# Studying language change through indexed and interlinked dictionaries

**Ore C.-E.[1], Grønvik O.[2]**

[1] *University of Oslo, Norway*
[2] *University of Bergen, Norway*

**Abstract**

In this paper we present our study how to use the Meta Dictionary of the Norwegian Language Collections to measure lexical stability in standard dictionaries across a timespan. The Meta Dictionary uses the lexical item as its core unit, expressing each lexical unit in a separate Meta Dictionary entry. The success of this model rests on having access to electronic versions of major and generally accepted dictionaries from the different stages of the orthography of a language. With this documentation it is possible to see, for instance, how much and which parts of the 1873 lexicon (Norwegian vernacular) is present in modern Nynorsk and Bokmål respectively, and whether this lexicon is present in its original orthography or not. This method for studies of the lexical development is comparable to remote sensing in archaeology and distant reading in literary studies. As an extended example of the application of the method we study a few issues related to the position of the pioneering lexicographers Ivar Aasen (1813-1896) and Hans Ross (1831-1912) in the description of Nynorsk, as shown in more recent lexicographical works, and in particular in two school dictionaries from 1954 and 1970 which border on being spellers.

**Keywords**: lexical item, lexicon, language change, dictionary, Meta Dictionary model, standard language, orthography, Norwegian

## 1 Lexical stability in standard dictionaries

How can one measure and document lexical and orthographic development in a written standard language? An obvious method is to start by comparing the selection of lexical items in a number of successive dictionaries. Although dictionaries do not document actual language usage, they represent what was thought essential vocabulary at the time of publication, in the form valid at the time. A systematic comparison of the lexical item inventory from a large number of dictionaries requires too much labour for manual execution, but becomes possible with the availability of systematized digital resources, that is, (retro) digitized dictionaries and a proper instrument for analysing them.

The instrument used in this study is the *Meta Dictionary* of the Norwegian Language Collections (Ore & Grønvik 2018). The *Meta Dictionary* is an electronic register of lexical items of Norwegian (Bokmål and Nynorsk), linking base forms with a wide range of usage examples and entries from 90 dictionaries. The *Meta Dictionary* also coordinates documentation of the vernacular with excerpts from literature in both the Norwegian Standard languages (Bokmål and Nynorsk). It has become a research tool, for instance for looking at change in standard orthographies through more than 150 years of documentation.

The *Meta Dictionary* format has also been adapted for the new editing and publication system of the *Dictionary of Old Norse Prose* (Ordbog over det Norrøne Prosasprog, ONP), see for example Johannsson & Battista (2014). For dictionary linking see also Møller, Troelsgård and Sørensen (2019). The German *Wörtebuch Netz* is an example of a coordinated dictionary collection (Wörterbuchnetz 2020).

Our aim is to use existing dictionaries and other materials in electronic form to document the history of language standardisation in Norway, in order to create a digital historical record of standard Norwegian orthography (Bokmål and Nynorsk) and its documentation in dictionaries.

## 2 Measuring lexical stability

Lexical stability in dictionaries and spellers can be measured by looking at the numbers of lexical items occurring in a series of (intentionally or unintentionally) normative documents, such as dictionaries and spellers. Lexical stability in dictionaries and spellers do not give certain information about usage, but they show what authors and publishers have regarded as necessary and desirable.

Measuring lexical stability in running texts is no more complex if the text is lemmatized; that is, for every word form (token) a base form (type/lexical item) has to be identified. Many European languages, among which Norwegian, have undergone changes in orthography and inflectional morphology over the last 150 years. For these languages, a lemmatizer optimized for the current version of the written standard, does not work well for older texts. There are several ways to construct lemmatizers, statistical, deep learning and, more traditionally, from full for registers. In the latter method a diachronic lemmatizer will need a set of full form registers each representing the orthography after a major language reform.

At present the *Meta Dictionary* registers for each lexical item a list of dated base forms which correspond to the orthography of a given period. Lexical stability in preserved and searchable text can to some extent be explored and measured through being searched with this register of dated base forms. The base forms reflect the orthography of their

time. This method would therefore probably be useful in dating the searched text.
A filtering of a text on the basis of a base form register, without inflected forms for each base form, would yield less reliable results in identifying and establishing a register of lexical items in the searched text. Results would depend on the character of the language in question. A text from a heavily inflected language with a great deal of heterogeneity would present greater difficulties than a very regular language with few inflected forms and few deviations from base forms.

