Christian-Emil Smith Ore/Oddrun Grønvik/Trond Minde

WORD BANKS, DICTIONARIES AND RESEARCH RESULTS BY THE ROADSIDE

Abstract Many European languages have undergone considerable changes in orthography over the last 150 years. This hampers the application of modern computer-based analysers to older text, and hence computer-based annotation and studies of text collections spanning a long period. As a step towards a functional analyser for Norwegian texts (Nynorsk standard) from the 19th century, funding was granted in 2020 for creating a full form generator for all inflected forms of headwords found in Ivar Aasen's dictionary published in 1873 (Aasen 1873) and his grammar from 1864 (Aasen 1864).

Creating this word bank led to new insight in Aasen (1873), its structure, internal organisation, and ambition level as well as its link to Aasen (1864). As a test, the full form list generated from this new word bank was used to analyse the word inventory of texts by Aa. O. Vinje, written in the period 1850–1870. The Vinje texts were also analysed using a full form list of modern standard Norwegian, to study the differences in applicability and see how Vinje's language relates to the written standard of modern Norwegian.

Keywords Dictionary and text analysis; full form systems; close reading of dictionaries

1. Introduction

Many modern European languages had their first written standard defined in the 19th century, but the standard may since have undergone substantial changes. This is the case for the Norwegian written standards. Until the second half of the 19th century, Danish was the only standard written language. The Nynorsk standard, based on the Norwegian vernacular, was introduced in the middle of the 19th century, but has been extensively revised. A history of standard revision causes modern tools for text analysis to be less well suited for older texts. Analysers must be adapted to both orthographic change and changes in inflectional morphology. This is a bootstrapping problem, since the creation of base forms with potential inflection forms requires analysis of text corpora, which in turn requires lemmatizers built from the same corpora.

Our solution to this problem for early Nynorsk is to use a central 19th century dictionary and grammar, both compiled by the Norwegian linguist and lexicographer Ivar Aasen (1813–1896). As a step towards a functional analyser for Norwegian texts (Nynorsk standard) from the 19th century, funding was granted in 2020 for creating a word bank and a full form generator for all headwords found in Ivar Aasen's Norsk Ordbog med dansk Forklaring 'Norwegian Dictionary with Danish definitions' (Aasen 1873) and Norsk Grammatik (Aasen 1864). The Aasen Word Bank was completed earlier this year.

Tools for making the full form generator were (1) Aasen (1864), (2) Aasen (1873), (3) the system and tool for Norwegian Word Bank (Norsk Ordbank), which is the main full form generator for modern Norwegian (separate ones for Bokmål and Nynorsk), see Hagen/Nøklestad (2010), Engh (2014), Grønvik/Ore (2014).

Creating the Aasen Word Bank led to new insight into Aasen (1873), its structure, organisation, and ambition level. The process and resulting findings will be discussed in this paper.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

