Ruth Vatvedt Fjeld and
Petter Henriksen:

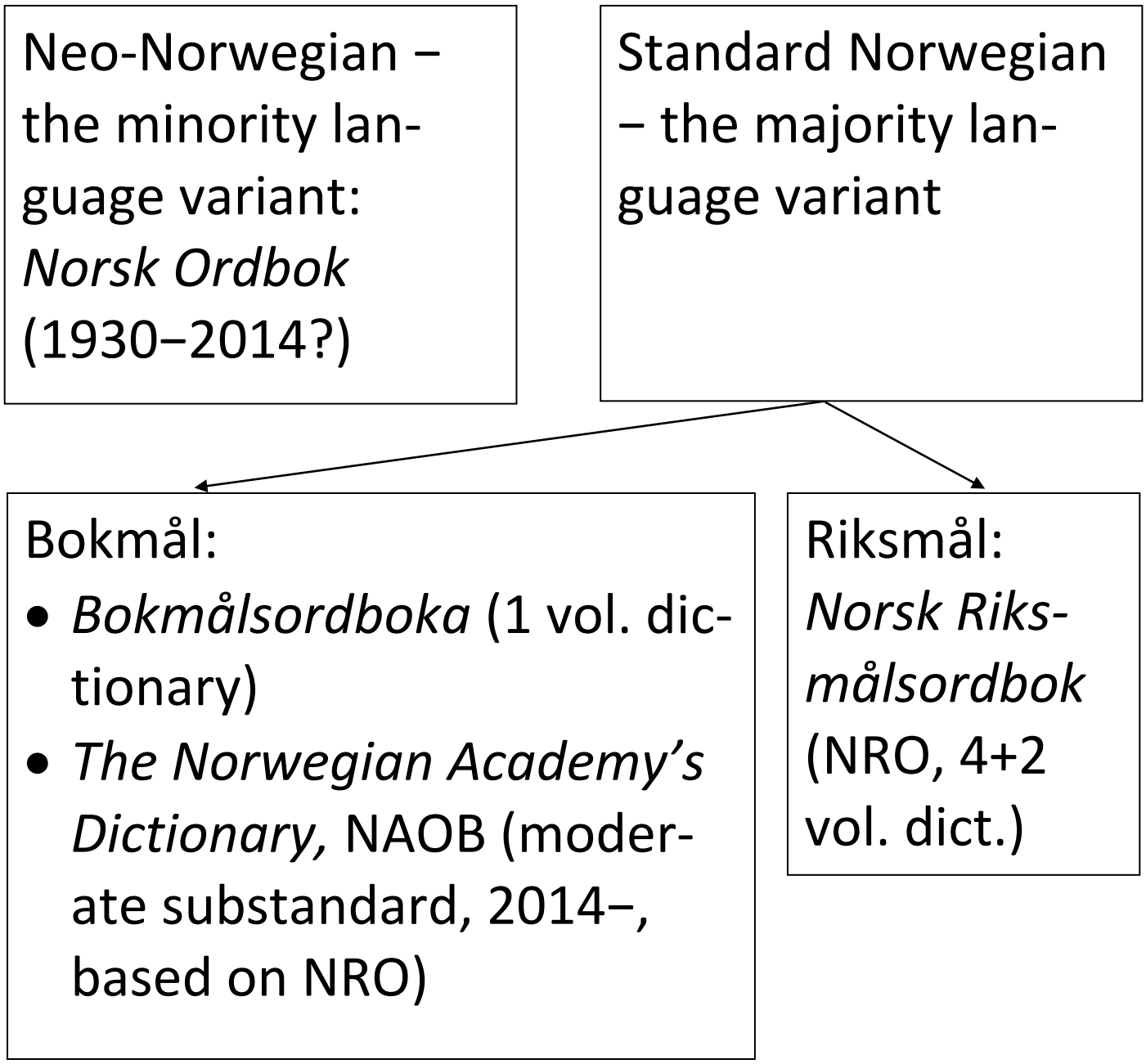
The BRO-project, a bridge in the wild, Norwegian linguistic land-scape