## 3 The instrument

The Meta Dictionary was originally designed and established in 1999 as a common index to the digitized source material for the Norwegian Dictionary project (NO2014). Each entry in the Meta Dictionary consists of a head and a body. The head is a list of base forms with information about POS, language, orthographic status and the time span for this status. The body of an entry consists of (hyper)links to separate source databases with examples of usage. See the Euralex 2018 paper by Ore & Grønvik (2018) and section 5 in this paper for a more detailed description of the sources.
The organizing principle is that each entry in the Meta Dictionary should correspond to one lexical item. Atkins and Rundell (2008: 163 f.) use the term 'lexical item' to denote everything that may deserve lexicographical treatment in a dictionary entry, including compounds, multiword expressions, symbols and abbreviations. We extend the concept to cover both diachronically and synchronically orthographic variants of base forms belonging to the same lexical item. Therefore, the Meta Dictionary method of comparing dictionaries at the level of lexical item is independent of orthographical variants of the headwords.
A lexical item that has been identified and described in a dictionary entry, and is generally accepted by language users, is likely to keep its place as a separate lexical item. In general, this holds true for Nynorsk. Two non-linguistic factors have nevertheless affected the issue of lexical identity for a few lexical items. They are (1) the policy of promoting orthographic approach between Nynorsk and Bokmål on the basis of vernacular word forms with a wide geographical and usage distribution (merging the verbs Nynorsk *ganga* + Bokmål *gå* to Nynorsk and Bokmål *gå* 'walk' is a case in point), (2) the more recent desire to use the dictionary entry head to inform language users of semantic alternatives (The Nynorsk nouns *lege* and *lækjar* 'medical doctor' have different forms and etymologies, and are different lexical items in linguistic terms, but can be found as alternative headword forms in dictionaries and spellers).
In the Meta Dictionary, pairs of type 2 are dealt with as separate lexical items, Transitions of type 1 are accepted. Derived forms, singly and in compounds, are treated as separate lexical items (a case in point are verbal adjectives, f.i. *–gjengen* (< Nynorsk *ganga* 'walk') and *–gått* (< Nynorsk and Bokmål *gå* 'walk').
The implication of using the lexical item through time, represented by the set of headword forms found for each particular lexical item in the standard orthography, is to aim for full homograph separation over time in the Meta Dictionary. The term 'homograph' is here used as in the editorial guidelines for *the Norwegian Dictionary (Norsk Ordbok)*, and is understood to mean 'headword with identical orthographic form to another headword' (Grønvik & Gundersen 2016: 78 ff.). The criteria for homograph separation in base form is that each homograph represents a different lexical item, identified by pronunciation, inflection, etymology and usage.
From a strictly synchronic view homography is (almost) identical to polysemy. From a diachronic point of view, the origin (etymology) is the discriminating feature for otherwise identical word forms. Distinguishing lexical items through a sequence of orthographic forms is not an exact science, and a complicating factor is that what at one point in time seems to be evidence that two word forms have different origins, does not preclude earlier common origins. The Norwegian adjective 'feig' has the two distinct meanings; (1) 'cowardly' from German 'feige' and (2) 'close to death' from Old Norse 'feigr', cf. English 'fey' with apparently the same Germanic root as the other two. In theory, the two different meanings of modern Norwegian 'feig' could be considered to be either homographs or two senses under the same lexical item. Most modern dictionaries of Norwegian chose the latter solution.
In our study the timespan is much more modest - only 150 years. Even so, full homograph separation can be difficult, as changing headword forms of different lexical items can change the pattern of homographs. Full homograph separation linked to materials therefore requires the sorting of materials according to context and meaning.
The Meta Dictionary was designed and funded to give a systematized access to the source materials for the NO2014 project. A large part of the source materials is found in the retro digitalized slip archive (3.2 million slips) comprising excerpts from a wide variety of sources. This digital slip archive was the basis for the construction of the Meta Dictionary. At this early stage the grouping of word forms followed the principles of lemma selection of the Norsk Ordbok, Heavily etymological, Norsk Ordbok grouped verbal derivatives (verbal nouns and adjectives) under the verb, without taking into account that a derived form might have developed into an independent lexical item. To some extent this grouping was used even for derived compound adjectives based on phrasal verbs. However, the empirical material we use is carefully linked to the Meta Dictionary, observing the principle of the lexical item with the exception that the material is not completely homograph separated. This latter fact should always be taken into consideration, but will have little or no influence on the results in the current study.
The success of the Meta Dictionary model rests on having access to electronic versions of major and generally accepted dictionaries from the different stages of the orthography of a language. The expression "generally accepted" implies 'accepted for use in education and official documents; used as a standard in examination systems'.
Of the two standard languages of Norwegian, Bokmål and Nynorsk, Nynorsk is at present the better equipped with major electronically available dictionaries (Ore 2020). The documentation of Nynorsk standard orthography rests on the following resources, all of which are linked to the Meta Dictionary (see Table 1).
With this documentation of Nynorsk 1873 – 2012 interlinked, it is possible to see - for instance - (1) how much and which

parts of the 1873 lexicon (Norwegian vernacular) is present in modern Nynorsk and Bokmål respectively, and (2) whether this lexicon is present in its original orthography, or not. It is also possible to see (3) what influence the vernacular can be said to have had on modern standard Norwegian (Bokmål and Nynorsk). This paper offers answers to these questions in relation to Nynorsk only. The data for Bokmål are not yet complete enough for a reliable analysis.
On the lexicographical level, this universe of dictionaries is also an excellent tool for studying the development in the focus of dictionary makers, that is, which lexical items are considered worthy of being described over time. The usefulness of the method is demonstrated and described in Ore (2020).

