Word banks, dictionaries and research results by the roadside

Oddrun Grønvik[1], Trond Minde[1], Christian-Emil Ore[2]

[1]University of Bergen, [2]University of Oslo

Euralex 2022

Mannheim 13.07.2022

The larger project

- Systematize heterogeneous source materials for lexical and lexicographic description of Norwegian language (cf. Euralex 2018 paper)
- Study the development of Norwegian orthography from 1860 onwards (cf. Euralex 2020 paper)
- Problem: Modern morpho-syntactic analyzers do not perform well on pre-1938 texts
- Develop a register of historical, inflected forms as a basis for a computer-based analysis of (literary) texts from 1860 to present
  - (e.g. Cost Action, CA16204: Distant Reading, novels 1859 – 1920)

- The unplanned bonus – the roadside findings:
  - increased insights in Ivar Aasen's lexicographical method.
  - These will not be dealt with in this presentation
  - They are covered in the published paper.

Content

- Methods
- The linguist and lexicographer Ivar Aasen
- The source material
- The word bank for 19th c. Nynorsk
- Test and results

Identifying historical orthographies 1

- Analyse corpora of selected and dated texts
  - Create a overview of lexical items and their realizations
  - Establish a tentative norm with base forms and inflected forms
- A standard method for a lexicographic description of a language

Identifying historical orthographies 2

Use documented orthographical norms

- Use relevant, authorized spellers and dictionaries closest in time (before) the period of interest
- Use grammars from the same period to create inflectional paradigms
- Can be used to study how far a given norm is used in texts

Drawback: The spellers and dictionaries usually have a low number of lexical items compared to real language, resulting in a low degree of coverage in the contemporary texts

Bonus: lexical items are identified more securely when supported by a definition

A register of lexical items with historical, inflected forms

Combine the two approaches

- Use the documented orthography (word lists and grammars)
- Analyse corpus of selected texts from the period of interest

The two approaches complement each other

- both approaches require a lot of work
- together they create useful tools for text analysis

Future plan:

- Use the result of the analysis of texts repeatedly to expand the lemma list
- Implement a compound analyser
- Syntax based disambiguater

A few words about Ivar Aasen (1813–1896)

A farmer's son and an autodidact, inspired by the development of comparative and historical linguistics

Studied

- the works of Rasmus Rask (1787–1832, Danish philologist and a principal founder of the science of comparative linguistics), and Jakob Grimm (*Deutsche Grammatik*)
- Old Norse, Latin, Greek and contemporary languages

In 1840, Aasen was tasked with

- documenting the Norwegian vernacular through dialect studies,
- testing the hypothesis that the Norwegian language stemmed from Old Norse and was not decayed Danish
- publish the results in the form of a grammar and a dictionary (Aasen 1848 and 1850)

Extended grammar (1864) and extended dictionary (1873)