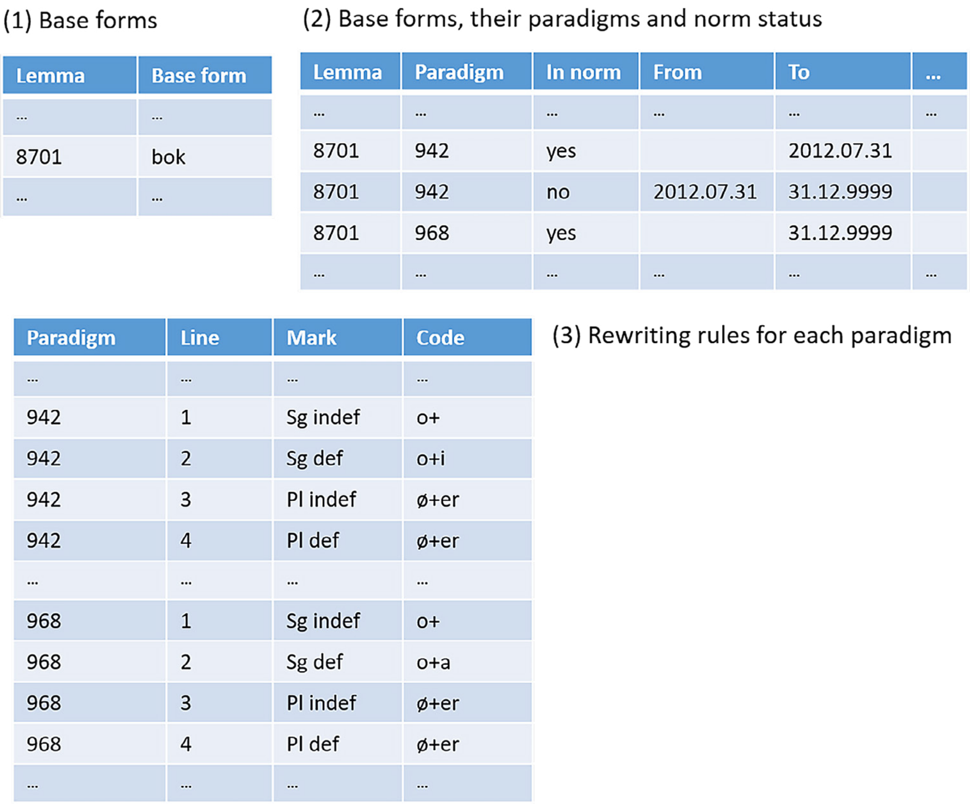
2. The word bank system

The word bank is a structure for storing information about words and their inflected forms. The fundamental idea is that a word (lexical item) is identified as the set of all possible inflected forms. In this model, the headword used in a dictionary entry as base form is a representative for the set of word forms. The Norwegian written standards of Nynorsk and Bokmål have been revised several times after 1900, including changes in inflections. For example, in the Nynorsk standard “bok” (book) had until 2012 the additional form “bok” (det. sg. dem.). Since 2012, the form “boka” (det. sg. fem) is the only accepted form. Figure 1 shows a simplified word bank structure. The basic structure consists of (1) a list of base forms (headwords), (2) a table with the information about which paradigm(s) a base form can follow, and (3) a set of rewriting rules (paradigms). Table 2 is the pivot of the word bank. For a given base form, each corresponding row identifies a paradigm, the orthographic status of the corresponding forms and the timespan of this status. The paradigm table shows examples of the rewriting rules used to generate the inflected forms. The ‘+’ is a wildcard character. The rewriting process runs as follows: The pattern in line 1 is used to find a match with a selected base form. In the example ‘+’ will be bound to ‘k’. Lines 2 to 4 are production lines, and the full form list will be “bok, boki, bøker, bøkene” and “bok, boka, bøker, bøkene”. There is an overlap of forms. If required, the list of inflected forms can be reduced to unique forms with information about the corresponding line in the paradigm. All paradigms linked to a base form will have the same number of lines and marks on them.

Each paradigm has additional information about part of speech (POS), prototypical base forms and comments. In the Aasen Word Bank project every paradigm is referred to the relevant paragraph in Aasen (1864).

The word bank system was developed almost 30 years ago in connection with the construction of a rule based morpho-syntactic tagger for the two modern Norwegian written standards Bokmål and Nynorsk. The rewriting system is based on IBM’s spellchecker, developed at the end of the 1980s. In the tagger project, base forms were linked to the corresponding entries in the general dictionaries Bokmålsordboka (Wangensteen 1986) and Nynorskordboka (Hovdenak et al. 1986). This was not required for the computational linguistic purpose of the taggers but has later turned out to be useful. The link between word bank entry and dictionary entry has made it possible to add a table of valid inflection forms to every entry in the online editions of these dictionaries. It is also possible to go from word bank entry to dictionary entry to check the definition(s) of a word, which is useful for separating homographs, etc.

Norwegian is a Germanic language with productive use of compounds, like for example German. A consequence is that the number of unique words (types) is for all practical purposes unlimited. A list of full forms can never be exhaustive. A solution to this problem is to use a so-called compound analyser, that is, a piece of software marking the border between possible elements. A compound analyser will be a future extension. We plan to test the software from the Oslo-Bergen tagger, see Hagen/Johannessen/Nøklestad (2000), Nøklestad (2022a, 2022b).



(1) Base forms

Lemma	Base form
…	…
8701	bok
…	…

(2) Base forms, their paradigms and norm status

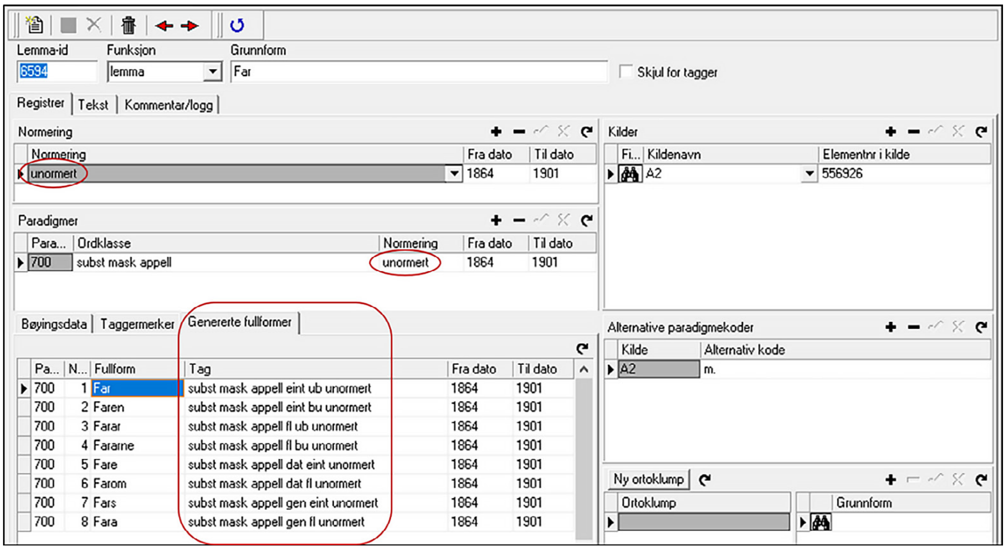
Lemma	Paradigm	In norm	From	To	…
…	…	…	…	…	…
8701	942	yes		2012.07.31	
8701	942	no	2012.07.31	31.12.9999	
8701	968	yes		31.12.9999	
…	…	…	…	…	…

(3) Rewriting rules for each paradigm

Paradigm	Line	Mark	Code
…	…	…	…
942	1	Sg indef	o+
942	2	Sg def	o+i
942	3	Pl indef	ø+er
942	4	Pl def	ø+er
…	…	…	…
968	1	Sg indef	o+
968	2	Sg def	o+a
968	3	Pl indef	ø+er
968	4	Pl def	ø+er
…	…	…	…

Fig. 1: Simplified structure of the word bank

The word bank structure has been used to create an inventory of Norwegian words with inflected forms and the history of changes in their orthographic status. The Word Bank is designed for modern Norwegian. It is a flexible structure, and by creating a new set of paradigm patterns based on Aasen (1864) and the headwords of Aasen (1873), it was possible to retro-create a word bank describing a language norm for the year 1873 – here termed the Aasen Word Bank.



