Comparing Orthographies in Space and Time through Lexicographic Resources

Christian-Emil Smith Ore[1], Oddrun Grønvik[2]
[1]University of Oslo, [2]University of Bergen
E-mail: c.e.s.ore@iln.uio.no, Oddrun.Gronvik@uib.no

Abstract

Many languages require an improved factual basis to facilitate computer-supported analysis of language variation and diachronic change. The material collections for the scholarly dictionaries of Norway serve as a platform for exploring the development and variation of Bokmål, the Norwegian written standard derived from Danish and modified towards the Norwegian vernacular through orthographic reforms that took place from 1901 to 2005. The development of modern Bokmål through usage should be analyzed by comparing corpora from different periods, lemmatized according to the then current orthography. This means building full form registers from time-bound orthographies. This plan is in process through digitizing orthographic dictionaries for Bokmål. The dictionaries are coordinated through the Dictionary Hotel, the electronic repository for retro digitized dictionaries and dialect collections at the Norwegian Language Collections, Bergen. At the lexical item level Bokmål and Nynorsk resources are coordinated through the Meta Dictionary, an electronic registry for the Norwegian lexicon. A common entry requires full identity in one headword form plus part-of-speech (POS). Preliminary results identify a core vocabulary for Bokmål of 6,900 lexical items, unchanged since 1938. More than 75,000 Meta Dictionary entries have a common identical form plus POS for Bokmål and Nynorsk. These numbers will increase when the Bokmål additions to the Meta Dictionary are quality controlled.

Keywords: dictionary, lexical item, full form register, computer assisted language analysis, corpus, lemmatizer, synchronic variation, diachronic change

1 Background, Problem, Task

The motivation behind the great historical dictionaries of the 19th and 20th centuries was to provide a full scholarly description of the single lexical item[1] through the dictionary entry, and through the entirety of entries, a scholarly description of whole languages through time. Their purpose was and is to replace speculation with facts, a task which overall has been completed successfully for the languages covered. The scholarly dictionaries have provided their language communities with reliable lists of lexical items, traced and identified through time, their formal qualities identified, their senses documented and ordered.

These laborious projects have been facilitated in the last decades through digitization, and many of them have been completed. In the process several lexicographical resources have been created, many of which have wider uses than dictionary production. Corpora, full form registers, spelling programs and syntactic taggers all stem from empirical linguistics and depend on lexicography to make sense of it by linking the sign – the lexical item – with sense description.

However, many languages still require a historical and comparative description to be properly documented. A case in point is Bokmål, the written standard of Norwegian which is used by the majority of the population.

1 The term “lexical item” is used as defined by Atkins and Rundell (2008:163 ff).

In a European context, Norwegian is a special case. Spoken Norwegian must today be considered one language, represented through a variety of non-standard dialects, and whether a spoken standard exists is an issue of debate. There are two written standards, Bokmål and Nynorsk, one stemming from Danish as spoken in Norway, the other from a comparative analysis of 19th century Norwegian dialects. The written standards have been modified, and in many respects brought closer to one another, through a series of orthographic reforms (from 1901 until 2012). The orthographic changes have affected all aspects of language, from sign inventory (*aa* > *å* in 1917) to syntax. The history of orthographic reform is well documented. There is however limited agreement on how the different orthographic reforms have affected actual usage, especially concerning the majority language, Bokmål. The discussion on how close or separate the two standards are, and how their separate and joint development has expressed itself in usage, remains ideologically slanted and easily gets emotional (cf. Vikør 2001: 53 f.).

Moving this discussion forward requires an improved factual base. The only way of finding out exactly how present-day Norwegian, especially Bokmål, has developed, is to analyze and compare Bokmål corpora from different periods in the 20th century. This analysis requires lemmatizers providing exact coverage of the orthography for both standard languages during the selected periods. Corpora, and the comparison of corpora, will show usage and changes in this. The orthography dictionaries compiled to put the orthographic reforms into practice show what the language reformers believed and hoped for. If lemmatizers are made on the basis of contemporaneous spellers, they will identify the vocabulary used in the corpora, but they will also show non-occurrence and divergence between the corpora and the orthography dictionaries, and ultimately contribute to our general knowledge of standardization processes.

The issue of improving the factual base for exploring the development of Bokmål has gained relevance because *Bokmålsordboka* and *Nynorskordboka*, the two Norwegian general-purpose monolingual dictionaries, are to be revised through a five-year project at the University of Bergen 2018 - 2023.[2] The dictionaries were first edited 1974-1986, and content revision has since then been minimal. Both dictionaries qualify as scholarly dictionaries, with evidence from the Norwegian Language Collections. *Bokmålsordboka* will be expanded to match *Nynorskordboka*, from roughly 65,000 to 100,000 entries. Corpus-based linguistic studies covering the whole of the 20th century will be an important support discipline. The entries of the dictionaries will be generated from the material index of the Language Collections and the Meta Dictionary (Stortingsmelding 1 (2017–2018)), as also done for the entries of *Norsk Ordbok* (Grønvik & Ore 2013: 254).

The largest historical corpus of Norwegian text is being built by the National Library. The Library is in the process of digitizing its entire collection of text printed in Norway. The resulting corpus at present comprises more than 1 million volumes of books, newspapers and so on, and currently has more than 50 billion tokens. All texts are linked to the National Bibliography with its metadata. The digitization process consists of scanning and automatic OCR. However, the large changes in Norwegian orthography in the 20th century cause problems for the OCR-process and the lemmatization. The only full form registries with broad coverage, the Norwegian Word Banks (Bokmål & Nynorsk), have a relatively short temporal coverage. Both word banks document the current orthography from about 1990 and until today, while the text to be analyzed goes back to the 18th century. This leads to mismatches, like search finds from about 1800 for the base form “telephone”. For reliable results, lemmatizers for Norwegian in previous orthographies are needed for both the written standards of the language. The question is how best to develop them.