## 4 Comparing lemma selection in dictionaries across a timespan

One can imagine the lexicon of a language as a vertical column consisting of fibres, where each fibre represents a lexical item. The column can be marked with years on the vertical axis. A horizontal cut at a given point will show the lexicon of the year in question. The ideal way of showing the lexicon of a given year is through a comprehensive dictionary using the orthography valid for that year, linked to earlier and later documentation of the same lexical item.
The model discussed here is based on the category system used in the Meta Dictionary. The Meta Dictionary uses the lexical item as its core unit, expressing each lexical item in a separate Meta Dictionary entry. A lexical item can have several headword forms in base form. The category schema of each headword form allow annotation by (1) language (Bokmål or Nynorsk), (2) part of speech (POS), (3) status within the orthography, and (4) start and end-date of a given status (Ore & Grønvik 2018).
The Meta Dictionary is not completely homograph separated - some entries comprise more than one lexical item. When comparing two dictionaries through shared lexical items in the Meta Dictionary we may get a slightly higher match than what is the true case. In the modern dictionary of Nynorsk, *Nynorskordboka* (2012), a little less than 9% of the headwords are marked as homographs. Therefore, there may be an uncertainty of around 5% in our results. This has no consequence for the tendencies shown. The most important aspect of this project at the present stage has been to create a practicable model with clear categories, which is true to the underlying linguistic and lexicographic description, is transparent and has general value (Ore 2016).
The model does not include the category “place”. For the standard language Nynorsk this is unimportant. Nynorsk is used in Norway only and has never been a first written language outside Norway. For Bokmål the case is different. The development of Bokmål starts as a history of deviation from standard Danish around 1900. To get a starting platform for the Danish of Norway, a general dictionary of 19th century Danish should be included in the Meta Dictionary materials. Only with this starting point will it be possible to document systematically the divergence of Danish in Norway from the Danish of Denmark and into Dano-Norwegian, resulting in today’s Bokmål. The standard language Nynorsk is based on the field work of the linguist and lexicographer Ivar Aasen in the mid 19th century, and became a written standard in parallel with Danish by an act of the Norwegian parliament in 1885.

## 5 Dictionary sources

In the present study we focus on Nynorsk, for which we have better digital resources. The major orthographic reforms of Nynorsk occurred in 1873, 1901, 1917, 1938, 1959, 1986 and 2012. The orthographic reforms have influenced lemma selection in dictionaries in the sense that editors will have felt obliged to give information on changes. The selected dictionaries were published close in time to the reforms in order to capture such changes.
Aasen and Ross (1873 - 1895) are scholarly pioneer dictionaries designed to portray the Norwegian vernacular, primarily from the country dialects used by the majority of the population 1840-1890. Vocabulary perceived by their informants as foreign, was not included in either dictionary. Aasen’s dictionary concentrated on the central, well documented vocabulary with derivations and some compounds. Ross includes more compounds and variants.
Skard 1903 (1901 orthography) is a bilingual school dictionary with Nynorsk headwords and Danish equivalents, with a focus on orthography and virtually no additional information. The Skard school dictionary (1. ed. 1901, expanded and corrected 1903) was the first series of school dictionaries covering Nynorsk.
The Norsk Ordbok Draft Manuscript (1917 orthography) was composed as a preliminary to Norsk Ordbok. Its lexical programme was to prepare for a scholarly dictionary covering the whole vocabulary of the Norwegian vernacular and written Nynorsk of all genres. The Draft manuscript is based on Aasen and Ross, some normative dictionaries (Schjøtt, Eskeland) from the beginning of the century, some large collections from specific dialects and a limited range of older, written sources of the vernacular from the Dano-Norwegian period 1550-1850. The Norsk Ordbok Draft Manuscript includes a fair amount of entries for imported vocabulary from Latin, Greek, French etc, but the specifically Norwegian vocabulary has priority.
Norsk Ordbok (1938 orthography) is at 330 000 lexical items almost three times the size of the Draft Manuscript. The special responsibility to present the original Norwegian materials is there, but in relation to lemma selection from written Nynorsk current general selection criteria were applied – i.e.no ejection of imported lexical items demonstrably in use in Nynorsk text.
Skard 1954 (1938 orthography) is the fifth edition of school dictionary originally published in 1922. It differs from Skard 1903 in size and scope, addressing the needs of the education system in general and of public administration.
Hellevik 1970 (1959 orthography) is the second series of school dictionaries to cover Nynorsk, and the first wholly monolingual series. Hellevik claims to give a "more complete and reliable image of actual usage in Nynorsk than any earlier (orthographic) dictionary", with explicit reference to the language collections ordered through the Meta Dictionary as a primary source.