Lemma-id: 6594
Funksjon: lemma
Grunnform: Far
□ Skjul for tagger

Normering

Normering	Fra dato	Til dato
unormert	1864	1901

Kilder

Fi...	Kildenavn	Elementnr i kilde
	A2	556926

Paradigmer

Para...	Ordklasse	Normering	Fra dato	Til dato
700	subst mask appell	unormert	1864	1901

Pa...	N...	Fullform	Tag	Fra dato	Til dato
700	1	Far	subst mask appell eint ub unormert	1864	1901
700	2	Faren	subst mask appell eint bu unormert	1864	1901
700	3	Farar	subst mask appell fl ub unormert	1864	1901
700	4	Fararne	subst mask appell fl bu unormert	1864	1901
700	5	Fare	subst mask appell dat eint unormert	1864	1901
700	6	Farom	subst mask appell dat fl unormert	1864	1901
700	7	Fars	subst mask appell gen eint unormert	1864	1901
700	8	Fara	subst mask appell gen fl unormert	1864	1901

Alternative paradigmekoder

Kilde	Alternativ kode
A2	m.

Ortoklump

Grunnform

Fig. 2: Aasen Word bank: The headword "Far m" 'father'. Status "unormert" ('non-standard') for headword form and paradigm respectively in red frames

3. A description of Aasen's Norwegian Grammar (1864)

3.1 The contents and focus of Aasen (1864)

Aasen (1864) is a work of 394 pages with the contents ordered in introduction, five sections and two addenda. At the micro level, Aasen (1864) has 399 numbered paragraphs. Most paragraphs have a general section and a section for comments ("Anm."). Details pertaining to individual lexical items are found under "Anm.".

Aasen started reworking his "Grammatik for det Norske Folkesprog" (Aasen 1848) in the mid-1850s, but soon saw the need for substantial change. Aasen (1864) became an independent dissertation on the structure of Norwegian as a sum of spoken varieties. It introduces a standardised orthography and morphology based on an analysis of speech. Aasen (1864) aims at documenting the unity, coherence, and independence of the linguistic structure of Norwegian, as a Nordic language, related to but separate from Danish and Swedish; using Old Norse as a touchstone, including only what could be documented from Aasen's synchronic collections.

Aasen (1864) is rich in content, brief in style, and has little textual redundance. Every piece of information is given once; contents are logically ordered, resting on the assumption that the book has been read from the beginning, and that readers remember what has been read.

The macrostructure of Aasen (1864) is new, compared to Aasen (1848). Its five sections deal with (1) phonology, (2) base word form, (3) morphology, (4) word creation, (5) syntax.

Section 2 in the Grammar, Base word form, is new to Aasen (1864). This section discusses syllable structure and tone, includes a table of consonant clusters before and after the root vowel, and comments on the endings of disyllabic words for each part of speech (§ 75). Section 2 also deals with root vowel change, with a discussion of Ablaut and Umlaut in Norwegian. The systematic changes in base word forms from Old Norse to synchronic Norwegian and the relationship between Norwegian and its closest cognate languages is set out.

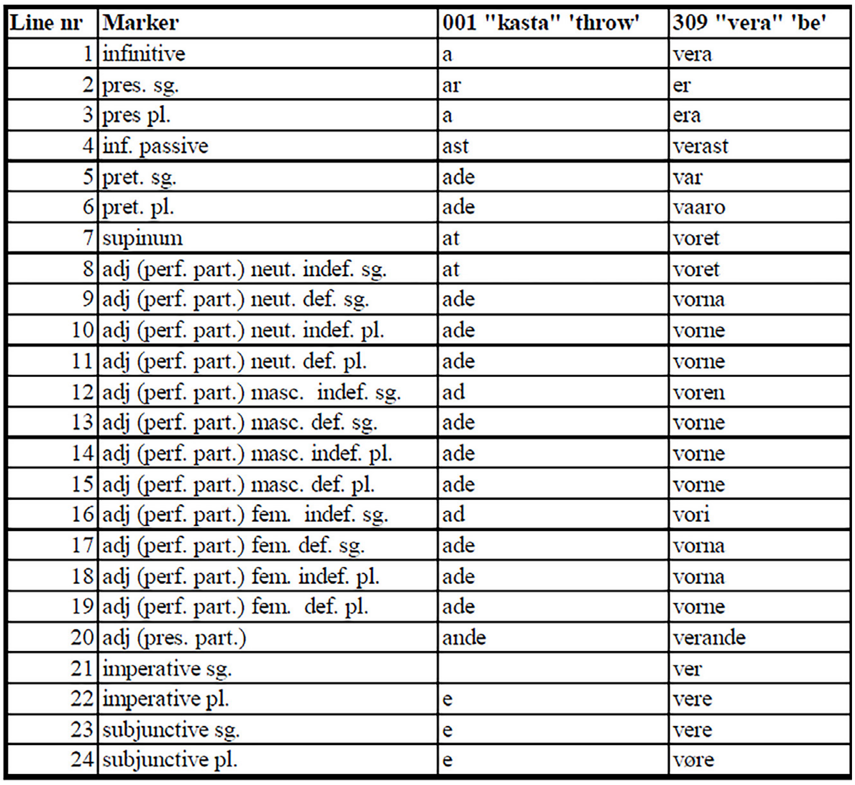
Much of the content of section 2 was inspired by a close study of Grimm's "Deutsche Grammatik" (1819–1837). In relation to Aasen (1873), section 2 sets out the framework for adapting headwords from Norwegian speech to model forms within a standard orthography, suggesting the categories of information to be expected in the treatment of individual words.