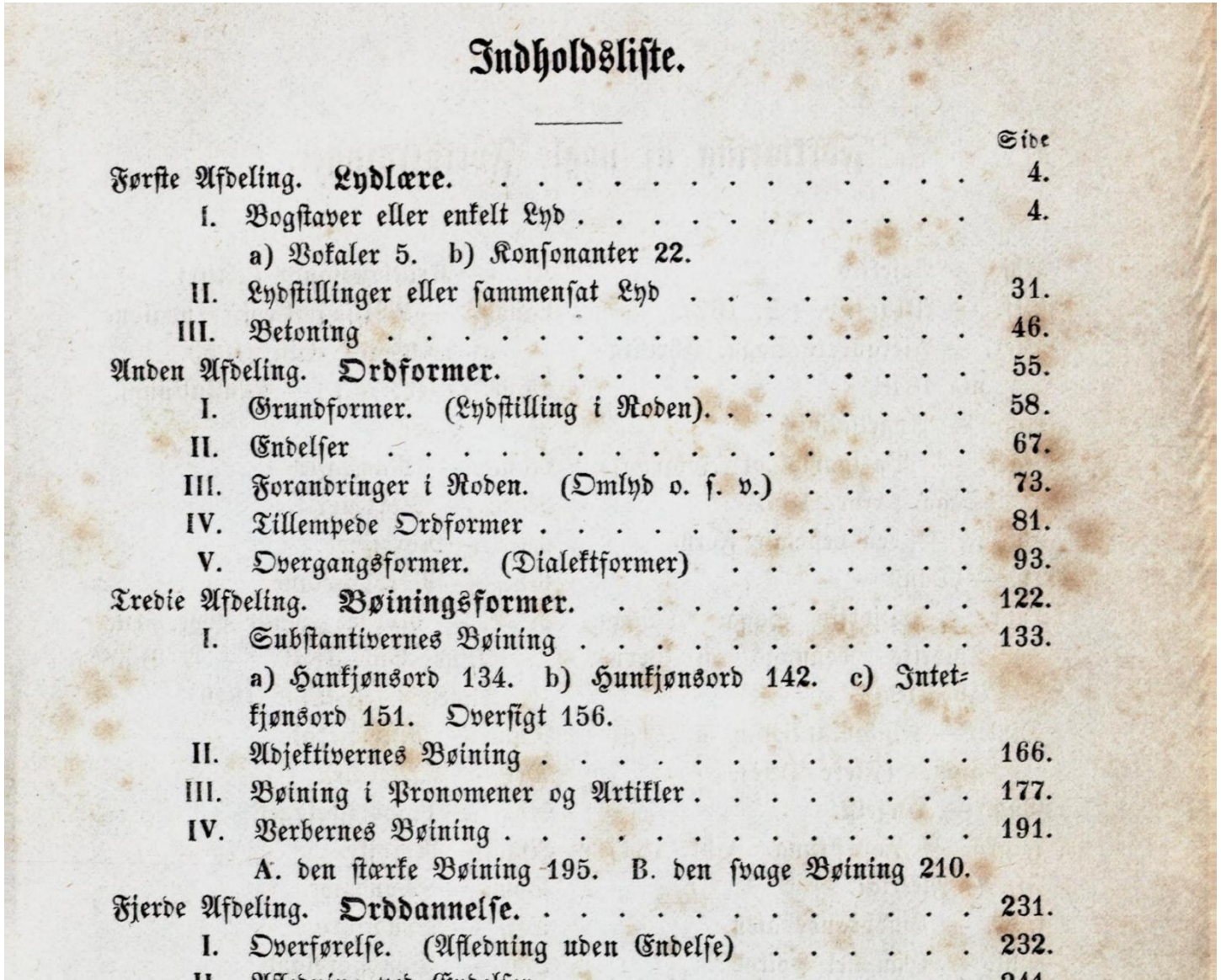
Source material – grammar 1864

1 Sound system

2 Word forms
- Base forms
- Suffixes
- Root changes
- Dialect forms

3 Inflectional forms
- Nouns
- Adjectives
- Verbs

4 Word formation

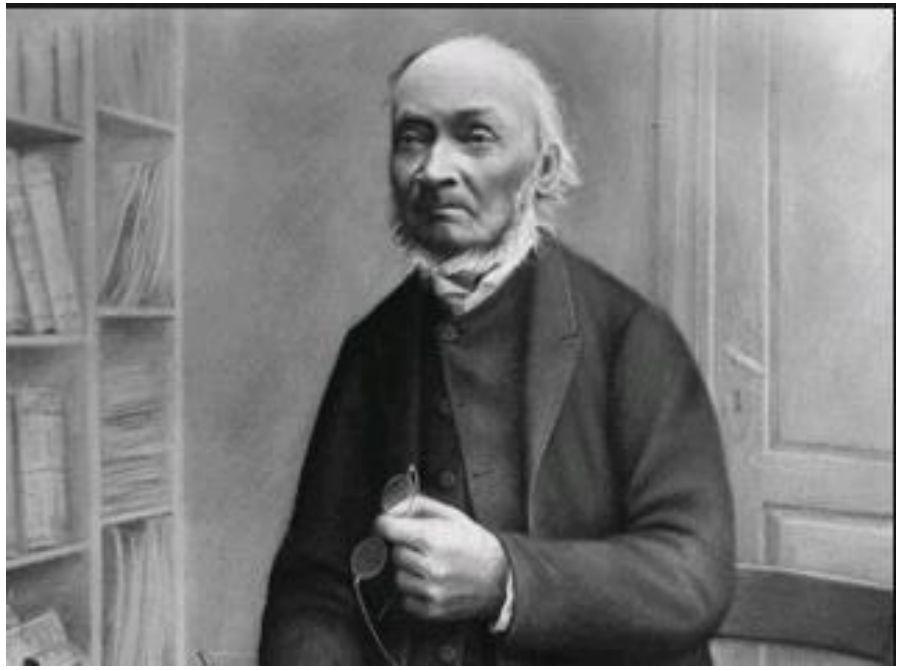


Indholdsliste.

