influx of words using the root *aske*. One of these new words is *askefast*, which means ‘ash stranded’ (stranded because of ash). This word was deemed the word of the year (Språkrådet, 2010) by Norwegians. This indicates some connection between large scale events and word-formation processes, the need for terminology that reflects the current events.

However, 10 years later most of these words have fallen out of use, and the few that are still in use refer to the event rather than acquire a new meaning. It is this specific reference that causes such words to rapidly disappear once the referenced event ends or fades from relevance. Because of this decline one may wonder if the *corona* related words will stay or if they too will disappear once the pandemic is far behind us. It seems more likely that the terms that are directly related to *corona* and covid will subside whilst terms that relate to for instance pandemics or lockdowns as a whole may stay, but only time will tell for certain. A number of studies have been conducted and may provide answers to how affected language may be by social situations. Some of these studies are mentioned and described in chapter 2.4.

2 Theoretical background

This chapter provides the theoretical background for my study on morphologically complex words from *The Coronavirus Corpus* (2019-). The following sections review the main concepts and terms related to word formation as well as word formation processes. Section 2.1 reviews the basic concepts and section 2.2 addresses the various word-formation processes in some detail using data from *The Coronavirus Corpus* (2019-) for illustration. Section 2.3 is concerned with productivity whilst section 2.4 elaborates on some of the earlier studies on similar topics.

2.1 Basic concepts in morphology

2.1.1 Inflection vs. Word-formation

In morphology there are two main processes to distinguish, i.e. word-formation and inflection. The former comprises derivation and compounding whilst the latter relates to creating new word-forms. The key difference between the two is their output. Any process involving inflection does not result in a new word, but a different word-form, which means that the word changes shape and acquires new syntactical meaning while its lexical content remains the same (Bauer at al., 2013, p.28). An example of this is tense, which in English is represented through inflection. For instance, adding the past tense *-ed* to *dream* results in its past tense form *dreamed* which has the same lexical meaning as *dream*. Some words, such as *covided* (verb) are a different word class than the perceived root *covid* (noun) despite that the only visual difference between the two words is the past tense inflectional suffix *-ed*. However, because inflection is unable to alter the category of a word and only creates new word forms it means a different process happened first. This process is conversion, which is covered later in section 2.2.4.1. Therefore, inflection is still discussed in this thesis as the presence of an inflectional suffix as in *covided* hints that another process has taken place prior to the suffixation.

Word-formation on the other hand results in a new word. In addition, sometimes this new word is of a different category than the original root, e.g. noun to adjective or adjective to intensifier. This is, however, not a requirement as some processes, such as compounding which combines two bases to form a new word without necessarily changing category. Word-formation involves a whole slew of different processes which contribute new words. These processes will be described in detail in section 2.2.

One such process is derivational affixation, where affixes attach to bases to form new words. Another process is known as conversion. This process involves a change of category without any formal marking, which means that visually the word does not necessarily change with an added affix or other constituent. Prefixes in English are all derivational, but suffixes may be either inflectional or derivational. Therefore, when a potential neologism is examined, it is important to determine whether the suffix is inflectional or derivational in order to determine the process involved. Bauer (2003) suggests three ways to tell if a suffix is derivational and inflectional. First and fastest is to determine if there is a change of category when comparing the base to the derived word. For instance, the affix *-al* may change a noun to an adjective which means it is a derivational affix. Secondly, the meaning of an inflectional affix tends to be regular in that every time a suffix such as plural *-s* is used, it always creates the same word-form, namely a plural variant. Derivational suffixes on the other hand do not necessarily create the same type of word with the same affix. Lastly, inflectional affixes are consistent in hosts, meaning that an inflectional affix that can attach to one member of a category can attach to all of them. This also means that if an inflectional affix that only attaches to verbs is found on what superficially appears to be a noun, then it is likely that the base underwent conversion prior to the suffixation. Derivational suffixes tend to be less consistent in their hosts in the sense that it may depend on stress rather than category.

2.1.2 Defining a new word

As the paper will examine how *corona* and *covid* are used in word-formation processes, it is important to define what a new word is. In order to consider a word to be "new" depends on if it has been attested in any meaningful respect before or if it has rarely ever been used. This is a dubious requirement in the sense that defining attestation is difficult, but it boils down to exploring dictionaries for its earlier use. Exploring dictionaries will not be done in this thesis as the main focus is not whether the word is a new word or not. Crucially, most terms created with either *covid* or *corona* as a base are going to be completely new because *covid* as a term did not exist prior to the pandemic whilst *corona* was rarely used in reference to the virus.

An important definition of "word" in this thesis is that the word is not designed by a corporation to be interesting or desirable, but that the word is created by normal individuals. This distinction is only relevant if there is in fact a difference between the words created by companies and normal people. There are in fact examples in the data that displays this, such as the word *covidnomics* which is likely a derivative formed by the noun *covid* + the suffix *-nomics* from *economics* and is made by a person in reference to the state of the economy after

covid. The government-created a similar word, *covidonomics,* which is likely a blend from *covid* and *economics* without its pre-antepenultimate syllable and the onset of the antepenultimate syllable. Other government words include *covidentify* referring to an app, *covishield* which refers to a vaccine and *coronabond* which is an idea for multiple countries to pool resources to minimize interest rates of loans.

Generally, the two terms used to describe new words are *neologism* and *nonce word*. Bauer et al. (2013, p.30) prefer to only use *neologism* and argue that the two words are not distinct from a morphological point of view. This is because usually a nonce word is defined as a word used but not institutionalized whilst a neologism is a new word that becomes a part of the community. According to this definition, then if a nonce word becomes commonly used and accepted it becomes a neologism. For instance, the term *quark* was a nonce word used in the novel *Finnegans Wake* (Joyce, 1939, p.383) but later became the term for the subatomic particle. This means that whether a word is a neologism or a nonce word does not depend on a characteristic of the word, but rather when it is examined. In the paper by Bakhmat et al. (2021) "nonce words" are listed as a mechanism that creates new words (Bakhmat et al., 2021, p.134), and while new words can be nonce words, the paper also lists new words as neologisms though the two definitions are one and the same according to Bauer et al. (2013). In this thesis, only the term neologism is utilised because determining if a word does not catch on, i.e. is a nonce word, would be difficult to determine through analysis of hapaxes. Additionally, all the words in the data are from news agencies and I would therefore consider the words institutionalised to some degree.

2.2 Word-formation processes

The main focus of this section is word formation processes. Section 2.2.1 focuses on compounding and 2.2.2 focuses on blending. Furthermore, 2.2.3 explores derivation and 2.2.4 focuses specifically on affixation. Lastly, section 2.2.5 focuses on other minor processes of derivation.

2.2.1 Compounding

The process of compounding is described by Bauer et al. (2013) "[…] as the formation of words through the combination of bases" (Bauer et al., 2013, p.431). An underlying issue relating to compounds is defining the ways in which to separate it from that of syntactic phrases. A syntactic phrase is a phrase that contains a head and modifiers or complements.

For instance, the Noun Phrase (henceforth NP) is a phrase with a noun as head, which can take modifiers and complements. There are multiple compound types that are important to take note of due to their difference in grammatical category as well as their appearance. The most common type of compound consists of two nouns, henceforth the N-N compound. Words such as *coronaracism* or *covidcard* are examples of N-N compounds. Another type of compound is the adjective noun compound (henceforth A-N compound) which involves an adjective as the first element and a noun as the second element, such as in *short wave*. In isolation these compounds may be hard to distinguish from NPs with a premodifier. Compounds can be formed with most syntactic categories, but some types of compounds do not occur through compounding.

Compounds which are not created through compounding, but other processes are generally known to be "non-canonical compounds" by Bauer et al. (2013) and tend to be in the form of prepositional compounds such as those formed with two prepositions (henceforth P-P compound, not to be confused with Prepositional Phrases), e.g. *into*, or formed with a noun and a preposition (henceforth N-P compound, not to be confused with Noun Phrases), e.g. *year-in* (Bauer et al., 2013, pp.452-453). One example of a process that results in non-canonical compounds is univerbation which fundamentally refers to two or more words merging together "due to their frequent adjacent co-occurrence in discourse." (Bauer et al., 2013, p.442). Those formed by regular means, i.e. the combination of two or more bases, are known as *canonical compounds*. The grammatical properties of these compounds are determined by the right-most element, which for this reason is often called the head of the compound (Bauer et al., 2013, p.443) i.e. *canonical compounds* are right-headed. With this in mind, I have decided to not differentiate the two types of compounds in order to gather varied data as well as to reduce the number of categories to analyse separately. Additionally, neither *covid* nor *corona* seem likely as constituents in non-canonical compounds.

There are different combinations possible in compounding where other elements than nouns make up the left- or right-most constituent. However, in this thesis, at least one of the constituents of every compound will be a noun because it is a requirement that the compound is created with either *covid* or *corona* as a constituent. This constituent can either be the head of the modifier of a compound which results in variation in compounds collected despite one that one constituent is always a noun.

In addition to defining compounds by their constituents, they are also separated into three types that depend on their orthography. In this thesis the terminology used by Bauer et al. (2013, pp.55-56, 432) is used when defining these types. The first type of compound is

known as ‘spaced compounds’. These types of compounds, such as *corona virus*, have a spacing that separates the two bases from each other while still operating as a compound, nonetheless. Testing if a word is a spaced compound can be done by trying to add an element in between, e.g. **corona English virus* (asterisk used to denote a string as ungrammatical), if the result is ungrammatical then it is a spaced compound. This type of compound can be more difficult to distinguish from syntactic phrases given the spacing. This is relevant for the A-N compounds mentioned above as they may look similar to a NP with a premodifier. The second type is called ‘solid compounds’ and refers to compounds written as one single word e.g. *coronaracism.* Solid compounds are considered uncontroversially to be compounds as they cannot be misread as syntactic phrases due to the lack of spacing. The third type is that of ‘hyphenated compounds’, which can be considered as a spelling compromise between spaced and solid compounds. As the name implies, this type of compound contains a hyphen between the two bases, e.g. *covid-control*, and like the solid compound cannot be mistaken for a syntactic phrase.

The orthography of compounds is not necessarily static, and some compounds may have multiple spellings. In addition, the orthography may gradually change from spaced to solid or hyphenated over time. Bauer et al. (2013, p.450) mention that the more lexicalized a compound is the more likely it is to favour a solid spelling and therefore more popular and frequent compounds are more likely to be spelled with solid spelling. This implies that as a compound becomes more lexicalized it may change spelling. However, other factors that affect this tendency is the length of the compound as well as if there occurs two identical consonants or vowels after each other (Bauer et al., 2013, p.450). Unfortunately, as of now the information discussed above is primarily relevant for the noun-noun compounds as this type of compound has received more attention in research.

2.2.2 Blends

The process of blending involves two (or more) bases that combine in order to form a new word through the deletion of phonetic and orthographic material from one or all of the bases involved. The blend *coronageddon* shows a combination of *corona* and the penultimate and ultimate syllable of *armageddon* where the pre-antepenultimate and antepenultimate syllable of *armageddon* is deleted. However, there are also blends like *covidiot* which appear ambiguous as to which base lost syllabic material as both *cov* + *idiot* and *covid* + *iot* are plausible.

Blending bears some similarity to compounding because the elements involved in the process are two (or more) bases that combine into a new word. But the difference is that these bases are combined after syllabic material has been deleted to form a shorter word. The size of the deleted material varies and depends on how many syllables are in each constituent as well as the size of the word. Syllables will be covered later in section 2.2.2.1.

Plag (2003) divides blends into two categories. The first category involves what Plag (2003) notes as "shortened compounds" or "clipped compounds" (Kubozono, 1990, p.2) and which are therefore "not true blends". The semantic properties of these blends are decided by the original right-hand element, i.e. the right-most element before the blending. For instance, *covidiot* does not refer to an entity that is both covid and an idiot but rather a type of idiot that refuses to follow covid-related restrictions. This is because the blend is headed by *idiot* which is the right-most element before the blending occurred as in *covid + idiot*. The other category features what Plag (2003) considers "true blends". Plag's (2003) main criterion for this category is that the two (or more) elements that make up the blend are both related to the meaning of the blend itself. In practice, this means that a blend "A + D" will not be a type of "D" but rather an "AD", for instance *boat + hotel* becomes a *boatel* which is both a boat and a hotel. The schema Plag (2003) provides for blends is that of "A B + C D => A D". However, Bauer et al. (2013) propose an additional schema of "A B + C D => AC" in order to include the types of blends in which the right element loses its final material rather than its initial material (Bauer et al., 2013, p.458). In light of these two schemas there are still blends which may be hard to analyse such as *covidiot* in which either *covid* is intact and the antepenultimate syllable of *idiot* is deleted, or the rime of the ultimate syllable in *covid* is deleted and *idiot* is intact. The reason either of the elements can be intact is because the B and C elements may be null. This does not mean that every blend follows this schema, but those which do not follow these schemas are much less frequent (Plag, 2003, pp.121-123). The definition of a shortened compound may seem to presuppose an earlier use of the compound in an unblended form, but this is not necessarily the case. A blend such as *covidiot* has close to zero entries that are not blends, which means it is unlikely that it was shortened from a compound.

2.2.2.1 Syllable structure

A syllable is made up of three elements. These elements are the onset, the nucleus and the coda. The nucleus and the coda make up the rime of a syllable, although the coda is optional, and it is therefore possible that a rime consists of the nucleus alone.

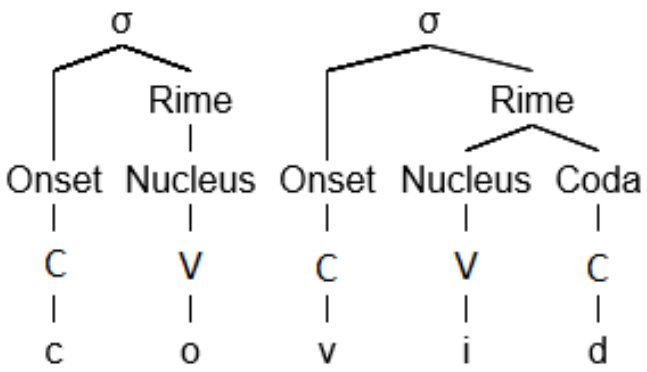

Figure 2.1 Syllable structure of *covid* (Plag, 2003)

As seen in Figure 2.1 above, *covid* consists of the two syllables *co* and *vid*. In the penultimate syllable the *c* is the onset and *o* is the nucleus, without any coda. The ultimate syllable has *v* as onset, *i* as nucleus and *d* as coda. The nucleus is always a vowel and appears to be the only obligatory part of a syllable. There are also more often multiple onsets than multiple codas in syllables, such as in *banana* where each syllable has an onset and a nucleus but no coda, or in *covid* where there are two onsets but only one coda (Plag, 2003, pp.81-82).

