Christian-Emil Smith Ore and Oddrun Grønvik

THE SPOKEN WORD AS REPRESENTED IN NORSK ORDBOK

Abstract Spoken language is the prerequisite of written standard languages for living language communities. Yet written sources dominate lexicographic description of standard languages, and awareness of dictionaries that specifically source speech seems limited. In *Norsk Ordbok* (*The Norwegian Dictionary*), and in the Language Collections on which the dictionary is based, oral materials are perceived as input to the national language Nynorsk, written or spoken. The purpose is integration into one whole, not a series of parallel lexical registers. Legitimacy is aimed at by explicit sourcing of linguistic information, whether from speech or literature. This paper looks at how speech is sourced within the entries of the dictionary *Norsk Ordbok* (NO), particularly at the sourcing of definitions. Explicit sourcing of speech in connection with definitions facilitates investigating the contribution of speech materials to *Norsk Ordbok* as a whole, and if and how the differences between speech and written text is reflected in *Norsk Ordbok*.

Keywords speech; dialect; vernacular; standard language; location; location source

1. Introduction

Speech-based dictionaries range from major multi-volume scholarly dictionaries (cf. *Dictionary of Finnish Dialects*, 2024, the *English Dialect Dictionary*, Wright, 1898–1905), to spontaneous web creations. A web search for "dialect lexicography" gives ca. 4 000 matches. Some ambitious modern digital dialect dictionaries are discussed in Euralex contributions (cf. Tier & Keumeulen, 2010). Yet speech sourced lexicography is generally seen as something quite different from mother tongue lexicography concerning a standard language.

Norsk Ordbok (NO) is an exception. The full title of NO is *Norsk Ordbok. Ordbok over det norske folkemålet og det nynorske skriftmålet* 'The Norwegian Dictionary'. Dictionary of the Norwegian vernacular and the Nynorsk written standard'. The aim of integrating the documentation of speech and writing motivates the editorial system of NO, from lemma selection to the details of source sorting and listing (Skard, 1932; Hellevik, 1956), but until recently, it has been impossible to examine the whole in order to see how well the result meets the original requirements.

This has to do with a change in production method. The first (paper) edition of NO was produced in two stages. The alphabetical section *a-h* was edited on paper in the period 1946–2002. The alphabetical section *i-å* was edited into a relational database under a consistent set of editorial rules in the project NO2014 (2002–2015). The last eight volumes were produced from this editorial system.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

In 2016, NO and the Norwegian Language Collections (Grønvik, 2015, p. 32) were moved from the University of Oslo to the University of Bergen. In 2019, the planned revision of the alphabetical section *a-h* started. The editorial system used for the alphabetical section *i-å* is still in use (though undergoing rewriting). The alphabetical section *a-h* of the printed dictionary is not complete in the dictionary database. Therefore, the editorial database is not well suited for analysing NO as a whole. Fortunately, the text of first two volumes were retro-digitized as formatted text files in 2004, and the editors' files for the rest of the alphabet section *a-h* are available as text files with a simple field mark-up. All the files have been thoroughly analysed and a given a semi-automatic xml mark-up at the level of detail found in the editorial system. This was first done in 2004 and then revised in 2023. As the alphabetical section *a-h* is under revision, the first edition has recently been published separately on the web as *Norsk Ordbok AH 2005* and *Norsk Ordbok IÅ 2016*. The latter is based on an xml-export from the editorial system of the alphabetical section *i-å* as it was in 2016. Both have been included in the extensive collection of background material underlying the revision of *a-h*.

NO can be accessed in several ways. It exists as a printed dictionary in 12 volumes. The editorial system is accessible for editors and others with special interests in the material. NO also exists as a web version that takes data directly from the editorial system, giving the current state of the contents. In addition, there are the fixed version *Norsk Ordbok AH 2005* and *Norsk Ordbok IÅ 2016*.

The whole of NO is now available for consistent analysis, and it is time to look at how NO fulfils its original, ambitious aim. The analyses referred to in this paper are based on the xml-texts for *a-h* and the editorial system (a relational database) for *i-å*.

2. The Terms 'Dialect' and 'Location'

The definition of the term 'dialect' may vary from one language community to another. In the study of the Norwegian spoken language, the following definitions are used:

Dialect = a geographically delimited language system (with a specific and defined phonology and morphology, to a lesser degree syntax and lexicon) (Sandøy, 1985, p. 68)

Norway lacks a spoken standard language ("received pronunciation") with the accompanying status and rules for usage that many other language communities have. In contrast, the space for and social acceptance of dialects is considerable (Netland & Opsahl, 2024). The closest one gets to a standard for speech is spoken versions of the two written standards Bokmål or Nynorsk, as used by hosts and announcers in news programmes and programme information in the national broadcasting corporation NRK. The varieties "spoken Bokmål or Nynorsk" covers word forms only, not intonation or accent. "In other programmes and in news items (reportage, commentaries, reports, interviews etc.) dialect may be used, including in national broadcasting" (NRK språket, 2007)

It follows that NO attempts to describe the Norwegian vernacular and the Nynorsk standard language in speech and writing, without assigning a limited register of genres or styles to the one or the other. Sections 6–8 deals with this issue in more detail. In this paper, “speech” is used in parallel with “dialect” in referring to dictionary entries built on oral materials.