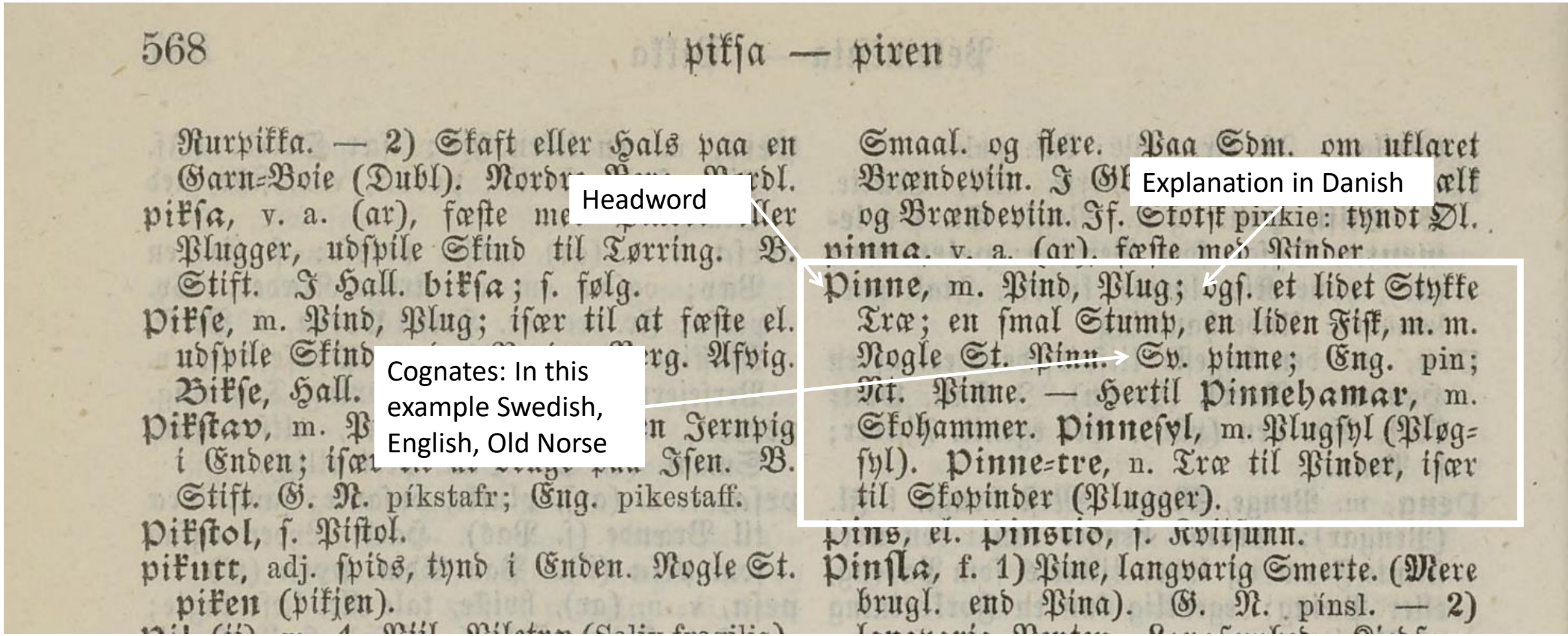
Ivar Aasen and word formation

Norwegian Grammar (1864)

- A chapter on word structure in Norwegian
- A prerequisite for a systematic treatment of the vocabulary and the inflectional system
- A treatise on the semantic potential of suffixes for derivation and inflection
- Aasen's grammars and dictionaries founded Nynorsk as a written standard and have defined all later research on the vernacular in Norway



Source material – dictionary 1873



568 pikſa — piren

Rurpikka. — 2) Skaft eller Hals paa en Garn=Boie (Dubl). Nordr[illegible]dl.
pikſa, v. a. (ar), fæſte me[illegible]er Plugger, udſpile Skind til Tørring. B. Stift. I Hall. **bikſa**; ſ. følg.
Pikſe, m. Pind, Plug; iſær til at fæſte el. udſpile Skind [illegible]rg. Afvig. **Bikſe**, Hall.
Pikſtav, m. P[illegible]n Jernpig i Enden; iſær [illegible] Iſen. B. Stift. G. N. pikstafr; Eng. pikestaff.
Pikſtol, ſ. Piſtol.
pikurr, adj. ſpids, tynd i Enden. Nogle St. **piken** (pikjen).