2 The Norwegian Language Collections, comprising lexicography, dialectology and onomastics, were moved from the University of Oslo to the University of Bergen in 2016, and are now a unit at the University Library.

2 Orthographies, Spellers and Digitization

There are two possible approaches to isolating orthographies within given periods: 1) Analyze a (selected, dated) corpus in terms of frequencies and distribution, and propose a list of lexical items on the basis of this analysis. 2) Start with the records of the orthographic standard of the selected period, as found in authorized (school) dictionaries, and then use grammars from the same period as a source for establishing full form schemas.

Either approach involves a lot of work, and both are useful; the corpus analysis will give the register and frequencies of forms to be identified (including plenty of homographs), while full form registers based on spellers and grammars will represent what was thought essential vocabulary at the time, in the form valid at the time. Without a valid full form register, automatic identification of word forms will be impossible.

Bokmålsordboka and *Nynorskordboka*, published together in 1986, were the first general purpose defining dictionaries for modern Norwegian. Earlier normative information on orthography is found in printed spellers, mostly for school, and in the official reports on suggested orthographic reforms. The most important reports cover the orthographic reforms of 1917, 1938, 1959, 2005 (Bokmål) and 2012 (Nynorsk). These are the natural pivot points for comparing before and after. For Nynorsk, ample and dated materials are present in the lexical index of the Language Collections, the Meta Dictionary, cf. section 3.1 ff. Additions to the Meta Dictionary at present therefore focus on covering the development of Bokmål.

2.1 Orthographic dictionaries and school spellers

Orthographic dictionaries are made to inform the public about the orthography of headwords and their inflected forms. In the Norwegian printed orthographic dictionaries and spellers of the 20th century, this information is given in a very compressed form. The first headword in base form is given in full, while derivations can be abbreviated to the word ending. Rows of compounds are often nested. Additional information, such as inflected forms, POS, sense, usage, multi-word expressions (MWEs), etymology, usage conventions are omitted, unless something more than the headword is needed to identify the headword to the user. Instances of all of the categories above can be found in school spellers, most often in an abbreviated form.

Judging by the look of the spellers, the user must be assumed to be an experienced mother tongue user, presumably a teacher, and getting children to understand and interpret the spellers correctly must have been part of their work.

The Norwegian school spellers are the work of individual compilers, often philologists with teaching experience. Unlike Sweden and Denmark, no central agency provided an authoritative orthographic dictionary. Each of the frequent orthographic reforms in the 20th century was based on recommendations from the Ministry of Church and School Affairs. The underlying reports and suggestions discuss principles and give examples. (cf. Den nye rettskrivning 1917: 24-25). Editors of school spellers were then tasked with fleshing out the recommendations and examples as best they could, by listing lexical items plus essential additional information. A typical example of a school speller is shown in Figure 1. School spellers needed a stamp of approval from the Ministry of Education before they could be used in school.

The first guidelines for orthographic dictionaries for school use were set up as house rules by the Norwegian Language Committee in the late 1960s. School spellers from before 1980 are geared towards saving space, omitting information that can be implied or assumed to be known by an adult literate

person. Some guidance concerning lexicographic conventions is normally found in the front matter, but no one would call these spellers explicit or learner-friendly. Conventions are expressed differently from speller to speller, and consistency levels vary.

2.2 Retro digitizing orthographic dictionaries

Retro digitizing 20th century orthographic dictionaries for Norwegian (Bokmål and Nynorsk) almost always involves interpretation doubts, because the original text is highly compressed and the punctuation ambiguous.

The sample shown below in Figure 1 is typical. A striking feature is the use of the typographical marker "/". This means that the string in front of the slash can be added to following strings starting with a hyphen, and the result will be a meaningful word form. The line "akt/e; -else, -en" is to be read as a list of three lexical items: *akt*, *akte*, *aktelse*. But a lexical item with a base form *akten* does not exist. The information that "-en" in this case is to be read as information on noun inflection for the preceding lexical item *aktelse*, indicating the masculine gender, can be found in the front matter. For *akt* (noun) and *akte* (verb) no POS information is given.

The slash does not mean that the string in front represents a lexical item. The line "aksj/e, -en, -onær" does not claim that there is a lexical item *aksj* – there is not. It means that "aksj/" can be added to "-onær" to render *aksjonær*. The "-en" in between is meant to say that *aksje* is a noun with the masculine gender.


ſ] A [alv

A

à (fem à seks)
abbed, -en
abbedi/sse, -en l. -a
abborr/en = åbor
abcbok, -a
abdi/kasjon, -sere
aber, en, et
abessin/ier, -sk
abnorm
abonne/nt, -ment
abonnere, -te
abrupt
absint

aeroplan, -er
affeksjon
affekt/asjon, -ert
affinitet
affisere, -te
affære, en
afgan/er, -sk
aften/er, -sang
aftens/bord, -mat
agat, -en
agè, holde i a.
agere, -te
agg/et (nag)
aggregat
aggressiv
agio

aksent
aksept/ere, -ert
aksidens
aksiom, -et
aksise, -en
aksj/e, -en, -onær
aksjon, -en
akt/e; -else, -en
akter, -skott, -speil, -ut
akten/for, -om
-aktig
aktrise, -en l. -a, -r
aktiv/itet
aktiv/um, fl. -a
aktor, -en, -er

allegori/sk
alle/gretto, -gro
allehelgensdag
allehånde
allemanns/eie, -venn
aller best osv.
allerede
alle sammen
alle slags, allsl.
allesteds, -nær-værende
alle tider, vegne
all/fader, -farvei l. -farveg
alli/anse, -ert

Figure 1 Eitrem 1939. The beginning of the section for the letter *a*.