Nynorskordboka (2005 orthography) was originally edited on the basis of the language collections now encompassed by the Meta Dictionary, in parallel with its sister volume, Bokmålsordboka. The two editorial teams drafted each their half of the alphabet and then swapped drafts, adapting each whole to the given standard language (Bokmål or Nynorsk), in order to avoid omissions and unnecessary differences. Nynorskordboka also had the rule that any word from the vernacular documented from three or more counties was to be included (Nynorskordboka 1986: VII).

| Year of Orthography | Year of publication | Author and title | Type | Number of lexical items |
|---|---|---|---|---|
| 1873 | 1873 | Ivar Aasen: *Norsk Ordbog*. | General dictionary | 38 711 |
| 1873 | 1895 | Hans Ross: Norsk Ordbog. | General dictionary, presented as an addition to Aasen (1873) | 49 220 |
| 1901 | 1903 | Matias Skard: Landsmaals-Ordlista | School speller | 12 000 |
| 1917 | 1991-1997 | Draft manuscript of *Norsk Ordbok* | General dictionary | 113 000 |
| 1938 | 1950 – 2016 | Norsk Ordbok 1-12 | Scholarly multivolume dictionary | 330 000 |
| 1938 | 1954 | Matias Skard; Nynorsk ordbok for rettskriving og literaturlesnad | Speller for use in education and administration. | 32 000 |
| 1959 | 1970 | Alf Hellevik: Nynorsk ordliste. Større utgåve | Speller for use in education and administration. | 29 000 |
| 1986 | 1986 | Nynorskordboka 1. ed. | General dictionary | 90 000 |
| 2012 | 2012 | Nynorskordboka web edition | General dictionary | 90 000 |

Table 1. Sources of Nynorsk Orthography 1873 – 2012.

## 6 Results and discussion

There are endless interesting byways to explore once dictionaries are properly interlinked and indexed at the lexical item level. In the rest of the paper we will give an extended example of how the Meta Dictionary method of comparing dictionaries can be used in practice.

We will study a few issues related to the position of the pioneering lexicographers Ivar Aasen (1813-1896) and Hans Ross (1831-1912) in the description of Nynorsk, as shown in more recent lexicographical works, and in particular in two school dictionaries which border on being spellers. The purpose is more to demonstrate what is possible to find out by using methods of sorting and grouping than to settle issues relating to the Nynorsk lexicon once and for all. Evidence from dictionaries and language collections cannot outweigh evidence from very large corpora. However, to move to that step, the base form history of each lexical item must be in place and a full form registry available for the whole.

This exploration of Nynorsk lexicography does not address issues arising from the rapprochement of Nynorsk and Bokmål in the course of the 20th century. Such comparisons must wait until the documentation of Danish as used in Norway is in place, together with materials showing the earliest deviations from the Danish of Denmark.

### 6.1 The dictionaries of Aasen and Ross and the later dictionaries in numbers and proportions

The dictionaries of Aasen and Ross, published in the second half of the 19th century, were the first description of the vernacular language in Norway. Table 2 gives an overview, showing to what extent the later dictionaries included lexical items from Aasen and Ross. Column 2 shows the scope of each dictionary by the number of lexical items given entries. Column 3 gives the percentage of the lexical items from Aasen and Ross with entries in the later dictionaries.

The size of the dictionaries differs. The school spellers (lines 2, 4 and 5) are smaller than the dictionaries of Aasen and Ross, while the scholarly dictionaries (lines 3, 6 and 7) are much larger. The numbers in the right column indicate the percentage of the lexical items in the dictionaries of Aasen and Ross which are brought forward. Skard (1903) contains 19 % of the lexical items found in dictionaries of Aasen and Ross. Hypothetically, it could contain at most 21 %. This means that Skard (1903) almost entirely consists of lexical items from Aasen and Ross. This is not surprising due to closeness in time.

Skard (1954) is almost 2.5 times the size of Skard (1903), and has a higher number of lexical items from the dictionaries of Aasen and Ross (17 410 lexical items). But the relative amount of lexical items from Aasen and Ross is smaller. Although more than 50% of the lexical items stem from Aasen and Ross, there is a clear tendency that the lexical items for the dictionaries of Aasen and Ross are less dominant, as so much other material is included. In Hellevik (1970) this tendency is much stronger. Here just 43% of the total (11 585 lexical items) stems from Aasen and Ross.

Two of the scholarly dictionaries, The Draft Manuscript of Norsk Ordbok and Norsk Ordbok itself, are much larger than the dictionaries of Aasen and Ross combined. They are also programmatically committed to include Aasen and Ross complete. Both cover 95% of the Aasen and Ross lexical items, i.e. all except some cross references.

The third dictionary, Nynorskordboka (1986), is a collegiate dictionary. The size and the focus on dialects allow for including a relatively large number of lexical items from Aasen and Ross.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| | Dictionary | Number of lexical items with entries | % of the total of lexical items in the dictionary stemming from Aasen and Ross | % lexical items in Aasen and Ross continued | % if all possible lexical items were continued |
| 1 | Aasen and Ross | 64 334 | - | - | - |
| 2 | Skard (1903) | 13 390 | 92 | 19 | 21 |
| 3 | N.O. Draft Manuscript | 95 908 | 64 | 95 | 100 |
| 4 | Skard (1954) | 31 864 | 55 | 28 | 51 |
| 5 | Hellevik (1970) | 26 201 | 43 | 18 | 42 |
| 6 | Nynorskordboka (1986) | 87 145 | 26 | 34 | 100 |
| 7 | Norsk Ordbok (2016) | 300 117 | 20 | 94 | 100 |

Table 2. Number (percentage) of lexical items in the first description of Nynorsk reoccurring in later dictionaries. The right column indicate the percentage if all possible lexical items (limited by space) reoccurred.