Section 4 (§ 241–300) discusses word creation in Norwegian, especially derivation from root forms or inflected forms, involving Umlaut, Ablaut, gender, and POS transitions. Aasen (1864) considers some frequently used word endings as compounding elements rather than derivational suffixes (§ 257–258). In the dictionary Aasen (1873), prefixes and suffixes regularly used in word formation have separate entries, including elements now used only in forming propria. The borderline between Aasen's view of derivation and compound can be explored further if the Aasen full form register is expanded with word forms from contemporary text, cf. chapter 5.

3.2 Aasen (1864) and the paradigm register

In the Aasen Word Bank each headword form is registered as it stands in Aasen (1873). Dialect forms (cross-referenced to other entries) that deviate from Aasen's standard orthography, are marked as non-standard headword forms, but may nevertheless have inflection paradigms. This is possible because the entry format for the Word Bank allows separate marking for the headword form itself and its paradigm. This system also allows standard headword forms to be equipped with non-standard paradigms. For an example see Figure 2.

The most important section of Aasen (1864) in relation to the Aasen Word Bank is section 3” Bøiningsformer” (‘Inflection Morphology’), comprising paragraph 151 to 240. Parts of speech (POS) are dealt with in sequence, starting with nouns and ending with verbs. Each POS has a general introduction; for nouns the issue of three linguistic genders comes first, then a discussion of linguistic gender versus biological gender. The categories of number, definiteness, and case in nouns are introduced as part of the inflection system for nouns. This system is used in relation to all word classes where inflection occurs. And since Norwegian is a Germanic language, nouns are also classified as “strong” (ending in a consonant) or weak (ending in a vowel).



Line nr	Marker	001 "kasta" 'throw'	309 "vera" 'be'
1	infinitive	a	vera
2	pres. sg.	ar	er
3	pres pl.	a	era
4	inf. passive	ast	verast
5	pret. sg.	ade	var
6	pret. pl.	ade	vaaro
7	supinum	at	voret
8	adj (perf. part.) neut. indef. sg.	at	voret
9	adj (perf. part.) neut. def. sg.	ade	vorna
10	adj (perf. part.) neut. indef. pl.	ade	vorne
11	adj (perf. part.) neut. def. pl.	ade	vorne
12	adj (perf. part.) masc. indef. sg.	ad	voren
13	adj (perf. part.) masc. def. sg.	ade	vorne
14	adj (perf. part.) masc. indef. pl.	ade	vorne
15	adj (perf. part.) masc. def. pl.	ade	vorne
16	adj (perf. part.) fem. indef. sg.	ad	vori
17	adj (perf. part.) fem. def. sg.	ade	vorna
18	adj (perf. part.) fem. indef. pl.	ade	vorna
19	adj (perf. part.) fem. def. pl.	ade	vorne
20	adj (pres. part.)	ande	verande
21	imperative sg.		ver
22	imperative pl.	e	vere
23	subjunctive sg.	e	vere
24	subjunctive pl.	e	vøre

Fig. 3: Form registers for regular verbs of the type “kasta” ‘throw’, and the unicum “vera” ‘be’

After the general POS introduction, individual inflection paradigms have a paragraph in which the recommended paradigm is shown in table form and commented on, in relation to (1) materials (speech forms), (2) Old Norse, and (3) Germanic cognates.

Aasen’s systematic mapping of the inflection system of Norwegian made the registration of paradigms a comparatively easy task. When doubt arose about the shaping of a paradigm, an extra paradigm was created, rather than forcing an uncertain interpretation as the only possibility. Several base forms have more than one paradigm, each paradigm set with the status that seems appropriate.

Aasen (1864) starts with a maximum model for each word class and deals early with the most complex cases. The result is comprehensive and well-ordered paradigms for (frequent and well-documented) irregular nouns and verbs, but also a fair amount of over-classification, e. g. for members of regular verb groups, cf. the paradigm of verb paradigm 001 (Fig. 3, middle column) where many of the adjectival forms of the past participle are identical.

4. A description of Aasen's Norwegian Dictionary (1873)

Aasen (1873) is a bilingual dictionary of 964 pages with 38.742 entries and a couple of addenda. Headwords are in Norwegian, definitions and editorial language in Danish. It is designed to present Aasen's Norwegian vocabulary collections, covering the whole language and all dialects, in an acceptable form for a new Norwegian written standard, referred to as "Landsmaal"[1] 'the language of the country' or "Folkesprog" 'the language of the people'. Aasen's first grammar (1848) and dictionary (1850) of the Norwegian vernacular had served to establish Norwegian as a modern and independent language and himself as a noteworthy comparative philologist in the field of Nordic (Germanic) languages[2]. Aasen (1873) was his magnum opus and is regarded as the foundation of Norwegian lexicography and dialectology.

Aasen (1873) is also the companion volume to Aasen (1864). In the dictionary preface, Aasen discusses criteria for exclusion and inclusion of lexical items and word forms, underlining the need for brevity and clarity at all costs, but says nothing about the organization of the dictionary as a whole or within the entry. Aasen's biographers have discussed purpose, size, and orthography, but the only one to comment on lexicographical aspects is Dagfinn Worren (2006), and his comments concern alphabetical ordering and definition formats.