Location = named place or area (where a dialect is used). When locations are explicitly stated in an NO definition, it can be assumed that the Language Collections at the University of Bergen hold documentation of a spoken usage of the word defined.

There are two referencing systems in use for referring to a place with regard to dialects, landscapes and (the area of) local municipalities. Landscape location is the older system, based on former administrative districts. Municipality location, referring to the lowest level in the administrative system, has been the preferred reference system for the study of the Norwegian vernacular since before World War II. The municipalities were based on the parishes as they were in 1837. This administrative partitioning of the country is consistent with the dialect borders. In the Language Collections at the University of Bergen, both location systems are used, but municipalities dominate in the more recent collections.

The use and acceptance of dialects in formal contexts in Norwegian society is accompanied by a strong interest in standard language and dialects among language users. It is important for NO to locate speech information through the entries, as users look for this information.

The location hierarchy of NO is a synthesis of the locations used by Norwegian dialectologists from 1870 and onwards. The oldest term, “landscape”, is in the middle (Table 1, level 4).

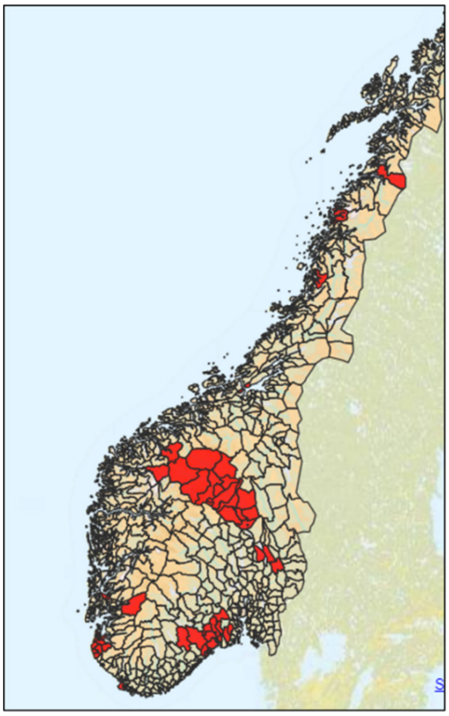
NO differs from previous Norwegian dictionaries in giving very detailed information about the location of usage. For example, the linguist Ivar Aasen in his dictionary (Aasen, 1873) did not go below “landscape”, which gave him a list of less than 200 possible locations, while NO uses about 900 locations, mostly the 750 municipalities (*kommune*) as they were in the 1940s, the most fine-grained administrative organisation of Norway at any time.

Table 1: The *Norsk Ordbok* location hierarchy

Level	Norwegian term	Number	English	Example
1	Kommune	750	Municipality	Nordre Land
2	Tvillingkommune	16	Adjoining municipalities with a joint name	Land
3	Del av landskap	6	Part of landscape	Nedre Romerike
4	Landskap	58	Landscape	Romerike
5	Del av fylke	5	Part of county	Vest-Oppland
6	Fylke	19	County	Oppland
7	Del av landsdel	6	Part of province	Flatbygdene på Austlandet
8	Landsdel	5	Province	Austlandet
9	Region	2	Region	Nordafjells
10	Land	1	Country	Noreg

In the electronic editorial system of NO, the location system is organized as a hierarchy of ten levels, starting from below with "kommune" ('municipality') and ending in the country as a whole "Noreg" ('Norway'), as shown in Table 1. When half or more units at one level are registered, editors can generate the higher level automatically. The sequence of locations is fixed, starting in the Southeast and ending in the North next to the border shared with Russia.

The location hierarchy is linked to a digital map of Norway with the municipal borders valid in the 1940s, and the locations of word forms, senses and usage examples can be shown in maps as in Figure 1.



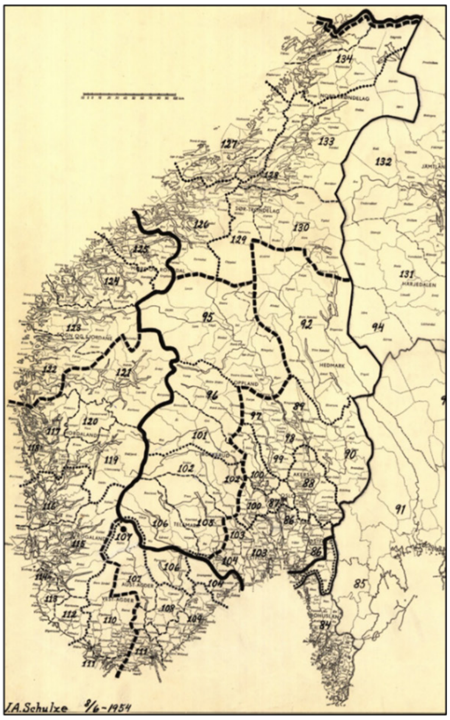


Fig. 1: Location sources (based on municipalities in 1947) for the NO entry IV *skrella* v. sense 3 'remove peel (from something)'