## 6.2 How much and which parts of the 1873 lexicon (Norwegian vernacular) is present in modern Nynorsk Dictionaries?

The first question in this chapter heading can easily be answered by inspecting the Meta Dictionary. There are 10 198 lexical items where a form of the headword has an entry both in the dictionaries of Aasen and Ross combined and in both of the two school dictionaries Skard (1954) and Hellevik (1970).

The second question in the chapter heading is more complex and can be reformulated as follows: To what extent do the lexical items in the dictionaries of Aasen and Ross that also occur in the post-war school dictionaries represent the central vocabulary of Nynorsk? The expression "central vocabulary" is found especially in second language teaching materials, but rarely defined. In this paper, "central vocabulary" is taken to mean the group of lexical items that are (deemed) essential to achieving mastery of a given language by virtue of their authenticity, meaning, relevance, frequency, stability in use across a timespan and their lexicogenetic potential. These indicators of centrality are based on a discussion in Fjeld and Vikør (2011:156 ff.).

The fact that an entry in an old dictionary reoccurs in a more recent dictionary does not in itself reflect the status of the word in the language at the time of selection for each dictionary. The Meta Dictionary entry does, however, link dictionary entries with usage examples from literature and from dialect collections, most of which were collected and published in the twentieth century, and with a majority of instances from the later period (post-1950). It is therefore possible to use the Meta Dictionary to look at the coverage of each lexical item in sources other than the dictionaries mentioned above.

When estimating the combined contribution of the dictionaries of Aasen and Ross to the lexical items of newer school dictionaries by means of the information found in the Meta Dictionary, it is relevant to look at the 10 198 joint lexical items in relation to following five categories:

- **Authenticity** here simply means independent proof that the word exists in the language outside the dictionaries. The Meta Dictionary shows that this requirement is met by all 10 198 lexical items – all have usage materials in addition to the dictionary entries, though not necessarily very many.

- **Frequency** cannot be measured directly through the Meta Dictionary, as it could in a lemmatized corpus. The number of registered instances (quotations and other lexical items of information) per entry gives an indication of centrality, since the range of sources is considerable (ca 5 000 literary sources, roughly 90 dictionaries, plus dialect archives and reference works). So looking at the number of instances behind each entry, compared to the Meta Dictionary as a whole, should give an indication of centrality. See table 3 below.

- **Stability across a timespan.** A high number of registered instances for an entry in the Meta Dictionary indicates stability across a timespan, since a high number from a short period of time in unlikely. But the best indicator of stability through time is the fact that almost all the 10 198 joint lexical items also occur in Nynorskordboka, edited 1974–1986, cf. table 2.

- **Word structure** is important in Norwegian, a Germanic language in which any simple (i.e. morphologically indivisible) lexical item can be used as a building block in complex lexical items (derivations and compounds). In general, simple lexical items tend to have more meanings and a higher lexicogenetic potential than lexical items with a complex structure. This tendency is strengthened if a simple lexical item also has high frequency in use. The Meta Dictionary headword forms are structure marked. It is therefore possible to see whether the joint lexical items have simple or complex headword base forms.

- **Spoken and written sources.** A lexical item documented both from spoken and from written sources will be more likely to hold a place in the central vocabulary of a language than a lexical item with only the one or the other, since lexical items with limited sources are more likely to be strongly marked in terms of style or subject field. In the case of the 10 198, they are all documented from the vernacular of the 19th century, through the dictionaries of Aasen and Ross. The majority of the additional sources registered in the Meta Dictionary come from literature, but it is possible that some of the 10 198 lexical items have speech sources only.

An analysis of lexical items as represented in the Meta Dictionary category system can give an indication on authenticity, frequency and word structure, and by implication say something about stability across a timespan. Estimating meaning and relevance of the lexical items common to Aasen and Ross and the school dictionaries, compared to the Meta Dictionary as a whole, would require a study of the materials behind each entry, which is beyond the scope of this paper.

| Number of instances | Aasen & Ross + Skard (1954) and Hellevik (1970) | Percent (of 10 198) | Aasen & Ross + Skard (1954) and Hellevik (1970) restricted to preserved headword. forms | Percent (of 7 483) | Total for all entries in the Meta Dictionary - Nynorsk |
|---|---|---|---|---|---|
| 1 (hapax legomenon) | | | | | 264 788 |
| 1 – 9 | 37 | 0.4 | 22 | 0.3 | 263 065 |
| 10 – 99 | 5 147 | 50.5 | 3 590 | 48.0 | 91 768 |
| 100 – 999 | 4 856 | 47.6 | 3 742 | 50.0 | 9 427 |
| >1000 | 158 | 1.5 | 129 | 1.7 | 241 |
| In all | 10 198 | 100 | 7 483 | 100 | 529 289 |

Table 3. Aasen and Ross lexical items grouped by number instances from other sources linked to the Meta Dictionary.