The process of blending discussed in 2.2.2 above is as mentioned very similar to compounding as both processes combine two or more constituents but differs as one or both of the constituents lose syllabic material when combined, i.e. *corona-demic* where *corona* is whole while *pandemic* lacks its antepenultimate syllable. Another example is the blend *coronapartheid* which consists of *corona* and *apartheid* without its antepenultimate syllable. The antepenultimate syllable in *apartheid* is an example of a syllable consisting of only of the nucleus without any onset or coda. The material lacking does not need to be an entire syllable as seen in *covspiracy* where *covid* is lacking the rime in the ultimate syllable and combines with *conspiracy* without its pre-antepenultimate syllable.

2.2.3 Derivation

Bauer et al. (2013) describe this word-formation process "[…] as the formation of words by combining affixes and bases" as well as "the operation of some non-combinatorial processes on a base" (Bauer et al., 2013, p.431). Non-combinatorial simply means that they are complex words not formed by adding affixes and bases together "as in a chain" (Plag, 2003, p.12) but through different means, such as changing category or reducing the syllabic material. According to Bauer (2003) derivation always leads to the creation of a new lexeme through the use of affixes, though it is not necessary for the output to be of a different syntactic category (i.e. noun, adjective) than the input. Additionally, prefixes are exclusively

derivational in English (Bauer, 2003, p.14), something which will be highly relevant when examining items in the corpus later on in this thesis.

2.2.3.1 Affixation

Affixes tend to affect the morphological characteristics of a base. As mentioned in 2.2.3 above, a derivational affix is an affix that creates a new word by attaching to a base. On the contrary, an inflectional affix only changes characteristics such as tense or plurality of the base and does not result in a new lexeme. As stated in 2.1.1 prefixes are always derivational, which ultimately means that it is suffixes that must be closely examined. When a derivational affix attaches to a base it creates a new word. This word has properties that depend on the morphological head. These properties can be gender (e.g. *-ess* as in *lion* (masculine) to *lioness* (feminine)), whether it is a count noun or not (e.g. *-hood* as in *knight* (count) to *knighthood* (non-count)) as well as if it is animate or inanimate (e.g. *-er* as in *love* (inanimate) to *lover* (animate)). For instance, according to Bauer (2003) the suffix *-ian* has the properties of animate and count and therefore any words the suffix combines with, such as *musician*, become animate and countable regardless of the properties of the other constituent. In most prefixed words it is the base that acts as the morphological head, except if the prefixed word already has a suffix attached to it. For derivational suffixes it is the right-most suffix that is the head as it determines the part of speech for the whole derivative (Bauer, 2003, p.179). For example, the adjective *covidy* consists of the noun *covid* with the adjectival suffix *-y* as the head as it changes the category of the base from a noun to an adjective. Crucially, it is not required that there is only one suffix per word because of recursive suffixation. Recursive suffixation refers to when a suffix attaches to a word which already contains a suffix, though not every suffix allows for recursive suffixation (Fabb, 1988; Hay & Plag, 2004).The word *covidization* consists of *covid* + the verbal suffix *-ize* to form the verb *covidize* on which the nominalizing suffix *-ation* (noted as *-ion* in Plag (2003)) attaches afterwards, forming *covidization* through recursive suffixation, with *-ation* as the morphological head of the derivative.

Affixation does not necessarily result in a category change but instead only alter the word's primary use. An example of this would be *covidity*, where the noun *covid* has combined with the nominalizing suffix *-ity* that creates nouns referring to a general property or state of an element (Plag, 2003, pp.91-92). The resulting derivative *covidity* refers to the level of covid that something has or is, as example [1] below:

[1] “the current state of ‘**covidity**’ in this state not only hasn’t encouraged us to drop our masks and eat in […]” (20-07-19 US)

Example [1] refers to the *covidity* of the state. Bauer et al. (2013) note that it is rare for the suffix *-ity* to attach to noun bases, and more frequently attaches to adjectives. However, *covidity* is an example of such an occurrence, which shows that a base can undergo derivation without category change. What is interesting is that this is the only occurrence of *covidity* in the corpus that does not refer to the clinical trials known as “*COVIDITY*”.
The adjective forming suffix *-y* combines with *covid* to form *covidy* as in the below example:

[2] “Maybe English children are just distinctively more *Covidy* than their Celtic cousins?!” (20-09-15 GB)

According to Urban Dictionary the word refers to items or behaviours related to *covid* in a negative sense.
Another adjective-forming suffix is that of *-ous* as in *covidious* in example [3] below:

[3] “So in these **Covidious** times let’s remember: […]” (20-04-06 NZ)

According to Bauer et al. (2013) the suffix *-ous* does not carry an inherent meaning aside from forming an adjective and is therefore a transpositional affix. That is, the meaning comes from the combination “of the base, the affix, and most importantly the context in which the form appears” (Bauer et al., 2013, p.314).

2.2.4 Non-affixational word-formation

This section is about the less frequent derivational processes that do not use affixes to form new words and may, for example, remove syllabic material (clipping) or change the category of a base without altering the base (conversion).

2.2.4.1 Conversion

Conversion is a process that changes a word’s category without a modification of a root. One such example is the word *must* which normally is only a modal verb but has undergone conversion to the noun *a must* which means “something a person has to see/do”. This is conversion because the category of the word was changed without any formal markers. While

the process of conversion results in a new word, the original word is not removed from the lexicon and may exist alongside the new word, hence why the modal *must* is still in use. However, this is not always the case, as the original word may fall out of use while the converted word stays, which is what happened to the modal auxiliaries, where the original verbs fell out of use. Another case of conversion can be observed in the verb *covided*. Superficially *covided* looks like a combination of the noun *covid* and the inflectional past tense suffix *-ed*. However, this suffix attaches exclusively to verbs and cannot change the category of the base it attaches to. This indicates that the noun *covid* has undergone conversion to the verb *to covid*, which is interesting because there are so far no entries where the verb *covid* is used in its infinitive. Which means that it is conversion, and the suffixation happens simultaneously, rather than the inflection occurring at a later stage. This also holds true with the progressive suffix -ing when added to *covid* to form *coviding* as in example [4] below:

[4] "'So… Are we still **coviding**?'" (20-07-09 US)

2.2.4.2 Abbreviation

Another derivational process without affixes is abbreviation. This process carries some similarity to blends because it entails multiple bases and some form of merging (Plag, 2003, p.126). The process will not be a part of the data in chapter 4 because searching for abbreviations in a corpus would be difficult given that it is impossible to know which letter of *covid* or *corona* that would be used as an element. The process is therefore better suited for a study where words that are already known to the researcher are examined. The process is described here nonetheless because it is the process which created the word *covid-19*. This abbreviation is made from the phrase "***co****rona* ***vi****rus* ***d****isease 20****19***" where the letters that make up the abbreviation are highlighted in bold. This form of abbreviation is considered an acronym as it is phonologically read as a word, rather than each of the initials read sequentially. Abbreviation differs from blending because blends generally omit only some of the content of each element while an abbreviation mainly uses the initials of each involved element and omits all other material. In addition, when abbreviations are formed it is the orthography of the involved words that is central rather than syllable deletion which is central to blends. However, Bauer (2003, p.47) mentions that when it is more than the initials of each base that is used to form the acronym they tend to merge into blends.

2.2.4.3 Clipping

Clipping is the reduction of phonetic material or orthographic material while the original meaning of the word is kept. Because the original meaning is kept it is not a process of derivation. An example of clipping is *rona* in which the antepenultimate syllable of *corona* is omitted entirely. The clipped element tends to co-occur with the definite article as in example [5] below:

[5] "All of the vaccines for **the rona**, are 100% effective at keeping people out of hospitals and alive." (21-04-16 US)

Bauer et al. (2013) note that clippings mainly result in monosyllabic items, but should the base consist of one unstressed syllable followed by a stressed one then it would generally result in disyllabic clippings such as *rona*. Thus, clippings with more than two syllables should be impossible. Lastly there are also some clippings which specifically target sub lexical morphemes and tend to occur in prefixed bases as well as compounds (Bauer et al., 2003, pp.402-403). This type of clipping results in an entire word, e.g. *corona* from *coronavirus* as *virus* can be viewed as a constituent. This clipping carries the exact same semantic meaning as *coronavirus*. The *corona* in *coronavirus* refers to the original Latin meaning of 'crown' (OED, s.v. /*corona*/). and does not have *coronavirus* as referent as opposed to the clipped *corona*. Therefore, the compound *coronavirus* is not made up of the clipped *corona* + *virus* and is not of interest to my empirical investigation in chapter 4. The same is true for *covid-19* as *covid* is a clipping of the word, though *covid* itself does not have another pre-existing referent.

2.3 Productivity

Bauer et al. (2013) state that some morphological processes are more frequently used than others, while some may not be used at all. In order to distinguish the frequency of these processes the term *productivity* is introduced. Bauer et al. (2013) introduce two main viewpoints on productivity which, while focusing on different aspects, are functionally equivalent.

The first viewpoint focuses on the constraints related to productivity and what elements an affix may attach to. For example, the nominalizing suffix *-al* would be described as attaching only to verbs with stress on the ultimate syllable, which assumes that the suffix may only appear on verbs. On the other hand, the second viewpoint focuses on what domain

the affix may attach to. Using the same example, the nominalizing suffix *-al* would be described as "productive only in the phonological domain which has stress on the final syllable of the base" (Bauer et al., 2013, p.578), which does not exclude the possibility of other categories than verbs as bases. The former "focuses on exclusion of the impossible […]" and the latter "[…] focuses on potential sites of inclusion" (Bauer et al., 2013, p.578).

Because *The Coronavirus Corpus* (2019-) collects data from newspapers it would reveal to some extent which words are deemed suitable as it is a journalist that chooses which words to write and experiment with. This means that the new words found in news articles likely have been carefully chosen to be functional and interesting.

Any mention of productivity in relation to the words that will be examined in chapter 4.1 and onwards is relative to the words in the data and not absolute. The reason productivity is not omitted, however, is that hapax legomena or generally low frequency tokens may point towards a more productive process involved, which means that if patterns were to arise in terms of the processes mentioned then closer examination may prove of interest (Bauer et al., 2013, pp.578-581). According to Bauer (2003) a hapax legomenon is "[…] a word which occurs once only in a particular text or corpus of texts" (Bauer, 2003, p.331). This means that any entry within a corpus with a frequency of one is a hapax legomenon.

Hapax legomena are useful in order to measure the productivity of a process as they are considered examples of the new words a process may create. This is because a process which is not productive (available) will mainly generate a few high-frequency tokens because the process only works with those bases, while a more productive (available) process will be able to create more words and will result in many low-frequency entries (Bauer, 2003, pp.86-87). The P-value as described in Plag (2003, pp.56-57) can be used to determine the general likelihood of hapax legomena in the data. A P-value is a number that describes the probability for encountering a hapax legomenon within a corpus, and more specifically within a type of process. The *P* stands for "productivity in the narrow sense" (Plag, 2003, p.57) and is used in the following formula:

$$[6]\ P = \frac{n_1^{aff}}{N^{aff}}$$

Which I modify for clarity, as it will be used for more than affixation:

$$[7]\ P = \frac{n_1^{process}}{N^{process}}$$

The numerator, $n_1^{process}$, is the number of hapax legomena created by the process, and is divided by the denominator, $N^{process}$, which is the total number of tokens created by the process. Where *process* is replaced by any process of interest, e.g. compounding, blending, conversion etc.. It is important to note that this number should not be taken at face value and may change considerably based on the size of a corpus.
For example, if the equation in example [7] results in a P value of 0,02, then it indicates that there is a 2% chance for an entry of the targeted process to be a hapax legomenon. If the process of blending has a P-value of 0,33 it means that every third blend token in the data is a hapax legomenon.

2.4 Previous studies

A study by Bakhmat et al. (2021) focuses on the neologisms that arose during the *coronavirus* pandemic. The study analyses different online dictionaries and their chosen "words of the year" in order "to trace lexical changes caused by the *coronavirus* outbreak and analyse newly coined lexemes" (Bakhmat et al., 2021, p.134). Additionally, they analysed multiple news articles related to the pandemic in order to elicit *corona*-related neologisms which they aptly named *coroneologisms* which itself is a blend. The main finding of the first part is that each of the four chosen dictionaries, Merriam-Webster, Collins, Cambridge and Oxford had a somewhat similar pattern, and that all of the dictionaries had in 2020 chosen pandemic related words of the year. The pandemic was deemed so impactful by the Oxford Dictionary that they picked several new words due to how chaotic the first year of the pandemic was (Bakhmat et al., 2021, p.135). In the second part of the study, which focuses on *coroneologisms*, they listed 52 different new words featuring *coron-*, *quaran-* and *covid-* as constituents.

The study by Alyeksyeyeva et al. (2020) focuses on the use of neologisms but more in depth on the processes that has occurred during the pandemic. The study mentions the existence of stages in terms of society and its "coronaspeak" as described by Thorne (2020) in an interview. These stages present a more concrete example as to how the language developed and proposed "medicalisation of our everyday vocabulary" (Alyeksyeyeva et al., 2020, p.204) as one of the earliest stages, which refers to how medical jargon, usually reserved to medical professionals, began to be used in everyday speech. One important note is that in this study the term "new coinage" is used in referring not only to new items but also to old items given new meaning.

The study by De Smedt (2020), focuses on compounds formed with *corona* or *korona* as constituent and the variation in spelling of *corona* with *c*- or *k*-. De Smedt (2020) points out that *korona* with *k*- originally refers to the corona of the sun only, and now also refers to the *coronavirus*. The primary spelling of *coronavirus* was with *c*- until the year 2020 where the use of *k*- rapidly increased. De Smedt (2020) notes that the number of types and tokens of *corona* compounds is steadily increasing, even towards the end of the timeframe in the study. In this study De Smedt (2020) also points towards an older study on *aske* compounds which is covered below.

Other relevant studies of new words formed during the pandemic are Al-Salman & Haider (2021), Akut (2020), Fitria (2021) and Simatupang & Supri (2020).

The following study is not about the *coronavirus* but is about the *ash*-compounds that were formed during the events following the eruption of Eyjafjallajökull 14th of April 2010. This is relevant because much like during the *coronavirus* pandemic, many neologisms were created that relate to the event. For instance, *askefast* 'to be stranded because of the ash'. These neologisms were formed through the word-formation processes, specifically compounding in the study below.

The study by De Smedt (2012) focuses on the *ash*-compounds formed during the events caused by the eruption. The goal of the study is to establish a quantitative analysis of these compounds and their occurrence in the Norwegian Newspaper Corpus. The study focuses on the period 14th of April until 23rd of May 2010 and examines the number of *aske*-compounds occurring every day of this period. In order to make certain that these compounds were indeed occurring more rapidly than before, the occurrences prior to 14th of April 2010 were examined in order to elicit the number of word forms already existing as well as their frequencies. The result was that there were 248 distinct forms after 2010 while only 26 prior to that. The main finding is that during these forty days there were a total of 2298 compounds using *aske*- of which 1368 entries were of *askesky* 'ash plume'. The new words started appearing during the second day of the events, which De Smedt (2012) states could potentially be related to the early reporting relating to the physical phenomena surrounding the eruption whilst the following coverage related more to the social effects of the eruption. The study finds a correlation between the height of the event and the number of new words appearing daily, with the new words declining as the effects of the eruption decline.