Fig. 2: Map ND001 from the Norwegian Dialect Atlas, showing the chief isoglosses for the dialect landscapes of Southern Norway, the East-West divide (thick black line) being the most important one

3. Editorial Rules in Norsk Ordbok for the Documentation of Speech

3.1 Origins

The written standard Nynorsk is based on the ground-breaking work by the Norwegian linguist Ivar Aasen (1813 – 1896). Based on his studies of the Norwegian vernacular through extensive fieldwork, Aasen wrote two grammars (Aasen, 1848; Aasen, 1864) and two monolingual dictionaries (Aasen, 1850; Aasen, 1873). The 1864 grammar and the 1873 dictionary are the foundation of the written standard Nynorsk that has been one of the two official standards since 1885.

The editorial rules for documenting the vernacular in NO have their background in principles laid down by Ivar Aasen in his 1864 grammar (Aasen, 1864). He emphasized the unity and coherence of the lexicon across dialect boundaries, which primarily express themselves in phonology and morphology:

> § 375. In the lexicon, or the use of the words themselves, the landscape languages are united to a much higher degree, so that one can hardly provide a significant collection of words particular to any single small district. (Aasen, 1864, p. 356) (authors' translation)

and further:

> § 385. To the Norwegian lexicon, we count all words used in the country as long as the form of each word may be termed Norwegian, and the word form does not collide with the old rules in the language for sound positioning and word forms. (Aasen, 1864, p. 366) (authors' translation)

In short: Any dialect word form can be expressed in the standard orthography of Nynorsk, which in turn can be used to coordinate all dialect forms of a given lexical item. This also means that all NO editors must be able to evaluate and standardize Norwegian dialect materials (Gundersen, 2026, p. 83).

3.2 Editorial Assumptions and Rules

The intention behind the treatment of speech materials in NO, as expressed in the editorial rule book (Gundersen, 2016 p. 380), is

- to give speech sources the same value as written (printed) sources get,
- to give special attention to the speech materials unique to the NO language collections (contributions from informants sent in 1930 – ca. 1980),
- to source speech information explicitly when the definition or usage example in question is specific to a given part of the country,
- to indicate speech usage if a word is generally used in speech rather than print, but not specific to a given area.

Material collection for NO started in 1930 as a national joint effort. Speech materials came from voluntary informants and from researcher collections, especially from phonologists and dialectologists. The older parts of the speech collections document Norwegian vernacular from the first part of the 20th century, while newer dialect sources tend to come in print. See Section 5 below.

Is the speech information in NO valid today? The dialect map of Norway has to some extent changed. Norway underwent considerable societal changes towards the end of the 19th century. Dialect collections from before 1900 are therefore since 2002 listed separately as "older sources", and not integrated in the (supposedly) synchronic dialect collections from after 1900 (Gundersen, 2016, p. 159). Location of usage taken from the dictionaries (Aasen, 1873; Ross, 1895) of Aasen and his closest colleague, Hans Ross, are used, but these sources are indicated by attaching an A or an R to the location name, e.g., 'bryddaup Tel A2' indicating that this form of *bryllaup* ('wedding') was located to 'Telemark' in Aasen (1873).

Speech sources for the central lexicon are registered in the dictionary database, but often suppressed in print. In his 1873 dictionary, which is based entirely on fieldwork, Aasen established the practice of listing the location of sources only for word forms, senses and usage examples that did not have nationwide or general coverage. Thus one finds location information only in 50% of the entries. NO continues this practice in relation to speech sources. Accordingly, entries from the central lexicon can have plentiful coverage in the speech collections, but definitions will not have location sources unless the sense itself is sparsely sourced. An example is the entry I *stein* m ('stone noun masc.'). Out of the six senses in the entry only one, no. 6, has location sources (6. 'floor in a fireplace; chimney base').

What value does NO assign to speech materials in lemma selection? NO does not in principle allow entries based on a single source. The basic requirement for writing an entry is sufficient materials to identify the headword form with POS, and give it a proper sense description, in the form of a definition with a hypernym and identifying additional features. Some informants provide all that, but it is still one single source. On the other hand, the dialect materials unique to the NO language collections carry especial weight. In the end, editors are trusted to make informed judgments based on the information available to them, and NO has 55 000 single source entries – about twenty percent. Fifteen percent of these entries have a location source, signifying that the entry information is drawn from speech.

The meaning of location names in the dictionary text is simply that the information found in the Language Collections has the location listed as its place of registration. Location sourcing does not entail a claim that the word form, sense or usage is unique to the location mentioned, nor does it mean that the word is not used in print. If further documentation turns up, it will be added to the Language Collections and in time to the entry in question.

