Margunn Rauset, Gyri Smørdal Losnegaard, Helge Dyvik, Paul Meurer, Rune Kyrkjebø, and Koenraad De Smedt

Words, Words!

Resources and Tools for Lexicography at the CLARINO Bergen Centre

Abstract: The CLARINO Bergen Centre, which provides scholars with access to digital language data and processing services, has in recent years provided substantial services to research and development in lexicography. This chapter describes the interplay between three major lexicography efforts and the centre. Easy access to large corpora in CLARINO and powerful tools for searching and analysing corpus materials help to secure an empirical foundation which far exceeds the lexicographical resources and possibilities available to lexicographers in Norway only a few years ago.

Keywords: CLARINO, lexicography, dictionaries, Norwegian, Bokmål, Nynorsk, corpora, treebanks, written standards

1 Introduction

With funding from the Research Council of Norway and a consortium of institutions, the CLARINO research infrastructure was established in the eponymous CLARINO project, which started in 2012. At present, four technical centres and two knowledge centres embody Norway's in-kind member contribution to CLARIN ERIC. One of these centres is the *CLARINO Bergen Centre*,[1] located at the University of Bergen, in co-operation with the Norwegian School of Economics.

[1] https://clarino.uib.no

Acknowledgements: The research infrastructure described in this chapter has been funded in part by the Research Council of Norway under grants 295700 (CLARINO+), 208375 (CLARINO) and 195323 (INESS) and by a consortium of institutions led by the University of Bergen. Revisjonsprosjektet and NO-AH are both fully funded by the Norwegian Government's Department of Culture. The NAOB dictionary project is funded by the Norwegian Government's Department of Culture (about 80%) and by a number of non-profit foundations and funds (about 20%).

Margunn Rauset, Gyri Smørdal Losnegaard, Helge Dyvik, Paul Meurer, Rune Kyrkjebø, and Koenraad De Smedt, University of Bergen, Bergen, Norway, e-mail: clarino@uib.no

https://doi.org/10.1515/9783110767377-021

Like other centres in the CLARIN distributed infrastructure, the CLARINO Bergen Centre provides scholars with access to language resources and tools through a repository and other services (De Smedt et al. 2016).

In recent years, the CLARINO Bergen Centre has started catering to research and development in lexicography in particular. The current chapter describes the interplay between three national lexicography efforts and the centre. Two of these, *Revisjonsprosjektet* and *NO-AH*, are located in Bergen, while the third project, *NAOB*, is governed by the Norwegian Academy for Language and Literature, located in Oslo. They will be described in more detail below.

The current drive in lexicographic activity in Bergen started in 2016, when 15 truckloads of digital and non-digitized language collections, including lexicographical materials and sources, were moved from Oslo to Bergen. With additional national funding, the University of Bergen began to establish itself as a hub for curating and extending these collections under the name *Språksamlingane* ('The Language Collections'). This name refers to the collections of dialects, place names and words that were built and maintained at the University of Oslo from the 19th century. Språksamlingane are based at the University of Bergen Library, steered at the strategic level by a committee led by the Department of Linguistic, Literary and Aesthetic Studies at the University of Bergen, and advised by a national board of experts. The Norwegian terminology portal *Termportalen*, developed with support from CLARINO since 2012, will also be hosted at Språksamlingane (Andersen and Gammeltoft 2022).

The bulk of the material transferred to Bergen consists of about 4 million records on paper cards, a large percentage of which are also digitized, and which have been employed in lexicographical work over the years. The University of Bergen was now faced with the challenge of running and maintaining the lexicographical databases. After the original Oracle system was up and running again, it was decided to reimplement the back-end for Språksamlingane. This is work in progress. A more urgent technical need arose, however, in 2018, when Revisjonsprosjektet got its go-ahead, namely the need for a versatile front-end and user interfaces for searching and revising the lexicographical data.

The technical and professional resources of CLARINO proved decisive for the ability of the University of Bergen to meet this challenge. Among the services provided by the centre, the following in particular provide an important foundation for the work described in this chapter:

1. *Corpuscle* is a corpus management tool providing access to plain text or tagged corpora, including audio and video with transcriptions (Meurer 2012a). It provides a powerful corpus search function based on efficient algorithms (Meurer 2020) and also produces word lists, collocations, and distributions. Its current holdings cover Norwegian and 15 other languages.

2. *INESS* is a treebanking platform providing access to treebanks in LFG, HPSG, dependency and constituency formats (Meurer et al. 2013; Rosén et al. 2012). Available treebanks cover more than a hundred languages and notably include *NorGramBank*, a large treebank for Norwegian, which will be further described below. Closely linked to INESS is the CLARIN Knowledge Centre on Treebanking, which provides expertise on treebanking construction, management, and exploration.

The remainder of this chapter is structured as follows. Section 2 describes NorGramBank as a CLARINO resource for all three lexicographical projects presented in this chapter. Section 3 discusses the relevance of Norwegian language policy for Norwegian lexicography. Section 4 introduces *Revisjonsprosjektet*, aimed at updating the Bokmål and Nynorsk dictionaries. Section 5 discusses work on the Norwegian Dictionary A to H (NO-AH) and Section 6 discusses work on the Norwegian Academy Dictionary (NAOB). The chapter is rounded off by a conclusion in Section 7.

2 NorGramBank: A resource for three lexicographical projects

NorGramBank is a Norwegian treebank, developed in the INESS project (2010–2017) at the University of Bergen (Dyvik et al. 2016) and now curated by CLARINO. It has been constructed through parsing with NorGram, followed by stochastic disambiguation of the parsing results, trained on a manually disambiguated subcorpus. NorGram is a manually written computational grammar for Norwegian within the framework of Lexical Functional Grammar (LFG). By 2017, NorGramBank comprised about 50 million words of analysed text (novels, children's books, non-fiction, newspapers, parliamentary debates, and some other genres). After the addition of more than 3,000 digitized fiction and non-fiction works, as requested by the NAOB project (see Section 6), the corpus now comprises about 160 million words of analysed text. These additional texts were made available to the CLARINO Bergen Centre in OCR-scanned form from the National Library after special permission to use copyrighted works had been obtained from the Norwegian government.

The LFG analyses in NorGramBank provide rich and detailed syntactic information about sentences, as well as some semantic information in the form of predicate-argument structures. The capacity to search for such information and sort the examples according to author, work, and other criteria is valuable for the development of dictionaries. The treebank provides information about the typical syntactic behaviour of a word (the adjectives modifying a noun, the functions of an adjective, the selected prepositions or argument structures of a verb, etc.), and it provides the means to find suitable examples from the literature. Having all this information at one's fingertips is clearly enticing to lexicographers.

Although the treebank query language *INESS Search* has a simple and intuitive syntax (Meurer 2012b), the complexity of the syntactic analyses may still lead to complex query expressions. In order to reduce this problem for the lexicographers, a template-driven "sketch" function has been developed (Rosén et al. 2020). A search template is a parameterized expression allowing the user to provide values for a selection of parameters, such as lemma forms or feature values, without engaging with the full search expression itself, and then run the query. Examples of the use of such templates will be given in the sections on the individual lexicographical projects.

3 Norwegian language policy and its relevance for lexicography

The language policies of Norway have had a clear impact on the development and publication of language resources. Norwegian has two official written standards – Bokmål and Nynorsk. The historical background to this situation is the union between Norway and Denmark, which lasted for 400 years and ended in 1814. Norwegian and Danish are closely related Scandinavian languages and the written language of the union was Danish, with its norm centre in Copenhagen. After Norwegian independence, two paths towards linguistic independence were established during the 19th century.

One path towards a Norwegian written standard, initiated by the poet and linguist Ivar Aasen, was based on the reconstruction of an idealized common ancestor of the most traditional rural dialects, especially in the Western part of the country. This standard, known as Landsmål and later as Nynorsk, had a rich literary development and was officially recognized as being equal with the existing Danish standard as early as 1885. It has later gone through some modernizing reforms.

The other path towards a Norwegian written standard was initiated by the school headmaster Knud Knudsen. It consisted in "Norwegianizing" the spelling and grammar of the existing Danish standard based on educated urban speech or spoken Riksmål, a variety which had its historical origin in a spoken Dano-Norwegian urban koiné that had been in use from the 17th century onwards. The programme was carried through by means of reforms starting in 1907, and the result, known as Riksmål and later as Bokmål, is now the dominant standard, used by about 88% of pupils.

Over the years several language reforms have been undertaken, which have had obvious consequences for lexicography and, more recently, for language technology applications such as spelling correction. A committee, appointed in 1964 with Professor Hans Vogt as its chair, proposed a body to protect and develop the Norwegian language, which resulted in the establishment of *Norsk Språkråd* ('The Language Council of Norway'), now *Språkrådet*.