Almost half the lexical items of the entire Meta Dictionary (Nynorsk) are represented by only one instance of lexical information, while the group of lexical items with 1000 or more instances represent 0.04 per cent of the total. In other words: the distribution of instances (similar to tokens in a corpus) corresponds to Zipf's law. The distribution of the instances found in Aasen and Ross combined and in Hellevik (1970) and Skard (1954) have a different distribution. Table 3 groups the lexical items from Aasen and Ross found in Skard 1954 and Hellevik 1970 by the number of instances found in the Meta Dictionary.
The structure of the Meta Dictionary makes it possible to get a precise count of instances for groups of lexical items. The 10 198 lexical items from Aasen and Ross present in Skard 1954 and Hellevik 1970 (table 3, column 2) have in all 1 768 636 instances In the Meta Dictionary, an average per entry of 173. The median is 98. The group of lexical items from the dictionaries of Aasen and Ross present in their original orthographic form in Skard and Hellevik have 1 359 683 instances, an average of 182 instances behind each lexical item, and the median is 103. The Meta Dictionary as a whole has roughly 3.5 million instances from Nynorsk sources. If the hapax lexical items are disregarded, the average number of instances per Nynorsk entry is 9. The lexical items originating from the dictionaries of Aasen and Ross belong to the group of Meta Dictionary lexical items with the highest number of instances and are consequently among the best documented ones. This fact suggests that the 10 198 lexical items from Aasen and Ross and found in Skard (1954) and Hellevik (1970) belong to the central vocabulary of Nynorsk – and this is even more so the case for the lexical items form the dictionaries in Aasen and Ross with preserved orthographic form.

## 6.3 The orthographic form of the Aasen and Ross lexical items - in the original orthography, or in a newer orthographic form?

The Meta dictionary entry is indexed by headword base forms belonging to a standard orthography of Bokmål or Nynorsk, with a starting date and a final date. The latter is set to 31.12.9999 for base forms within the current orthography. This artificial final date is not shown in the web interface. Since there have been several revisions of the orthography, the entry head can look like this:

| |
|---|
| køyrar m (Nynorsk, 1873-)<br>kjørar m (Nynorsk, 1959-2012) |

Figure 1. The form *køyrar* ('driver' noun, masculine) is spelt as it was in Aasen 1873;
the form *kjørar* was a permitted form from 1959 until 2012, but is no longer part of the Nynorsk orthography.

| POS | (1) Aasen and Ross Lexical items in Skard (1954) and Hellevik (1970) (100 %) | (2) Subset of lexical items with present day orthography (73.5 %) | (3) Subset of simple (non-compounds) lexical items with present day orthography (60.8 %) |
|---|---|---|---|
| In all | 10 259 | 7 516 | 6207 |
| Adjective | 1750 | 1213 | 858 |
| Adverb | 228 | 123 | 97 |
| Conjunction /Subjunction | 11 | 11 | 10 |
| First part of compound | 1 | 1 | I |
| Interjection | 6 | 2 | 2 |
| Noun, no certain gender | 8 | 3 | 3 |
| Noun fem. | 1576 | 797 | 632 |
| Noun masc. | 2667 | 2022 | 1683 |
| Noun neut. | 1561 | 1084 | 780 |
| Numeral | 26 | 13 | 12 |
| Preposition | 71 | 33 | 15 |
| Pronoun | 39 | 32 | 29 |
| Verb | 2631 | 2286 | 2182 |

Table 4. Lexical items common to Aasen and Ross, Skard (1954) and Hellevik (1970) distributed according to POS. (1) in all, (2) with preserved orthography from Aasen and Ross in the base form, (3) with preserved orthography and simple word structure. The total sum is smaller than the sum of lexical items with POS due to the fact that some nouns having more than one gender.

Several of the orthographic revisions were motivated by an expressed need for modernity, the implication being that Aasen's orthography is - and was - out of date. It is therefore interesting to see how many of the lexical items from the dictionaries of Aasen and Ross, included in Skard (1954) and Hellevik (1970), where the original orthographic base form is preserved in today's Nynorsk orthography. The numbers are as shown in table 4.
The Meta Dictionary has POS as a category. The POS distribution in included in table 4. What is striking in this table is the size of the verb group, compared to nouns and adjectives. Verbs have both derived adjectival forms (participles) and nominal forms (suffix-derived verbal nouns), which makes them lexicogenetically powerful.
The lexical items found in Aasen and Ross AND in Skard 1954 and Hellevik 1970 are so well represented in the Meta Dictionary that it seems reasonable to consider them part of the central vocabulary of Nynorsk. This assumption is supported by the fact that 73.5 % of the total have preserved their orthographic form. 60 % are have a simple word structure – they are not compounds, and if they are derivations they are most likely derived by suffix (f.i. –*leg* adj. '-ly') Some orthographic changes have been minimal. In 1917, nouns with the feminine gender and ending in the vowel -*a* had the ending changed to -*e*. This change probably partly explains the fact that only half of the feminine nouns – the ones ending in a consonant – have preserved their original orthographic form. However, such detailed checking is outside the scope of this paper.