1 The language situation in Norway

Most nations have their hands full compiling one national dictionary. Norway, with its 5 million inhabitants, needs two, due to two official variant forms of Norwegian:

- Bokmål/Standard Norwegian
- Nynorsk/Neo-Norwegian

2 The dictionary situation in Norway



Riksmål – Private standard, linguistically conservative forms, from the Danish heritage
Bokmål – Official standard, most conservative and radical forms

3 The BRO-initiative

- To gather and unite the different standards that are not Neo-Norwegian in one dictionary.
- BRO = ‘bridge’, an acronym for ***B****okmålets og* ***R****iksmålets* ***O****rdbase*, 'The vocabulary database of standard Norwegian'.

4 The fundament for building a bridge

2005 – New official standard for bokmål, incorporating most *riksmål*-forms, which paved the way for cooperation in documenting the vocabulary in one common dictionary project.

5 The BRO-project rests on two piers

5.1 From the Academy side

- *Norsk Riksmålsordbok* (NRO, 1937-1995)

- *NAOB* (The Norwegian Academy's Dictionary (2014?–)

NAOB is a revision and modernisation of NRO:

- Published solely on the Internet
- Semantical updating
- Simplification of complex hierarchies of meanings
- Adjustment of politically or socially obsolete viewpoints
- Quotes from recent Norwegian literature
- Addition of new entries

NAOB's content is xml-coded, parsed by a stringently lexicographical DTD, resulting in articles with an easily understandable publishing interface. Given sufficient funding, NAOB will be published in 2014. NAOB will offer a state-of-the-art technical platform for future updating.

5.2 From UiO's side

- *Bokmålsordboka*
- *Norsk Ordbank,* a list of the standard bokmål vocabulary with full flexion forms
- *Leksikografisk bokmålskorpus (LBK),* a balanced, electronic text corpus aiming at 100 mill. words

6 Challenges in bridging the gap

- The Section for Bokmål Lexicography, funded by the Norwegian government, is committed to documenting the whole scope of

official bokmål. *Bokmålsordboka* includes all forms in the official standard, as well as combinations of morphological variant forms in derivations and compounds, with no indications of stylistic value or usability (= descriptive dictionary).

- The Norwegian Academy for Language and Literature, partially funded by the Norwegian government, seeks in NAOB to publish an alternative-free moderate substandard of bokmål, in addition to documenting the forms used in Norwegian literature back to the early 1800-hundreds (= descriptive/normative dictionary).

The two parties' backgrounds and ideologies present the challenge of incorporating both conservative and radical forms in one coherent dictionary inside a common lemma list.

7 Solution

Issues of normativity and usability will be solved by means of corpus examination of the use of different forms and mix of forms. The frequency in use will be found in LBK, and the number of hits will be given in the dictionary.

BRO will be given an interface which enables users to choose variant forms and substandards as preferences for entry forms and morphological variants.

7.1 Lemmatization of variant forms

NAOB uses the most common variant in written, formal contexts as lemma sign form (labeled the NAOB norm), i.e. the entry word for the main article, and sets up links from alter-

native forms. *Bokmålsordboka* lemmatizes all forms in their alphabetical order.

The difference can be exemplified by the Norwegian word for bridge, **bro**, and its parallel form **bru (**with a different stylistic value). In an all-inclusive bokmål dictionary without a recommended subnorm, they would be presented as variants of the same lemma in their rightful alphabetical place, with definition by the first word lemmatized.

An unmarked way to organize these articles, is like this:

bro el. **bru** + full article
bru el. **bro** + full article

7.2 Noun gender and noun morphology

Nouns have three genders in bokmål: masculinum, femininum and neutrum.

In riksmål femininum is only accepted in a handful of typically “Norwegian” words, like *jente, øy, bikkje*.

Our example of *bro/bru* also is affected in this respect. Here lies one of the main differences between bokmål and riksmål.