There is a good reason for this lack of interest. Detailed analysis of a dictionary presupposes that the contents can be classified and counted, a major task, and not one to be undertaken manually. The Aasen Word Bank facilitates analysis by numbers and reveals Aasen's working method in greater detail than what has been possible until now.

A nineteenth century dictionary is primarily a running text. To break up entries into ordered categories via a detailed lexicographic database would be forcing contents into a straitjacket and corrupting the result. It is better to respect the category system of the author, which in a dictionary means breaking the text into entries, isolating headwords with their POS, and then annotate entry contents. In this analysis, the aim has been to identify all word forms that Aasen himself presents as lexical items, whether in the standard orthography or in dialect form only.

4.1 Entry types, numbers, and contents

The entry formats of Aasen (1873) can roughly be divided into main (content) entries and cross-reference entries. The total number of lexical items (entry headwords and additional base forms with POS) extracted by machine analysis is 43.194. About 6000 additional head-

[1] The word form "Landsmaal" as a term was first used by Aasen in Aasen (1864 § 341).

[2] Aasen was well known to German philologists. The 1848 grammar and the 1850 dictionary were reviewed by Theodor Möbius in Gersdorfs Repertorium in February 1851, and both Grammar and Dictionary were sent to Möbius by Aasen himself immediately after publication. Aasen (1873) was reviewed by K. Maurer (1873) and F. Liebrecht (1874).

words are so far identified by ongoing manual control, the total being likely to end at 50–53 000. This means that many entries cover more than one lexical item.

Each main entry gives a description of its headword according to Aasen's criteria. Inflected forms and a selection of speech variants are listed, as are forms from other Germanic languages and Old Norse. Aasen also illustrates the lexicogenetic potential of each headword by giving information about lexical items connected to the headword by form, i.e. derived forms, compounds and multiword expressions. A note on headword form as the first element in a compound is often included. Definitions are supported by synonyms or near-equivalents.

Some lexical items listed within entries of other headwords have their own entries, but not all, cf. Figure 4. Lexicographers of Norwegian after Aasen have tried to extract every lexical item in Aasen (1873), irrespective of the word form's placing as headword or within the entry text (see for example Grunnmanuskriptet (1935/1997)). The aim of the analysis in the Aasen Word Bank is the same. Doubtful cases are checked against the Norwegian language collections. A word form found as an independent lexical item, will be included. Doubtful cases particularly concern lexical items where the orthographic form may include spaces or hyphens.

(1) **Aar**, f. (Fl. **Aarar**), Aare, Redskab at roe med. G.N. ár (Eng. oar). Afvigende Fl. Aarer (Tel. og fl.). I Sammensætning sædvanlig med "a", som dog tildeels skulde være "ar"; saaledes: **Aara(r) blad**, n. Aareblad. **Aaraburd**, m. 1) den maade hvorpaa man bevæger Aarerne; 2) Rum til at røre Aarerne (Sfj.). **Aaradrag**, n. enkelt Drag eller Træk med Aarerne. **Aararlask**, s. Lask. **Aararlom** (oo), m. Grebet paa en Aare. **Aaraløysa**, f. Mangel paa Aarer.

Fig. 4: The entry for "Aar f." 'oar' lists six compounds, which can be seen as nested entries, or supplementary information to the headword Aar. The entry comments on the infix variation a/a(r)/ar. The Aasen Word Bank includes entries for compound prefixes plus infix.

4.2 Grouping headwords round a definition

In many cases Aasen's word harvest consists of several word forms from different parts of the country which are full synonyms, but not dialect variants of the same base form. The result may be an entry as shown in Figure 5, which is centred around a definition. Here, the structurally simplest form is selected as headword for the entry, while the other four are listed after the introduction "Ogsaa kaldet" 'also called'.

(2) **Kjøta**, f. Kjødside, Indside paa Skind eller Huder. Hard. Ogsaa kaldet: **Kjøkka** (Kjøtka), Hall., **Kjøtska**, Buskr. Ellers: **Kjøtrosa** (o'), f. B. Stift, Nordl. **Kjøtroslid** (i'), f. Sæt.

Fig. 5: The entry "Kjøta f." 'inside of (animal) hide'

Two of the other word forms are cross-referenced to this entry, the others are not mentioned anywhere else. In this type of entry, Aasen approaches the thesaurus entry format, which he knew from Roget's work. More than 600 entries include a synonym section similar to the entry above. This entry format came to attention through the Aasen Word Bank format, which makes it possible to sort out entries linked to multiple (standard and non-standard) base forms.

4.3 Types of cross-reference entries in Aasen (1873)

The number of cross-reference entries in Aasen (1873) is about 4800 – 12,4% of the total number of entries (38.742). This is a very high share compared to later dictionaries describing a more standardized version of Nynorsk, where the proportion of cross-reference entries lies around 6% (the number is taken from the editorial database, dated 2013).

Most dictionaries indicate the status of a headword by the entry format. An entry consisting of a headword and a cross-reference to another headword, can be assumed to be less important, and most often a non-standard form of the target for the cross-reference entry. In Aasen (1873), this assumption would be a fallacy. The Aasen Word Bank has through its structure and comment system allowed a more detailed mapping of Aasen's use of the cross-reference system, carried out as part of the manual control. Cross-references are of two kinds, those which only point to a target headword (ca. 4100) and those which contain minimal other information about the headword, most often a Danish equivalent (ca. 700).