3 Methodology

The methodology used in this thesis is that of a corpus study using *The Coronavirus Corpus* (Davies, 2019-) as well as the self-compiled *ash-corpus* mentioned in chapter 2. Section 3.1 describes the main corpus; section 3.2 explains how the different types of word-formation are defined. Section 3.3 focuses on the data collection itself from the main corpus and the procedure relating to it. Section 3.4 is about productivity and the role of hapax legomena in the corpus as well as their relevance. Section 3.5 introduces the *ash-corpus* and its use as well as the method of collection.

3.1 The main corpus

The main corpus used is *The Coronavirus Corpus* (2019-) which is a monitor corpus, i.e. a corpus that is constantly updated in regular intervals and takes in whatever information is of interest to the creator of the corpus. This corpus collects data every night from news articles that contain at least two occurrences of the words *coronavirus, covid* or *covid-19* or contain terms related to disease and spread. This results in a large corpus that grows over time, but also a corpus that does not change any of its information after it has been collected.

One flaw that is not inherently the fault of the corpus is that while language change may occur within anyone in the general populace, the corpus itself focuses on news articles exclusively. This excludes the more innovative language use of the younger generations found on social media (Tagliamonte, 2011) which may have an entirely different distribution. However, if the form is attested in the corpus, then it has been used within a news article which can be considered a relatively formal form of media compared to social media.
While the study itself will be of a quantitative nature there will be a qualitative aspect to it in terms of closely analysing some of the forms found in the dataset. It is necessary to closely analyse some entries in the dataset because some may appear to have been formed through word-formation processes, while it actually is a different process that is behind the word. In some cases it may be that the formatting of *The Coronavirus Corpus* (2019-) is the reason that a word looks like e.g. a derivative or a compound. The problem with this is that aside from analysing every single entry in context, there is not necessarily any efficient way of finding entries that stem from formatting. In addition, there may be cases where the process that formed the word is not one of the word-formation processes examined in this thesis, but a different process entirely such as univerbation (Bauer et al., 2013). What makes corpus studies ideal for the study of word-formation processes is the number of hapax legomena. A hapax legomenon is "[…] a word which occurs once only in a particular text or corpus of

texts" (Bauer, 2003, p.331). These types of entries can be indicative of the productivity of the processes (Bauer, 2003, pp.86-87), as well as which of the two bases, *corona* and *covid,* are favoured when it comes to the respective word-formation processes they partake in. This is because a large number of hapax legomena may imply that the two bases, *corona* and *covid,* partake in multiple different word-formation processes.

3.1.1 Types and tokens.

Two important terms used in corpus linguistics are *type* and *token.* The term *type* refers to a unique entry in a corpus. For instance, *covidiots* and *covidsphere* are two different *types*. The term *token* refers to the frequency of each *type*, i.e. how many occurrences a type has in the dataset.

For instance, if the output looks as in Figure 3.1 below then it is interpreted as the *type coronageddon* has six *tokens*:

ALL FORMS (SAMPLE): 100 200 500 WORDS	FREQ
CORONAGEDDON	6

Figure 3.1 Search output

The search does not group the result based on *lemmas*, which means that *coronageddon* is its own entry alongside *coronageddons* as well as any other inflected form. One reason for not using lemmas is that the corpus appears oversensitive to grouping by lemmas. For example, using the search string "covid*" outputs three lemmas: "covid-19", "covid" and "[]" where the latter lemma appears to be nothing, because the context page shows no highlighted entries. A drawback of not using lemmas is that it increases the number of entries and may result in an entry having less tokens because the filter skips the other word-forms.

It is important to note that no program is used to find word-formation processes within *The Coronavirus Corpus* (2019-), and that it is my own analysis that informs the annotation of the data and which process that created the entry. Because some processes are harder to spot than others it is necessary to establish criteria in order to identify the different word-formation processes. These criteria are described in the following section.

3.2 Identification of word types[1]

The Coronavirus Corpus (2019-) cannot be filtered directly for the word-formation processes of interest. Therefore, it is necessary to introduce requirements related to the different word-formation processes to filter the output of the different search strings. The following sections are dedicated to describing these requirements. If an entry fulfils one of these requirements it will be considered as the result of only one of the processes. However, if an entry fulfils the requirements for multiple processes (e.g. *corona vac*, clipping and compounding) then it will be annotated according to whichever process happened last, e.g. the word *corona vac* is compounding despite the clipping of *vaccine* to *vac*. 3.2.1 describes compounding, 3.2.2 focuses on blends while 3.2.3 is concerned with affixation. Section 3.2.4 describes conversion whilst 3.2.5 describes the criteria for clipping.

3.2.1 Compounding

As described in section 2.2.1, a compound is a word consisting of (at least) two bases, e.g. *coronavirus* which consists of the two bases *corona* and *virus*. Because of this, the first requirement is that a compound has to consist of at least two bases that are either spaced (e.g. *covid hospital*), connected by a hyphen (e.g. *covid-delayed*), or written solid (e.g. *coronacases*). However, if the word in question is potentially ambiguous between a spaced compound or a syntactic phrase, as is the case with AN strings (e.g. *full covid*), then it is important to dispel this ambiguity. One test is the insertion of an adjective or affix in-between the two bases (e.g. **full red covid*). Should the result not function as a word at all then the word in question is a compound. A problem may arise when making criteria for A-N compounds, however, as they may appear similar to that of NPs with adjectives as modifiers such as *short wave* (NP) as in "a wave that is short" and *short wave* (compound) as in "a short wave radio". A-N compounds written solid or hyphenated are considered to be lexicalized words rather than phrases in this thesis (Bauer et al., 2013, p.451).

3.2.2 Blends

Blends do not prove difficult to find and extract due to their very particular forms (e.g. *covidient*, ***covid*** + *obed**ient***). The main requirement will be to extract items which follow one of the schemata discussed in section 2.2.2. The schemata, which were introduced by Plag

[1]Abbreviation will not be considered as a category in this thesis. Because the thesis seeks out connections between *covid* and *corona* as elements of word-formation it makes little sense to use this category, especially considering that covid itself is an abbreviation.

(2003) and Bauer et al. (2013), are as follows: A B + C D => AD or AC. A and B refer to the syllabic material in the first word in the blend whilst C and D refer to the syllabic material in the second word in the blend. There is also a possibility of zero elements, that one part of the blend is a word. This is shown in the examples below, where the whole word is in bold.

[1] **covid**-preneur as in
"Year of the **COVID**-preneur: […]" (21-04-08 US)
[2] **covid**iot[2] as in
"Don't be a **Covid**iot, stay home." (20-04-24 GB)
[3] **corona**geddon as in
"Talk of **corona**geddon this winter is not without foundation." (20-03-21 ZA)

In example [1] A is *covid* and B is a zero element whilst C is the pre-antepenultimate and antepenultimate syllables, and D is the penultimate and ultimate syllables in *entrepreneur* with the addition of a hyphen to connect the two constituents. Example [2] and [3] are similar because A is the whole first word, B is a zero element, C stands for the antepenultimate syllable of the second constituent and D stands for the penultimate and ultimate syllables in both examples. These letters refer to syllabic material in monosyllabic blends, i.e. blends that either create a word with only one syllable or where one of the constituents is only one syllable. When the blend and both constituents have two syllables or more, then A, B, C and D will correspond to either entire syllables, multiple syllables or zero elements as in example [1], [2] and [3] above. Plag (2003) distinguishes between "true blends" and shortened compounds, which are discussed in detail in chapter 2.2.2. In this thesis, however, I will not distinguish between true blends and shortened compounds but subsume them under the same category. This is done because it is not the semantics of the blends that is of importance to the thesis, but rather their use in newspaper articles.

3.2.3 Derivational affixation

Derivational affixation is a straightforward process to identify. For example, if a word with *covid* or *corona* is identified with any prefix (e.g. *noncorona*, *post-covid*) it is automatically considered derivational affixation because "all prefixes in English are derivational" (Bauer, 2003, p.14). For a more detailed discussion of derivational affixation see section 2.2.3.

[2] It is also possible to say idiot is the whole word as in *cov**idiot***

Suffixes can be either inflectional or derivational, and can be distinguished based on if the suffix results in a new word with additional meaning compared to the original base and/or changes the word class of its base (i.e. noun => verb, e.g. *covidize* where the verbal suffix *-ize* makes a verb from *covid*). If the suffix does not do either of the two, but rather creates a new word form of the base then it is inflectional (e.g. *covid* => *covids*) rather than derivational. The reason it is not a requirement that the word class changes is because some affixes produce more abstract words within the same class, e.g. *covidity* where the nominal suffix *-ity* derives abstract nouns from nouns, and refers to some form of degree of something. Should the word class change (e.g. adjective to noun), then the process is exclusively derivational affixation.

3.2.4 Conversion

The process of conversion can at times be elusive because the output of the process is identical to its input. However, there are revealing cases such as *covided* or *coronaed* where *covid* and *corona* bear a verbal inflectional suffix. Since verbal inflections only attach to verbal bases, *covid* and *corona* must have undergone conversion prior to suffixation. It is in this way that inflection is relevant to this thesis. Here are some examples of the verbs in context:[3]

[4] "We're all **COVIDed** to death." (21-02-01 CA)
[5] "I have been **covided**!" (21-04-27 AU)
[6] "[…] I will be absolutely doing anything I can to not get **coronaed**" (20-03-03 GB)
[7] "[…] it looks well and truly '**coronaed**'" (20-03-10 AU)

While examples [4-7] are all verbs, conversion is not limited to creating verbs from nouns. For example, most derivational affixes attach only to bases of a specific category, so if an affix is attached to a base that it normally does not attach to then it is likely that conversion has occurred.

3.2.5 Clipping

Clipping, much like blends, is simple to identify as the output of the process is irregular and normally does not correspond to pre-existing lexemes. For instance, *rona* does not correspond to earlier words but corresponds to the penultimate and ultimate syllables in *corona*. Clipping

[3] These examples are taken from the unfiltered data, as no entry of conversion was present after filtering.

is likely to be the least frequent process in the data given that *covid* does not seem likely to be clipped, nor does it seem to be a lot of variation possible with *corona*. Additionally, compared to the other processes listed here which uses constituents such as affixes or bases, except for conversion, clipping uses only the base *corona* and *covid* and is therefore limited in how many clippings can be produced.

3.3 Data collection

My corpus consists of data extracted from *The Coronavirus Corpus* (2019-) using the search strings in Table 3.1 below. The formulation of the search strings employed is informed by the types of word-formation processes that *corona* and *covid* partake in. While these search strings may yield duplicates it is important to note that it is the overall result across search strings that matters the most and not how many processes are present in each search string. The search strings are used to gather data randomly, and as efficiently as possible given the time constraints of this thesis. The wildcard function, *, in *The Coronavirus Corpus* (2019-) is especially useful for this research because it returns entries where the asterisk is replaced by either nothing, which returns only *corona* or *covid*, or it is replaced by any word or other elements. If the asterisk immediately follows *corona* or *covid* (i.e. *corona*, covid**) the search string will return hyphenated elements in addition to words written solid such as derivatives and compounds. While there are three different compound types, it would not be useful to examine how many compound types each search string outputs because the only search strings that allow for variance in spelling are those where the asterisk is written solid with the base (e.g. *covid*, *corona* and *cov**). This is because when the asterisk is separated from the base by a space (e.g. *corona *, covid **) the only relevant result is spaced compounds.

Table 3.1 Search strings for data collection

Search string	Main Word-formation processes targeted	Examples
corona*	compounding and affixation	"coronatracker"
corona *	compounding	"corona tax"
*corona	compounding	"post-corona"
* corona	compounding	"village corona"
cor*	blending	"coronaccessories"
covid*	compounding and affixation	"covidism"
covid *	compounding	"covid warriors"
*covid	compounding	"zero-covid"
* covid	compounding	"common covid"
cov*	blending	"covexit"

Note that there is no specific search string for converted elements, i.e. elements resulting from conversion, in Table 3.1. The reason for this is the lack of morphological marking in conversion which would make it next to impossible to notice conversion without further affixation. Example [8] below shows *covid* converted to an adjective:

[8] "A very COVID homecoming" (20-06-21 CA)

The use of *covid* as in example [8] would be counted as a token under the type *covid* rather than a separate type. Therefore, the unmodified converted types could only be located through exploring extended context. However, the searches that favour solid compounds will also potentially yield converted items. An example of this is *coviding*[4] as in example [9] and [10] below:

[9] "So... Are we still ***COVIDing***?" (20-07-09 US)
[10] "I must say, this ***COVIDing*** has messed up my summer" (20-08-17 US)

In the first example the noun *covid* has been converted to a verb with an attached present participle suffix, while the second example is *covid* converted to a verb + the nominalizing -*ing* suffix. Therefore, example [9] is a converted verb with an inflectional ending while

[4] This entry is only present in the corpus and not the collected data and is purely used as an example.

example [10] is a noun derived from a converted verb. Crucially, example [10] is a case of derivational affixation but is also counted as conversion because it attaches to *covid*. Converted elements such as example [9] and [10] are easy to spot in the data because both suffixes attach to verbs while *covid* is ordinarily a noun. This indicates that conversion has happened prior to suffixation.

In relation to the search string "cor*" I am aware that this is not a syllable and that "co*" is the antepenultimate syllable of *corona*. However, if I used "co*" as a search string the output could additionally contain *covid* bases as well as more unrelated words than with "cor*". Additionally, it is then similar in size to "cov*" and is therefore more consistent as this search string is also more than one syllable. Therefore, I decided to keep the onset of the next syllable in both search strings as this would separate the two bases.