4. NO Speech Sources

The Language collections used by NO comprise:

1. Slip archives, originally on paper, digitized and available on the web. The current volume is at about 3.5 million slips. It has been estimated that the number of slips covering speech number about 0.5 million from ca. 700 informants.

2. The Dialect Synopsis – detailed form information on the pronunciation of about 1600 word forms, with the location system that NO uses.

3. The Norwegian Dialect Atlas – 596 maps showing isoglosses relevant to Norwegian dialects, created 1950–1980, digitized 2005–2007.

4. Printed dialect dictionaries. There are about 500 of them in the NO bibliography, and still more exists (Nes, 1986). Some are available only as paper books, some are also accessible via the National Library electronic bookshelf, some are transcribed and available to editors in the NO collections of electronic texts (primarily transcripts of pre 1900 texts). A remote aim is to have them all added to the Dictionary Hotel (Ordbokshotellet), a portal of dictionaries and spellers used as sources for NO.

5. The Dictionary Hotel (Ore & Grønvik, 2018) currently holds 80 dictionaries, 69 of which are dialect dictionaries with ca. 225 000 entries in all. Sixteen out of nineteen counties are represented. Northern Norway dialects have been prioritised, with 22 dictionaries in the Dictionary Hotel.

6. Speech corpora (NDK; LIA) were not available in time for the first (paper) edition of NO, which was completed in 2015. Norsk Dialektkorpus (NDK) is a dialect corpus collected 2009–2012, transcribed in a phonetic version and standardized to Bokmål (Johannessen et al., 2009, p. 74). This has caused the speech vocabulary to be misrepresented, as standard Bokmål rejects a number of much used dialect words and word forms, which entails translation to the correct Bokmål word. The speech corpus LIA Norsk v. 1.1 consists of transcriptions of dialect recordings (1939–1996) archived at Norwegian universities. LIA is standardized to Nynorsk and has in its first text version 3.5 million tokens and 1274 speakers from 226 places in Norway (Hagen & Vangsnes, 2023, p. 124). The LIA Norsk corpus was published in 2019 and is available from Språkbanken (the Language Bank) of Norway's National Library.

The collections listed above have been put together for different purposes, and are diverse in contents, organization and metadata. They support each other; the Dialect Synopsis and the Norwegian Dialect Atlas provide a general framework covering historical phonology and geographical distribution of linguistic features, while the Dictionary Hotel show a cross-country spectrum of forms and meanings indexed in standard Nynorsk form. The dialect dictionaries portray the lexicon of individual dialects. The slip archives give invaluable information, but are also the most heterogeneous in contents and metadata.

The Speech corpora NDK and LIA Norsk will be useful in documenting basic vocabulary, and especially pragmatic speech markers and the give and take of casual conversation (Askeland, 2017, p. 75), where the older collections have little to offer. The lemma inventory of the speech corpora has not yet been described in detail.

5. Quantities – Numbers of Entries and Definitions With Location of the Sources

The practice of accepting speech sources and integrating documentation from speech and writing makes NO one of a kind among the scholarly dictionaries for Nordic languages. It is therefore of interest to find out how the speech materials affect the whole.

A number of assumptions are current about the inventory of the spoken lexicon that lacks documentation in published text. The chief assumption is that speech materials deviate from the norms of the proper (standard) language, and are therefore unimportant (Sandøy, 1993, p. 11). At the same time, a number of empirically based fields of knowledge (from botany to zoology) depend on users' knowledge and terminology, whether documented in writing or not. The same applies, surely, to lexicography and linguistic fieldwork; the NO position is that documenting speech is important if the aim is to provide a valid portrait of a language through mother tongue lexicography.

The following sections will deal with how information from speech and writing is integrated at entry level, in terms of quantities and distribution, and will finally look at what sort of information the speech materials offer that is not documented in the written sources.

6. Speech Materials in Quantities and Proportions

Entries form three groups according to source type. These are firstly entries with only literary sources, secondly entries with both literary and spoken sources, thirdly entries with only spoken sources. Speech source distribution is shown in Table 2. Source type distribution is shown in Table 3.

NO entries can have information about dialect location in three parts of the entry: (1) the word form and etymology section that follows the entry head, (2) as sources for definitions, (3) as sources for usage examples. This paper focusses on source location for definitions. Within an entry, the senses are organized in a standard tree structure. The general structure is as follows: Some of the nodes (numbered senses) in the tree structure are just placeholders used to group the more finely graded senses and contain no definitions. For senses with definitions, the definitions are given as a semicolon separated list in the printed dictionary and as a list of definition fields in the database. Each definition can have one or more references to literary sources and/or references to places of the spoken sources.

As mentioned in Section 4.2, general countrywide spoken usage of a headword is not explicitly sourced. Table 2 shows that 48 percent of all entries have one or more speech sources. 51 percent of all entries have literary sources only, or lack any explicit source. Entries for the letter *c q w x z* could have been excluded, as these consonants are not used in Norwegian orthography except in imported vocabulary, but the number of entries for these letters is so small that the outcome would not have been affected.