## 6.4 What influence can the vernacular be said to have had on modern standard Nynorsk?

This discussion is limited to the presence of Aasen and Ross in the two school dictionaries Skard (1954) and Hellevik (1970), with a side glance at Nynorskordboka.
The dictionaries of Aasen and Ross represent 70 years of fieldwork in the Norwegian vernacular. This paper shows that a substantial part of smaller modern dictionaries consists of lexical items also found in the dictionaries of Aasen and Ross. A majority has preserved the original orthography, and a majority has a simple word structure, offering opportunities for easy re-use in new and complex word forms. It seems reasonable to conclude that the spoken language, both in standard form and in dialect form, has been included in the linguistic toolbox of modern Nynorsk.

## 6.5 Modernising the selection of lemmas – how does Hellevik (1970) differ from Skard (1954)?

This essay in investigating dictionaries through sorting and grouping will round off by taking a closer look at the two school dictionaries, to see at what points they resemble each other and at what points they differ.
In the preface to his dictionary Hellevik (1970) claims to have a more modern lexicon to offer than other contemporary

school dictionaries. Below, in tables 5 and 6, the selection of lexical items unique to either dictionary is looked at and compared at some points. A full comparison of the two lists of lexical items would involve comparing categories which the Meta Dictionary does not cover, i.e. pronunciation, inflection and sense description. The comparison below is limited to POS, with observations on some derivational suffixes.
Table 3 shows that Skard (1954) has 55 % of its lexical items in common with the combined dictionaries of Aasen and Ross, while the Hellevik (1970) has 43 % in common with the combined dictionaries of Aasen and Ross. Relatively speaking and in absolute numbers Skard (1954) is closer to Aasen and Ross in its lemma selection than Hellevik (1970) is.

| Dictionary | Lexical items | Lexical items from Aasen and Ross | Shared lexical items Aasen and Ross | Shared lexical items | Number of unique lexical items |
|---|---|---|---|---|---|
| Skard (1954) | 31 864 | 17 421 | 10 198 | 18509 | 13 355 |
| Hellevik (1970) | 26 201 | 11 498 | 10 198 | 18509 | 7 692 |

Table 5. A comparison of contents between Skard (1954) and Hellevik (1970).

Table 5 shows that Skard (1954) and Hellevik (1970) have 18 509 lexical items in common, 58 % of Skard (1954), close to 67 % of Hellevik (1970). More than half of the shared lexical items, 10 196 are also found in the dictionaries of Aasen and Ross. The chief differences between them must therefore be found in the parts of the dictionaries that are unique to either. Hellevik has the smaller number of unique entries.

| POS | Hellevik (1970) | Skard (1954) |
|---|---|---|
| Unique entries in all | 7 692 | 13 355 |
| Adjective | 1 344 | 2 771 |
| Adverb | 280 | 137 |
| Abbreviation | 9 | 10 |
| Compound first part | 41 | 6 |
| Interjection | 13 | 7 |
| Conjunction /subjunction | 14 | 5 |
| Derivational prefix | 9 | 13 |
| Preposition | 61 | 32 |
| Pronoun | 9 | 7 |
| Noun (no certain gender) | 23 | 6 |
| Noun feminine gender | 790 | 3 392 |
| Noun masculine gender | 2 693 | 3 134 |
| Noun neuter gender | 2 005 | 2 208 |
| Name (place, person) | 265 | 611 |
| Numeral | 9 | 2 |
| Symbol | 4 | |
| Verb | 1 136 | 1 165 |

Table 6 The POS distribution in Hellevik (1970) and in Skard (1954) for the lexical items unique to either dictionary.

Table 6 shows the POS distribution in Hellevik (1970) and in Skard (1954). As some nouns have more than one gender, the sum of POS occurrences is slightly larger than the total of unique lexical items. What is notable in Skard (1954) is the high number of adjectives and of nouns with feminine gender. Unique to Skard (1954) is a large number of nouns which are verbal derivatives with the suffix *–ing*, feminine gender in Nynorsk. What is notable in Hellevik (1970) is the comparatively high number for the function POS – especially adverbs and prepositions – and the special entries for the

first part of compounds, since nesting is not used. A closer look shows that Hellevik (1970) has included a number of compound and phrasal adverbs and prepositions, much used in everyday language but largely unnoticed by grammarians. The high number of entries for first parts of compounds found in Hellevik (1970) may have a basis in a post war focus on systems of word formation – lexicogenesis - in the standardisation of written Norwegian (both Nynorsk and Bokmål).

## 7 Conclusion

Comparing the inventory of lexical items through a series of very different dictionaries becomes a history of lemma selection for 150 years. The dictionaries have different sources, and differ in size and planned usage. However, once a lexical item has been identified and described in a dictionary, it will not disappear as a lemma candidate. The inclusion or exclusion of a documented lexical item therefore expresses a choice on the part of the editor. The results of the comparison discussed in this paper will therefore show what direction lemma selection for Nynorsk dictionaries has taken over the years.