The simple type is used for linking dialect forms to standard forms (3), inflected forms to base forms (4), alternative standard forms to the main entry form (5):

(3) **abakleg**, s. avbakleg.

(4) **fraus**, s. frjosa.

(5) **andleg**, s. andeleg.

But often the headword form fails to match the target form linguistically, as in this example:

(6) **frestalle** (fleste), s. fleire. ‘several’

There is no way the headword form “frestalle” can be a form of “fleire”. The explanation is found under the entry “fleire”:

(7) “Paa Sdm. **frestalle**, for flest-alle.” ‘In Sunnmøre **frestalle**, for “flest-alle’”

The expanded version would be “In one region of Norway, the multiword expression “flest alle” is used to express the notion ‘several’”. Cross-references in Aasen (1873) are also used as in encyclopaedias, meaning ‘explanation or more information will be found under the target entry’.

A look in later dictionaries and the digital Norwegian language collections confirms that the standard forms “flest alle” and “flestalle” are listed as a separate lexical item. In the Aasen Word Bank an entry for “flest-alle” has been added with links to the two entries “frestalle” and “fleire”. This means that all cross-reference entries must be checked manually, to ensure that each headword is correctly marked in the Aasen Word Bank as standard or non-standard.

Cross-references can also serve to draw attention to the most widely attested speech form. For some lexical items, Aasen for system reasons chose a disyllabic headword form, although the dominant spoken form is monosyllabic. In the dictionary (Aasen 1873) – the opening of the entry for “Fader” as standard form:

(8) **Fader**, m. (Fl. **Feder**), Fader (pater). Lyder mest alm. **Faer**, **Far** (som i Svensk og Dansk); … .

The most widely used speech form “Far” is cross-referenced to “Fader m” (and another entirely different headword).

(9) **Far**, m. s. Fader og Fare.

The simple form “Far” then turns up in compounds for relationship words, with uncertain status information. Orthographic reformers after Aasen who preferred a more orthophone approach, would therefore often find their work done for them in Aasen (1873), with information on extent and usage of the non-standard forms.

4.4 To what extent is Aasen (1873) a normative dictionary?

The headwords of main entries in Aasen (1873), taken together, are generally considered a proposal for a Nynorsk standard orthography. Although Nynorsk had been used in books and journals since the 1850s, its literary corpus was still very limited in 1873, and the orthography was heterogenous. Aasen knew very well that he was putting forward a proposal, not a decree. He could hope for acceptance, but not command it. Aasen’s authenticity principle dictated that he would not include a word unless he himself had heard it or had information on form, sense and usage confirmed from more than one contemporary source whom he trusted. His language collections – now in the National Library – start in the 1830s and gain force in the 1840s. In addition to his own collections, he received several manuscripts from others. From these, he used what he could verify, if he trusted the informant.

In the introduction to Aasen (1873), Aasen states that although there is some literature written in forms of modern Norwegian, it is not included in the materials used for the dictionary (p. XII). His dictionary is a presentation of spoken, contemporary Norwegian, properly verified.

Aasen’s textual corpus was transcribed speech, in dialect form. Modern Norwegian yet had no written standard expression. The model he used for synthetizing dialect forms into a proposed standard orthography was to find an orthographic form which would link speech forms in a systematic way. The consonant digraph “rn” as syllable ending is the standard example. In Aasen’s day, “rn” had been replaced in speech by /dn/ (sound differentiation) or /nn/ (assimilation). But “rn” would explain both speech forms, had the support of use in standard Danish and Swedish, and was used in Old Norse, so “rn” was introduced, e. g. in “Bjørn” ‘bear’.

Aasen set out to document the whole language. This means that his collections are solid and multi-sourced for frequent vocabulary, but also rich in thinly documented and doubtful cases. His approach to identifying dialect forms with a standard form, and putting the standard form into its proper linguistic context, is revealed by studying entry types and entry format.

4.5 Aasen (1873) – dictionary type and mode of analysis

In his discussion of bilingual dictionary types, the lexicographer Ladislav Zgusta outlines three subgroups “with a remarkably outstanding concentration upon some purpose” (Zgusta 1971, pp. 304 f.). One of them he terms the “ethnolinguistic bilingual dictionary” constructed for languages with little or no written literature. In such cases, the defining language is often another, well-established literary language, used to bridge a culture gap. A dictionary of this type is designed to introduce as a written standard a language existing only in the vernacular and might well be termed a pioneer dictionary.

Aasen (1873) is a pioneer dictionary, designed to introduce a written standard for the Norwegian vernacular (Grønvik 1992). Several of the dictionary features discussed above become methodically rational in this perspective. His materials are heterogeneous in terms of form and place of origin – the dictionary must convince its users that the language is cohesive and well designed. The well-documented central vocabulary is the backbone of the dictionary. Words that are not so well attested, must be guided into the language system through cross-referencing and by being included as support material in more comprehensive entries. Cross-referencing irregular inflection forms guides users to the proper entries. By organising his dictionary in this fashion, Aasen is also able to throw light on the word creation system of Norwegian, especially by establishing the prefix form of compounds.

The least conventional feature, seen from present-day lexicographical practice, is the type of entry that groups linguistically unrelated synonyms round a definition, without giving the headwords in the group conventional entries, or even cross-references. Aasen (1873) is a semasiological dictionary, but this type of entry belongs in an onomasiological dictionary. Aasen was deeply interested in onomasiology and left a manuscript for a Norwegian thesaurus published posthumously (Norsk Maalbunad 1925).