Several laws ensure the continued use of Nynorsk and Bokmål with equal status. Since 1980, *Mållova* ('The Language Standard Act') regulates the use of the two written standards in the public sector, and all pupils learn Bokmål as well as Nynorsk at school. In 2009, a parliamentary white paper *Mål og meining* ('Language/goal and meaning', an intended ambiguity) aimed at securing the position of Norwegian in a digitizing society and proposed the establishment of the Norwegian Language Bank to provide language resources supporting language technology. The Language Bank is now one of the CLARINO centres. A more recent parliamentary white paper *Humaniora i Norge* ('The Humanities in Norway') acknowledges the important role that CLARINO is playing in language research and technology.

Finally, on 25 March 2021, *Språklova* ('The Language Act') set out an extensive policy to secure the equal status of the two written standards, but also to protect minority languages such as Sami and Norwegian Sign Language.[2] Furthermore, the proposition underlying this law points out the importance of Språksamlingane and of Termportalen, the latter developed through CLARINO. It also mentions the three lexicographical projects described below as important contributions to the Norwegian language. All of this underlines the historical context and the political importance of the current lexicographical work and the role that CLARINO is playing in Norway. The following sections will discuss the three lexicographical projects in some detail.

[2] Prop. 108 L (2019–2020) Lov om språk (Språklova), adopted by the Norwegian Parliament on March 25, 2021, based on the following proposal: https://www.regjeringen.no/no/dokumenter/prop.-108-l-20192020/id2701451/.

4 Updating of the Bokmål and Nynorsk dictionaries

Revisjonsprosjektet ('The Updating Project'), is an update of the medium-size dictionaries for modern Bokmål and Nynorsk.[3] One proposal from the above-mentioned Vogt committee was to establish a lexicographical department at the University of Oslo (UiO). Neither Bokmål nor Nynorsk had practical handbook-size dictionaries affordable for regular language users, and compiling such would be the first major task for the new department. The compilation of *Bokmålsordboka* ('the Bokmål dictionary') and *Nynorskordboka* ('the Nynorsk dictionary') in parallel was in itself considered as a tool to build an atmosphere of mutual respect and a recognition of the two written standards with equal official status.

The first printed editions of these dictionaries were published in 1986. One group of lexicographers had been working on *Bokmålsordboka* and another on *Nynorskordboka* since 1974, as a cooperation between the Department of Lexicography at UiO and The Norwegian Language Council. According to the initial plan, both dictionaries should cover modern Bokmål and Nynorsk as used in literature and the media. In addition, they should each have around 600 to 700 pages and 60,000 entries with the same structure and information categories (Kulbrandstad 1976: 8). The editorial staff of *Nynorskordboka* wrote the manuscript of the letters *a*–*k* and *v*, whereas the editorial staff of *Bokmålsordboka* compiled the entries between *l*–*u* and *w*–*å*. They then exchanged manuscripts and thus could benefit from each other's work (Landrø and Wangensteen 1986: v).

Nevertheless, the dictionaries ended up with distinct features and several differences. The most striking difference is that *Nynorskordboka* has around 90,000 entries, whereas *Bokmålsordboka* has 65,000 entries. One reason for this difference is that Bokmål, unlike Nynorsk at the time, already had other comparable dictionary resources. The lexicographers working with *Nynorskordboka* argued that it was important to manifest the close relation between the dialects and written Nynorsk, so they included lemmas documented in use in three Norwegian counties (or two in Northern Norway), even though rarely used in written texts. *Nynorskordboka* thus describes written and oral vocabulary, whereas *Bokmålsordboka* documents written language. *Nynorskordboka* also includes more compound words than *Bokmålsordboka*. On the other hand, the latter contains more loan words from Danish and German (Hovdenak 2014: 234; Worren 1998: 63).

In later editions of both dictionaries, spelling and inflection have been updated according to the official standards. Some new lemmas were added, but most of

[3] http://www.uib.no/revisjonsprosjektet

the articles stayed unchanged since the 1986 edition. A thorough content update, based on new material in many genres, was therefore much needed. Whereas the latest printed editions are fairly dated (Hovdenak et al. 2006; Wangensteen 2005), the two dictionaries have also been available as an online edition via a common web interface since 1994.[4] This common portal is extensively used by pupils and the general public, while the app *Ordbøkene* on iOS and Android, available since 2017, has also become quite popular. On average the web page and the app have a combined total of 160,000 hits a day. When users see entries in the online dictionaries in the default side-by-side view, the differences become more noticeable than the lexicographers of the printed editions could foresee. The change of medium makes the need for synchronous updating more visible.

With these editions as a starting point, both dictionaries are being updated in Revisjonsprosjektet, a project carried out from 2018 until 2024 at the University of Bergen, in cooperation with the Language Council of Norway. The project has three main aims. The first goal is to make the dictionaries more similar in structure and coverage, as far as possible. The dictionaries aim to document common language use in the written varieties Bokmål and Nynorsk, and as a principle all entries should be found in both dictionaries if the lemma is used in both varieties. The second goal is to check whether definitions, examples of usage and fixed expressions are in line with present-day language use, defined as the period from the 1970 until today. The third is to supplement the dictionary with new words and meanings that have entered the language (Rauset 2019: 169).

The digital language resources provided by CLARINO are of great help with respect to all three goals. As the project has progressed well by now and the technology has been sufficiently developed, we can report on experiences from our changed lexicographic practice in the remainder of this section.

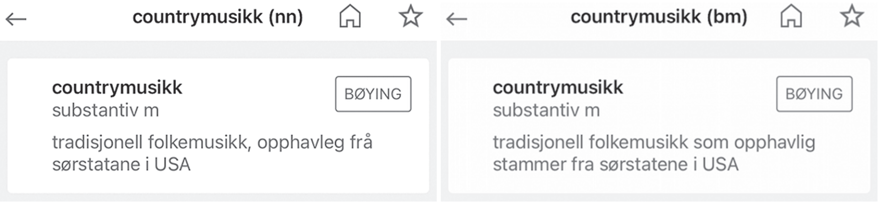
In 2018, *Corpuscle-Lex* was developed as an extension of the above-mentioned Corpuscle. It is a bespoke online environment for lexicographers in which corpus search and dictionary management are integrated in a single web-based environment, thereby improving the workflow considerably. Corpuscle-Lex provides search in up to 12 online Norwegian corpora simultaneously. With more than 2.8 billion words[5] from a variety of sources and genres, this is the largest corpus collection for Norwegian that lexicographers have ever had available at their fingertips. Simultaneous search in user-selected corpora is enabled thanks to previous work on metadata and search algorithms in CLARINO.

[4] Both dictionaries are available online at https://ordbokene.no.

[5] Nynorsk comprises 185 million words (6.5%) in this collection and Bokmål 2.65 billion (93.5%).

In addition to corpus search, Corpuscle-Lex has an interface to the existing dictionaries and *Ordbanken* ('the Norwegian Word Bank'). Crucially, it also has a dictionary editing tool in which several dictionaries can be edited side by side. Finally, the system also has a direct communication link to the Language Council, through which normalization issues can be addressed in a very efficient way. The lexicographers at the University of Bergen decide which entries to include, but if in doubt, the Language Council in Oslo has the final say when it comes to how Norwegian words are spelled and inflected.

Work on the dictionary entry *countrymusikk* ('country music') can illustrate various aspects of this lexicographic practice. Figure 1 shows screenshots from the app showing the original entries for this word in both language varieties.



countrymusikk (nn)

countrymusikk
substantiv m
tradisjonell folkemusikk, opphavleg frå sørstatane i USA

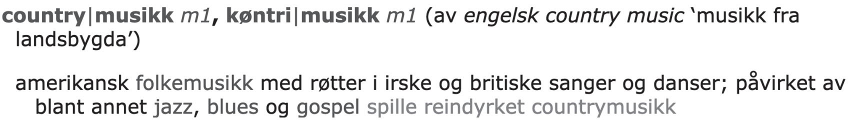
countrymusikk (bm)

countrymusikk
substantiv m
tradisjonell folkemusikk som opphavlig stammer fra sørstatene i USA

Figure 1: Original entries for *countrymusikk* with a single spelling as shown in the app. The label *bm* is bokmål, *nn* is nynorsk. The explanation is "traditional folk music which originally stems from the southern states in the USA".

The 12 corpora in Corpuscle-Lex document that some language users spontaneously have Norwegianized the spelling of *country* to *køntri*. The search for words matching the regular expression "`countrymusikk.*|køntrimusikk.*`" gives 1,282 hits, 27 (2%) of which are *køntrimusikk*, a form which until recently was not accepted. Furthermore, searching for "`køntri.*`" gives 378 hits, all of them related to the music genre. Based on the results from Corpuscle-Lex, although not great in numbers, the Language Council has defined *køntri* and all of its compounds as a part of the official standard for both Bokmål and Nynorsk. The revised version of the dictionaries therefore includes both *countrymusikk* and *køntrimusikk*, as shown in Figure 2. The lexicographer has also updated the definition and added etymological information and an attested example, based on an authentic example in the concordance, to illustrate a typical use of the lemma.

One of the most frequently used tools in Corpuscle-Lex is the "Word list" (with frequencies) which the lexicographers can generate from a regular expression search in the corpora they consider expedient. The corpus managing tool is very flexible, and based on what they are looking for, the lexicographers include or exclude annotated or unannotated corpora, oral or written corpora, corpora with texts in Bokmål or Nynorsk, corpora with specific genres, and so on. The word list function makes it easy to evaluate whether the existing entries are the most relevant in an updated version of the dictionaries.