Of all sense sections with definitions, 46 percent have location of source. Of the single definition fields, 36 percent have location of source. The smaller percentages for sense sections and definition fields make sense; most entries with a definition with location information will often have sense sections with definitions without such information, and many sense sections have a mixture.

Table 2: Number of entries, senses and single definition fields per letter and percentage of the respective totals

	Entries		**Senses**		**Definition fields**	
Letter	**Total**	**with location**	**Total**	**with location**	**Total**	**with location**
A	5 584	38 %	8 000	36 %	10 261	28 %
B	15 643	40 %	22 859	41 %	28 860	31 %
C	204	0 %	229	0 %	270	0 %
D	8 369	44 %	14 035	44 %	17 632	33 %
E	5 945	31 %	8 549	32 %	11 576	24 %
F	27 284	32 %	42 109	30 %	54 656	23 %
G	20 751	46 %	34 611	43 %	44 982	32 %
H	19 080	49 %	31 447	47 %	42 190	34 %
I	4 599	33 %	7 490	27 %	9 186	23 %
J	3 983	46 %	5 669	43 %	6 537	37 %
K	26 157	52 %	39 959	51 %	48 505	43 %
L	15 513	48 %	22 497	47 %	28 278	39 %
M	15 157	46 %	19 709	47 %	24 686	39 %
N	7 337	48 %	10 021	48 %	12 473	39 %
O	7 045	44 %	9 854	41 %	12 594	33 %
P	10 441	40 %	13 723	43 %	17 131	35 %
Q	13	0 %	16	0 %	19	0 %
R	13 233	55 %	18 756	57 %	24 016	46 %
S	47 933	58 %	72 637	56 %	92 202	44 %
T	16 751	57 %	25 898	52 %	33 723	41 %
U	6 953	50 %	10 958	44 %	15 259	32 %
V	11 723	55 %	17 968	49 %	23 709	38 %
W	48	0 %	48	0 %	53	0 %
X	20	0 %	23	0 %	24	0 %
Y	574	59 %	916	54 %	1 341	37 %
Z	39	0 %	40	0 %	44	0 %
Æ	339	44 %	520	40 %	690	31 %
Ø	1 118	60 %	1 663	54 %	2 186	41 %
Å	1 952	62 %	2 904	57 %	3 696	45 %
Total	**293 788**	**48 %**	**443 108**	**46 %**	**566 779**	**36 %**

We have also had a look at the relative frequencies of definitions with literary sources, spoken sources and both. Table 3 gives an overview. Entries without any references to spoken or printed sources are excluded from the table. Hence the totals are smaller than in Table 2.

Table 3 deals only with entries and definition fields that are explicitly sourced. Definitions that are generally valid are not included in the total, cf. Section 4.2. Distribution quantities for written sources and speech sources in entries and definitions balance each other, with some more explicit sourcing for speech than from written sources.

Most definition fields with location sources list few locations. 93 percent list from 1 to 9 locations, 7 percent list from 10 to 99 locations, while only 107 definition fields list more than 100 locations for one definition.

Table 3: Entries and definition fields with literary and location sources, grouped by source type. Distribution shown in percent of total number.

Entries					**Definition fields**			
Letter	**with loc. and/or lit.**	**with lit. only**	**both loc. and lit.**	**with loc. only**	**with loc. and/or lit.**	**with lit. only**	**with loc. and lit.**	**with loc. only**
A	3 427	38 %	22 %	40 %	4 495	37 %	15 %	48 %
B	10 265	39 %	18 %	43 %	13 788	35 %	12 %	52 %
C	27	100 %	0 %	0 %	29	100 %	0 %	0 %
D	5 804	37 %	22 %	41 %	8 670	32 %	15 %	53 %
E	3 505	47 %	21 %	32 %	4 837	43 %	15 %	42 %
F	17 931	51 %	19 %	29 %	23 496	47 %	15 %	38 %
G	15 543	39 %	23 %	38 %	22 206	36 %	18 %	46 %
H	14 999	38 %	26 %	36 %	22 056	36 %	20 %	44 %
I	3 135	51 %	19 %	30 %	4 171	50 %	15 %	35 %
J	3 313	44 %	24 %	32 %	4 364	44 %	19 %	37 %
K	20 828	35 %	25 %	40 %	31 006	32 %	19 %	49 %
L	11 905	37 %	26 %	37 %	17 115	36 %	20 %	44 %
M	11 107	38 %	25 %	37 %	14 833	36 %	20 %	44 %
N	5 252	33 %	27 %	40 %	7 152	31 %	21 %	47 %
O	4 934	38 %	25 %	38 %	6 674	39 %	19 %	43 %
P	7 243	42 %	18 %	39 %	9 777	38 %	14 %	48 %
Q	3	100 %	0 %	0 %	3	100 %	0 %	0 %
R	1 0103	28 %	27 %	45 %	14 604	25 %	21 %	54 %
S	37 069	25 %	27 %	47 %	53 702	24 %	21 %	56 %
T	13 097	27 %	27 %	46 %	18 298	25 %	21 %	54 %
U	5 017	30 %	28 %	42 %	6 808	29 %	23 %	48 %
V	9 025	28 %	29 %	43 %	12 217	27 %	23 %	50 %
W	22	100 %	0 %	0 %	23	100 %	0 %	0 %