The Aasen Word Bank was launched because it is needed in analysing early Nynorsk text. But analysing the dictionary through the Word Bank system has also brought new insights in the dictionary as it stands, in Aasen's lexicographical method (which he never discussed in any context), and above all made the contents of the dictionary more accessible and therefore possible to evaluate. It is to be hoped that the Aasen Word Bank will contribute to a wider scholarly interest in and use of Aasen (1873).

5. A first application of the Aasen Word Bank as an analytic tool

The Nynorsk written standard of today is based on Norsk Grammatik (1864) and Norsk Ordbog (1873), two major works by Ivar Aasen. In 1885, the Norwegian Parliament accepted Nynorsk as a second written standard in addition to Danish.

The works of the Norwegian author Aa. O. Vinje (1818–1870) are early examples of texts written in a form of Nynorsk, with a degree of standardisation. Aasen (1873) was published three years after the death of Aa. O. Vinje, so Vinje cannot have used it. But Vinje and Aasen were in close contact for many years. It is generally assumed that Aasen advised Vinje on his orthography (Aasen 1957), although final choices will have been Vinje's own. The most important choices on inflection morphology are published in a small supplement to his journal Dølen (Vinje 1859).

We have used text written in Norwegian from the collected works of Vinje (Vinje 1916–1921) as a first test case for the form list produced form the Aasen word bank. Vinje's primary literary output is journalism and essays. In the mid-19th century, the tradition of long novels was not developed in Norway, in contrast to what we find in France and UK. The collection of Vinje's works is not large, only 750.000 running words.

	Word forms found	Unique word forms found
In both	51.8977	10.968
Only in Aasen Word Bank	35.371	4.572
Only in the Nynorsk Word Bank	44.409	7.837
Total found in either and/or both	598.757	23.377
Not found in neither	79.312	24.493

Table 1: The resulting numbers from the analysis of Vinje's writings

From the Aasen Word Bank a list of all full forms was generated, approximately 290.000 unique word forms or 520.000 unique triples (word form + POS + information about inflection categories). A similar list with recommended forms generated from the word bank of modern Nynorsk consists of 415.723 unique word forms and 585.046 triples (word form + POS + information about inflection categories). The Aasen Word Bank has 45.000 base forms, and the modern Word Bank has 120.000. As we have seen, the Aasen norm has a much richer inflectional system, which explains the relatively larger number of word forms in the Aasen word bank.

A small program identifies the word forms in the running text and checks for each if it is an inflected form in the Aasen Word Bank and/or an inflected form of modern Norwegian. Table 1 shows the resulting numbers. As mentioned in section 2 the word bank does not have a system for compound analysis. Consequently, the number of matches is lower than it could be.

The total number of unique word forms is 47.870. Of these, 26.917 occur once, while 157 word forms occur 500 times or more. A brief comment on the most frequent word forms: 137 of them are found in both word banks, nine only in the Aasen Word Bank, three only in the Nynorsk Word Bank, and seven word forms are specific to Vinje. Of the nine in the Aasen Word Bank only, six are inflection forms where the spelling has been changed after 1873, the last three base forms. The three in the Nynorsk Word Bank only are inflection forms changed after 1873. Some of the word forms specific to Vinje can today be seen as signature forms ("ikki" 'not' (instead of Aasen's "ikkje"), "ero" 'are' (plural form used in Aasen 1864, but replaced with the form "era" in Aasen 1873)). Others reflect his linguistic environment, and the occurrence of Danish text in his writings (the Danish form "ikke" 'not', the preposition form "af" which he used in Norwegian and Danish irrespectively). Others just seem to show different orthographic habits from what Aasen was to recommend in his dictionary three years after Vinje's death.[3]

About half, 49%, is found in one or both of the word banks used. 51% is not found in either word bank. A finer classification has been done on the word forms of the letter a (1798 instances) Below we comment briefly on word forms starting with a- and found 1) in both word banks, 2) only in the Aasen Word Bank, 3) only in the Nynorsk Word Bank, 4) in neither word bank.

1) The unique word forms from Vinje that are found in both word banks comprise the core vocabulary of Nynorsk, with inflected forms. The number of compounds is relatively

[3] Vinje wrote "kver" 'every', Aasen "kvar"; "altid" vs. "alltid"; "up" vs. "upp".

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

small. Imported vocabulary is not there, because Aasen did not include it in his dictionary. There are very few doubtful cases, mainly because Aasen's orthographic norm is very clearly defined and made explicit by the Aasen Word Bank.

2) The word forms from Vinje found only in the Aasen Word Bank are mainly inflected forms that are no longer part of Nynorsk orthography, e. g. regular adjectives ending in "-ad" (now reduced to "-a").

3) Of the word forms from Vinje found only in the Nynorsk Word Bank, about 40% are forms of imported (e. g. non-Germanic) vocabulary (not included in Aasen's dictionary, but essential in journalistic text), 17% name forms, ca. 10% then classified as Danish word forms, since included in Nynorsk standard orthography. The rest, about 33%, are word forms of headwords that are either consistent with Aasen's orthography, but missing in his dictionary (25%), or specific to Vinje's personal and much more heterogeneous orthography (8%). Vinje wrote, printed, and sold twice a week, and when in doubt he seems to have chosen word forms consistent with his dialect from Western Telemark, some of which have found their place in modern Nynorsk.