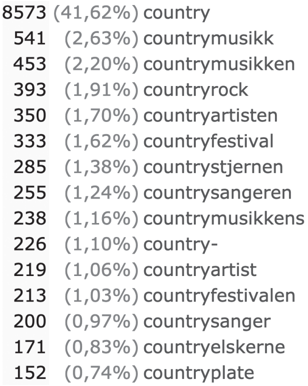


country|musikk *m1*, **køntri|musikk** *m1* (av *engelsk country music* 'musikk fra landsbygda')

amerikansk folkemusikk med røtter i irske og britiske sanger og danser; påvirket av blant annet jazz, blues og gospel spille reindyrket countrymusikk

Figure 2: Revised entry for *countrymusikk/køntrimusikk* in *Bokmålsordboka*.

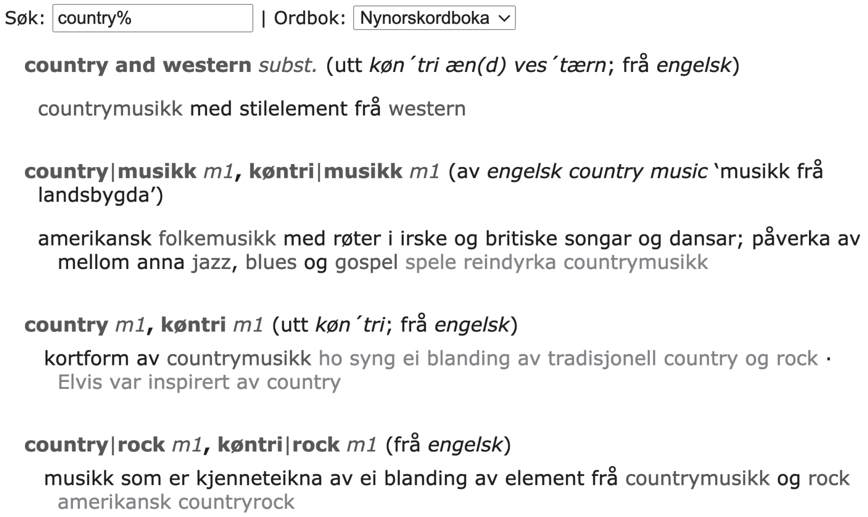
In the old version of the dictionary there were only two entries including the word *country*: *country and western* and *countrymusikk*. However, the word list generated by searching for "country.*" in all 12 corpora in Corpuscle-Lex gives 20,600 hits and 2,004 unique forms, showing that this is a highly productive word in Norwegian. Figure 3 shows the most frequent matches. Based on this word list and on collocations from the corpora, the lexicographer chose to compile two more entries, and the updated dictionaries now have four entries including the word *country*, as shown in Figure 4 from *Nynorskordboka*.



8573	(41,62%)	country
541	(2,63%)	countrymusikk
453	(2,20%)	countrymusikken
393	(1,91%)	countryrock
350	(1,70%)	countryartisten
333	(1,62%)	countryfestival
285	(1,38%)	countrystjernen
255	(1,24%)	countrysangeren
238	(1,16%)	countrymusikkens
226	(1,10%)	country-
219	(1,06%)	countryartist
213	(1,03%)	countryfestivalen
200	(0,97%)	countrysanger
171	(0,83%)	countryelskerne
152	(0,74%)	countryplate

Figure 3: The most frequent of 2,004 words matching "country.*" in Corpuscle-Lex.

The word list is our most efficient tool to identify both neologisms and lemmas that could and maybe should have been included in the dictionaries a long time ago (Lyse 2020: 219). So far, 5,200 new entries have been added to *Bokmålsordboka* and 5,000 to *Nynorskordboka*. Among these are relatively newly imported words in Norwegian, many from the IT domain, such as *backup*, *batch*, *bugg/bøgg* ('bug') and *dokkingstasjon* ('docking station'), along with words referring to new concepts in a Norwegian context, such as *abaya* (a garment), *bilkollektiv* ('car share'), *delingsøkonomi* ('sharing economy'), *designerdop* ('designer drug'), *droneangrep* ('drone strike'), *elsparkesykkel* ('electric scooter') and *koronavirus* ('coronavirus'). Among lemmas with a longer history in Norwegian, but with new dictionary entries, we can find *allmennlege* ('general practitioner'), *alpint* ('alpine skiing'), *badetøy* ('swimwear'), *brukerorientert* ('user-oriented'), *CO_2-utslipp* ('CO_2 emission') and *didjeridu* ('didgeridoo'). Compounds with *atom-* have become a part of everyday speech since the dictionaries first were published in 1986, but there are 10 new compounds in the updated versions, including *atomavfall* ('nuclear waste') and *atomstridshode* ('nuclear warhead').



Søk: country% | Ordbok: Nynorskordboka

country and western *subst.* (utt *køn´tri æn(d) ves´tærn*; frå *engelsk*)

countrymusikk med stilelement frå western

country|musikk *m1*, **køntri|musikk** *m1* (av *engelsk country music* 'musikk frå landsbygda')

amerikansk folkemusikk med røter i irske og britiske songar og dansar; påverka av mellom anna jazz, blues og gospel spele reindyrka countrymusikk

country *m1*, **køntri** *m1* (utt *køn´tri*; frå *engelsk*)

kortform av countrymusikk ho syng ei blanding av tradisjonell country og rock · Elvis var inspirert av country

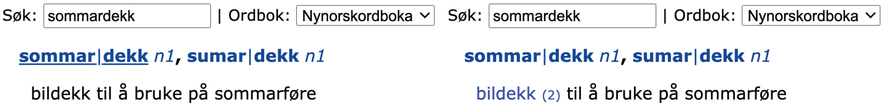
country|rock *m1*, **køntri|rock** *m1* (frå *engelsk*)

musikk som er kjenneteikna av ei blanding av element frå countrymusikk og rock amerikansk countryrock

Figure 4: New and updated entries starting with *country* in *Nynorskordboka*.

All lexicographers in Revisjonsprosjektet are working with both *Bokmålsordboka* and *Nynorskordboka*, and unless the corpora and spelling rules indicate that there are real differences in language use between the two written standards, entries are created or updated in parallel. So far, 5,700 new entries have been compiled in the smaller *Bokmålsordboka* because there were parallel existing entries in *Nynorskordboka*, whereas 2,200 new entries have been compiled in *Nynorskordboka* based on existing entries in *Bokmålsordboka*. The large corpus collection in Corpuscle-Lex and the corpus based methodology in the project makes it easier to identify the differences between the two standards. As a result, the updated selection of entries, both those that are found in only one of the dictionaries and those that are found in both, reflects modern language use to a higher degree than before. Quality is further assured thanks to the interface supporting the sharing of articles with colleagues and with the Language Council of Norway.

Since digital dictionaries allow cross-references to other entries by establishing hyperlinks, attention has been paid to making this process easy and accurate. The word *bildekk (2)*, marked in blue in the updated definition on the right in Figure 5, is such a cross-reference. The process of such linking is exemplified by the editing window for the entry *sommardekk/sumardekk* ('summer tire') in *Nynorskordboka*, shown in Figure 6. The definition contains a word *bildekk* with two meanings in Norwegian ('car deck' or 'car tire'), which motivates a link to the correct meaning in this context. By adding an @ in front of *bildekk* in the definition and choosing the intended meaning *(2)* from a popup menu, an appropriate link is made.



Søk: sommardekk | Ordbok: Nynorskordboka

sommar|dekk *n1*, **sumar|dekk** *n1*

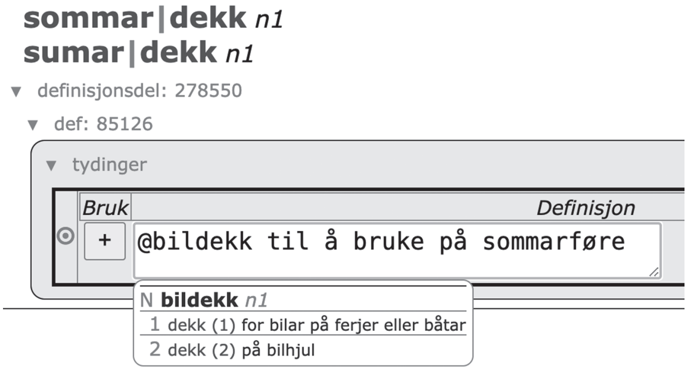
bildekk til å bruke på sommarføre

Søk: sommardekk | Ordbok: Nynorskordboka

sommar|dekk *n1*, **sumar|dekk** *n1*

bildekk (2) til å bruke på sommarføre

Figure 5: Original (left) and updated (right) entries for *sommardekk/sumardekk* ('summer tire') in *Nynorskordboka*.



sommar|dekk *n1*
sumar|dekk *n1*

▾ definisjonsdel: 278550

▾ def: 85126

▾ tydinger

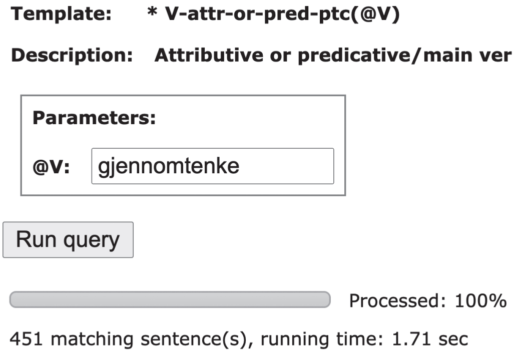
Bruk	Definisjon
+	@bildekk til å bruke på sommarføre

N **bildekk** *n1*
1 dekk (1) for bilar på ferjer eller båtar
2 dekk (2) på bilhjul

Figure 6: Editing tool showing the linking of the definition of *sommardekk/sumardekk* ('summer tire') in *Nynorskordboka* to the intended meaning (2) of *bildekk* ('car tire').