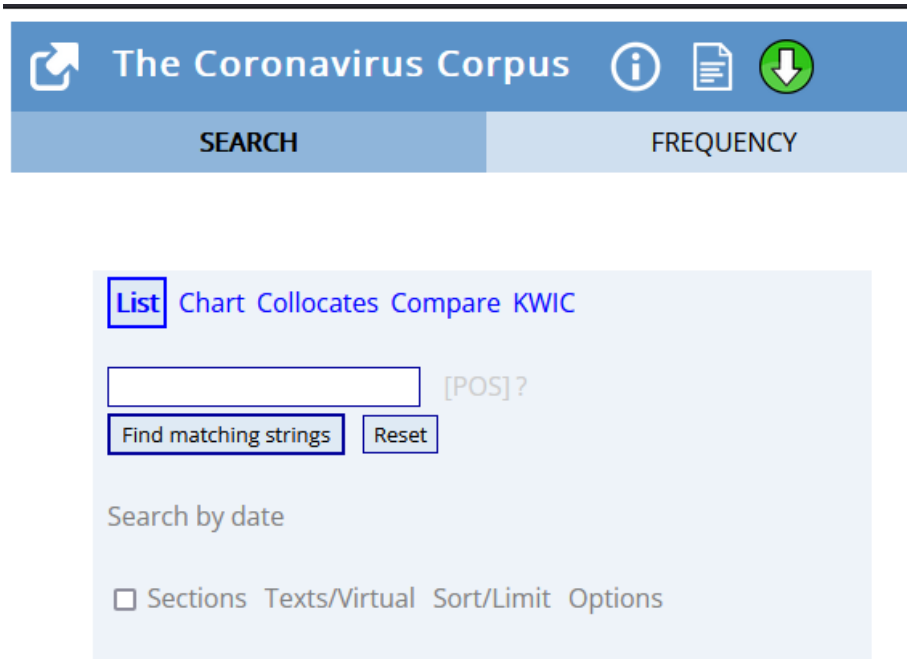

Figure 3.2 Search mask in *The Coronavirus Corpus* (2019-)

Figure 3.2 above shows how the search mask in *The Coronavirus Corpus* (2019-) is structured. In this bar a search string is typed in and then additional modifiers may be selected to sort or limit the results in different ways by examining the tabs *Sections*, *Texts/Virtual*, *Sort/Limit* as well as *Options*.

The main tabs relevant for my empirical study are *Sections* and *Options*. *Sections* allows for the specification of either where the data comes from, i.e. country of origin, or the date it was published, either month or specific day. It is possible to select both at once but only date will be considered in this thesis. Time was chosen because examining the change over time is interesting and will yield a suitable number of types to examine. The reason

countries were not filtered in addition to timeframes is because it would run the risk of resulting in less data as the search would be very specific. Because of the focus on timeframes, it is important to limit the time to a manageable size. Knowing this, I selected the half a year long timeframe 01.2020-06.2020 because it marks the first half year during the pandemic and is when most of the words are first created. The second timeframe is the same as the first except one year later, 01.2021-06.2021, because it may show which words stayed and which words fell out of use during this year. This allows for a comparison of the frequencies and the word-formation processes present in each of the timeframes to find out how big of a change, if any, occurred to the vocabulary during these years.

The *Options* tab allows the user to specify different thresholds, such as limiting the maximum number of hits, as well as sorting or grouping the results. The results may be grouped based on lemmas (e.g. *corona* and *coronas* would go under the lemma *corona*, rather than counted as two separate entries with frequencies) or grouped based on words where every entry would be separate (e.g. *covid* and *covids* are separate entries with separate frequencies). It is also possible to turn on case sensitivity, which would be used for a more orthographical study where capitalization matters. During every search, the number of hits is capped at 4000 in the *options*, although this turned out to be unnecessary after contacting the owner of the corpus, Mark Davies, as he confirmed that for any one search there is a hard limit of 4000 entries maximum. No other part of *Options* is used in the searches as it is considered unnecessary with more variables given the time constraint at hand.

The feature KWIC search (KeyWord In Context) could in theory be ideal for this type of research because it results in a list of contexts for the keyword and would therefore be easier to analyse. However, in practice when KWIC is used it results mainly in entries containing *coronavirus* and *covid-19* because of the high frequency of these words. The reason for this is that the KWIC function shows the context of all the entries within a type while only allowing 2000 lines of context. Therefore, KWIC would be more ideal if exploring the use of *coronavirus* or *covid-19* in specific contexts was the focus. Thus, the more ideal method is through the list function which allows the researcher to specify a maximum number of types.

The consequence of this upper limit of 4000 entries is that every day as new words are added the hapax legomena in the corpus which occur at the bottom are pushed down beyond these 4000 words which in turn means that collecting data over multiple days could lead to different results. Because of this, all of the data is collected on the same date, 06.01.2022, in

order to guarantee that new entries will not be a factor in the comparison. The date of data collection is the day I decided to collect the data and otherwise is not significant.

The output yielded by the search strings specified in Table 3.1 is copied and pasted into an excel document. In this document the data is formatted to include only the entries and their frequencies as well as their entry numbers in the corpus. The formatted data is copied into a formula created in excel (as seen in figure 3.3 below). The formula is made to allow the user to pick out every nth entry in the document and display this in a separate location alongside its frequencies. The part named "How many total" displays how many entries that have been extracted from the original set while "Every N" is where the researcher decides on the interval for data collection. After this process is over the word types are copied into another section where the data is manually analysed and annotated (see section 3.3.1). Finally, another formula extracts the frequency of each annotated word and displays this based on the codes annotated. These frequencies are used to create tables of each word-formation process by base and by search string.

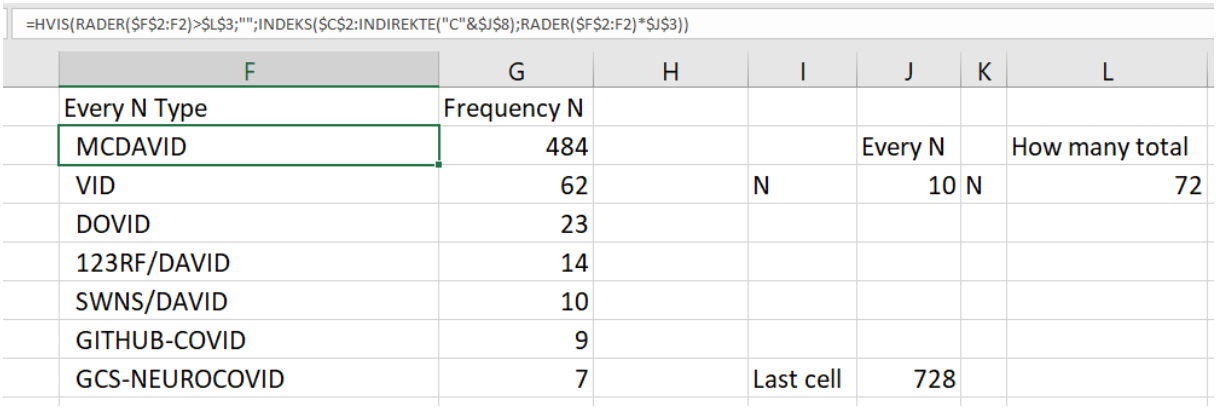

Figure 3.3 The formula for calculating every nth entry

As Figure 3.3 shows it is every 10th entry that is included because I consider it the upper limit in terms of how much data I can analyse for this thesis. At most it results in 400 entries per search string. Because of the inclusion of only every 10th entry it is important to note that the results are random and small changes in entry number may falsely indicate a change in the frequency of use. This means that any resulting data should not be considered indicative of trends of change but rather point towards possible areas of interest for further research.
A separate reason for choosing to take every 10th entry is because the number of hapax legomena may say a lot about the productivity of a process (Bauer, 2003, pp.86-87) and after 10-20 entries there tends to be a considerably drop in the frequency of every type, which means that a large portion of the collected data are hapax legomena.

The entries which do not match any word-formation process using *covid* and *corona* as bases are excluded from the list of features. These entries will, however, be used for calculating the percentage of the relative frequency per total number of items. Other excluded items are hashtags as their structure is not representative of word-formation. For example, *#tech4covid* where *for* is substituted with the number 4 because they are homophonous. In addition, constructions which have periods and slashes, e.g. *doctor.covid* or *science/covid*, are ignored as it is not possible to determine whether they are a deliberate spelling, a mistake or a formatting error occurring during data collection.

3.3.1 Coding of entries

All relevant entries are manually coded with respect to the word formation process involved: Clipping (**Clp**), blends (**Bld**), affixation (**Aff**), conversion (**Cnv**) and lastly compounding is divided into three subcategories following the form of the compound. Solid compounds, those written without any spacing, (**SoC**), spaced compounds (**SpC**) and finally hyphenated compounds, which are compounds connected by a hyphen, (**HyC**). The point of most interest is finding out if either of the bases are favoured and if so which of the two bases are favoured by which of the word-formation processes. Furthermore, it is of interest which of the processes are most frequent overall, hence the need to create categories for not only base type but also process type.

Some entries may be confusing at first glance or appear questionable. These entries are either annotated by a hyphen if they are hard to designate or confusing or marked by the code of the process followed by a question mark (e.g. “HyC?”) if it appears to be related to the process but requires more context. Because of this, the data will be examined twice where the first examination is dedicated to quickly assessing and labelling everything that is clearly created by a word-formation process while also labelling the questionable or related entries. Afterwards, the second examination is performed, where only the entries marked with question marks and hyphens will be thoroughly examined through the use of extended context. For example, *corona visited* at first glance could be a spaced compound and could be an adjective which describes somewhere that *corona* has been and affected. However, upon closer inspection its only use is related to *corona* as a subject and *visited* as a verb. Another is the word *covidy* which appears to be *covid* + the adjective forming suffix *-y*, but upon further examination it turns out that entries in the filtered data refer to a name *covidy* as in example [11] below:

[11] “We would have preferred a nonthreatening word that rolls off the tongue, like **COVIDY**” (20-02-12 US)

Crucially, *covidy* also exists as an adjective, though outside the timeframe, shown in example [2] in chapter 2, repeated here as example [12] below:

[12] “Maybe English children are just distinctively more ***Covidy*** than their Celtic cousins?!” (20-09-15 GB)

This works the other way as well as some words may appear to not be created by a word-formation process but when examined further, are actually from word-formation processes. An example of this is *long covid* which appears to be *covid* modified by the adjective *long*, but upon closer examination there are entries written solid, *longcovid* which indicates that it is in fact a compound. Another example is the blend *covidivorc*, which appears to be spelt poorly but upon examining the context it turns out that it is *covidivorcées* and that the corpus was not able to code for the é and instead segmented the word into *covidivorc ? es*. This made the corpus register it as two separate words. Knowing this, the entry is edited to reflect what it is supposed to look like as this is a formatting mistake rather than a spelling mistake.

Once every entry is coded and collected, the different frequencies will be displayed in order to compare these. The words with *corona* or *covid* as constituents will be compared with each other to determine which base, if any, is the most productive within each of the word-formation processes in the data. Additionally, the frequency of the different word-formation processes will be compared to one another across words using *corona* and *covid* as bases to see which process is the most frequent regardless of base. These frequencies are separated based on the two timeframes from which the data has been collected, 01.2020-06.2020 and 01.2021-06.2021. This is done in order to compare how the frequencies have changed from the first half year of the pandemic to the third half year of the pandemic. As such the comparison may also account for the increase/decrease or sudden occurrence of certain forms. This will be useful information as it may show how the progression of the pandemic affects the language use in daily life albeit only within the data and not necessarily real life. As described by Alyeksyeyeva et al. (2020), the first step of this pandemic in terms of linguistics was a medicalization of everyday language, meaning that there may be more medically related terms occurring in the first timeframe than in the last. It is, however, unlikely to be multiple medical terms present in the data. This is because *corona* or *covid* will

be present in every entry, which leaves little room for medical terms such as *key workers*, *incubation period*, *asymptomatic* unless they contain *covid* or *corona* as in *coronapositive*, *corona-quarantine* and *covid-diagnosed*.

3.4 The *ash-corpus*

As mentioned, I compiled my own *ash-corpus* from NRK[5] articles on the volcanic eruption in 2010. The reason that I did not use the Norwegian Newspaper Corpus, is firstly because a study by De Smedt (2012) has utilised this corpus. Secondly, the Norwegian Newspaper Corpus does not collect data from NRK, and it is therefore new data that is contributed through this thesis. These articles were found in a dedicated section on NRK about the volcanic eruption.

The data is collected by copying the contents of the NRK articles and pasting the data in a .txt file to analyse it. An issue with the event section of NRK is that not every article that is relevant has been added which means that it is not the full range of articles from this event. However, finding every article that has been skipped would take considerably longer time and is therefore not feasible as it would require a thorough examination of the entirety of NRK from 2011 to 2010.

The corpus has data from ca. 265 articles with a total of 108731 words of which 1073 are *ash*-related words. The extracted .txt file is searched for words created from word formation using the software AntConc which is used to sort the data through search strings. The search string used is that of *aske** where the asterisk, *, is a wildcard function which returns any word which either has *aske* 'ash' such as *askefast* 'ash stranded' or only *aske* and its inflections. This search will then return every compound as well as inflected form of *aske*. The two most frequent words, *aske* 'ash' and *askesky* 'ashplume' were both removed from the overall data because the former is not a compound, and the latter is a word frequently used prior to this event. These two words account for 896 tokens which is high compared to the third frequent word, *askefast* 'ash stranded', which has 29 tokens.

The reason "aske *," "* aske" and "*aske" is not used as search strings is because after a preliminary search, none of these resulted in any relevant entries.

Before creating a list over frequencies, it is necessary to manually purge the duplicates found

[5] NRK stands for "Norsk rikskringkasting", generally translated to Norwegian Broadcasting Corporation. NRK is Norway's largest media corporation and was established in 1933.

in the NRK articles. This must be done because NRK has references and links to other articles within an article. These references would not attest multiple uses of *ash*-compounds because they sometimes use the same wording as the title of the articles referenced.
If these entries were kept in the *ash-corpus*, then it would increase the frequencies of *ash*-compounds without reflecting the actual use of the word in news articles. The purged dataset is then searched using the search string "aske*" and counting the lines on which certain forms occur. The data is grouped by lemmas which means that, for example, plural nouns such as *askeproblemer* 'ash problems' and definite nouns such as *askeproblemet* 'the ash problem' count as two tokens of the compound *askeproblem* rather than two different word forms. All ash-related words that are not from a word-formation process are excluded from the frequencies of the ash-words but is kept in the total frequency in order to calculate percentages. When the frequencies as well as the P-values have been calculated they are entered into an excel spread sheet. Every occurrence of *ash* + another constituent is added to the dataset, meaning the data collection is not randomized, though no relevant entry is excluded. This is possible only because the *ash-corpus* is small when compared to *The Coronavirus Corpus* (2019-) where it is necessary to reduce the amount of data to analyse.
The *ash-corpus* differs from *The Coronavirus Corpus* (2019-) in a number of ways. The biggest contrast is that of where the data has come from. The *ash-corpus* has been compiled from NRK news articles exclusively and there are no other websites involved aside from three articles from yr.no which were posted to NRK (Yr, 2010). Additionally, it is only the elements in the event section that has been added as this is readily available data. Because of this, the data only reflects NRK and no other news company in Norway. While this results in a less varied dataset, the alternative would be to compile data from multiple news outlets, which would take longer time and is therefore not feasible for the current thesis. The second biggest contrast is about the data itself, which stems from the word-formation that has occurred using *aske* as a base. Compounding is the only process that has been found within the data while all other processes mentioned in the theory chapter has been found in *The Coronavirus Corpus* (2019-). Because NRK is a government funded news company it does not need to rely on "clickbait", i.e. eye-catching titles or misleading articles, in order to have more subscriptions or website visits. The result of this is articles that are not as heavily embellished and is more likely to reflect everyday language use.