Table 1 shows a simple instance of extracting lexical items from a school speller. Three lines of text contain five base forms of six different lexical items, one of them a noun with two genders.

Table 1. Example from Figure 1 of headword extraction and added POS (Eitrem 1939).

Text	Inflection	Extracted lexical item	Added POS	Sense
abdi/kasjon, -sere		abdikasjon	noun, masculine	abdication
abdi/kasjon, -sere		abdisere	verb	abdicate
aber	en, et	aber	noun, masculine or neuter	disadvantage
abessin/ier, -sk		abessinier	noun, masculine	Abyssinian
abessinsk		abessinsk	adjective	Abyssinian
abessinsk		abessinsk	noun, gender masculine	Abyssinian

In the sample above, all base forms lack explicit POS, and one, *abessinsk*, can double for two lexical items. The addition of a base form in the current orthography and POS for all represents a reasonable interpretation, but it is still an interpretation. Therefore, it is important to present the text itself in context in the electronic version, with the preceding and following entries shown. Users need to be able to compare what was printed with what is claimed to be a true interpretation.

From the point of view of digitization three concerns emerge: 1) to preserve the text of the original document as carefully as possible, so that it can be presented as it was; 2) to extract electronically the full register of lexical items with such supporting information as there is; 3) to supply essential and missing information from reliable contemporary sources.

These aims necessitate careful encoding and tagging, based on the conventions of the original document.

3 The case of *Norsk rettskrivningsordbok* versus the School Spellers

The purpose of discussing the inclusion of *Norsk rettskrivningsordbok* in the Dictionary Hotel is to highlight the process of computer assisted tagging of a dictionary.

Norsk rettskrivningsordbok (Sverdrup 1940) is the largest orthographic dictionary attempted for Norwegian Bokmål. It still exists under the title *Tanums store rettskrivningsordbok* and has been through numerous expansions and revisions. The original compiler was Jakob Sverdrup (1881–1938), professor of Germanic philology at Det Kongelige Fredriks Universitet (now the University of Oslo). In the introduction to *Norsk Rettskrivningsordbok*, it is stated that Sverdrup and his co-editor, Sandvei, drew the materials from the lexicographic excerpt collections in existence at the University, but also from catalogues and registers of all sorts, such as goods registers from the major retailers.

The orthography reform of 1938 was a major one, aimed at bringing Bokmål and Nynorsk closer together. It caused lasting excitement and resentment, especially among Bokmål users, since it went far in giving word forms from the Norwegian vernacular equal status with the traditional Danish-based forms. The vernacular forms were often identical with existing Nynorsk ones, but the influence of orthophonic ideals was stronger for Bokmål than for Nynorsk. The 1938 reform therefore includes forms like *selle* verb ‘to sell’ (equal to what is found in Swedish standard language) in addition to the traditional Bokmål form *selge*.

With more than 175,000 entries, Sverdrup’s orthographic dictionary is the most comprehensive documentation of the 1938 orthographic reform for Bokmål, and therefore a valuable measuring point for assessing the influence of orthographic reform on usage. Using it as a basis for a lemmatizer will provide a before-and-after separation mark for Bokmål text. It is also probable that a high proportion of the word forms never will be found in any corpus.

Table 2. Text sample from Sverdrup (1940). Paragraphs numbered for reference.

1	**adalhending**, -en; -er, -ene (helrim).
2	**adalin** (sovemiddel).
3	**adam** (fra Bibelens Adam); den gamle Adam; i Adams drakt.
4	**adamitter** pl. (sekt); adamittisk adj., n. -.
5	**adams˽barn**, ˽drakt, ˽eple, ˽fiken, ˽hjerte, ˽natur, ˽slekt, ˽sønn, ˽tre, ˽ætt.
6	**adapsjon**, -en; -er, -ene (tilpasning); adapsjons˽evne, ˽form o. fl., adaptere, -te, -t (tilpasse).

Table 2 shows a sample of the transcribed text from Sverdrup (1940), set up in table format here for reference. This text has all the categories expected in a defining dictionary, but the presentation

is far more compact and the ordering unpredictable, as only the categories deemed essential for headword identification are included in each entry. First headwords are set in bold. Meaning (1, 2, 4, and 6) and etymological information (3) both appear in parentheses. POS information appears as abbreviated inflection forms (1, 4, and 6) or an abbreviation (4). MWEs are rendered in plain text (3), after semicolons. Derived headwords in base form are rendered in plain text (4, 6). Compounds are nested (5, 6).

The four school spellers included in the Dictionary Hotel and linked to the Meta Dictionary have from 13,000 to 24,000 entries. These are basic school spellers, and the vocabulary covered can therefore be expected to overlap, and also give guidance as to what school authorities saw as the central vocabulary from 1938 to 1986. With the exception of the initial headword in each paragraph, the organization of information is as unpredictable, and the range of categories as wide, as in Sverdrup (1940).

4 The Dictionary Hotel

The Dictionary Hotel is the central repository of the Language Collections for searchable electronic dictionaries. It was created in 2005 as a part of the databases with background material for the Norsk Ordbok (Norwegian Dictionary) project, and linked to the material index Metaordboka (Meta Dictionary). The purpose was to create a common framework for retro-digitized dictionaries. It is an electronic library for mostly retro-digitized dictionaries, glossaries and other collections of lexical information, and is equipped with a portal for searching them in parallel and showing aligned search results (Tvedt et al. 2007). The present contents are 60 dialect dictionaries and word collections, plus a couple of large dictionaries. The school spellers and Sverdrup (1940) are stored in the Dictionary Hotel. The collection is expanded when possible, depending on capacity. The Dictionary Hotel is analogous to the German Wörterbuchnetz (Wörterbuchnetz 2018); see also Moulin and Nyhan (2014) for a discussion of this.