X	11	100 %	0 %	0 %	11	100 %	0 %	0 %
Y	507	33 %	34 %	34 %	710	30 %	27 %	42 %
Z	18	100 %	0 %	0 %	20	100 %	0 %	0 %
Æ	274	45 %	30 %	25 %	371	42 %	23 %	35 %
Ø	903	26 %	29 %	45 %	1 211	25 %	24 %	51 %
Å	1 687	28 %	30 %	42 %	2 267	27 %	23 %	50 %
Total	**216 954**	**35 %**	**25 %**	**40 %**	**304 914**	**33 %**	**19 %**	**48 %**

7. Words and Senses From the Vernacular – a Sample

The number of definition fields with speech sources only is close to 150 000. As a sample, we have selected the definition fields sourced from the same locations as the 69 dialect dictionaries in the Dictionary Hotel, with 12 595 lines drawn from 11 183 different entries.

In the sample of definitions, grammatical markers and other abbreviations have been removed to facilitate sorting by first word in the definition proper. Headword form with POS will show the lexicon profile of the sample. A set of definitions will also reflect the definition format register for the dictionary in question, which in turn indicates type of word and sense.

Speech is assumed to differ from writing in using more verbs and in using fewer compounds, especially compound nouns. If this assumption is correct, definitions sourced from speech materials should be found under a group of headwords reflecting their speech origins.

The POS distribution in the sample is 17 % definitions for adjectives and adverbs, 64 % nouns and 15 % verbs. In NO as a whole, the distribution is different: the dictionary has 15 % adjectives and adverbs, 76 % nouns and 7 % verbs. In the two standard dictionaries for Bokmål and Nynorsk, entries for verbs are between 9 and 10 % of the total. The almost double number of verbs in the headword list of the sample is therefore significant.

NO distinguishes between compounds on one side and simple and derived word forms on the other. In NO 75 % of headwords are compounds, 25 % simple or derived headword forms. NO has 9 600 entries for multi-word expressions (MWE), corresponding to about 4 % of the total number. Word type distribution in the sample headwords is 22 % simple word forms, 25 % derived word forms, 49 % compound word forms, and 4 % MWE. The significant difference lies in the proportion of compounds: 75 % in NO as a whole, 49 % in the sample of definitions with speech sources only. It seems that simple or derived word forms play a much larger part in speech than in writing. The headwords of the sample definitions strengthen the general view of word types in speech versus writing.

Headword selection by informants and editors of dialect dictionaries tends to be weighted towards content POS. Function words (prepositions, conjunctions, subjunctions, determinatives etc.) are almost the same in Bokmål and Nynorsk, and common to all speakers of Norwegian. Function words are rarely mentioned in lexical documentation of vernaculars, while content words (nouns, adjectives and adverbs, verbs, MWEs) dominate. The tendency in the sample analysed here corresponds with what is thought to be important differences between the lexicon of speech and the lexicon of writing.

Another assumption concerns word origins, or etymology. Written language is taken to have more lexical items imported from other languages ("hard words"), whereas the speech lexicon is supposed to stay closer to the basic lexicon, learnt early and known to everyone in a language community. In Norway, imported lexical items have arrived from a number of languages, and the Greek and Latin influence via English (and formerly French) is strong in written language and the media. In contrast, there are few imported lexical items in the sample, either in headword simple forms or in derivation endings. Simple word forms are monosyllabic or disyllabic and have the phonological structure expected for Norwegian base forms. A majority of adjectives are derived; the dominant endings are *-al, -en, -ig, -leg, -sam, -ut*, all of which belong to the Norwegian standard repertory.

The formats of full definitions correspond to the POS of the headword. The sample shows no different use of definition formats from entries with mainly literary sources.

An NO definition can start with an extension in brackets and a grammatical marker before the definition itself. A full definition has a hypernym with additional distinguishing features. NO also uses synonym definitions where several words have an identical meaning. The less used word forms are defined with the more frequent word, which has a full definition. The total of synonym definitions in NO is 48 643, about 16 per cent of all sourced definitions fields. In the sample 4250 definitions are synonym definitions, 33,7 per cent, or a third. This is reasonable; the sample is drawn from a limited register of speech sources. It is noteworthy that the sample also holds the main definitions for 12 plant species. This shows that the NO written sources do not document all standard plant names for species – yet.

A full definition can be seen as a status marker for the headword in question; if the headword in a given sense has a full definition and also full synonyms, it is the most important word form in its peer group. A species name used in schoolbooks will have a full definition; less used names for the same species are often synonym defined, i.e., with a single word. The standard Norwegian word for the plant dandelion, "løvetann" is linked to 114 entries through synonym definitions. The web version of NO shows the list of synonyms under the heading "Tilvist frå" ('cross referenced from').