3.5 Productivity

In this thesis, the productivity of a process is considered in relation to the other processes (or lack thereof) within the corpus data. For example, if *corona* + derivational affixation is found more often than *corona* + compounding, then the process of affixation with *corona* as a base is considered more productive than compounding. Crucially, it is not raw frequency of every type that is important but rather the number of different types that result from the same process. The productivity measured in this thesis should not be taken as a general measure of productivity for these processes overall but is used in this paper to help determine which process is the most frequent within the data. Measuring productivity also helps determine which of the two bases is the most productive in the different processes because it creates a value which can be compared. This value is known as the P-value and describes the likelihood that a token of a given process is a hapax legomenon (see section 2.3 for a more detailed explanation) within a corpus (Plag, 2003). The equation in chapter two, example [7], is repeated here as example [13]:

[13] $P = \frac{n_1^{process}}{N^{process}}$

This value is acquired by dividing the total number of hapax legomena formed by a word-formation process by the total number of tokens from the same process. Two separate P-values will be calculated for each word-formation process as there are two bases that can take part in each process. The reason for calculating the P-value is because it will allow for a comparison between the different word-formation processes as well as between the two bases, *covid* and *corona*. This is possible because productivity can generally be measured by the number of hapax legomena produced by a process, which is reflected in the likelihood that a token of a process is a hapax legomenon. This comparison is directly related to the research questions, as it will help answer which of the processes are the most productive and potentially which of the two bases are favoured, if any at all.

The results of this empirical study will be compared to the results found in the *ash-corpus*. In addition, the compounds of both *corona* and *covid* are compared to the *ash* compounds, where similarities and differences between the formations will be examined and discussed.

The data from both *The Coronavirus Corpus* (2019-) and the NRK event section relate to an event affecting people where the former event is a worldwide pandemic caused by the

coronavirus and the latter relates to a volcanic eruption on Iceland which stopped flight traffic in many northern countries (e.g. Norway, Iceland, Denmark, Sweden), potentially leaving people stranded abroad. While the events themselves are of different scales, there are similarities in the neologisms created because of them (e.g. *corona-free*, *askefri* 'ash-free',) as both events affected daily life.

3.6 Limitations of the corpora

Exploring word-formation processes through corpora comes with issues and limitations that in some cases cannot be excluded entirely as a factor. The best option is then mitigating these issues as much as possible. The primary issue relates to locating the different word-formation processes in the collected data. In section 3.2 I proposed a set of key features to look for, but these features are not entirely flawless and there may be occurrences where for example spaced A-N compounds may be excluded or an Adjective Phrase may be included as a compound when it is not. Additionally, some structures may at first glance appear to be compounds with a verb such as *corona visited* which may appear to be *corona* + participle but is actually a subject and a verb (e.g. *corona* visited our country). Therefore, it is important to examine the ambiguous entries closely in context. This thesis is inclusive and therefore does not differentiate between canonical and non-canonical compounds when annotating. Mistaking compounds for syntactic phrases is an issue that only applies to compounding alone and even then, it is only relevant for spaced compounds as words produced by most other word-formation processes are not possible to mistake for other constructions due to their distinct forms. Solid compounds and hyphenated compounds are not possible to mistake for phrases due to their orthography. An issue related to the search strings is the variation in results each string can result in. The search strings where the asterisk is separated by a space preceding the base (* *corona*) or following the base (*covid* *) will only result in spaced compounds and no other element of word-formation.

As mentioned in section 3.3, there is a maximum of 4000 types which means that even if there are more types, they would not be accessible. An issue that arises with this is the types listed at the bottom of the corpus, the hapax legomena, vary from day to day meaning that despite the use of fixed dates there is still variation as to which types are displayed and even the frequency of the types. Two problems arise from this. Firstly, it means that all data must be collected the same day to remove the possibility of collection time as a factor. This also means that the data collection must be done over again if some issue is discovered as recollecting only parts of the data on another day will yield different results. Secondly, it

means that the data collection part is harder to repeat in future studies as there will be variation in which types will be present. As of now I have found no way to remove this factor, nor do I understand why the data is shifted from day to day, as well as increasing or decreasing in frequency.

Also mentioned in section 3.3 is that duplicates may arise because the output of search strings may overlap. This is not very likely to be a problem because it would require all of the duplicated types to be filtered in. However, duplicates may also arise because of the corpus itself somehow collecting duplicate entries.

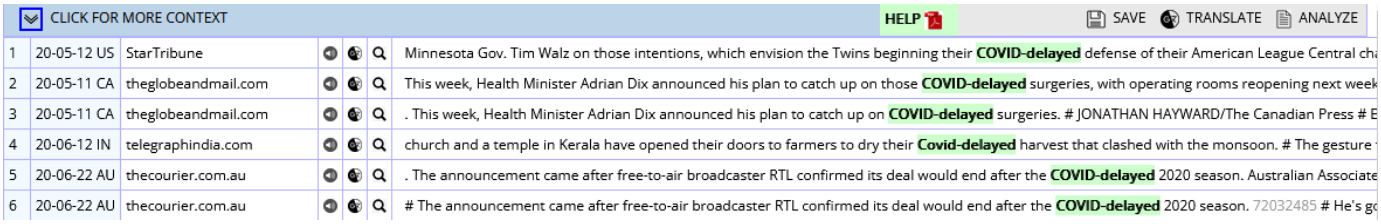

Figure 3.4 Duplicated entries

Figure 3.4 above shows that for the compound *covid-delayed* there are at least two duplicate entries shown in example [14] and [15]:

[14] "[…] Health Minister Adrian Dix announced his plan to catch up on those *COVID-delayed* surgeries […]" (20-05-11 CA)
[15] "RTL confirmed its deal would end after the *COVID-delayed* 2020 season." (20-06-22 AU)

The entries directly below either of these examples is exactly the same aside from some textual deviations that are not shown in the extended context. The scale of this thesis does not allow for cleaning up these duplicates because it is coincidental whether they are noticed or not. For instance, the duplicates of *covid-delayed* were only noticed because it was used as an example.

An issue related to the *ash-corpus* that has been compiled is that it contains exclusively compounds. This means that any form of comparison can only be done between *corona* or *covid* compounds and *ash*-compounds. Additionally, there is only one type where *aske* is the head of the compound, namely *vulkanaske* 'volcanic ash' and its definite form and is the only form where *aske* appears in a hyphenated compound *vulkan-aske*. However,

despite the limited possibilities of comparison, the *ash-corpus* is still useful for comparing the compounds found in the two corpora.

There may be multiple reasons as to why only compounding is represented in the *ash-corpus*. One possible reason is that the collected corpus is too small, and that NRK might be more reserved towards other word-formation processes. The Norwegian Newspaper Corpus used by De Smedt (2012) is considerably larger than the *ash-corpus* and could potentially include more word-formation processes than compounding, although a new study would have to be conducted.

The more likely reason is that the term *aske* may not be productive in other processes than compounding, which seems very plausible because there is not a single other process represented in the data, which seems unlikely considering that there are 108731 words in the corpus and 1073 tokens directly related to *aske*. Holmes & Enger (2018) mentions that out of the word-formation processes, it is compounding that is most frequent in Norwegian with derivation as the second most frequent process. Processes such as conversion would be hard to find, but with the lack of any infinitive markers or inflectional endings in the dataset it will be ruled out in this thesis. Because of this, the comparison does not necessarily reflect reality but may show potential areas where future research may apply.

4 Results and Discussion

The data presented in this chapter underwent the randomization process described in chapter 3, where only every 10th type has been extracted and further annotated. The data therefore do not describe the main corpus as most high frequency items are not included because they typically appear above the tenth entry in the corpus. Any word that uses *coronavirus* and *covid-19* as constituents as well as other words without *corona* and *covid* as constituents is not considered relevant for this study. With all this in mind the data examined as to what trends may emerge rather than purely examining frequencies, i.e. the number of hapax legomena counts more than the overall frequency of words.

In the analysis I will use the term *efficiency*, which in this thesis refers to if a search-string results in many or few word-formation processes compared to unrelated words. If a search string results in a sizeable percentage of word-formation processes compared to unrelated processes, then it is considered a highly efficient search string. The search strings that are found to be inefficient will not be altered underway, so the term may help explain some of the significant differences in the data as a problem with the structure of the search string rather than a lack of processes. To calculate the efficiency of a search string, the percentage of relevant types to total number of types is calculated. For instance, the search string **covid* output five relevant types and 19 non-relevant types in the 2020 data, which is an efficiency of 21%.

4.1 The *corona* and *covid* data

The tables in this section aside from Table 4.11, Table 4.12 and Table 4.13 all have the same layout. In the tables the number in parenthesis is the number of hapax legomena of the coded type i.e. the number in parenthesis is not separate from the total outside the parenthesis. For example, the solid compound has [7 (3)] tokens where three of the seven tokens are hapax legomena. The number of types also contains hapax legomenon. However, it is unnecessary to display the number of hapax legomena in parenthesis here as well because the number is the same for tokens and types. This is because a hapax legomena is a single token and therefore also a single type. Note that the tables are coded with white for 2020 data and grey for 2021 data within the search string specified in the title. All data is taken from 01.01-31.06 in each year, denoted as 01-06 in the tables. It is important to note that while it is only the first half of 2020 and 2021 that is examined it is referred to as “year 2020” and “year 2021” or only 2020 and 2021. Lastly, no table is made for year 2022 although some select data will be examined to see if some trends continue as expected or drop off.

The number of *uncoded* types and tokens is also included in addition to the words created by word-formation processes. *Uncoded* refers to any type that did not fit the criteria for word-formation processes, words with spelling mistakes (e.g. *covidiotof*) and brand-related words or names (e.g. *covidshield*). Usually, the number of uncoded types is larger than the number of morphologically complex words by a considerable margin. Lastly the total number of types in every table is the number of coded types added to the number of uncoded types.

In these sections, the output of the search strings is discussed briefly and compared across the search strings. The most important comparison is between Table 4.11 and 4.12 but comparing the two time periods of each search string will show how the frequencies change.

4.1.1 covid tables

This section has the tables that uses the search-string based on *covid* and its shortened form *cov*.

Table 4.1 covid*

	2020 – 01-06		2021 – 01-06	
Code	Types	Tokens	Types	Tokens
Affixation	0	0	0	0
Blending	1	136	1	1(1)
Conversion	0	0	0	0
Clipping	0	0	0	0
Solid Compound	5	7 (3)	6	26 (3)
Hyphenated Compound	40	79 (25)	66	929 (32)
Spaced Compound	0	0	0	0
Uncoded	114	911 (67)	128	391 (82)
Efficiency	30%		37%	
Total	160	1133	201	1347

The output of the search string *covid** shows that compounding has the most types, and that most of the hapax legomena are from hyphenated compounds. Blends have the second highest number of types which is what was expected. However, there is a considerable difference between blends with only 1 in 2020 and 1 in 2021 and compounds with 45 and 72, respectively.

In the course of a year there has been a significant increase in the raw number of types and tokens in hyphenated compounds. However, around half of the tokens represent two

types, *covid-secure* with 434 tokens and *covid-affected* with 166 tokens. The most frequent compound is *covid-secure*, which might relate to the pandemic moving towards the approval and administration of the vaccine. WHO revealed plans relating to the first deliveries of the vaccine on the 22nd of January (WHO, 2021). Here are some examples of the entries found in Table 4.1 with the word-formation process behind the example in parenthesis:

[1] "[…] it can save you from becoming a **Covidiot**." (20-04-28 IN) (Bld)
[2] "[…] accused of being **CovidNazis** for daring to tell […] citizens to stay at home […]" (20-05-17 GB) (SoC)
[3] "[…] your university may choose to reopen it with extra **COVID-precautions** […]" (20-05-17 IE) (HyC)
[4] "As one of the thousands of '**COVIDivorcées**' I can relate." (21-03-17 US) (Bld)
[5] "[…] we must brace ourselves to live in this new **Covidsphere**." (21-03-31 IN) (SoC)
[6] "[…] a number of **Covid-specialty** hospitals have been designated […]" (20-06-16 HK) (HyC)

Table 4.2 *covid

	2020 – 01-06		2021 – 01-06	
Code	Types	Tokens	Types	Tokens
Aff	3	7 (1)	3	110 (2)
Bld	0	0	0	0
Cnv	0	0	0	0
Clp	0	0	0	0
SoC	0	0	0	0
HyC	2	3 (1)	3	3 (3)
SpC	0	0	0	0
Uncoded	19	56 (13)	27	63 (17)
Efficiency	21%		18%	
Total	24	66	33	176

There are overall few types from the search string **covid* and only two word-formation processes are represented. The search string **covid* is the only configuration that allows prefixes as well as solid and hyphenated compounds with *covid* as the right-most element. The similar search string, * *covid*, only allows spaced compounds and no prefixes.
Below are some examples from the entries in the table:

[7] "[…] right-wing **PRO-COVID** crowd will gather in Hartford […]" (20-04-20 US) (Aff)
[8] "[…] the **during-Covid** part." (20-04-15 NZ) (HyC)
[9] "[…] 'it felt very **un-covid** and normal, finally,' Kirby said." (21-03-26 US) (Aff)
[10] "This will take the number of such Jumbo Field Hospitals as **mega-Covid** care facilities from 7 to 10 […]" (21-04-12 IN) (HyC)

There are more derivatives than compounding overall in Table 4.2, but the sample size is considerably smaller than e.g. Table 4.1. The similar number of hapaxes in affixation [1 and 2] and compounding [1 and 3] could mean that the productivity of the two processes is similar in the search string **covid*. The examples above are all hapax legomena from the two processes found in Table 4.2. The processes blending and clipping are not favoured by this type of search string. This is because **covid* does not lack any syllabic material as opposed to the search string *cov** which consists of the penultimate syllable of *covid* with the onset of the ultimate syllable.

Table 4.3 cov*

	2020 – 01-06		2021 – 01-06	
Code	Types	Tokens	Types	Tokens
Aff	0	0	2	5 (0)
Bld	3	5 (2)	1	1 (1)
Cnv	0	0	0	0
Clp	0	0	0	0
SoC	4	5 (3)	7	16 (3)
HyC	41	158 (22)	60	1384 (31)
SpC	0	0	0	0
Uncoded	167	6029 (92)	203	8643 (120)
Efficiency	23%		25%	
Total	216	6198	272	10047

Similar to Table 4.1, there is a considerable increase in hyphenated compounds from the year 2020 to 2021 in the dataset. Additionally, the process of compounding is the most dominant process with 45 and 67 types in 2020 and 2021, respectively. Affixation has [0 and 1] while blends have [3 and 1] respectively. As with Table 4.1, blending is the process with the second highest number of types. This number declines in the next timeframe from three to one

despite the overall increase in total types which could mean that blends become less frequent. The search-string *cov**, lacks the nucleus and coda of the ultimate syllable which means the search string allows for more blends and clippings than the other search strings. Here are some examples from the entries in the table:

[11] "This **covspiracy** has been around a long-time!" (20-03-13 GB) (Bld)
[12] "We are so doomed now **COVIDIDIOTS** […]" (20-06-21 IN) (SoC)
[13] "USA TODAY has confirmed that a **COVID-coordinator** has been hired" (20-06-16 US) (HyC)
[14] "[…] has described this as the '**Covidization** of research' […]" (21-01-18 US) (Aff)
[15] "Year of the **COVID-Preneur** […]" (21-04-08 US) (Bld)
[16] "[…] cooking up a delectable festival feast for **COVIDstruck** patients" (21-04-12 IN) (SoC)

The search string favours blends and clippings as these are the only processes of word-formation which may lack material such as example [11] and [15] above. Despite that the search string favours blends and clippings there are only four types of blends and no clipping as opposed to 112 types of compounds across both timeframes. Given these numbers, it is safe to assume that *covid* is not often used as a base in blending when compared to compounding.