Each item in this library is stored as an xml-document with inline mark-up. The major purpose is to facilitate the study of lexical information. Therefore, the texts are encoded according to TEI dictionary format, which provides for dividing each text into a series of smaller text chunks or entries. For each entry, one or more headwords are marked. These are used as linking points to the common index the Meta Dictionary (see section 5 below).

Each original dictionary or collection is considered a unique document in its own right. Standardization levels vary, and all have their editorial idiosyncrasies. Therefore, every entry has an added layer containing one or more standard base forms with POS which is used in linking the individual dictionary to the corresponding Meta Dictionary entry. The standard language used for the dialect materials is Nynorsk. This is in accordance with language documentation practice in Norwegian philology since dialect studies started in the 19th century. It should be mentioned that different layers of interpretation are somewhat tricky to express in inline encoding. For an alternative approach, see Bouda and Cysouw (2012).

Every document in the Dictionary Hotel has its standard language set to either Bokmål or Nynorsk. In linking a document to the Meta Dictionary, this means that entries in a dictionary marked as "Bokmål" will be linked to a Meta Dictionary entry with an identical base form plus POS marked as "Bokmål", and the corresponding procedure for documents set as Nynorsk documents.

4.1 Encoding decoded information

From an everyday point of view, encoding dictionary texts may seem a straightforward task. But the requirement is scholarly reproducible results based on the application of transparent methods.

Therefore, each dictionary must be treated as a unique document, analyzed and given a mark-up documenting the analyzer's understanding of the author's intentions and a decoding of the information found in text. The text itself must not be corrupted in the process.

Dictionaries can be very complex texts with plenty of ambiguities and lacunae. As shown above, this is true even of school spellers, a this genre shows a wide variation. In the introduction to the encoding of dictionaries in the TEI P5 guidelines (TEI 2018) two important issues are highlighted:

First, because the structure of dictionary entries varies widely both among and within dictionaries, the simplest way for an encoding scheme to accommodate the entire range of structures actually encountered is to allow virtually any element to appear virtually anywhere in a dictionary entry. It is clear, however, that strong and consistent structural principles do govern the vast majority of conventional dictionaries. […]

Second, since so much of the information in printed dictionaries is implicit or highly compressed, their encoding requires clear thought about whether it is to capture the precise typographic form of the source text or the underlying structure of the information it presents. Since both of these views of the dictionary may be of interest, it proves necessary to develop methods of recording both, and of recording the interrelationship between them as well. […] (TEI P5 Guidelines, Chapter 9 Dictionaries)

Both aspects are important in the Dictionary Hotel. Ideally, our goal is to preserve the typographic form at least to a degree that documents the analyzer's interpretation of the contents, that is, the information about the lexical items.

Some simplification of the original typography or character inventory may be necessary, for example removing word stress markers or changing the typeface from Black Letters (Fractur) to Antiqua, if the typeface is irrelevant to the entry structure. The encoded text can then be used to reproduce a text with a layout close to the original. In this way we follow the basic principles for modern electronic text philology, a practice recommended for all retro digitization of dictionaries.

4.2 Preparations for encoding the text

The encoding and proofreading of the text is important in preparing for text analysis and tagging The following is a run-through of the procedure established for the Dictionary Hotel:

(1) Consider the encoding a preparation for structural analysis to get the tagging as correct as possible, while preserving the original text. (2) Check the character set. A dialect dictionary may contain characters not found in Unicode to represent particular sounds (the Norwegian sound transcription alphabet Norvegia has no ISO standard). If a character of this kind is found then it must be given an established substitute for the encoding, preferably an entity. In other cases, a particular character may be used both to indicate a sound quality and a change of category in the text. This is easily interpreted by a human reader, but not by a computer program, and should be handled in connection with the encoding. (3) Decide on a system for handling the relationship between typography and field content. Typography can be over-specific or ambiguous or both. Italics for both cross references and deviant dialect forms are an example of a common occurrence. (4) Encode paragraphs as continuous strings, do not try to reproduce the line breaks of the printed page. (5) Proofread meticulously, as a missing character or space may throw the automatic tagging off course and cause failure to pick up embedded lexical items or cause strings of compounds to come out with the wrong first part.

4.3 Text analysis and mark-up

The most interesting step comes when the encoding is done. Analyzing the text contents: what constitutes a headword, POS, inflection, description of meaning or usage examples? In dictionaries made

by scholars the formats tend to be well defined and based on the scheme used in (old, printed) Latin dictionaries. As in these dictionaries, a head word's POS markers are often given indirectly by listing (often abbreviated) inflected forms. In spellers this practice tends to be the rule.

The dictionaries and glossaries which are candidates for the Dictionary Hotel comprise thousands of entries. A manual mark-up is therefore impossible. The mark-up is done by a script developed in an iterative process: finding general patterns, analyzing the results by the use of standard KWIC-tools, refining the script and adding exceptions until the result is satisfactory; see Christmann (2001) for a description of the similar process behind the digital *Grimm's Deutsches Wörterbuch*. Printed dictionaries are idiosyncratic, and the scholarly and philological requirement of respect for the original text makes it doubtful that there is much to gain from trying to develop an analyzer based on deep learning (that is, the use of a neural network). Depending on the required level of analysis, the process may consume several person-weeks. The resulting script documents the analysis and should be kept for future use and consultation. Manual encoding by search and replace in a text editor is not recommended. The process corresponds to the construction of a hand-tagged training corpus and the development of a statistically-based corpus analyzer. The main difference is that one has to carry out the process for each dictionary.