Smaal. og flere. Paa Sdm. om uklaret Brændeviin. I Gl[illegible]ælk og Brændeviin. Jf. Storſk pinkie: tyndt Dl.
pinna, v. a. (ar), fæſte med Pinder.
Pinne, m. Pind, Plug; ogſ. et lidet Stykke Træ; en ſmal Stump, en liden Fiſk, m. m. Nogle St. Pinn. Sv. pinne; Eng. pin; Nt. Pinne. — Hertil **Pinnehamar**, m. Skohammer. **Pinneſyl**, m. Plugſyl (Pløg=ſyl). **Pinne=tre**, n. Træ til Pinder, iſær til Skopinder (Plugger).
Pins, el. **Pinstid**, ſ. Kvitſunn.
Pinſla, f. 1) Pine, langvarig Smerte. (Mere brugl. end Pina). G. N. pinsl. — 2)

Headword

Explanation in Danish

Cognates: In this example Swedish, English, Old Norse

Source material – in numbers

Norwegian Dictionary (Aasen 1873)

- Digitized as a formatted text in 1996
- Semi-automatic analysis and a TEI markup in 2016
- 42 000 headwords identified and marked up
- 10 000 additional dialect forms identified inside the text of the entries
- Linked and published in the Meta dictionary system (cf Euralex 2018)

The digital edition was used to create a skeleton for the 19th c. Nynorsk Word Bank

Norwegian Grammar (Aasen 1864)

- Electronic text exists, but no computer based analysis
- contains inflection schemas for all inflected POS subgroups
- reference point for paradigm creation

The word banks for modern Norwegian

Originally based on IBM's spell checker system from 1990:

Base forms

- Linked to the relevant entries in various editions of dictionaries
  - the Norwegian monolingual dictionaries *Bokmålsordboka* and *Nynorskordboka*

Inflectional patterns (paradigms)

Links between base forms and paradigms

- Status in a norm
- Time span for the status

POS categories are in accordance with

- the Norwegian reference grammar (1997)
- The recommendation of the Norwegian Language Council/Directorate for Education and Training (2005)

The Ivar Aasen word bank, 19th c. Nynorsk

- A historic-orthographic registry of inflected forms, to be used as a tool for analysing texts from the second part of 19th c.
- Uses the structure of the word banks for modern Norwegian:
  - base forms linked to Ivar Aasen's dictionary (1873)
  - Inflection patterns (paradigms) with references to the grammar (1864)
- The register is freely available to all (CC BY 4.0), like the other parts of the Norwegian Word Bank

The overall structure of a modern word bank

(Sample: ‘bok’ = ‘book’)

(1) Base forms

Lemma-id	Base form
…	…
8701	bok
…	…

Lemma-id	Para-id	Norm	From	To	…
…	…	…	…	…	…
8701	942	yes		2012	…
8701	942	no	2012	9999	…
8701	968	yes		9999	…
…	…	…	…	…	…

(2) Base forms, their paradigms and norm status

(3) Rewriting rules for each paradigm

Para-id	Line	Feature	Rewriting rule	Example
…	…	…	…	
942	1	sg indef	o+	bok
942	2	sg def	o+i	boki
942	3	pl indef	ø+er	bøker
942	4	pl def	ø+ene	bøkene
…	…	…	…	
968	1	sg indef	o+	bok
968	2	sg def	o+a	boka
968	3	pl indef	ø+er	bøker
968	4	pl def	ø+ene	bøkene
…	…	…	…	

Para-id	POS	Details	Example	
…	…	…	…	…
942	subst fem appell	Omlyd O/Ø	bok	…
968	subst fem appell	Omlyd O/Ø	bok	…
…	…	…	…	…

(4) Detailed information about each paradigm

Aasen-norm – extended noun paradigm

(Samle: ‘bok’ = ‘book’)