Table 4.4 covid *

	2020 – 01-06		2021 – 01-06	
Code	Types	Tokens	Types	Tokens
Aff	0	0	0	0
Bld	0	0	0	0
Cnv	0	0	0	0
Clp	0	0	0	0
SoC	0	0	0	0
HyC	0	0	0	0
SpC	235	2433 (109)	253	13795 (44)
Uncoded	97	1975 (53)	147	3150 (60)
Efficiency	71%		63%	
Total	332	4408	400	16945

The only morphologically complex word that the search string *covid ** can output is spaced compounds as it is the only morphologically complex word present in this thesis that allows space between the two constituents. An interesting observation is that in comparison to the earlier tables, there are more coded items than uncoded items, which means that most of the search string's output was spaced compounds. The search string has an efficiency of 71% in 2020 and 63% in 2021. Additionally, the number of coded items overall is higher than in earlier tables. This considerable number of spaced compounds is consistent with what Bauer et al. (2013) show in their Table 19.1 (Bauer et al., 2013, p.450). The table refers to the proportion of spellings in relation to compounds and shows that two of the corpora have more than 65% spaced compounds whilst the third has 28% (Bauer et al., 2013, p.450).
While the total number of types increased by 68, the number of spaced compounds only increased by 18 from 235 to 253 tokens. In addition, the number of hapax legomena dropped from 109 to 44. Here are some examples of the spaced compounds found in the table:

[17] "The fact they put a curfew into place is one of the more obviously unrelated **COVID tactics** they've deployed […]" (20-06-01 US) (SpC)
[18] "OPI was hardly spared the **COVID scythe**" (21-05-23 US) (SpC)

The minor increase in types combined with the decrease in hapax legomena may indicate a decrease in productivity for spaced compounds in this search string. For instance, the compound in example [17] is only present in the 2020 data.

Table 4.5 * covid

	2020 – 01-06		2021 – 01-06	
Code	Types	Tokens	Types	Tokens
Aff	0	0	0	0
Bld	0	0	0	0
Cnv	0	0	0	0
Clp	0	0	0	0
SoC	0	0	0	0
HyC	0	0	0	0
SpC	0	0	0	0
Uncoded	320	3693(168)	400	14211 (74)
Efficiency	0%		0%	
Total	320	3693	400	14289

As with Table 4.4, none of the other involved word-formation processes result in spaced words and are therefore not present in the data. Despite this, there are no spaced compounds in the data which means that the search string has 0% efficiency. A possible explanation for this is that *covid* is not very productive as the second base in a word. This explanation could apply to Table 4.2 as the search string **covid* has the second lowest efficiency of 21% and 18%. On the other hand, *covid* as modifier or base for suffixes seems more productive as *covid** and *covid ** has the highest efficiencies. The two search strings, c*ovid** and *covid **, have 30% and 71% respectively in 2020 and 37% and 63% respectively in 2021. Example [19] and [20] are taken outside the filtered data to show how this type of compound may appear:

[19] "[…] hopes that '**Peak COVID**' has come and gone […]" (20-06-08 US) (SpC)
[20] "[…] It appears to be more contagious among children than **common COVID**." (21-02-09 US) (SpC)

4.1.2 *corona* tables

This section contains the tables that uses the search-string based on *corona* and its shortened form *cor*.

Table 4.6 corona*

	2020 – 01-06		2021 – 01-06	
Code	Types	Tokens	Types	Tokens
Aff	0	0	0	0
Bld	0	0	2	6 (0)
Cnv	0	0	0	0
Clp	0	0	0	0
SoC	6	21 (4)	4	10 (1)
HyC	22	51 (15)	5	5 (5)
SpC	0	0	0	0
Uncoded	108	1222 (58)	59	376 (35)
Efficiency	21%		16%	
Total	136	1284	70	398

While Table 4.1 shows an increase in overall frequency from 2020 to 2021 Table 4.6 shows a decrease in types and tokens. It appears that both *corona** and *covid** search strings favour

compounding, with hyphenated compounds as the majority. The total number of compounds declined from 28 in 2020 to only 9 in 2021. This is more than three times less despite that the total number of types was reduced by only half. The total number of blends did increase from zero to two types, although none of the six tokens is a hapax legomenon, indicating that the process is not productive in this search string. However, as with Table 4.1, compounding has the most types overall in the data. Here are some examples from the data:

[21] "[…] young people and families taking what many were calling a **coronavacation**." (20-03-20 US) (SoC)
[22] "The same thing that's wreaking **corona-havoc** in other Latin American countries." (20-05-21 US) (HyC)
[23] "But now we've come to the global **coronacession** […]" (21-03-11 AU) (Bld)
[24] "[…] use '**coronasutra**' positions to reduce face-to-face contact […]" (21-04-10 US) (SoC)
[25] "[…] irrespective of **corona-inspired** unemployment, […]" (21-01-28 US) (HyC)

Table 4.7 *corona

	2020 – 01-06		2021 – 01-06	
Code	Coded types	Tokens	Coded types	Tokens
Aff	0	0	1	6 (0)
Bld	0	0	0	0
Cnv	0	0	0	0
Clp	0	0	0	0
SoC	0	0	0	0
HyC	0	0	0	0
SpC	0	0	0	0
Uncoded	12	27 (7)	4	5 (3)
Efficiency	0%		20%	
Total	12	27	5	11

There are fewer overall types and tokens from this search string than with **covid.* The 2020 data had no tokens that correspond to the word-formation processes of interest. In the 2021 data there is a minimal increase from zero affixed bases to one. The efficiency in 2021 was 20% due to the low sample size, but without a single hapax legomenon in the data. Overall, this shows that solid or hyphenated compounds with *corona* as head are less favoured than

solid or hyphenated compounds with *covid* as head. This implies that *corona* rarely takes premodifiers such as prefixes or other bases. The only morphologically complex word in the filtered data is *non-corona*, which is given more context below:

[26] "[…] too much staff and resources into coronavirus and neglecting the **non-corona** patients […]" (21-01-28 US) (HyC)
[27] "In **non-corona** times, every day there are around 2,000 students here." (21-01-27 US)

Table 4.8 cor*

	2020 – 01-06		2021 – 01-06	
Code	Types	Tokens	Types	Tokens
Aff	0	0	0	0
Bld	1	1 (1)	2	2 (2)
Cnv	0	0	0	0
Clp	0	0	0	0
SoC	9	50 (3)	2	4 (1)
HyC	18	28 (12)	7	23 (3)
SpC	0	0	0	0
Uncoded	326	20565 (166)	252	13901 (127)
Efficiency	8%		5%	
Total	354	20641	264	13931

Table 4.8 also shows a drastic reduction in the number of types as well as the overall frequency of tokens. The only increase was in the number of hapax legomena created by blending from one to two, but otherwise solid compounds decreased from nine to two types and hyphenated compounds from 18 to seven types. Looking at the Tables so far, it becomes clear that there is an overall trend that *corona* related search strings decline in types and tokens created by compounding from 2020 to 2021, and only blending and affixation have had an increase, although marginal. This reduction can also be seen in Table 4.6 and 4.7. The search string *cor** lacks the nucleus of the penultimate syllable as well as the entire ultimate syllable. Like Table 4.3, the lack of material in the search string *cor** should favour blends, but compounding is still the most frequent process. Here are some examples with context from the data:

[28] “Even without the ‘**corona-demic**’, hospitals and clinics would be gutted if doctors stayed at home […]” (20-03-10 AU) (Bld)
[29] “**Coronabias** and the fat tail” (20-04-22 IE) (SoC)
[30] “[…] any newspaper is likely to be **corona-covered** […]” (20-04-23 AU) (HyC)
[31] “[…] the time warp has featured an element of ‘**coronastalgia’**, if you will.” (21-01-11 US) (Bld)
[32] “The Covid package is wrapped into a $2.3 trillion, almost 5,600-page ‘**coronabus**’ bill” (21-01-06 US) (SoC)
[33] “[…] the contribution of **Corona-Volunteers** in prevention and rescue of Covid-19 has been commendable.” (21-05-21 IN) (HyC)

Examples [28] and [31] are two of the three blends in the data. Example [28] is a shortening of the compound *corona pandemic* to *corona-demic*, and [31] refers to missing the quarantine life.

Table 4.9 corona *

	2020 – 01-06		2021 – 01-06	
Code	Coded types	Tokens	Coded types	Tokens
Aff	0	0	0	0
Bld	0	0	0	0
Cnv	0	0	0	0
Clp	0	0	0	0
SoC	0	0	0	0
HyC	0	0	0	0
SpC	115	882 (56)	51	230 (24)
Uncoded	61	624 (34)	34	457 (17)
Efficiency	65%		61%	
Total	176	1506	86	688

Table 4.9 shows a decline in spaced compounds with *corona* as left constituent from 115 to 51 types. The opposite occurred with the spaced compounds with *covid* as *base* in Table 4.4. This decline follows the observed trend mentioned under Table 4.8 where compounding with *corona* as base decline in frequency from year 2020 to 2021. The opposite occurs with processes using *covid* as the base. The only type of word-formation process allowed by the

search string *corona* * is spaced compounds as none of the other processes allow for a spacing between the base and a different element. Here are some examples from the data:

[34] "'don't touch Fido's **corona ball**, it's covered in dog slobber'." (20-03-30 AU) (SpC)
[35] "The stock rebounded 163% from its lowest level after the **Corona crash** last March." (21-02-03 NZ) (SpC)

Example [35] is one of the hapax legomena in the data and refers to the economic crash caused by *corona*.

Table 4.10 * corona

	2020 – 01-06		2021 – 01-06	
Code	Coded types	Tokens	Coded types	Tokens
Aff	0	0	0	0
Bld	0	0	0	0
Cnv	0	0	0	0
Clp	0	0	0	0
SoC	0	0	0	0
HyC	0	0	0	0
SpC	3	12 (0)	1	1 (1)
Uncoded	162	1399 (88)	94	504 (55)
Efficiency	2%		2%	
Total	164	1406	95	505

Table 4.10 has a less noticeable reduction in the word-formation processes due to the small number of types, but the total number of types has been more than halved from 2020 to 2021. However, the total number of hapax legomena increased from zero to one in the same time span, meaning that the process increased slightly in productivity.
When Table 4.10 is compared to Table 4.9 it is clear that *corona* is used more frequently as a modifier than a head for the spaced search strings. This is also the case with the search strings with the search strings that are written solidly with the asterisk, e.g. *cor** and *corona** have a higher type and token count than **corona*. This is similar to what was noted in Table 4.5 with *covid,* which indicates that neither of the two bases are favoured by prefixes nor as the head of compounds. Below are some of the few entries output of this search string:

[36] “It is similar to **past Corona**.” (20-04-19 ZA) (SpC)
[37] “[…] he’s not coming here during **peak corona**.” (21-01-13 US) (SpC)

4.1.3 Data across search strings

The following two tables include all information that the tables in section 4.1.1 and 4.1.2 have. In addition, the tables feature the total number of compounds as well as the percentage of hapax legomena to coded types. The P-value for the total number of compounds was not calculated by adding the individual P-values from the three different compounds, but by using the standard equation. The hapax legomena percentage does not include the uncoded types as these do not represent any word-formation processes.

Table 4.11 Overall frequency across search strings for covid

All search strings	2020 – 01-06			2021 – 01-06		
Code	Types	Tokens	P	Types	Tokens	P
Aff	3	7 (1)	0,143	5	115 (2)	0,017
Bld	4	141 (2)	0,014	2	2 (2)	1,0
Cnv	0	0	N/A	0	0	N/A
Clp	0	0	N/A	0	0	N/A
SoC	9	12 (6)	0,500	13	42 (6)	0,143
HyC	83	243 (48)	0,198	130	2317 (67)	0,029
SpC	235	2433 (109)	0,045	253	13795 (44)	0,003
Comp. Total	327	2686 (163)	0,061	395	16153 (116)	0,007
Covid total	334	2831 (166)	0,059	402	16270 (120)	0,007
Uncoded	715	18393 (392)	0,021	905	26580 (353)	0,013
HL%	50%			29%		
Total	1049	21225		1307	42850	

Table 4.11 shows that there is a drastic increase in hyphenated compounds from year 2020 to 2021 with an increase from 83 to 130 types while the number of hapax legomena increased from 48 to 67. The P-value decreased from 0,198 to 0,029 in the same timeframe. The P-value, as discussed in section 2.3, refers to the likelihood that a token from a word-formation process is a hapax legomenon. The P-value for affixation in 2020 0,143 means there is a 14,3% chance that a token of affixation is a hapax legomenon. For example, if there were 1000 tokens of affixation then 143 of these would be hapax legomena. The percentage denoted as HL% is the percentage of types that are hapax legomena of all word-formation processes. This measure ignores the uncoded elements as the main interest is to see how many

hapax legomena created by word-formation processes are in the data. The HL% shows that 50% of all types in the 2020 data are hapax legomena compared to the 29% in 2021. This shows a decline in overall productivity when *covid* is used as a constituent, but also that in 2020 half of all entries were new words. The process of compounding is the most frequent process in this table with 327 types in 2020 and 395 types in 2021 as well as 163 and 116 hapax legomena, respectively. There is a considerable jump from the most frequent process to the second most frequent process as blends have four types in 2020 and two in 2021 while affixation has three types in 2020 and five types in 2021. It seems that after the pandemic has lasted a year, there is less variety in the words created because while the number of types and tokens increases from 334 and 2831 to 402 and 16270 respectively, the number of hapaxes decrease from 166 to 120. The P-value also significantly decreases from 0,059 to 0,007, which is because the P-value becomes smaller as the number of tokens increases and the number of hapaxes decrease. Because this data is collected from news sites it may be a change in coverage of the pandemic is the reason, although a different type of analysis would be required to confirm this.