Another useful resource is NorGramBank, which was introduced in Section 2. In this treebank, one can search for complex syntactic constructions and their frequencies, which is useful for finding typical uses of words in constructions. The lexicographers in Revisjonsprosjektet use templates in NorGramBank to show usage and frequency. The template *V-argframes(@V)* is useful both for finding the most common uses of a verb (valency frames, common prepositions, or particles) and possible reflexive use of the verb. The templates *ADJ-attrib-or-nominal(@ADJ)* and *V-attr-or-pred-ptc(@V)* yield frequencies of adjectival and nominal use of participles, thereby providing empirical grounds for the possible creation of a separate entry for a derived adjective.

As an example, consider the verb *gjennomtenke* ('think through') which had a single entry in *Bokmålsordboka*. With the help of the template *V-attr-or-pred-ptc(@V)*, shown in Figure 7, it was found that the attributive use of the participle *gjennomtenkt* ('well thought out') was higher than its verbal use, cf. the frequencies displayed in Figure 8. Consequently, the entry was split so that a separate entry for the attributive use of *gjennomtenkt* ('well thought out') was established, as shown in Figure 9.



Template: * V-attr-or-pred-ptc(@V)

Description: Attributive or predicative/main verb function of a participle

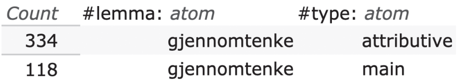
Parameters:

@V: gjennomtenke

Processed: 100%

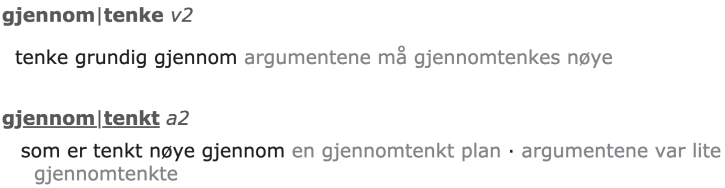
451 matching sentence(s), running time: 1.71 sec

Figure 7: Search for *gjennomtenke* ('think through') with a template in NorGramBank.



Count	#lemma: *atom*	#type: *atom*
334	gjennomtenke	attributive
118	gjennomtenke	main

Figure 8: Frequencies of usage of the past participle of *gjennomtenke* ('think through').



gjennom|tenke *v2*

tenke grundig gjennom argumentene må gjennomtenkes nøye

gjennom|tenkt *a2*

som er tenkt nøye gjennom en gjennomtenkt plan · argumentene var lite gjennomtenkte

Figure 9: Update through separate entries for *gjennomtenke* ('think through') and *gjennomtenkt* ('well thought out').

5 Norwegian Dictionary A to H (NO-AH)

The second project, also at Bergen, is NO-AH, the revision and update of *Norsk Ordbok*,[6] a comprehensive dictionary in twelve volumes with around 330,000 entries, which provides an exhaustive account of the vocabulary of Norwegian dialects and the written language Nynorsk.

Norsk Ordbok (NO) as a lexicographic project started in 1930. Aiming to account for both spoken Norwegian and the then relatively young written language Nynorsk, and building on Ivar Aasen's documentation of the dialects in his dictionary (Aasen 1873), NO would rely on two types of material: citations from the Nynorsk literature and material from the dialects (Vikør 2018: 29). Besides the historical and variational aspects, NO has an emphasis on documentation, including the principle that information about usage, origin, and geography must be linked to source materials. The NO archives and language collections are thus an essential part of the dictionary as a lexical resource. Most of the paper files and archives were digitized in the 1990s by the Unit for Digital Documentation (EDD) at the University of Oslo. EDD also developed the lexicographic monitor corpus *Nynorskkorpuset* (the Nynorsk corpus) as a new empirical basis for NO. These resources were later transferred to Bergen in the context of establishing Språksamlingane. The printed dictionary was finalized under the project Norsk Ordbok 2014, which lasted from 2002 to 2016 with increased funding, more staff, and the digitization of the editorial process. This project also produced a partial digital edition spanning the letters *i* to *å*.

The current project, NO-AH, started in 2019. The main objective is to update the letters *a* to *h*, which is the oldest part of the dictionary, compiled prior to digitization, and thereby complete the digital edition. A second goal is to provide stable and up-to-date resource management. The ambition is to create a dynamic system of interconnected databases, complete with facilities for update and extension. CLARINO is involved in both content update and resource management. New interfaces are being developed in cooperation with CLARINO, with Corpuscle-Lex as an integrated part of the dictionary writing system. NO-AH also benefits from CLARINO services and expertise in activities involving agreements, licensing, and providing standardized metadata descriptions.

Nynorskkorpuset is a valuable source of lexicographic evidence for NO. The current version has more than 100 million words and texts from 1866 onward. Most of these are newer materials: about 85% of the texts were published after 1975 and 75% after 2000. The corpus is extended annually with texts from the publishing company Det Norske Samlaget and other sources. CLARINO assists with

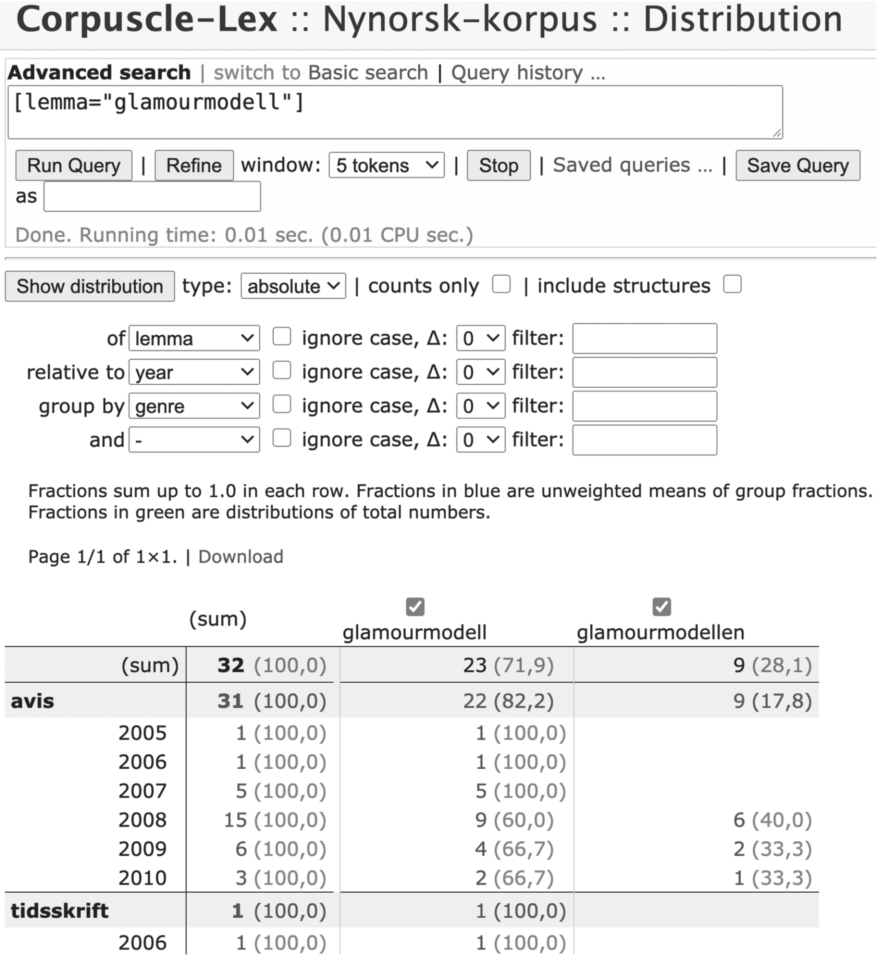
[6] http://norsk-ordbok.no

rights clearance, the design of licensing agreements, and making the materials searchable on the Corpuscle and Corpuscle-Lex platforms. These provide an easy way to get an overview of the current corpus contents, such as words, lemmas, metadata categories, and grammatical annotation. Inspecting the lemmas in a particular alphabet section facilitates the identification of words that have not been included in the dictionary so far.