Bokmålsordboka contains approximately 10.700 feminine nouns, all needing double gender marking and double morphological tables:

bru el. **bro** f1 el. m1

f1 means femininum class 1, inflected according to this pattern:

brua	broa
bruer	broer
bruene	broene

m1 means masculinum class 1, inflected:

broen	bruen

broer	bruer
broene	bruene

NAOB only presents the masculine inflection for *bro.* The LBK-corpus shows that all forms are used in modern written Norwegian.

7.3 Verb morphology

Bokmål allows alternative ways of tempus inflection of verbs, especially in the weak forms:

kaste (throw) *–kastet–kastet* ("moderate")
or
kaste–kasta–kasta ("radical")

BRO will present full morphology for all verbs, accompanied by frequency information from the LBK-corpus.

7.4 Compounds and derivations

All word forms with variants in bokmål may be given with frequency information for all inflected forms:

	LBK	NoTa
bortkasta	20	5
bortkastet	179	2
bortkastede	24	1
bortkastete	3	0

Here the written corpus frequency may be compared with frequency in NoTa – a corpus of spoken Norwegian, indicating that the so-called radical variant is most common in spoken Norwegian, and the conservative variant is most common in written Norwegian.

7.5 Multiword expressions

Bokmålsordboka has three multiword expressions listed for the concrete meaning of *bro/bru*:

1. *b-a over elva* ("the bridge over the river")
2. *bygge b-* ("build a bridge")
3. *bryte alle b-er (bak seg)* ("break all bridges")

1 shows the semantic meaning of the lemma.
2 is a semantic collocation.
3 is more of an idiom than an example.
These will be marked as different information types.

Another task is identifying and lemmatizing the most frequent expressions. The corpus shows that *bryte alle broer* has only one hit in LBK, while the idiom *brenne alle broer* (burn all bridges) has 19 hits and would have been

more adequate in a dictionary documenting common use.

In BRO a set of grammatical collocations will be documented for all relevant lemmas found semiautomatically by means of corpus analysis of LBK . A statistical analysis through DeepDict Lexifier (Bick 2009) shows that the most common collocations containing *bro*, are:

V+N-collocation: *bygge, passere, sprenge, krysse, brenne en bro*

A+N-collocation: *sukkenes bro, usynlig bro, provisorisk bro, smal bro, gigantisk bro*

N+PP postmodifiers: *bro over kløft, elv, avgrunn, vollgrav*

8 BRO as a product of Norwegian language history and national identity

The wide norms for orthography and morphology in Norwegian are a heritage from the

Danish linguist Rasmus Rask, who launched the so called orthophonic principle in orthography in 1826. Rask's principle had a great impact on Norwegian language planning, as we can see from the first Norwegian standardized norm from 1862. Since Norway is a large, rugged and sparsely populated country, it has developed more dialect variation than most other nations. To unify as many dialects as possible inside one standard, the first norm from 1907 was made relatively open for orthographic variant forms.

The dialects or spoken written standards also came in use among the establishment in the central cities and through the 20th century, identified as upper class sociolects. Which forms you chose inside the open norm showed which social class you belonged to.

Consequently, determining which forms were accepted inside the norm, was not only a lin-

guistic issue, it was also a general political issue that made way for the famous – and to other nations mystifying – Norwegian language struggle.

Norway has been a community with small differences in wealth and status, and we have had much less of an aristocracy to set its indisputable upper class sociolect as a standard norm for orthography and phonology, than most other European linguistic societies. In Norwegian, all dialects are generally accepted in all ways of life, and pronunciation is seldom given in dictionaries, to avoid any discrimination or elitism. Many poets use their own dialect in their lyrics, arguing that their dialects are their "heart language", enabling them to express their inner, personal attitudes in the best way.

To edit a common dictionary presenting all norms and subnorms in such a linguistic cli-

mate is a difficult task. In light of this we hope to make an interface for the dictionary where the users may make their own choices between several substandards before entering the dictionary. This, of course, must not prevent the interface from showing all variants, among other things a most useful tool for teachers when correcting their pupils' work.

It is said that Norway is a society of individualists where all members are equal. This might seem like a utopia, but the BRO-dictionary will try to satisfy all lexicographical requirements to embody this typically Norwegian form of national identity.