Table 4.12 Overall frequency across search strings for *corona*

All search strings	2020 – 01-06			2021 – 01-06		
Code	Types	Tokens	P	Types	Tokens	P
Aff	0	0	N/A	1	6 (0)	0
Bld	1	1 (1)	1,0	4	8 (2)	0,250
Cnv	0	0	N/A	0	0	N/A
Clp	0	0	N/A	0	0	N/A
SoC	15	61 (7)	0,115	6	15 (2)	0,133
HyC	38	74 (27)	0,365	12	28 (8)	0,286
SpC	119	955 (56)	0,059	52	222 (25)	0,113
Comp. Total	172	1090 (93)	0,085	70	265 (35)	0,132
Corona total	173	1091 (94)	0,086	76	280 (38)	0,136
Uncoded	670	23837 (353)	0,015	443	15268 (253)	0,0166
HL%	54%			50%		
Total	843	24928		519	15548	

Table 4.12 shows that there has been an overall decrease in all types and tokens of compounding, whilst there has been a slight increase in the types of blends from one to five types and derivatives from zero to one. The data show that compounding is the most frequent process across both bases and that blending is the second most frequent process. Out of the

three types of compounds it is the solid compound that is the least frequent. This might show that it is more difficult to create solid compounds with either of the two bases despite that the structure is the least ambiguous alongside hyphenated compounds (Bauer et al., 2013, pp.431-432). Bauer et al. (2013, p.450) state that the more lexicalized a compound is the more likely it is to be written as a solid compound. This is likely why there are more spaced and hyphenated compounds in the data than solid compounds as the data only reflects one and a half years in the pandemic which may be too short for the neologisms to become lexicalized. In every table thus far *corona* as a base in compounding has declined in number of types and tokens consistently. It could be because there is less data in the 2021 timeframe. However, the number of coded types is not always reduced in the same degree as total number of types. For example, in Table 4.6 the total number of types is reduced by half from 136 to 70 while the other elements are not halved but reduced by other factors. The number of coded hapaxes drop from 19 to six and the number of coded types drop from 28 to 11. This uneven reduction shows that in this table the decline in types and hapax legomena is not caused by the reduction in sample size alone.

Table 4.13 Overall frequency across constituent

All search strings	2020 – 01-06			2021 – 01-06		
Code	Types	Tokens	P	Types	Tokens	P
Aff	3	7 (1)	0,143	6	121 (2)	0,017
Bld	5	142 (3)	0,021	6	10 (4)	0,400
Cnv	0	0	N/A	0	0	N/A
Clp	0	0	N/A	0	0	N/A
SoC	24	73 (13)	0,178	19	57 (8)	0,140
HyC	121	317 (75)	0,237	142	2345 (75)	0,032
SpC	354	3388 (165)	0,049	305	14017 (69)	0,005
Comp. Total	499	3776(256)	0,068	465	16417 (151)	0,009
Total of all	507	3922 (260)	0,066	478	16550 (158)	0,010
Uncoded	1385	42231 (745)	0,018	1348	41848 (606)	0,015
HL%	51%			33%		
Total	1892	46153		1826	58398	

Table 4.13 shows the overall frequencies of the different word-formation processes regardless of constituent. The data in this table shows that it there is a decline in productivity of all word-formation processes from 0,066 to 0,009 from 2020 to 2021. Considering that the number of tokens increase while the number of types and hapaxes decrease, it is likely that more word

forms have become normalized and that there is less variety in the neologisms used by news agencies in the year 2021.

4.2 The ash data

In this section the data related to the *ash-corpus* is displayed and will be discussed.

Table 4.14 Ash data

	Types	Percentage (Types)	Tokens	Percentage (Tokens)
Ash-words (excluding hapax legomena)	30	53 %	150	85 %
Hapax legomena	27	47 %	27	15 %
Solid compounds	56	98%	176	99,5%
Hyphenated compounds	1	2%	1	0,5%
Total ash-words	57	100%	177	100 %
P-value for ash-words	0,153			
Total in corpus	10416		108731	

The words *aske* ‘ash’ and *askesky* ‘ashplume’ are not considered as types in this *ash-corpus* because they were often used prior to the volcanic eruption in 2010 and are therefore excluded from the data. There are other terms that have existed prior to the eruption, e.g. *askelag* ‘ash layer’ and *askespredning* ‘ash spread’, but these are not as frequent as the two aforementioned words and are therefore included. The most frequent word in the data other than *aske* and *askesky* is that of *askefast* ‘ash stranded’ with 29 tokens. This is more than twice as many tokens than the next frequent word *askeproblem* ‘ash problem’ with twelve tokens. The word *askefast* is likely more frequent because it represents a new and unfamiliar situation where families were stranded in foreign countries because of the eruption. The word *askefast* co-occurs with the verb *å sitte* ‘to sit’ to form a verb phrase *å sitte askefast* ‘to be ashstranded’ which differs from other words like *askeproblem* which is a noun.
The only relevant hyphenated compound found is that of *aske-erstatning* ‘ash-recompense’ which is a hapax legomenon. There is a high percentage of hapax legomena in the collected dataset at 47% of the 57 types.

The *ash-corpus* contains more raw data than the data collected from *The Coronavirus Corpus* (2019-) but despite this only has 57 *ash*-related types. The reason the *ash-corpus* has more types and tokens than the data extracted from *The Coronavirus Corpus* (2019-) is because the data collected from the NRK event section was not filtered for *aske* until after the

collection whilst the data collected from *The Coronavirus Corpus* (2019-) was filtered through the use of search-strings prior to the collection which limited the output to only *corona* and *covid* related types and tokens.

The most significant difference between the two corpora is that the only word-formation process represented in the *ash-corpus* is compounding, consisting of one hyphenated compound and 56 solid compounds. This lack of variance possibly indicates that *aske* is not very productive in the other processes, at the very least in those words caught by NRK and their writers. It is possible that a different result would be found if data is collected from social media and other forums, but this would require a separate study with considerably more time. It is also possible that the other processes are not as frequently used in Norwegian. Holmes & Enger (2018) describe Norwegian word-formation, but only lists compounding and derivation as primary processes with no discussion of blending (Holmes & Enger, 2018, pp.454-490). In addition, Holmes & Enger (2018) mention that compounding is likely an "even more productive aspect Norwegian word-formation than is derivation […]" (Holmes & Enger, 2018, p.456). In the case of the minor processes, Holmes & Enger (2018) state that conversion is likely rarer in Norwegian than in English although they note that it is "common for a verb stem and a noun stem to be related without any affixation" (Holmes & Enger, 2018, pp.483-484). The process of clipping is also not mentioned in their discussion of the processes, and it may be concluded that these processes are less frequent in Norwegian than in English, which is also suggested by the results found in this thesis. The reason that the main output of compounding in the data is solid compounds is because in Norwegian every compound is written solid (or potentially hyphenated) and not spaced (Holmes & Enger, 2018, p.458).

4.3 Discussion of the data

The following sections are dedicated to discussing the data found in section 4.1 and 4.2. Section 4.3.1 focuses on the *covid* and *corona* data, section 4.3.2 looks at the ash data alongside the *corona* and *covid* data. Section 4.3.3 discusses data that was not in the filtered dataset but found in the unfiltered data.

4.3.1 *Covid* and *corona* data

Compounding is the most frequent word-formation process overall and judging purely by raw numbers it is *covid* as base that is favoured in both time frames. The constituent *covid* has a

total of 334 types in 2020 which is close to two times more than *corona* with 173 types, and 402 types in 2021 which is more than five times the number of types for *corona* with 76 types. The only form of compounding where *corona* is favoured as base is for solid compounds, but only in the 2020 data with 15 types and seven hapaxes while *covid* has nine types and six hapaxes in the same timeframe. In blending the favoured constituent is *covid* in 2020 with four types and two hapaxes, but *corona* is favoured in 2021 with four types and two hapaxes.

If we judge based on P-values, the productivity of a process, then the result is different. According to P-value in Table 4.13 the process that is most productive in the data across constituents is blending with 0,400 in 2021 and derivational affixation with 0,143 in 2020. While hyphenated compounds has a higher P-value in 2020 of 0,237 it is part of the process *compounding* rather than representing the entire process. The constituent favoured in compounding overall is *corona* with a P-value of 0,085 in 2020 and 0,132 in 2021 where *covid* has 0,061 in 2020 and 0,007 in 2021. Affixation favours the use of *covid* in both timeframes with P-values of 0,143 and 0,017 in 2020 and 2021 respectively while *corona* has no P-value in 2020 as there are no derivatives and a P-value of 0 in 2021 due to a lack of hapaxes. The last attested process, blending, favours *corona* in the 2020 data with a P-value of 1, though there is only one type, while *covid* has 0,014. In the 2021 data it is *covid* that is favoured with a P-value of 1, where *corona* has a P-value of 0,250. The constituent that is favoured in the data overall is *corona* as the constituent has a P-value of 0,086 in 2020 and 0,136 in 2021 while *covid* has a P-value of 0,059 in 2020 and 0,007 in 2021. These P-values are across word-formation processes, but despite this are similar to the P-value of compounding for each of the constituents, which shows how much this process affects the measure of productivity in this thesis.

I found it surprising that derivational affixation and blends had a very low number of types across the constituents, with only nine types of derivatives and eleven types of blending in the data, with three and seven hapaxes respectively. In addition, there was no representation of conversion and clipping in the filtered data. The lack of clipping in the data is not surprising given the restrictions of the data-collection as no search string is able to exclusively target clipping. Using strings like "*rona" which lack the initial syllable (here the antepenultimate) results in clipping. However, this search string resulted in only one type and therefore would more likely just skew the results in addition, creating a corresponding string for *covid* would be difficult as "*vid" does not seem to be a functional clipping.

As mentioned in section 4.1, there seems to be a trend in this data that the processes that use *corona* as constituent declines in the number of both types and tokens from the year 2020 to 2021. On the other hand, the processes that use *covid* as constituent increase in number of types and tokens in the same timeframe. An interesting observation is that while the number of types increase for spaced compounds with *covid*, the number of hapax legomena decrease from 109 to 44. Because of the increase of tokens and decrease of hapax legomena the P-value decreases considerably, from 0,045 to 0,003 which is about 14 times less. This might be because the 2021 data featured considerably more types of *covid*, i.e. the hapax legomena start further down in *The Coronavirus Corpus* (2019-). As mentioned in chapter 3.3, the corpus can only display a maximum of 4000 types, which means that more hapax legomena will not be detected if the search string outputs more than 4000 types. The method of data collection could not be a factor in the decline of hapax legomena in spaced compounds because the method was designed to collect more hapax legomena.

4.3.2 *Ash, corona* and *covid*

Comparing the three constituents and the words formed with them, it becomes clear that the semantics of some of the words are similar. For instance, *askekaos* 'ash chaos' and *covid chaos* both refer to some form of chaos caused by the modifying constituent. Likewise, *askerammede* 'ash affected' and *covid-affected* refer to the impact that the event has on someone or something. Below are some instances in context.

[1] "[…] the **COVID-affected** countries" (20-03-11 IN)
[2] "Se bilder fra de **askerammede** gårdene på Island" (25.06. 2010 NRK)
View pictures from the ash affected farms on Iceland

Example [2] is copied from a hyperlink on NRK, which is why the infinitive form of *se* 'to see' is used.

Table 4.15 P-values for compounds

Dataset	Types	Tokens (Hapax legomena)	P-value
covid-compounds (2020)	327	2686 (163)	0,061
corona-compounds (2020)	172	1090 (93)	0,085
covid-compounds (2021)	395	16153 (116)	0,007
corona-compounds (2021)	70	265 (35)	0,132
ash-compounds	57	177 (27)	0,153

The P-value for *ash*-compounds is considerably higher than most of the P-values. The second highest P-value is 0,132 for the *corona*-compounds in 2021. This implies that for every token of an *ash*-compound there is a 15,3% chance that one token is a hapax legomenon, while there is a 13,2% chance for *corona*-compounds. For tokens of *covid*-compounds in 2021 this chance is only 0,7% despite that the hapaxes make up 29% of all types with *covid*. This means that while the overall number of *covid*-compounds increased from 2020 to 2021, the productivity decreased considerably. The opposite is true for *corona*-compounds, where the P-value in 2020 is 0,085 which nearly doubles in 2021 at 0,132. All of this means that *aske* is the most productive constituent for compounding out of all the constituents based on the P-value.

The *ash-corpus* consists of 10416 types and 108731 tokens whilst the filtered data taken from *The Coronavirus Corpus* (2019-) consist of 3718 types and 104551 tokens. 57 types from the ash corpus were ash-related words and 985 types from the filtered data from *The Coronavirus Corpus* (2019-) were words created with *corona* or *covid* as a constituent. Crucially, the sample size of the *ash-corpus* is limited to NRK articles and may therefore not paint the full picture as to the productivity of this constituent. The considerable difference in *ash*-related types and tokens compared to *covid* and *corona* related ones is expected due to the considerable difference in scale, with the volcanic eruption affecting the Nordic countries primarily, whilst the *coronavirus* pandemic affects close to the entire world in some way. Therefore, there are more news agencies that take up the *coronavirus* as a topic.

4.3.3 Some observations in the unfiltered data

The term "unfiltered data" does not refer to all the data present in the corpus, but it refers to the data collected from the two timeframes without applying the filter of every tenth entry. It is therefore still a limited sample that only examines 12 months total and not the entire duration of the pandemic. The unfiltered data does have most processes of word-formation but as it has not been examined thoroughly (as it was not the goal of this thesis) it is uncertain whether clipping is present outside of compounds. Conversion is present in the unfiltered data, however, with for instance *covided* and *coronaed* as in the examples below:

[3] "'We're all **COVIDed** to death.'" (21-02-01 CA)
[4] "I have been **covided**!" (21-04-27 AU)
[5] "[…] doing anything I can to not get **coronaed**." (20-03-03 GB)
[6] "[…] for the time being it looks well and truly **'coronaed'**" (20-03-10 AU)

These examples are clearly cases of conversion as the inflectional past tense suffix *-ed* only attaches to verbs. Both bases, *corona* and *covid*, are normally nouns and therefore must have been converted to verbs in order for the suffix to attach.
In the unfiltered data there is a trend with the search strings **covid* and **corona* where the most frequent elements are the prefixes *anti-*, *pre-*, *post-*, and *non- covid*/*corona*. All of these result in a prefixed word that can be used as an adjective. Below are some examples from the unfiltered data:

[7] "[…] but it could be a while before returning to **pre-COVID** life." (20-03-20 US)
[8] "[…] items such as counterfeit masks and **anti-corona** sprays." (20-03-31 GB)
[9] "[…] in the **post-corona** world." (21-01-27 US)
[10] "[…] a field hospital was deployed to handle **non-covid** care." (21-06-21 IE)

As with the filtered data it is *covid* based words that has the highest token count, e.g. *pre-covid* at 3584 tokens as opposed to *pre-corona* at 16 tokens in the 2021 data.