The update of NO must account not only for new additions to the vocabulary and the present usage of words, but also for words and senses that may already have become outdated or obsolete. Vikør (2018: 19) describes the documentation in NO of the word *glamour* (Vikør et al. 2002: 321), which has two entries: one for the simplex word and one for the word as part of a compound. There is one example of a compound, *glamour-boy* ('poster boy', 'advertising object'), dating back to 1975. This compound, however, returns zero results in Nynorskkorpuset. Although it seems that the word is no longer in use, the entry will be kept, unchanged, for historical documentation. On the other hand, new compounds with glamour have emerged since 2002. Querying the corpus for words starting with *glamour*, we find that *glamourmodell* ('glamour model') is the most frequent compound. To get a first impression of the development and use of this word, the "Distribution" tool in Corpuscle-Lex can be used to show all occurrences of the lemma relative to year and genre. The resulting overview in Figure 10 shows that the compound appears in corpus texts from 2005 onward, that it seemed to reach a peak around 2008, that it does not occur after 2010, and that it is found mainly in newspapers. There are 23 instances of the lemma in Nynorskkorpuset. Extending the search to the other corpora in Corpuscle-Lex increases the number to 56 instances, limited to the period 2005 to 2011. Although the word is not very frequent in these corpora, it is well-documented in that period. The word *glamourmodell* is thus a candidate for documentation as a compound in NO.

Information from syntactic searches in NorGramBank (described in Section 2) is particularly useful in the lexical description of words with many senses and which occur with high frequency in the sources. NorGramBank allows for targeted queries that can provide evidence for colligations (syntactic collocations). The query template *N-argofverbs(@N)* retrieves information about verbs having a particular noun as its argument 1 or 2. This was used for the noun *bane* ('roadway', 'railway', 'track', 'course', 'bane'). Figure 11 shows the top query results.[7] In the instances where verbs take *bane* as their ARG1 (typically the subject), the verb normally appears after the

[7] In this case we have chosen to search in all Norwegian texts, not only Nynorsk. The Nynorsk part of NorGramBank is relatively small, and searching the entire treebank improves the chances of getting enough results to work with. The results should be treated as "seeds" to be followed up by more targeted queries in relevant treebanks.


Advanced search | switch to Basic search | Query history ...
Run Query | Refine
Stop | Saved queries ... | Save Query
Show distribution
Page 1/1 of 1×1. | Download

Corpuscle-Lex :: Nynorsk-korpus :: Distribution

```
[lemma="glamourmodell"]
```

window: 5 tokens

as

Done. Running time: 0.01 sec. (0.01 CPU sec.)

type: absolute | counts only □ | include structures □

of lemma □ ignore case, Δ: 0 filter:
relative to year □ ignore case, Δ: 0 filter:
group by genre □ ignore case, Δ: 0 filter:
and - □ ignore case, Δ: 0 filter:

Fractions sum up to 1.0 in each row. Fractions in blue are unweighted means of group fractions. Fractions in green are distributions of total numbers.

	(sum)	☑ glamourmodell	☑ glamourmodellen
(sum)	**32** (100,0)	23 (71,9)	9 (28,1)
avis	**31** (100,0)	22 (82,2)	9 (17,8)
2005	1 (100,0)	1 (100,0)	
2006	1 (100,0)	1 (100,0)	
2007	5 (100,0)	5 (100,0)	
2008	15 (100,0)	9 (60,0)	6 (40,0)
2009	6 (100,0)	4 (66,7)	2 (33,3)
2010	3 (100,0)	2 (66,7)	1 (33,3)
tidsskrift	**1** (100,0)	1 (100,0)	
2006	1 (100,0)	1 (100,0)	

Figure 10: Distribution of the lemma *glamourmodell* ('glamour model') per year and grouped by genre: *avis* ('newspapers') and *tidsskrift* ('magazines').

noun, and therefore is listed in the column to the right ('C-arg1of'). This is the case with the most frequent combination, *bane* + *være* ('be'). Verbs with bane as their ARG2 (normally object) appear in the left column ('A-arg2of'), the verb normally preceding the noun. This is the case with the second most frequent combination, *ha* ('have') + *bane*.

The published NO entry for *bane* (homograph II) has six main senses, as shown in the facsimile in Figure 12. The first three senses (marked with arabic numerals) correspond to the following approximate English counterparts:

1. (communication, transport) levelled road: roadway, railroad, track
2. (sports) indoor or outdoor site made or reserved for activity: field, course, pitch
3. (movement, direction) trajectory, orbit, course

Several verbs are primarily used with one of these senses. The following verbs (including particle verbs and prepositional verbs)[8] in the list (part of which is shown in Figure 11) tend to colligate with the respective senses as follows:

1. with *bane* as the subject: *komme* ('come'), *gå* ('go'), *stoppe* ('stop'); as the object: *ta* ('take'), *vente*på* ('wait for'), *gå*av* ('get off'), *gå*på* ('get on'), *bygge* ('build');
2. as the object: *gå*av* ('get off'), *gå*på* ('get on'), *komme*på* ('come onto'), *være* ('be'), *bli* ('become'), *bygge* ('build');
3. as the object: *studere* ('study'), *følge* ('follow'), *estimere* ('estimate'), *beskrive* ('describe'), *påvirke* ('influence').

Moreover, some verbs dominate in multiword expressions, such as *bringe på bane* ('bring on track'), *være på bane* ('be on track'), *skygge banen* ('stay away'), and *tenke i [. . .] baner* ('think along [. . .] lines'). Searches with other templates support these findings and make it clearer what is stable and what is variable in such expressions. The empirical data also support promotion of the meaning 'transportation by railroad', which perhaps was not as much used in the middle of the 1900s, but which is now very common, as is evident from some of the collocations that were found.

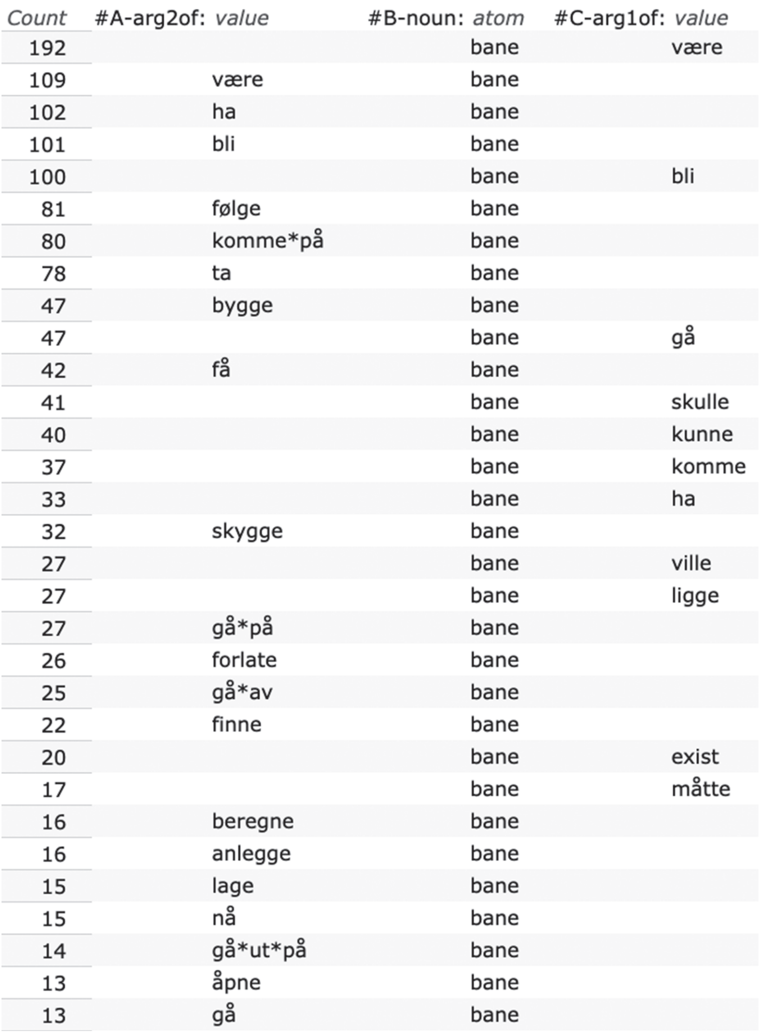
6 The Norwegian Academy Dictionary (NAOB)

The third project is located in Oslo under the auspices of The Norwegian Academy for Language and Literature[9] and consists in the further development of NAOB ('The Norwegian Academy Dictionary'), the most comprehensive dictionary for Bokmål, comprising around 225,000 lemmas with detailed information about semantics and idioms. The descriptions are exemplified with many citations from literature in several genres from a little before 1830 until today.[10] NAOB is freely

[8] The * in the example verbal predicates is not the Kleene star, but marks a composition of a verb and a selected preposition.

[9] https://www.detnorskeakademi.no

[10] Thus it includes about 80 years of literature from the Dano-Norwegian period; see Section 3. In comparison, the Swedish dictionary SAOB describes the period from the 1520s until the time of editing, and the Danish ODS the period from around 1700 until 1955.