Plans for the future include developing a full form generator for the older orthographies – which will involve expansion of the existing schemas to include plural forms of verbs and the dative case of nouns, and, of course, adding materials on Bokmål in order to facilitate the use of the Meta Dictionary as a tool in tracing the development of the majority written standard of Norway.

## 8 References

Aasen, I. (1873). *Norsk Ordbog med dansk forklaring*: Christiania: Mallings Boghandel, 1873.

Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford. Oxford University Press.

Bergenholtz, H., Cantell, I., Fjeld, R.V., Gundersen, D., Jónsson, J.H. & Svensén, B. (1997). *Nordisk leksikografisk ordbok*. Skrifter utgitt av Nordisk forening for leksikografi. Skrift nr. 4. Oslo. Universitetsforlaget.

*Bokmålsordboka* (2019). Accessed at: http://ordbok.uib.no [30/05/2020]

*Draft manuscript of Norsk Ordbok (Grunnmanuskriptet–1940)* (1997). Accessed at: http://usd.uib.no/perl/search/search.cgi?appid=59&tabid=993 [30/05/2020]

Eskeland, S. (1919). Framandordbok. med tyding og rettleiding um lesemaaten til 8-9 tusen av dei vanlegaste framandordi. Kristiania. Norli.

Fjeld, R.V., & Vikør, L.S. (2011). *Ord og ordbøker*. Kristiansand. Høyskoleforlaget.

Grønvik, O. (1980). Framandorda og norsk språkutvikling i nyare tid. I: *Sprog i Norden*, 1980, s. 39-60. Accessed at: http://ojs.statsbiblioteket.dk/index.php/sin/issue/archive [30/05/2020]

Grønvik, O. & Gundersen, H. (2016). *Redigeringshandbok for Norsk Ordbok 2014*. Accessed at: http://no2014.uib.no/eNo/tekst/redigeringshandboka/redigeringshandboka.pdf [30/05/2020]

Hellevik, A. (1970). Nynorsk ordliste. Større utgåve. Med fornorskings-tillegg og liste ovr forkortingar. Oslo. Det Norske Samlaget.

Johannsson, E.T., & Battista, S. (2014). A Dictionary of Old Norse Prose and its Users — Paper vs. Web-based Edition. In Abel, A., Vettori, C., Ralli, N. (eds) *Proceeding of the XVI EURALEX, The User in Focus; 15-19 July 2014, Bolzano*, Bolzano, Eurac Research, 2014, ISBN: 978-88-88906-97-3. Accessed at: http://euralex.org/category/publications/euralex-2014/ [30/05/2020].

Møller Svendsen, M.-M., Troelsgård, T. & Sørensen, N. H. (2019). *Salmesang og superordbog: ordbogslinkning i praksis* (Hymn Singing and a Super Dictionary: Dictionary Linking in Practise), presentation at NFL 2019, to be published in *Leksikografi 15, abstract available at: https://www.helsinki.fi/sv/konferenser/15-konferensen-om-lexikografi-i-norden/program-och-abstrakt* [30/05/2020]

Norsk Ordbok (2016). Norsk ordbok – Ordbok over det norske folkemålet og det nynorske skriftmålet, 1-12. Oslo, Samlaget 1950–2016

Nynorskordboka (1986). *Nynorskordboka – Definisjons og rettskrivingsordbok*. Oslo. Det Norske Samlaget.

*Nynorskordboka* (2019). Accessed at: http://ordbok.uib.no [30/05/2020]

Ore, C.-E. S. (2016). Gamle ordbøker og digitale utgaver. I *Nordiske Studier i Leksikografi 13 Rapport fra 13. Konference om Leksikografi i Norden København 19.-22. mai 2015*. Nordisk forening for leksikografi 2016 ISBN 978-87-992447-6-8. pp. 203-216

Ore, C.E. S. (2020). Å ta Hans Ross på ordet. Ross' ordbok i relasjon til Aasens med Metaordboka som verktøy. In *Nordiske Studier i Leksikografi 15* (under publication)

Ore, C.-E. S., Grønvik, O. (2018). Comparing Orthographies in Space and Time through Lexicographic Resources. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Znanstvena založba Filozofske fakultete 2018 ISBN 978-961-06-0097-8. pp. 159-172

Ross, H. (1895). *Norsk ordbog: Tillæg til "Norsk ordbog" af Ivar Aasen*. In Ross, Hans: *Norsk ordbog*. Universitetsforlaget. Oslo. 1971.

Schjøtt, S. (1909). *Dansk-norsk ordbog*. Kristiania. Aschehoug.

Skard, M. (1903). *Landsmaals-ordlista godkjend til skulebruk. Norsk rettskrivningslæra II*. Andre utgaava, gjennomsedd og auka. Kristiania. Aschehoug

Skard, M. (1954). *Nynorsk ordbok for rettskriving og litteraturlesnad*. 5. utgåva ved Vemund Skard. Oslo. Aschehoug.

*Wörterbuchnetz* (2020). Accessed at: http://www.woerterbuchnetz.de [30/05/2020].