The output from the analyzing script is an xml-encoded text where the result of a scholarly interpretation is represented by the mark-up. In a well-structured dictionary the entry format normally has a standard structure: one or more base forms acting as headwords, POS, spelling variants and the hierarchy of definition and citations. All the information is already present in the original text, though the presentation may require a fair amount of interpretation, cf. Figure 1 and Tables 1 and 2 above. There are, however, some challenges frequently encountered in dictionaries, and especially school spellers for Germanic languages, such as Norwegian:

Entry organization is not entirely alphabetical: the system of creating compounds causes editors to organize dictionaries into nests, with an introductory entry for the initial part, and then a series of entries for derivations and compounds.

Finding the POS: Since the last part of a compound is usually an independent word with a separate entry, information about inflection and POS is often omitted in entries for compounds. For spellers these minimal compound entries are the norm. In such cases, the complete compound is the first headword of the nest plus the relevant second part, and these can be stored together in an attribute. The POS is more problematic, since one has to find a single headword corresponding to the second part, which in many cases is not possible due to the high frequency of compounds constructed from more than two parts, or homograph candidates for the second part of the compound.

As a consequence, there will at the end always be a significant number of identified headwords without POS information. In order to link a tagged dictionary to the Meta Dictionary, POS information must be added to every headword semi-automatically or manually. This information is not a part of the original dictionary. To keep the original and added information separate, an extra level is added. The result will be a "dictionary" where all headwords will have POS information either from the original analyzed dictionary or from other sources. The original analysis is put into a dictScrap element and can be extracted and displayed by the application of, say, xslt-transformations.

(1) <entry xml:id="SverdrNR_orig000031" n="nest_no_1"><form><orth>abelmoskus</orth><-gramGrp><pos> m</pos></gramGrp></form><dictScrap><form type="headword" orig="a-belmoskus">abelmoskus</form> bot.</dictScrap></entry>

5 Linking Base Forms and Materials

The access point of scholarly language analysis is always the word form in context, whether the raw materials are registered as sound or text. Registered raw materials must be easily accessible by solid criteria and have adequate metadata, including source information. The only functional linking of base forms on the one hand with materials on the other hand is to organize both around the lexical item as a fixed point. The lexical item has to have an established identity, and there must be sufficient materials to prove POS, inflection and at least one established sense. This information is found in the major scholarly dictionaries, and also in the traditional well-ordered collections underlying them.

In both Bokmål and Nynorsk, a lexical item can have more than one base form within a given orthography, and is almost certain to have more than one if the time span covered by the materials is long enough – something which is true of any language. The written and spoken expressions of the base forms attached to one particular lexical item can also vary considerably synchronically (through different dialect forms) and in the written standard (through orthographic variants).

In order to compare two closely related orthographies, an index is needed which allows alignment of the base forms of each orthography for the same lexical item.

5.1 The Meta Dictionary – headword forms and POS

For Norwegian, there is one language tool particularly suitable for indexing Bokmål and Nynorsk in parallel – the Meta Dictionary. This is an electronic registry for the Norwegian lexicon which allows linking lexicographical evidence to several base forms through one node, the Meta Dictionary entry, representing the lexical item. This node or entry may contain additional information about the lexical item, and in this way the Meta Dictionary is different from the Dictionary Hotel or the German Wörterbuchnetz (see also Ore 2000 and Ore & Ore 2010 for a shorter description in English). The Meta Dictionary format has also been adapted for the new editing and publication system for the Danish Dictionary of Old Norse (*Ordbog over det Norrøne Prosasprog*, ONP), see the related discussion in Johannsson and Battista (2014).

Since the Meta Dictionary indexes materials according to the headword form of the lexical item, the POS register is limited to forms found in headwords, i.e. the base forms of nouns, adjectives, adverbs and verbs. Nouns can be assigned gender and can have the singular or plural form. Adjectives and adverbs can be marked with degree (positive, comparative, superlative). Verbs are listed in the infinitive form. All other POS are single forms in Norwegian. This means that all pronouns have separate entries in the Meta Dictionary, irrespective of their syntactic function.

A modern language index organized around the lexical item must also handle categories with limited recognition in the lexicological literature. The written standard vocabulary contains abbreviations and symbols, and these are POS-marked as such. Prefixes are divided into two groups in Norwegian lexicography, the preformatives marked “pref” and those that represent the first part of compounds (including joining infix) marked “føreledd” ‘first section’. The Meta Dictionary also has entries for propria marked as such. The most commonly used MWEs found as headwords in orthographic dictionaries also have Meta Dictionary entries with the POS depending on syntactic function.

5.2 The Meta Dictionary entry

The entry head allows indexing attached materials with (1) standard language (Bokmål or Nynorsk); (2) several base forms per language form; (3) standardization status for each base form; (4) start and end date for each status; (5) POS; (6) segmentation. This entry format allows for building a diachronic

base form register by linking lexical resources and indexing them per entry for one or both of Bokmål and Nynorsk. Base forms which are or have been standard forms are used as entry headwords, while the sources express the range of forms found in spoken and written language over time. The base form schema in the Meta Dictionary is close to the format used in the Word Bank entries, but the Meta Dictionary has no information on inflection, or rules for generating full form schemas.

The Meta Dictionary entry body consists of references – hyperlinks – to various source databases. Most of these show usage examples or give definitions, but there are also materials only showing occurrence (in a given place or at a given time), pronunciation, POS, word formation potential (derivations, compounds) and so on. A typical Meta Dictionary entry is shown in Figure 2 below. The left part shows the list of links to the materials, with different icons for different source types. The middle part shows the entry tree, with segmented forms. To the right the schema for each base form is shown, with orthographic form, language choice, and status in the orthography, with start and end dates.