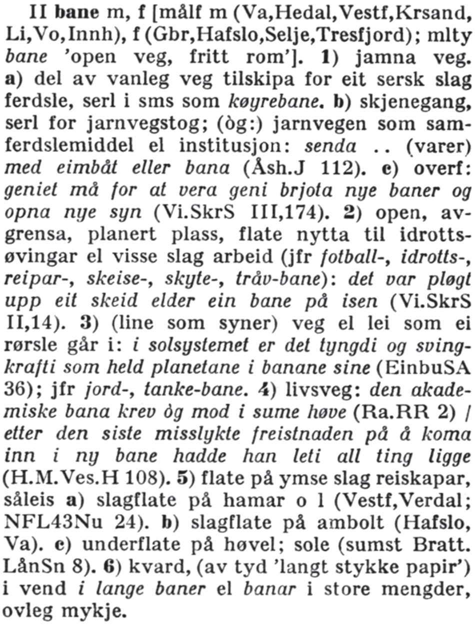
Count	#A-arg2of: *value*	#B-noun: *atom*	#C-arg1of: *value*
192		bane	være
109	være	bane	
102	ha	bane	
101	bli	bane	
100		bane	bli
81	følge	bane	
80	komme*på	bane	
78	ta	bane	
47	bygge	bane	
47		bane	gå
42	få	bane	
41		bane	skulle
40		bane	kunne
37		bane	komme
33		bane	ha
32	skygge	bane	
27		bane	ville
27		bane	ligge
27	gå*på	bane	
26	forlate	bane	
25	gå*av	bane	
22	finne	bane	
20		bane	exist
17		bane	måtte
16	beregne	bane	
16	anlegge	bane	
15	lage	bane	
15	nå	bane	
14	gå*ut*på	bane	
13	åpne	bane	
13	gå	bane	

Figure 11: Top frequencies of verb occurrences with *bane* as argument 1 (‘C-arg1of’) or argument 2 (‘A-arg2of’).

available online[11] and is not published in book form. On average there are 70,000 searches per day from 30,000 unique users.

NAOB, which came online in 2017 and was officially launched on January 24, 2018, is a product of thorough revision, modernization, and extension of an older

[11] http://naob.no



II **bane** m, f [målf m (Va,Hedal,Vestf,Krsand, Li,Vo,Innh), f (Gbr,Hafslo,Selje,Tresfjord); mlty *bane* 'open veg, fritt rom']. **1**) jamna veg. **a**) del av vanleg veg tilskipa for eit sersk slag ferdsle, serl i sms som *køyrebane*. **b**) skjenegang, serl for jarnvegstog; (òg:) jarnvegen som samferdslemiddel el institusjon: *senda* .. (varer) *med eimbåt eller bana* (Åsh.J 112). **c**) overf: *geniet må for at vera geni brjota nye baner og opna nye syn* (Vi.SkrS III,174). **2**) open, avgrensa, planert plass, flate nytta til idrottsøvingar el visse slag arbeid (jfr *fotball-*, *idrotts-*, *reipar-*, *skeise-*, *skyte-*, *tråv-bane*): *det var pløgt upp eit skeid elder ein bane på isen* (Vi.SkrS II,14). **3**) (line som syner) veg el lei som ei rørsle går i: *i solsystemet er det tyngdi og swingkrafti som held planetane i banane sine* (EinbuSA 36); jfr *jord-*, *tanke-bane*. **4**) livsveg: *den akademiske bana krev òg mod i sume høve* (Ra.RR 2) / *etter den siste misslykte freistnaden på å koma inn i ny bane hadde han leti all ting ligge* (H.M.Ves.H 108). **5**) flate på ymse slag reiskapar, såleis **a**) slagflate på hamar o l (Vestf,Verdal; NFL43Nu 24). **b**) slagflate på ambolt (Hafslo, Va). **c**) underflate på høvel; sole (sumst Bratt. LånSn 8). **6**) kvard, (av tyd 'langt stykke papir') i vend *i lange baner* el *banar* i store mengder, ovleg mykje.

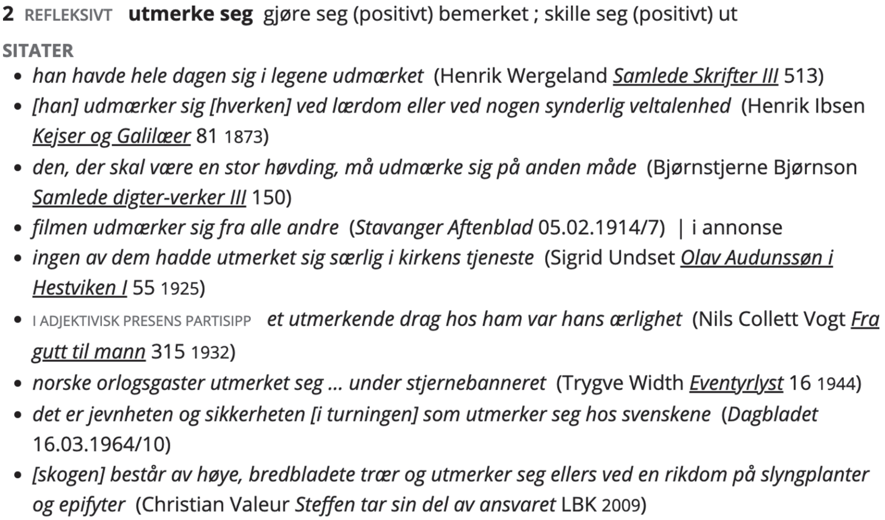
Figure 12: Entry II for *bane* in NO (scanned).

dictionary, *Norsk Riksmålsordbok*, a six-volume dictionary whose first volume appeared in 1937. The literary citations, counting more than 300,000 from 6,500 sources at the time of NAOB's launch, are a central part of a documenting dictionary, providing evidence for the semantic and grammatical descriptions in the dictionary entries. An important part of the revision, modernization and extension leading to NAOB has been updating the literary citations with further examples from more modern and more varied literature. This process still goes on within the limits of modest grants, and this is primarily where the CLARINO Bergen Centre and its treebank NorGramBank[12] come in.

As an example we may consider a NorGramBank search template which was used when a dictionary entry turned out to miss a meaning. The verb *utmerke* ('distinguish') is especially common as a reflexive verb *utmerke seg* ('distinguish oneself'). The dictionary only gave examples where this meant to distinguish oneself positively, as shown in Figure 13.

The editors had reasons to assume that the expression can also be used to describe distinguishing oneself in a negative way. Now, the treebank NorGramBank does not allow sorting examples according to word senses, but it may also

[12] Cf. Section 2.



2 REFLEKSIVT **utmerke seg** gjøre seg (positivt) bemerket ; skille seg (positivt) ut

SITATER

- *han havde hele dagen sig i legene udmærket* (Henrik Wergeland Samlede Skrifter III 513)
- *[han] udmærker sig [hverken] ved lærdom eller ved nogen synderlig veltalenhed* (Henrik Ibsen Kejser og Galilæer 81 1873)
- *den, der skal være en stor høvding, må udmærke sig på anden måde* (Bjørnstjerne Bjørnson Samlede digter-verker III 150)
- *filmen udmærker sig fra alle andre* (*Stavanger Aftenblad* 05.02.1914/7) | i annonse
- *ingen av dem hadde utmerket sig særlig i kirkens tjeneste* (Sigrid Undset Olav Audunssøn i Hestviken I 55 1925)
- I ADJEKTIVISK PRESENS PARTISIPP *et utmerkende drag hos ham var hans ærlighet* (Nils Collett Vogt Fra gutt til mann 315 1932)
- *norske orlogsgaster utmerket seg … under stjernebanneret* (Trygve Width Eventyrlyst 16 1944)
- *det er jevnheten og sikkerheten [i turningen] som utmerker seg hos svenskene* (*Dagbladet* 16.03.1964/10)
- *[skogen] består av høye, bredbladete trær og utmerker seg ellers ved en rikdom på slyngplanter og epifyter* (Christian Valeur *Steffen tar sin del av ansvaret* LBK 2009)

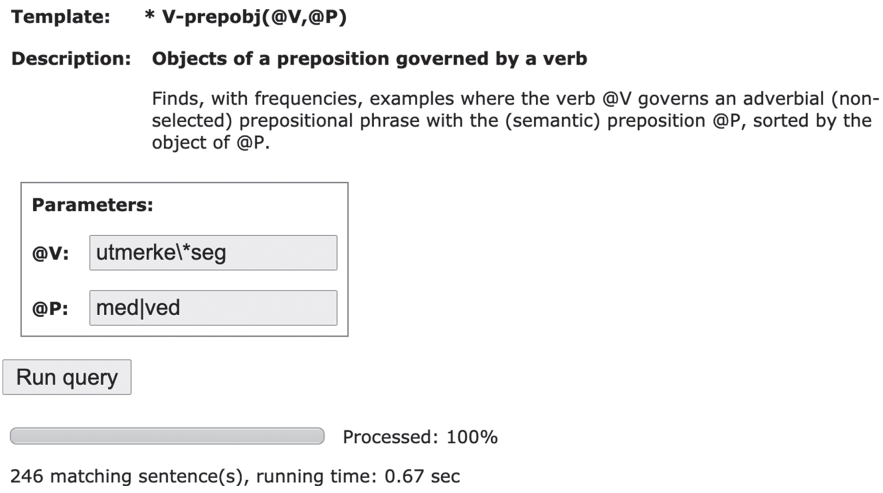
Figure 13: Part of the NAOB entry for the verb *utmerke*, only describing ‘distinguishing oneself’ in a positive way.

be helpful to sort them according to which words occur in specific syntactic positions around the target word. Specifically, the verb *utmerke seg* typically occurs with a prepositional phrase with *med* (‘with’) or *ved* (‘by’), specifying what something is distinguished by. Examples sorted according to what someone or something is distinguished by (expressed by a verb or a noun) can be found by using the template *V-prepobj(@V,@P)*, which allows the user to specify one or more verbs and one or more prepositions, as shown in Figure 14.