4) Of the word forms found in neither word bank, a much larger sample should have been analysed than there has been time for, for the word creation system for Norwegian gives great weight to frequently used particles. Vinje often used the preposition word form "af" where Aasen chose "av" 'of'; this choice affects ca. 300 compounds. It can however be said with certainty that the larger groups of word forms specific to Vinje will belong to the following types: a) Danish word forms (from quotes, and shorter texts in Danish). b) Imported word forms, many of which today have an orthography adapted to Norwegian. c) Word forms consistent with Aasen's orthography, but not found in Aasen's dictionary. d) Word forms specific to Vinje, i. e. word forms spelt by the ear and often reflecting Vinje's dialect basis.

There is also a sprinkling of (quoted) word forms from modern languages and Latin. Vinje was proud of his command of English, and this group seems to be the largest.

The conclusion must be that Vinje and Aasen mostly agreed on what their written (standard) Norwegian should look like. In the frequently used word forms, the number of deviations between Vinje and Aasen are few, but they show up because of their frequency. As for the rest, Vinje was a writer and journalist navigating in uncharted waters, his concern was to make sure he got read. He used and made the words he needed, and when in doubt, it seems that Norwegian speech was his compass.

References

Aasen, I. (1848): Det norske Folkesprogs Grammatik. Kristiania.

Aasen, I. (1850): Ordbog over det norske Folkesprog. Kristiania.

Aasen, I. (1864): Norsk Grammatik. Kristiania.

Aasen, I. (1873): Norsk Ordbog. Med dansk Forklaring. Kristiania.

Aasen, I. (1925): Norsk maalbunad. Samanstilling av norske ord etter umgrip og tyding. Oslo.

Aasen, I. (1957): Brev og dagbøker. B. 1. Brev 1828–1861. [ed. Reidar Djupedal]. Oslo.

Engh, J. (2014): IBMs leksikografiske prosjekt for norsk 1984–1991. In: Maal og Minne 106 (1): pp. 67–101. http://ojs.novus.no/index.php/MOM/article/view/225. (last access: 25-03-2022).

Grimm, J. (1819–1837): Deutsche Grammatik. Göttingen

Grunnmanuskriptet (1935/1997): https://usd.uib.no/perl/search/search.cgi?tabid=993&appid=59.

Grønvik, O. (1992): The earliest dictionaries of Nynorsk in the light of present day dictionary typology. In: The Nordic Languages and Modern Linguistics 7. Proceedings of the Seventh International Conference of Nordic and General Linguistics in Tórshavn, 7–11 August 1989, Vol. I. Føroya Fróðskaparfelag (Annales Societatis Scientiarum Færoensis, Supplementum XVIII). Tórshavn.

Grønvik, O. (2016): The lexicography of Norwegian. International handbook of modern lexis and lexicography. Berlin/Heidelberg, pp. 1–34.

Hagen, K./Johannessen, J. B./Nøklestad, A. (2000): A constraint-based tagger for Norwegian. In: 17th Scandinavian Conference of Linguistics, Volume I, no. 19. Odense Working Papers in Language and Communication.

Hagen, K./Nøklestad, A. (2010): Bruk av et norsk leksikon til tagging og andre språkteknologiske formål. LexicoNordica (17). https://tidsskrift.dk/lexn/article/view/18624 (last access: 25-03-2022).

Hovdenak, M. et al. (1986): Nynorskordboka. definisjons- og rettskrivingsordbok. [3. ed. 2006]. Oslo.

Nøklestad, A. (2022): The Oslo Bergen Tagger. https://github.com/noklesta/The-Oslo-Bergen-Tagger (last access: 25-05-2022).

Nøklestad, A. (2022): The Compound analyzer software. https://github.com/textlab/mtag (last access: 25-05-2022).

Ore, C.-E. S. (2016): Gamle ordbøker og digitale utgaver. Nordiske Studier i Leksikografi 13. Rapport fra 13. Konference om Leksikografi i Norden København 19.–22. mai 2015. København, pp. 203–216.

Ore, C.-E. S. (2020): Å ta Hans Ross på ordet: Ross' ordbok i relasjon til Aasens med Metaordboka som verktøy. Nordiska studier i lexikografi 15.Rapport från 15 konferensen om lexikografi i Norden. Helsinki, pp. 253–264.

Roget, P. M. (1852): Thesaurus of English words and phrases classified so as to facilitate the expression of ideas and to assist in literary composition. London.

Venås, K. (1996): Då tida var fullkomen. Ivar Aasen. Oslo.

Vinje, Aa. O. (1859): Det norske Landsmaals vigtigaste Bøygningsformer. Bilag til Dølen nr. 30, 1859.

Vinje, Aa. O. (1916–1921): Skrifter i Samling I–V. Oslo.

Walton, S. J. (1996): Ivar Aasens kropp. Oslo.

Wangensteen, B. (1986): Bokmålsordboka. Definisjons- og rettskrivningsordbok. [3. ed. 2005]. Oslo.

Worren, D. (2006): Molbech som mønster for Aasen. In: Nordiske Studier i Leksikografi 8. Nordisk Forening for Leksikografi, pp. 391–406.

Contact information

Christian-Emil Smith Ore
Department of Linguistics and Scandinavia Studies, University of Oslo
c.e.s.ore@iln.uio.no

Oddrun Grønvik
Department of Linguistic, Literary and Aesthetic Studies, University of Bergen
Oddrun.Gronvik@uib.no

Trond Minde
Department of Linguistic, Literary and Aesthetic Studies, University of Bergen
Trond.Minde@uib.no

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.