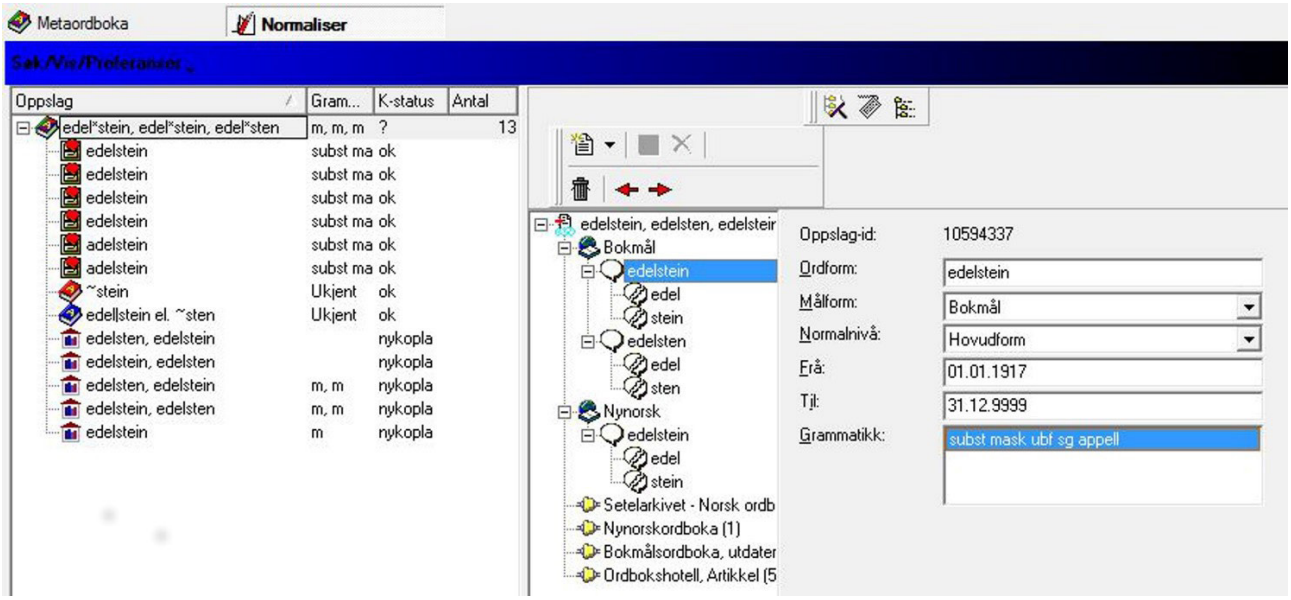

Figure 2. The Meta Dictionary entry, as shown in the editor, for the noun *edelsten* el *edelstein* m (Bokmål), *edelstein* (Nynorsk) 'jewel'.

The orthographic differences between Bokmål and Nynorsk may consist of one character only, but that one character is important. The alignment criterion is therefore full identity in base form and POS. Since the Meta Dictionary has a time scale, it is possible to show the transition from one orthographic form to another, cf. Figure 3. The Meta Dictionary is available and freely searchable on the web.

sjåfør m (Bokmål, 1917-)
chauffeur m (Bokmål, 1901-1917)
sjåfør m (Nynorsk, 1917-)
chauffeur m (Nynorsk, 1873-1917)

Figure 3. The Meta Dictionary entry for *sjåfør*, noun, masculine, with the previously valid form *chauffeur*, plus information about written standard and year.

5.3 Meta Dictionary Contents

The Meta Dictionary coordinates materials in standard language from different periods and transcriptions of spoken material. Nynorsk at present has a better coverage than Bokmål. Primary sources of usage for Nynorsk comprise excerpts (text, image) with metadata, or lines from corpus concordances. Secondary sources comprise electronic versions of scholarly dictionaries for Nynorsk, including major works covering the orthographies of 1873, 1917, 1938 and the current one of 2012. These dictionaries have from 40,000 to 300,000 entries, and the orthographies for Nynorsk lexical items are therefore reasonably well covered. Primary sources for Bokmål at present comprise a neologism archive (*Nyordsarkivet*, excerpts from a range of sources, ca 1970 – 1995). The secondary sources are dictionaries and spellers, ranging in size from ca 13,000 to 180,000 entries and covering the orthographies of 1938, 1959 and 2005. Many of the dictionaries and spellers are collected in the Dictionary Hotel (see section 4 above).

The Meta Dictionary was initially created as an index for the Nynorsk language collections, linking standard Nynorsk and dialect information from 1600 until today (Ore 2000). The purpose was to assist the editing of *Norsk Ordbok* (completed 2016). Through the Meta Dictionary, the Norwegian Language Collections for Nynorsk and Norwegian dialects were indexed with one base form in the Nynorsk 1938 orthography with traditional forms, plus POS. Homograph separation beyond base form plus POS has always been possible, but was not attempted in the Meta Dictionary while *Norsk Ordbok* was in production.

The Meta Dictionary is now used as a tool of coordination between Bokmål and Nynorsk. Coordination at the level of the lexical item is a long step forward towards solid ground for analysis and comparison of Bokmål and Nynorsk at different times during the 20th century and up to the present.

Linkage of a resource to the Meta Dictionary is automated; if the new entry in the resource finds a Meta Dictionary entry with the same base form and POS, it gets linked, if not it gets a new Meta Dictionary entry. The status system in the Meta Dictionary tells the human moderator – a trained lexicographer – where changes have been made, and changes are quality checked before being approved with or without adjustment.

6 Preliminary results

The Bokmål materials so far linked to the Meta Dictionary are limited, but some findings are nevertheless worthy of note.

6.1 The correlation between Bokmål and Nynorsk in the Meta Dictionary

The Bokmål additions to the Meta Dictionary have been made in the last 18 months, and there has not been time for much manual alignment, as Sverdrup (1940) is a fairly recent addition. But there are some interesting figures. Table 3 below shows contents in the Meta Dictionary after the uploading of Sverdrup 1940.

Table 3. Results after adding Sverdrup 1940 to the Dictionary Hotel, with linkage to the Meta Dictionary.