Figure 15 shows a small section of the query output, alphabetically sorted by the prepositional objects. The word *mangel* in the fifth and sixth rows means ‘lack’, which makes it a likely place to find a negative meaning of the verb. Clicking on the *Ottar Brox* row then displays the relevant examples, as shown in Figure 15. The first of the two examples (meaning ‘He will also distinguish himself by lack of consistency in his chosen actions’) is suitable and may be selected by clicking on “Copy”, which yields information about the example in an XML format, shown in example (1), which can be directly inserted into the NAOB database.

(1)
```
<sitatledd><sitat>Han vil også utmerke seg ved mangel på konsistens i sine
handlingsvalg, og at en ikke kan forutsi hva han vil finne på å gjøre.
</sitat><kilde><forf>Ottar Brox</forf> <verk>Hva skjer i Nord-Norge? :en studie
i norsk utkantpolitikk</verk> <ref>39</ref>

<urn>https://urn.nb.no/URN:NBN:no-nb_digibok_2013071208165</urn></kilde></
sitatledd>
```



Template: * V-prepobj(@V,@P)

Description: **Objects of a preposition governed by a verb**

Finds, with frequencies, examples where the verb @V governs an adverbial (non-selected) prepositional phrase with the (semantic) preposition @P, sorted by the object of @P.

Parameters:

@V: utmerke*seg

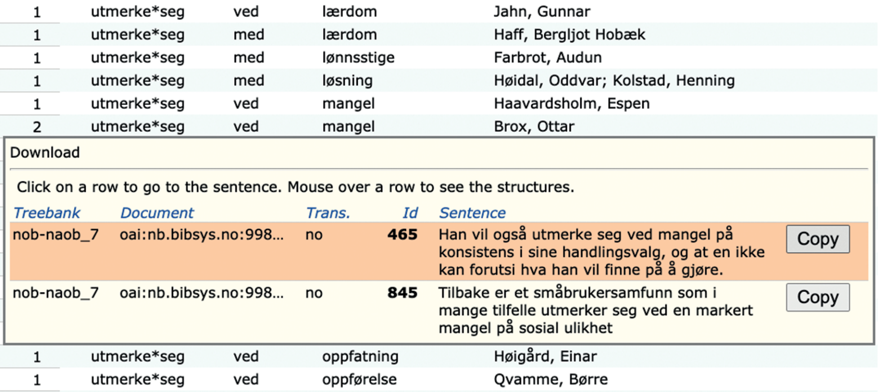
@P: med|ved

Processed: 100%

246 matching sentence(s), running time: 0.67 sec

Figure 14: The search template *V-prepobj(@V,@P)* with parameter values filled in by the user.

Using the template to collect examples like this resulted in the extension of the *utmerke* entry in Figure 16, where the Brox quote occurs as the third example. Clicking on the underlined book title in the dictionary entry will bring the user to the relevant scanned page of the book in the National Library, part of which is shown in Figure 17.



1	utmerke*seg	ved	lærdom	Jahn, Gunnar
1	utmerke*seg	med	lærdom	Haff, Bergljot Hobæk
1	utmerke*seg	med	lønnsstige	Farbrot, Audun
1	utmerke*seg	med	løsning	Høidal, Oddvar; Kolstad, Henning
1	utmerke*seg	ved	mangel	Haavardsholm, Espen
2	utmerke*seg	ved	mangel	Brox, Ottar

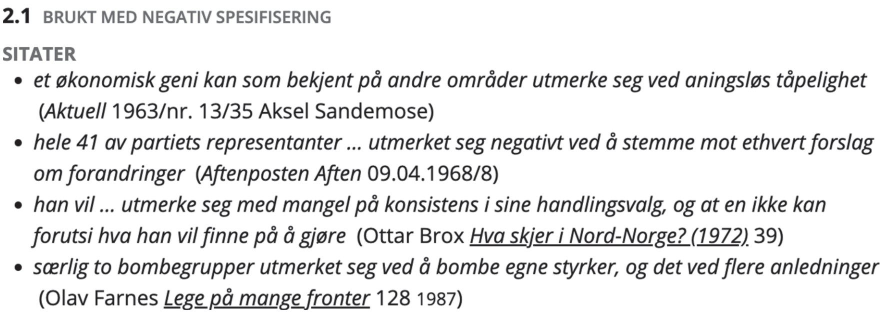
Download

Click on a row to go to the sentence. Mouse over a row to see the structures.

Treebank	*Document*	*Trans.*	*Id*	*Sentence*	
nob-naob_7	oai:nb.bibsys.no:998...	no	**465**	Han vil også utmerke seg ved mangel på konsistens i sine handlingsvalg, og at en ikke kan forutsi hva han vil finne på å gjøre.	Copy
nob-naob_7	oai:nb.bibsys.no:998...	no	**845**	Tilbake er et småbrukersamfunn som i mange tilfelle utmerker seg ved en markert mangel på sosial ulikhet	Copy

1	utmerke*seg	ved	oppfatning	Høigård, Einar
1	utmerke*seg	ved	oppførelse	Qvamme, Børre

Figure 15: Numbers of matching patterns with authors; examples from the author Ottar Brox are inspected.

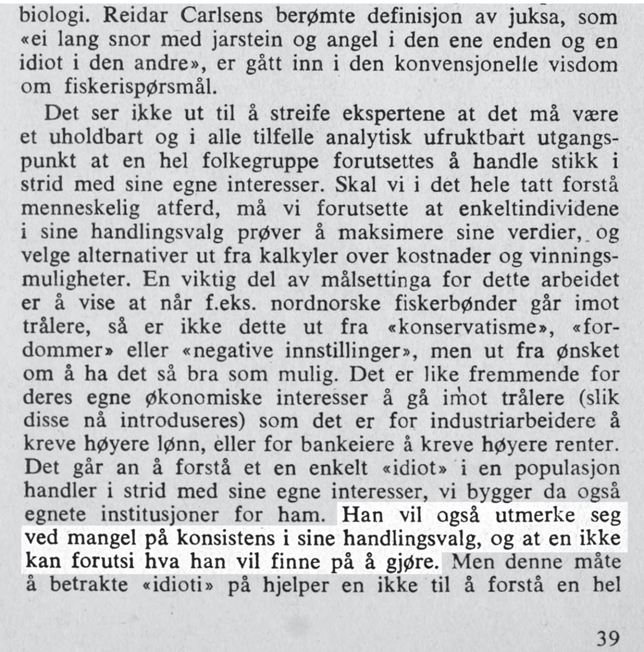


2.1 BRUKT MED NEGATIV SPESIFISERING

SITATER

- *et økonomisk geni kan som bekjent på andre områder utmerke seg ved aningsløs tåpelighet* (*Aktuell* 1963/nr. 13/35 Aksel Sandemose)
- *hele 41 av partiets representanter … utmerket seg negativt ved å stemme mot ethvert forslag om forandringer* (*Aftenposten Aften* 09.04.1968/8)
- *han vil … utmerke seg med mangel på konsistens i sine handlingsvalg, og at en ikke kan forutsi hva han vil finne på å gjøre* (Ottar Brox *Hva skjer i Nord-Norge? (1972)* 39)
- *særlig to bombegrupper utmerket seg ved å bombe egne styrker, og det ved flere anledninger* (Olav Farnes *Lege på mange fronter* 128 1987)

Figure 16: Part of the updated NAOB entry for the verb *utmerke*, listing examples of 'distinguishing oneself' in a negative way.



biologi. Reidar Carlsens berømte definisjon av juksa, som «ei lang snor med jarstein og angel i den ene enden og en idiot i den andre», er gått inn i den konvensjonelle visdom om fiskerispørsmål.