8. Meanings and Fields of Knowledge

The meanings and fields of knowledge reflected in the sample are too diverse for comparing proportions at the present stage, but some observations can be made.

The headwords in the sample refer to concrete objects rather than abstracts. The evidence comes from definition extensions and the use of adjectives qualifying hypernyms.

The initial extension in brackets uses the formula “(om …)” ‘concerning …’. It is found in ca. 200 sample definitions. In the extension, very few abstract nouns are found in what follows “om”. In NO as a whole, one finds f. i. “om framferd, handling, kjensle, mengde, regel” (‘concerning behaviour, action, emotion, quantity, rule’). Hypernyms for classes of concepts are also frequent, f. i. “om dyr, landskap, reiskap” (‘concerning animals, landscape, tools’). These are not found in the sample, with one exception “om arbeid” ‘about work’.

A large number of words refer to objects distinguished by their size, shape or any other measurable quality. Definitions starting with the adjective opposites *stor – liten, små* (‘large – small’)*, lang – kort, stutt* (‘long – short’)*, tunn – tjukk* (‘thin – thick’), *tung – lett* (‘heavy – light’) are frequent. The definitions refer to the physical qualities of objects or living beings.

Some topic areas come forward as important in the sample, that are less visible in NO as a whole. Such topics are landscape, its shape and usefulness; wood and stone as (pieces of) objects and material; and equipment for sorting, treating and transporting natural produce.

These sense sections can be very specific, and their communicative value requires context. An example:

«**steinslag** stein som har slegi mot bartre og grodd inn i treet»

‘**stoneblow** stone that has hit a pine tree and grown into the tree’

This is an important combination of headword and definition if you plan to fell trees in a steep and scree-exposed landscape, as felling a tree with «steinslag» could start a scree. Norway is a country where life is shaped by topography, and wood, especially timber, has been a major source of activity and income at least since the early Middle Ages. But there is no Norwegian terminological dictionary for wood and woodwork, except for an addendum in a book on log building (Strømshaug, 1997). At a guess, the terminology of wood is integrated into the Norwegian language to the point where it is not thought of as a separate field of expertise. The alternative possibility is negligence and omission.

Many definitions refer to persons and animals, and characterize appearance, behaviour and habits. The main impression from these definitions is that they represent assessment of qualities or usefulness rather than personal and moral judgments. Negative judgments can shine through, partly in defining language and partly in headword form. If a compound ends in -*fant* (masculine) ‘(young) man (without a settled home)’, the definition is unlikely to be complimentary. The sample, comprising 11 000 entries, has 38 headwords with the element *fant*; NO as a whole has 471 out of around 300 000 headwords, suggesting that this characterisation is more used in speech than in writing.

9. Conclusion

This article gives an overview of the way spoken source material is used and documented in *Norsk Ordbok* (NO). We also looked at the entries of NO via the source system in order to see if there are visible and measurable marks of origin separating information from speech sources from information from written sources. Significant differences are found in relation to word typology, distribution of POS, the use of definition formats and subject matter.

This paper is a preliminary investigation. Results suggest that more can be found, both in content and in method development. The major dictionaries with their wealth of linked headwords and definitions have a part to play in ensuring natural language a place in a digital future where both speech and writing need correct interpretation and use.

References

Aasen, I. (1850). *Ordbog over det Norske Folkesprog* (Dictionary of the Norwegian Vernacular).

Aasen, I. (1864). *Norsk Grammatik* (Norwegian Grammar). P. T. Mallings Forlagsboghandel.

Aasen, I. (1873). *Norsk Ordbog med dansk forklaring*. (Norwegian Dictionary with Explanatitons in Danish) Christiania. Mallings Boghandel.

About the Norwegian Dialect Synopsis (Målføresynopsisen). Retrieved May 27, 2024, from https://www.edd.uio.no/synops/work/synops_ext/Opplysningstekstar_eng.html

Askeland, A. (2020). *Serdu*: eit attersyn og eit nytt perspektiv (*You see*: a Review and a New Perspective). In O. Almenningen, O. Grønvik, L. S. Vikør, & D. Worren (Eds.), *Leksikografen frå Leksvika* (pp. 75–82). Novus forlag.

Dictionary of Finnish dialects. Retrieved January 30, 2024, from https://www.kotus.fi/en/dictionaries/dictionary_of_finnish_dialects

Grønvik, O. (2015). Language Documentation in a Globalised World. In J. Ragnar Hagland, & Å. Wetås (Eds.), *Ivar Aasen ute og heime – om moderne språkdokumentasjon etter Ivar Aasen*. (pp. 28–53). Skrifter 2015 nr 1.