		Bokmål	Nynorsk
1	Entries in the Meta Dictionary per language	361 006	545 766
2	Unique headwords (base forms)	359 159	545 862
3	Unique combinations of headword plus POS	371 351	557 627
4	Entries with headwords in Bokmål or Nynorsk only	258 509	443 269
5	Entries with headwords in both Bokmål and Nynorsk	102 497	102 497
6	Entries with more than one headword per language	10 273	11 720

The Meta Dictionary is dynamic. Since the alignment criteria for automatic attachment to existing entries are strict (see section 5.3 the end), there will be changes as the manual alignment process gets under way. The numbers of lines 1 and 4 will decrease a little as minor divergencies get sorted; the numbers of lines 5 and 6 will increase.

The number of unique headwords (line 2) is higher than the number of entries (line 1) for both Bokmål and Nynorsk, because some entries have more than one headword (cf. line 6). The most striking piece of information is found in row 5 – between 1938 and today, which shows that more than 75,000 Meta Dictionary entries have one or more head word forms that are identical for Bokmål and Nynorsk, which is counter-intuitive to many Norwegians. This alignment does not cover inflection morphology. To some extent this sameness between headword forms has been reduced after 1980, but there are still a high number of examples, almost twice the overlap between *Bokmålsordboka* and *Nynorskordboka*.

Table 4. Overlap between the Bokmål school spellers in the Dictionary Hotel, and overlap with *Bokmålsordboka* and Sverdrup 1940.

	Sources	Number of entries	Number of (shared) headwords	Headwords found in Lexicographic Bokmål Corpus (1985 – 2005)	Hits in% of number of (shared) headwords
1	School speller 1939 I	13,046	13,119	10,380	79%
2	School speller 1939 II	17,336	18,179	13,351	73%
3	School speller 1959	17,361	17,788	14,218	79%
4	School speller 1973	23,950	25,168	20,099	80%
5	Bokmålsordboka 1986 – 2005	71,142	75,590	61,228	81%
6	Sverdrup 1940	175,948	179,000	75,500	43%
7	Sverdup + BOB		48,600	40,900	84%
8	All school spellers		7,676	7,159	93%
9	All school spellers + *Bokmålsordboka*		7,423	7,036	95%
10	All school spellers + Sverdrup (1940)		7,365	6,906	94%
11	All school spellers + BOB + Sverdrup		6,903	6,806	95%

In Table 4 explores the contents of the school spellers in the Dictionary Hotel and Sverdrup (1940) and *Bokmålsordboka*, in terms of numbers of entries and headwords. The table also shows the co-occurrence of headwords in the school spellers, and in the school spellers and Sverdrup (1940) and *Bokmålsordboka* (2005). These publications give information about the orthography of Bokmål over 80 years. A headword list of 6,906 items is common to all, and it seems safe to guess that this is a core list of essential vocabulary.

Table 3 also shows the results of testing the headword lists of the Bokmål school spellers (of 1939, 1959 and 1973) against the Lexicographic Bokmål Corpus (LBK), a 100-million-token corpus covering the period 1980 – 2005, and containing text from a variety of genres. The school spellers admittedly have a limited vocabulary, but even so, the rate of hits from the individual school spellers at 70 – 80% suggests that the orthographic changes from 1938 until today cannot have been dramatic. Line 11 shows the hit rate for headwords found in all school spellers, in Sverdrup (1940) and *Bokmålsordboka*, with a hit rate in the LBK of 95%, which seems to confirm the core list status.

6.2 Possibilities and future goals

What remains to be seen is how and to what extent the different groups of headwords have been used. In order to find out, full form registers for the different orthographies must be developed.

First the Norwegian case: the Language Collections at the University of Bergen has expandable full form registers for Bokmål and Nynorsk – the Word Banks. The next step should therefore be: (1) linking headwords of the Meta Dictionary to the proper Word Bank, thus equipping them with inflection paradigms for the current orthographies; (2) adding missing dated inflection paradigms to the Word Banks back to 1938 – probably few would be needed – and adjusting the Word Bank entries accordingly. With these steps taken, it will be possible to make exact and detailed examinations of all text going back to 1938, and also to examine the use of 1938 forms in text from before the orthographic reform. One could then follow on with documenting the orthography before and after 1917 and back to 1901 (Bokmål) and 1873 (Aasen's *Norsk Ordbog*).

Second, the general case: All languages have a history of language standardization. What happened to the orthography of English from 1700 to 1800? Is it possible to measure the effects of Doctor Johnson's dictionary of 1755 by comparing corpora from before 1750 with corpora from after 1755? At present, there is a widespread assumption that language standardization is something that happens, more or less spontaneously, if the language community is large and literate enough. But these assumptions have not been critically examined by direct examination of text. A Doctor Johnson full form register, with a tagger, would thus be very welcome, together with a Meta Dictionary for English.

7 Conclusion

Our aim has been studying how one can use lexicographic resources to measure language variation and change, and what sort of tool can give reliable results in this context. These questions arise when considering the state and history of the Norwegian written standards, Bokmål and Nynorsk. By "tool" one should not only focus on concrete software applications, and in this paper we have discussed the methods and requirements for setting up a research environment, describing a full-scale test. The major benefit for researchers and the public alike is that the contents of the dictionaries and spellers discussed in this paper are accessible for whatever searches one cares to make. Endless questions can now be met with precise answers, which in turn can be critically examined.

A less scholarly motive for getting this done is the fact that full form registers for Bokmål and Nynorsk for around 500,000 lexical items, expressed in taggers for corpora, would contribute to making Norwegian easier to use in the various language technology applications needed in a society increasingly dependent on electronic support.

References

Aasen, Ivar (1873): *Norsk Ordbog med dansk Forklaring*. Fjerde uforandrede udgave, (1918). Kristiania: Vestmannalaget og Cammermeyers forlag.