Det ser ikke ut til å streife ekspertene at det må være et uholdbart og i alle tilfelle analytisk ufruktbart utgangspunkt at en hel folkegruppe forutsettes å handle stikk i strid med sine egne interesser. Skal vi i det hele tatt forstå menneskelig atferd, må vi forutsette at enkeltindividene i sine handlingsvalg prøver å maksimere sine verdier, og velge alternativer ut fra kalkyler over kostnader og vinningsmuligheter. En viktig del av målsettinga for dette arbeidet er å vise at når f.eks. nordnorske fiskerbønder går imot trålere, så er ikke dette ut fra «konservatisme», «fordommer» eller «negative innstillinger», men ut fra ønsket om å ha det så bra som mulig. Det er like fremmende for deres egne økonomiske interesser å gå imot trålere (slik disse nå introduseres) som det er for industriarbeidere å kreve høyere lønn, eller for bankeiere å kreve høyere renter. Det går an å forstå et en enkelt «idiot» i en populasjon handler i strid med sine egne interesser, vi bygger da også egnete institusjoner for ham. Han vil også utmerke seg ved mangel på konsistens i sine handlingsvalg, og at en ikke kan forutsi hva han vil finne på å gjøre. Men denne måte å betrakte «idioti» på hjelper en ikke til å forstå en hel

39

Figure 17: Excerpt of the scanned page at the National Library of Norway containing the citation from Ottar Brox, accessed from a hyperlink in the updated NAOB entry.

7 Conclusion

Lexicographical work and language technology tools and resources are mutually dependent. On the one hand, a suitable lexicon is paramount for the development of natural language processing applications (Rosén 2014). On the other hand, corpora and related natural language processing tools and resources provide a wealth of information on patterns of lexis (Hasselgård, Ebeling, and Ebeling 2013) and can strongly support lexicographical work, as documented in this chapter.

For three ongoing national lexicographical projects in Norway, the CLARINO Bergen Centre has been providing access to data, tools, and know-how. The update of NO-AH is still in an initial phase, while the other projects are well under way. Although lexicographical resources or applications were not explicitly mentioned in the 2012 project plan that established CLARINO, experience shows that the data, tools and practices in CLARINO are adaptable to the needs of modern lexicography.

Lexicographers are highly dependent on source materials. Corpus resources at the CLARINO Bergen Centre have been made available through the INESS treebanking platform and through the Corpuscle corpus management and search tool. Both systems were further developed to better serve emerging needs. Corpuscle was extended specifically for lexicographical work as the bespoke platform Corpuscle-Lex. To make the advanced search facilities of INESS easier to use and more amenable to the needs of lexicographers, the search interface was adapted and augmented with query templates providing word sketches. Training in Corpuscle-Lex and INESS is given to all new dictionary editors.

Taken together, the infrastructure provides tools and services that simply did not exist before CLARINO, thereby improving a situation with fragmented source materials and unsolved copyright and technical issues. The work carried out within CLARINO with respect to harmonizing data formats and resolving restricted licenses has facilitated and increased the efficiency of the lexicographical work. Easy access to large materials in CLARINO and tools for analysing these data secures an empirical foundation which far exceeds the lexicographical resources and possibilities available only a few years ago. When language resources from Språksamlingane and other sources were included in our current lexicographic practice, best practices from CLARIN were also adopted. CLARIN license agreement templates are employed and if necessary adapted in order to include, deposit, curate, and deploy such resources for academic as well as dictionary development purposes.

The adaptability of the CLARINO infrastructure has been an enabling force for the tight integration of the CLARINO corpus tools with lexicographical editing tools, which makes for an efficient workflow. This would not have been possible without support from the experienced workforce in both Språksamlingane and CLARINO. A strategy must be set out to manage and sustain that combined work force in the future.

Bibliography

Aasen, Ivar. 1873. *Norsk Ordbog med dansk Forklaring [Norwegian dictionary with Danish definitions]*. Christiania: Mallings Boghandel.

Andersen, Gisle & Peder Gammeltoft. 2022. The role of CLARIN in advancing work in terminology: The case of Termportalen – the national terminology portal for Norway. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

De Smedt, Koenraad, Gunn Inger Lyse Samdal, Rune Kyrkjebø, Hemed Ali Hemed Al Ruwehy, Øyvind Liland Gjesdal, Victoria Rosén & Paul Meurer. 2016. The CLARINO Bergen Centre: Development and Deployment. *Linköping Electronic Conference Proceedings* 123: 1–12.

Dyvik, Helge, Paul Meurer, Victoria Rosén, Koenraad De Smedt, Petter Haugereid, Gyri Smørdal Losnegaard, Gunn Inger Lyse & Martha Thunes. 2016. NorGramBank: A 'Deep' Treebank for Norwegian. *LREC Proceedings*, pp. 3555–3562.

Hasselgård, Hilde, Jarle Ebeling & Signe Oksefjell Ebeling. (eds.) 2013. *Corpus perspectives on patterns of lexis*. Studies in Corpus Linguistics no. 57. Amsterdam/Philadelphia: John Benjamins.

Hovdenak, Marit. 2014. Bokmålsordboka og Nynorskordboka gjennom ein generasjon [The Bokmål dictionary and the Nynorsk dictionary spanning a generation]. *Nordiske studier i leksikografi* 12: 229–246.

Hovdenak, Marit, Laurits Killingbergtrø, Arne Lauvhjell, Sigurd Nordlie, Magne Rommetveit & Dagfinn Worren. (eds.) 2006. *Nynorskordboka: definisjons- og rettskrivingsordbok [Nynorsk dictionary: definition and spelling dictionary]*. 4th edn. Oslo: Det Norske Samlaget.

Kulbrandstad, Lars Anders. 1976. Norsk handordbok [Norwegian practical dictionary]. *Språknytt* 2: 7–8.

Landrø, Marit Ingebjørg & Boye Wangensteen. (eds.) 1986. *Bokmålsordboka: definisjons- og rettskrivningsordbok [Bokmål dictionary: definition and spelling dictionary]*. 1st edn. Oslo: Universitetsforlaget.

Lyse, Gunn Inger. 2020. Ut med 'adamsslekt' og inn med 'arveprinsesse'? Leksikografiske metodar i revisjonen av Bokmålsordboka og Nynorskordboka [Throw out 'Adam's kin' and take in 'princess heir'? Lexicographic methods in the revision of the Bokmål dictionary and the Nynorsk dictionary]. *Nordiske Studier i Leksikografi* 15: 215–224.

Meurer, Paul. 2012a. Corpuscle: A new corpus management platform for annotated corpora. In Gisle Andersen (ed.), *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*, Studies in Corpus Linguistics no. 49, 31–49. Amsterdam/Philadelphia: John Benjamins.

Meurer, Paul. 2012b. INESS-Search: A search system for LFG (and other) treebanks. *The proceedings of the LFG Conference*, pp. 404–421.

Meurer, Paul. 2020. Designing efficient algorithms for querying large corpora. *Oslo Studies in Language* 11 (2): 283–302.

Meurer, Paul, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, Gunn Inger Lyse, Gyri Smørdal Losnegaard & Martha Thunes. 2013. The INESS Treebanking Infrastructure. *Linköping Electronic Conference Proceedings* 85: 453–458.

Rauset, Margunn. 2019. Bokmålsordboka og Nynorskordboka: Einegga, toegga eller siamesiske tvillingar? [The Bokmål dictionary and the Nynorsk dictionary: One-egged, two-egged, or Siamese twins?]. *LexicoNordica* 26: 155–175.

Rosén, Victoria. 2014. Språkteknologiens behov for leksikalsk informasjon [Language technology needs for lexical information]. *Nordiske studier i leksikografi* 12: 13–41.

Rosén, Victoria, Koenraad De Smedt, Paul Meurer & Helge Dyvik. 2012. An Open Infrastructure for Advanced Treebanking. In Jan Hajič, Koenraad De Smedt, Marko Tadić & António Branco (eds.), *META-RESEARCH Workshop on Advanced Treebanking*, 22–29. Paris: ELRA.

Rosén, Victoria, Helge Dyvik, Paul Meurer & Koenraad De Smedt. 2020. Creating and exploring LFG treebanks. *The proceedings of the LFG Conference*, pp. 328–348.

Vikør, Lars S.. 2018. *Inn i Norsk Ordbok: Brukarrettleiing og dokumentasjon [Into the Norwegian Dictionary: User guide and documentation]*. Oslo: Det Norske Samlaget.

Vikør, Lars S., Olaf Almenningen, Reidar Bø, Oddrun Grønvik, Arnbjørg Hageberg, Tor Erik Jenstad, Laurits Killingbergtrø, Magne Myhren, Sigurd Nordlie, Gudrun Dahler Vik & Dagfinn Worren. (eds.) 2002. *Norsk ordbok: Ordbok over det norske folkemålet og det nynorske skriftmålet [Norwegian dictionary: Dictionary of the Norwegian spoken language and the Nynorsk written language]*. Volume 4. Oslo: Det Norske Samlaget.

Wangensteen, Boye. (ed.) 2005. *Bokmålsordboka: definisjons- og rettskrivningsordbok [Bokmål dictionary: definition and spelling dictionary]*. 3rd edn. Oslo: Universitetsforlaget.

Worren, Dagfinn. 1998. Om å avgrense eit ordtilfang: Soga om målføreorda i Nynorskordboka [About delineating a vocabulary: The saga of dialect words in the Nynorsk dictionary]. In Ruth Vatvedt Fjeld & Boye Wangensteen (eds.), *Normer og regler. Festskrift til Dag Gundersen 15. januar 1998*, 59–70. Oslo: Kunnskapsforlaget.