4.3.3.1 Some observations in the corpus as a whole

This data is still related to the timeframe of interest, 01-06 in the respective year, but in addition a quick search has been done in the year 2022 to see how the frequency of the two constituents *corona* and *covid* has developed.
An interesting change from 2020 to 2021 is that *covid* became considerably more frequent from 65882 tokens to 239922 while the use of *covid-19* declined from 1166535 tokens to 921025. This change is likely the reason that the number of types and tokens in Table 4.12 increased considerably from 2020 to 2021. In fact, a quick examination of the same timeframe in 2022 shows that the use of *covid* continues to increase to 276105 whilst the frequency of *covid-19* decreases to 506052. A similar change happened to *coronavirus* where the token count declines from 1155534 in 2020 to 311893 in 2021. However, *corona* also decreases in frequency in the same timeframe, from 25812 in 2020 to 9332 in 2021. The same timeframe in 2022 is no different, where *coronavirus* has dropped to 104723 tokens and *corona* 2064 tokens. This decline in use corresponds to the overall reduction seen in Table 4.12 and seems likely to continue as the pandemic goes on.
The adjective *covidian* is an interesting derivative. It is made up of *covid* plus the adjectival suffix *-ian*. This suffix was present in the unfiltered data but has a token count of 2 and 8 in

2020 and 2021, respectively. In 2022, however, this word has increased to 23 tokens. This word is an adjective used to describe someone that has tested positive for covid as in example [11] below:

[11] “Was I exposed to a **Covidian** co-worker?” (22-01-27 US)

Interestingly, the *-ity* suffix was applied to *covidian* in 2021 to form *covidianity* which is referring to a *covid* related religion in the same way as “christianity”. This derivative has two tokens in 2021 but no occurrences in 2020, nor in 2022 shown in example [12] below:

[12] “Rob Slane suggests that 2020 saw the birth of a new religion, ‘**Covidianity**.’” (21-01-07 LK)

This derivative is interesting because the stem *covidian* is an adjective referring to something *covid* related, while the stem of Christianity, Christian, may refer to someone who practices Christianity i.e. the element that *-ity* attached to.

5 Conclusion

This thesis analysed data from *The Coronavirus Corpus* (2019-) in order to examine how word-formation processes use the two constituents *corona* and *covid* when creating new words. Additionally, a comparative self-compiled *ash-corpus* was created to explore differences and similarities between the word-formation occurring in two different languages. In order to guide this analysis, the following research questions were formulated:

1. Is one of the two constituents *corona* and *covid* favoured by word-formation processes?
1a. If one constituent is favoured over the other, which word-formation process favours which constituent?
1b. Which word-formation process is the most productive in the data overall?

2. What differences and similarities exist between the analysis of the *ash-corpus* and the analysis of data from *The Coronavirus Corpus* (2019-)?
2a. Which word-formation process, if any, is the most productive in the *ash-corpus*?
2b. Are there any differences in terms of productivity between the two analyses?

The corpus study conducted in this thesis was done in order to answer these questions. Search-strings were used in order to streamline the data collection when searching in *The Coronavirus Corpus* (2019-), after which the data was further filtered to only include every tenth entry. This data was then annotated according to the relevant word-formation processes and the context of some entries was examined to remove ambiguity. The word-formation processes used in the analysis in this thesis are derivational affixation, compounding, blending, clipping and conversion.

The self-compiled *ash-corpus* consists of 265 articles collected from the Norwegian news agency NRK. This data was collected from their event section dedicated to the volcanic eruption in Iceland on the April 2010 due to its widespread effect on the northern countries of Europe. The articles in the event section span around 1 year from 2010 to 2011. Each article was pasted into a document and analysed using the program AntConc in order to find words as well as purge duplicates. In total this corpus contains 10416 types and 108731 tokens of which 57 types and 177 tokens are *ash*-words.

5.1 Findings

The following sections will answer the research questions mentioned above.

5.1.1 *corona* and *covid* data

Both of the constituents, *corona* and *covid*, are favoured by various word-formation processes. The only processes that do not favour either is clipping and conversion due to a lack of data. Crucially, these processes do exist both within *The Coronavirus Corpus* (2019-) and the unfiltered data, but because of the filter were not collected. The P-value for compounding is higher for *corona* than for *covid* which means that compounding is one of the processes that favour *corona* as constituent. Derivational affixation is the opposite where the P-value is higher for *covid* and zero for *corona*. Blending favours *corona* in the 2020 data but *covid* in the 2021 data which implies there may have been a shift in the overall popularity of the two constituents. Interestingly the P-value for each of the constituents when favoured was 1, which means that every token was a hapax legomena in the data. Though the token count for *corona* in 2020 was 1 and for *covid* in 2021 it was 2, which is not very large.
The most frequent process in the data overall is compounding while the second most frequent process was blending, which is consistent with the findings of Alyeksyeyeva et al. (2020, p.34). While compounding is the most frequent process in the data, it is blending that is the most productive process. As is seen in the data, a high frequency means it is less likely that the process has a high P-value.

Because the P-value is higher for *corona* in general it means that the productivity of any word-formation process using *corona* as constituent is high. The P-value is at the highest in 2021 at 0,134 which means that there is a 13% chance of finding a hapax legomenon amongst all tokens with *corona* as constituent. Essentially it means that out of the two constituents, *covid* and *corona*, it is *corona* that is favoured and most productive when creating new words. This is expected considering that the overall frequency of *covid* has increased over time whilst *corona* has decreased. It is likely that this difference in popularity means that more *covid* words become institutionalized and results in less hapax legomena.

5.1.2 *corona, covid* and *aske* findings

The key difference between the *ash-corpus* and *The Coronavirus Corpus* (2019-) is that the former corpus contains only compounding. Crucially, *The Coronavirus Corpus* (2019-) was filtered and further reduced to only collect every tenth entry yet still contained compounding,

blending and derivational affixation while the *ash-corpus* was not filtered at all yet only contains compounding. This implies that, at least for NRK, *aske* as a constituent is only productive when used to form compounds and does not partake in any other word-formation process.

Because the *ash-corpus* only contains compounding it is a given that this process is the most productive. In addition, compounds with *aske-* as a constituent are the most productive compounds in the data overall. As shown in Table 4.14 *ash*-compounds has a P-value of 0,153 whilst the second highest P-value, which belongs to the *corona* compounds in 2021, is 0,132. To compare, the highest P-value of *covid* compounds is in 2020 at 0,061. This shows that it is more likely that a given token of an *ash*-compound is a hapax legomenon than a given token of a *corona* or *covid* compound. However, the raw number of hapax legomena formed during the pandemic is considerably higher than the number of hapax legomena found in the NRK event section. This can be attributed to the difference in scale of the two events, where the eruption affected mainly the Nordic countries (e.g. Norway, Denmark, Sweden, Iceland) as well as Britain whilst the pandemic is world-wide with few countries unaffected by the virus. Because of the internet, even if a country was unaffected by the virus, they have been affected by the media coverage from other countries or from social media platforms. In addition, the virus has caused some economic problems, hence the rise of words such as *coronanomics*, *coronacrash* as well as *coronacredit* etc..

What becomes clear from the comparisons is that people have created new words to relate to these events and the ways the event has affected their daily lives. One of the most popular words in the data, *covidiot*, is an example of the attitude taken towards those who do not respect curfews or covid guidelines. The most popular *ash*-compound is *askefast* 'ash stranded' and refers to someone stranded at an airport or in a country unable to travel home because of the ash, something that a lot of people experienced during this eruption.

Considering that *ash*-neologisms have fallen largely out of use today it seems likely that *coroneologisms* will also fade away over time once the *coronavirus* pandemic calms down and the virus and lockdowns do not affect anyone. This is because the majority of neologisms formed during either of these events have very specific references, e.g. *covspiracy* refers to conspiracy theories related to covid and *askekrav* 'ash claim' refers to having a claim because of something *ash*-related. Once there are no covid related guidelines or ash problems then there will be no need for either word.

5.2 Contributions

To my knowledge, no study before has compared the *ash* word-formation that occurred during the volcanic eruption in 2010 with the *corona* and *covid* word-formation that is currently occurring to the same extent as this thesis. Measuring the productivity of not only the word-formation processes using *corona* and *covid* as bases but also the processes that use *aske* as a constituent creates a more in depth understanding of these processes and illustrates the possible areas where future research can apply. De Smedt (2020) used some examples to refer to a similarity between the pandemic and the eruption but did not compare the two events and examine the word-formation of both.

I hope the current thesis has shown possible areas of further research relating to word-formation processes in relation to events such as pandemics or similar that affect the general populace negatively. As is shown in this thesis, there are similarities in the words created in both analyses.

I also hope the thesis shows that there are similarities in the word-formation that happens across languages, but also that there are differences in which process is most frequent or is considered rare. This is highlighted in the comparison between the English and Norwegian data as not all processes were attested. For instance, only compounding was present in the Norwegian data whilst three out of five processes of interest were present in the English data.

5.3 Potential areas of future research

If I had more time, I would have expanded the searches with either more diverse search strings, or thorough analysis. For example, instead of every tenth entry I could have collected every fifth or, with enough time, I could have analysed 4000 entries per search string. At that scale it would be possible with more certainty in the analysis as well as less chance at skipping relevant entries. As mentioned earlier, the increase or decrease of hapax legomena between the two time periods can be attributed to the collection of only every tenth entry. Doing another study seeks to include conversion and clipping more efficiently would also be interesting, as the two processes were present in the data but were not collected because of the filter. This would be a direct consequence of analysing more entries than every tenth as suggested above. In addition, focusing exclusively on conversion or clipping or including abbreviations could be interesting as well.

Performing an in-depth comparison between the compounds formed with *corona* and *covid* to the compounds formed with *aske* could show similarities or differences between the

formulation of compounds in English speaking countries and Norwegian. This could be interesting because similarities were found in the current thesis, despite that the *ash-corpus* was used as a small part of the thesis.

When enough time has passed it could be interesting to do a study on which words are still used once the pandemic has subsided. Due to the nature of *The Coronavirus Corpus* (2019-) it will likely collect data up until the point when the pandemic is no longer relevant, or when news articles do not include the terms that the corpus searches for. Comparing the frequency of the different word-formation processes over time could show how many neologisms have a chance of staying. It would then be possible to see if there are certain features a word requires in order to continue its use once its primary reference is less relevant. Additionally, it would be possible to see if one type of word-formation process is more likely to create a long-lasting word.

Performing a study with a bigger focus on the comparison of word-formation processes across languages could prove interesting, because this thesis found next to no variation in Norwegian while there were multiple different entries in English. While Enger & Holmes (2018) state that compounding is the most frequent word-formation process in Norwegian, it does not mean that it should be the only one possible to affect *aske* as a constituent. Therefore, expanding the search to use more news corporations could increase the chances of finding more processes. For instance, *askete* 'ashy' should be a possible derivation to describe something as being ash-like such as "en askete lukt", 'an ashy smell', but was not found in the NRK data. Examining difference in word-formation using *corona* and *covid* as constituent in Norwegian and English would be an interesting study, as it would make it easier to find differences and similarities between which word-formation processes are used in the two languages.

References

Akut, K. B., (2020). Morphological analysis of the Neologisms during the COVID-19 Pandemic. *International journal of English Language Studies*, *2*(3). 01-07 DOI:10.32996/ijels.2020.2.3.11

Al-Salman, S., Haider, A. S. (2021) COVID-19 trending neologisms and word formation processes in English. *Russian Journal of Linguistics 25* (1). 24–42. DOI: 10.22363/2687-0088-2021-25-1-24-42

Alyeksyeyeva, I. O., Chaiuk, T. A., Galitska, E. A. (2020) Coronaspeak as Key to Coronaculture: Studying New Cultural Practices Through Neologisms. *International Journal of English Linguistics*, *10*(6). DOI: 10.5539/ijel.v10n6p202

Bauer, L., (2003) *Introducing Linguistic Morphology* (2nd ed.) Edinburgh University Press

Bauer L., Lieber R., Plag I. (2013) *Oxford reference guide to English morphology* Oxford University Press

Buchstaller, I., Mearns, A. (2018) The Effect of Economic Trajectory and Speaker Profile on Lifespan Change: Evidence from Stative Possessives on Tyneside. In S., Jansen, N., Braber, (eds.) *Sociolinguistics in England*. London: Palgrave, 215–241. DOI: 10.1057/978-1-137-56288-3_9

Corona, n.1 . (n.d.). In *Oxford English Dictionary Online*. Retrieved from https://www.oed.com/view/Entry/41771?rskey=6wAxfT&result=1

Davies, Mark. (2019-) *The Coronavirus Corpus*. Available online at https://www.english-corpora.org/corona/

De Smedt, K. (2012) Ash compound frenzy: A case study in the Norwegian newspaper corpus. G., Andersen (Ed.). *Exploring newspaper language: Using the web to create and investigate a large corpus of modern norwegian.* ProQuest Ebook Central https://ebookcentral.proquest.com/lib/bergen-ebooks/detail.action?docID=869351

Enger, H-O., Holmes, P. (2018) Norwegian: A comprehensive grammar

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

Folkehelseinstituttet (FHI). (2020, January) Fakta om koronaviruset SARS-CoV-2 og sykdommen covid-19. Retrieved from: https://www.fhi.no/nettpub/coronavirus/fakta/fakta-om-koronavirus-coronavirus-2019-ncov/?term=&h=1

Fitria, T., N. (2021) Word formation process of terms in COVID-19 pandemic. *Leksika 15*(1), 18-26. DOI: 10.30595/lks.v15i1.9248

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(3), 565–596. http://www.jstor.org/stable/4048097

Joyce, J. (1939) *Finnegans Wake*. London: Faber and Faber

Kubozono, H. (1990) Phonological Constraints on Blending in English as a Case for Phonology-Morphology Interface. In B. Geert, & J. van Marle (Eds.), *Yearbook of Morphology 3*. 1-20. Berlin, Boston: De Gruyter Mouton. DOI: 10.1515/9783112420744

Lawson, R. (2020) *Coronavirus has led to an explosion of new words and phrases – and that helps us cope* https://theconversation.com/coronavirus-has-led-to-an-explosion-of-new-words-and-phrases-and-that-helps-us-cope-136909

NRK (2010, 14th of April - 2011, 9th of July) Retrieved from: https://www.nrk.no/emne/vulkanutbruddet-pa-island-2010-1.7081019

Plag, I., (2003), *Word-formation in English* Cambridge University Press

Simatupang, E. C., Supri, I. Z. (2020). Compound words that occur during the global pandemic covid-19: A morphosemantic study. *English Review: Journal of English Education*, *8*(2), 291-298. DOI: 10.25134/erjee.v8i2.2824

Språkrådet (2010, 24th of December). Oskefast/Askefast *Ordet for året 2010* Retrieved from: https://www.sprakradet.no/Vi-og-vart/hva-skjer/Aktuelt-ord/Ordet-for-aret-Oskefastaskefast/

Thorne, T. (2020). Spotlight on COVID: *Pandemic language and the role of linguistics*. King's College London. News Centre. Retrieved from https://www.kcl.ac.uk/news/spotlight-on-covid-pandemic-language-and-the-role-of-linguists-1

World Health Organization (WHO) (2021, 22nd of January) Retrieved from: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19-22-january-2021

Yr (2010, 16th of April) Retrieved from: https://www.yr.no/artikkel/na-blir-det-askeplask_-1.7083120