Atkins, S. & Michael R. (2008): *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Bouda P., Cysouw M. (2012) Treating Dictionaries as a Linked-Data Corpus. In: Chiarcos C., Nordhoff S., Hellmann S. (eds): *Linked Data in Linguistics*. Springer, Berlin, Heidelberg, 978-3-642-28248-5, Accessed at: https://doi.org/10.1007/978-3-642-28249-2_2 [30/03/2018].

Christmann, R.(2001); Books into Bytes: Jacob and Wilhelm Grimm's Deutsches Wörterbuch on CD-ROM and on the Internet. In: *Literary and Linguistic Computing*, Volume 16, Issue 2, 1 June 2001, Pages 121–133, Accessed at: https://doi.org/10.1093/llc/16.2.121 [18/052018].

Den nye rettskrivning. Regler og ordlister (The new Orthography. Rules and Word Lists). Utarbeidet ved Den departementale rettskrivningskomite. Kristiania. Det Mallingske Bogtrykkeri 1918. Fastsatt ved kgl.res. 21. desember 1917. Accessed at: http://www.sprakradet.no/Spraka-vare/Norsk/Faksimilebiblioteket /Den_nye_ rettskrivning_1917 [25/03/2018].

Eitrem, H. (1939): *Rettskrivning 1938. Regler for skoler og privat bruk*. Oslo: Fabritius.

Grønvik, Oddrun & Ore, Christian-Emil Smith (2013): What should the electronic dictionary do for you – and how?, In Iztok Kosem; Jelena Kallas; Polona Gantar; Simon Krek; Margit Langements & Maria Tuulik (ed.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*, Eesti Keele Instituut. ISBN 978-961-93594-0-2. pp. 243 – 260. Accessed at: http://eki.ee/elex2013/proceedings/eLex2013_17_Gronvik+Ore.pdf [17/05/2018].

Hovdenak, M., Killingbergtrø, L., Lauvhjell, A., Nordlie, S., Rommetveit, M. & Worren, D. (2006): *Nynorskordboka* (4. utg.). Oslo: 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/ [30/03/2018].

Krogsrud, T. & Seip, D.A (1959): *Rettskrivningsregler. Rettskrivningslære og ordliste*. Etter «Ny læreboknormal 1959». Oslo: Cappelen.

Lange, A. (1939): *Norsk rettskrivingsordliste*. Oslo: Tiden Norsk Forlag.

LBK (2011) Leksikografisk bokmålskorpus. Universitetet i Oslo. Accessed at: http://www.hf.uio.no/iln/ tjenester/kunnskap/samlinger/bokmal/tilgang-korpus [25/03/2018].

Metaordboka. Språksamlingane, Universitetet i Bergen. Accessed at: http://usd.uib.no/perl/search/search.cgi? appid=7&tabid=57 [25/03/2018].

Moulin, C. & Nyhan, J. (2014): The Dynamics of Digital Publications, An Exploration of Digital Lexicography. In: Davidhazi, P. (ed) *New publication cultures in the humanities: exploring the paradigm shift*, Amsterdam University Press, 2014, ISBN: 90-485-1971-3.

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

Nyordsarkivet. Bokmål. (The Neologism Archive. Bokmål) Språksamlingane, Universitetet i Bergen. Accessed at: http://usd.uib.no/perl/search/search.cgi?appid=245&tabid=3521 [20/03/2018].

Ordbokshotell. Språksamlingane, Universitetet i Bergen. Accessed at: http://usd.uib.no/perl/search/search. cgi?appid=118&tabid=1777 [25/03/2018].

Ore, C.-E., (2000): Metaordboken – et rammeverk for Norsk Ordbok. In: *Nordiska studier i lexikografi* 5. Göteborg: Nordisk forening for leksikografi, s. 250—270.

Ore, C.-E. & Ore E. (2010) Re-linking a Dictionary Universe or the Metadictionary Ten Years Later In: *Conference Abstracts, King's College London, London, July 7 – 10, 2010* Published by Office for Humanities Communication Centre for Computing in the Humanities, King's College Digital Humanities 2010, London.

Stortingsmelding 1 (2017–2018) *Nasjonalbudsjettet* 2018. Kap. 326 post 75 Tilskudd til ordboksarbeid. (Parliamentary Report 1 2017-2018. National Budget 2018. Ch. 26 subsection 75 Allocation to Lexicography) Accessed at: https://www.regjeringen.no/no/dokumenter/prop.-1-s-kud-20172018/id2574640/sec2#match2 [12/10/2017].

Sverdrup, J. (1940): *Norsk rettskrivningsordbok*. Bokmål. Oslo: Tanum.

TEI (2018) *Text Encoding Initiative, Guidelines to Text Encoding (P5)*. Accessed at: http://www.tei-c.org [30/03/2018].

Torvik, I. (1973): *Ordliste for alle. Bokmål*. Oslo: Universitets-forlaget.

Tvedt, L.J, Lien, E. & Eide, Ø. (2007): Ordbokshotellet – varig lagring og formidling av norske ordsamlingar. In: Arboe, T. (red.), *Nordisk dialektologi og sociolingvistik, Foredrag på 8. Nordiske Dialektologkonference, Århus 2006*. Peter Skautrup Centret for Jysk Dialektforskning, Aarhus Universitet, s. 379–388.

Vikør, L. (2001): *The Nordic Languages*. Their Status and Interrelations. Oslo. Norsk språkråd. 3. ed.

Wangensteen, B. (2006): *Bokmålsordboka*. Definisjons- og rettskrivingsordbok. (3. utg.). Oslo: Kunnskapsforlaget.

Wörterbuchnetz (2018) Accessed at: http://www.woerterbuchnetz.de [30/03/2018].