Gundersen, H. (2016). *Redigeringshandbok for Norsk Ordbok 2014* (Editorial Guidelines for Norsk Ordbok 2014). Retrieved May 27, 2024, from http://no2014.uib.no/eNo/tekst/redigeringshandboka/redigeringshandboka.pdf

Hagen, K., & Vangsnes, Ø. A. (2023). LIA-korpusa – eldre talemålsopptak for norsk og samisk gjort tilgjengelege (The LIA Corpora – Older speech recordings of Norwegian and Sami made accessible). *Nordlyd* Vol. 47 No. 2 (2023). *Struktur, ideologi og mangfald*. https://septentrio.uit.no/index.php/nordlyd/article/view/7157/7836

Hellevik, A. (1956). *På skattegraving i eige mål* (Treasure Hunting in one's own Language). Syn & Segn. 1956 Vol. 62.

Johannessen, J. B., Priestley, J., Hagen, K., Åfarli, T. A, & Vangsnes, Ø. A. (2009). The Nordic Dialect Corpus – an Advanced Research Tool. In K. Jokinen, & B. Eckhard (Eds.), *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*. NEALT Proceedings Series Volume 4. https://aclanthology.org/W09-4612/

LIA Norsk v. 1.1. (Norwegian dialect corpus v. 1.1). Retrieved May 27, 2024, from https://www.nb.no/sprakbanken/en/resource-catalogue/oai-tekstlab-uio-no-lia-norsk/

Målføresynopsisen (Dialect Synopsis for the Norwegian Dialects). Retrieved May 27, 2024, from https://usd.uib.no/perl/search/search.cgi?appid=145&tabid=2165

Nes, O. (1986). *Norsk dialektbibliografi* (Norwegian Dialect Bibliography). Novus Forlag.

Netland, R., & Opsahl, T. (2024). Dialekt in *Store Norske Leksikon på snl.no.* Retrieved May 27, 2024, from https://snl.no/dialekt

Nordisk Dialektkorpus (NDK) (Nordic Dialect Corpus). Retrieved May 27, 2024, from https://tekstlab.uio.no/nota/scandiasyn/

Norsk dialektatlas (Norwegian Dialect Atlas) Kartene. Retrieved May 27, 2024, from https://usd.uib.no/perl/search/search.cgi?appid=210&tabid=2338

Norsk Ordbok. Ordbok over det norske folkemålet og det nynorske skriftmålet (*The Norwegian Dictionary.* Dictionary of the Norwegian vernacular and the Nynorsk written standard). 1966–2016. Vol. 1–12. Oslo. Det Norske Samlaget.

Norsk Ordbok. Retrieved May 27, 2024, from https://alfa.norsk-ordbok.no/

Norsk Ordbok AH 2005 (https://usd.uib.no/perl/search/search.cgi?appid=251&tabid=3581)

NRK-språket. Språkreglar (nynorsk) 2007. (National Broadcasting Language. Guidelines (Nynorsk)). Retrieved May 27, 2024, https://sprak.nrk.no/rettleiing/nynorsk/

Ordbokshotellet (The Dictionary Hotel). Retrieved May 27, 2024, from https://usd.uib.no/perl/search/search.cgi?appid=118&tabid=1777

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* (pp. 159–172). Znanstvena založba Filozofske fakultete.

Ross, H. (1895). *Norsk Ordbog*. Tillæg til "Norsk Ordbog" af Ivar Aasen (Norwegian Dictionary. Supplement to "Norsk Ordbog" by Ivar Aasen). Kristiania.

Sandøy, H. (1985). *Norsk dialektkunnskap* (Norwegian Dialectology). Novus Forlag.

Sandøy, H. (1993). *Talemål* (Spoken Language). Novus Forlag.

Setelarkivet til Norsk Ordbok (The slip archive of Norsk Ordbok). Retrieved May 27, 2024, from https://usd.uib.no/perl/search/search.cgi?appid=8&tabid=436

Skard, S. (1932). Norsk Ordbok. Historie – plan – arbeidsskipnad. (Norsk Ordbok. History – Plan - Organisation). In H. U. Karlsen, L. S. Vikør, & Å. Wetås (Eds.), *Livet er æve, og evig er ordet: Norsk ordbok 1930–2016* (pp. 40–67). Det Norske Samlaget.

Strømshaug, K. (1997). *Lafting. Emne og omgangsmåte* (Log Building. Materials and Handling). Landbruksforlaget.

Tier, V.d., & Keymeulen, J.v. (2010). Towards the completion of the Dictionary of the Flemish Dialects. Ghent University (Belgium). In A. Dykstra, & T. Schoonheim (Eds.), *Proceedings of the XIV EURALEX International Congress. 6–10 July 2010.* Fryske Akademy – Afûk.

Wright, J. H. (Ed.) (1981). The English Dialect Dictionary. Vol. I–VI. Oxford University Press. Retrieved January 20, 2024, from https://www.uibk.ac.at/en/disc/blog/english-dialect-dictionary-online/

Aasen, I. (1848). *Det Norske Folkesprogs Grammatik* (Grammar of the Norwegian Vernacular). Det Kongelige Norske Videnskabers Selskab. Werner & Comp.

Contact information

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

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

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.