P-id	Linje	Rewriting rule	feature	explanation	Inflected form
942	1	o+	sg indet	singular indet	bok
942	2	o+i	sg det	singular det	boki
942	3	ø+er	pl indet	plural indet	bøker
942	4	ø+erna	pl det	plural det	bøkerna
942	5	o+enne	dat sg	dativ singular	bokenne
942	6	o+om	dat pl	dativ plural	bokom
942	7	o+ar	gen sg	genitiv singular	bokar
942	8	o+a	gen pl	genitiv plural	boka

P-id	POS	Details	Example	Justification	
…	…	…	…		…
					…
942	subst fem appell	OmlydV O/Ø	bok	Aasen 1864 §171	…
…	…	…	…		…

What is achieved?

In the first part of 2021 Trond Minde and Oddrun Grønvik examined the entire set of headwords in Aasen 1873

As a result we have a complete word bank for the ‘Aasen-norm’

- 42 901 base forms
- 468 372 inflected forms
- 412 paradigms (inflectional schemes)

The corresponding numbers for the modern Nynorsk standard (2012)

- 105 600 base forms
- 558 500 inflected forms
- 580 paradigms (inflectional schemes)

A first test of coverage

Aasmund Olavsson Vinje (1818–1870)

- Norwegian writer, journalist and lawyer
- In close contact with Ivar Aasen

Material: Collected writings I – V

- Danish texts in vol. I were removed
- Mostly factual prose

A simple test

- Every token (occurrence of a word form) in the texts was checked against a list of full forms from the Aasen-norm (1864) and from the Nynorsk-norm (2012)

Results of the test

	All word forms (tokens)		Unique word forms (types)	
In both:	518 977	77 %	10 968	23 %
Only in Aasen:	35 371	5 %	4 572	10 %
Only modern Norwegian (2012):	44 409	7 %	7 837	16 %
Totals, found:	598 757	88 %	23 377	49 %
In neither:	79 312	12 %	24 493	51 %
Running word forms	678 069		47 870	

The central (most frequent) vocabulary of Vinje is well covered by the Aasen Word Bank.
The most important expansion potential lies in the word forms found in neither word bank.

Vinje word forms outside the Aasen norm

A.O.Vinje's prose is journalistic and related to events of the day.

- Danish word forms (often in quotations): *af*, *jeg, efter*
- Non-Germanic imported word forms: *nationale, encyclopedi, epikureisk*
- names of persons and places: *Napoleon, Bretland, Aftenbladet*
- word forms consistent with Aasen's orthography, but not in the dictionary (mainly compounds, some derived forms): *ferdastav, folkeavrøysting, gatestein*
- vernacular word forms specific to A.O. Vinje: *ikki, somykit, eigong*

The missing word-types in the Aasen Word Bank follow from

- Aasen's mandate of documenting the Norwegian vernacular
- the concentration on the basic vocabulary and word formation system

What we have achieved and future work

- The full form list based on Aasen 1873 and 1864 is a good starting point for analysing vocabularies from the Norwegian vernacular
- The Vinje test shows a good coverage of Aasen 1873 for the most frequent vocabulary (and mostly non-compounds)
- Vinje used imported words and create many compounds. These are not in Aasen 1973. Partly found in the modern Word Bank
- Germanic languages have productive compounding and it is well known that the number of word forms in a corpus is much larger than the number of headwords in any dictionary
- Next step – analyser for compounds and a syntactic tagger based on a extended word list and corpus consisting of early Norwegian texts (second half of the 19th c.)

Aasen-word bank

- The database
  - https://usd.uib.no under the category ‘Språk’
- The list of full forms
  - https://www.edd.uio.no/prosjekt/
- The dictionaries of I. Aasen (1874) og H. Ross (1895)
  - https://www.edd.uio.no/aasen_ross/aasen_ross.html
- The dictionary of I. Aasen (1873), facsimile
  - https://www.nb.no/items/84ec2f60981ea9b9e23a8973d3ad030a
- The grammar of I. Aasen (1864), facsimile
  - https://www.nb.no/items/6246e4cae7243d6ad5672e9ea66796d1

Literature

Aasen, I. (1848): *Det norske Folkesprogs Grammatik.* Christiania. Trykt hos Werner & Comp. [book]
Aasen, I. (1850): *Ordbog over det norske Folkesprog*. Kristiania. Det Kongelige Norske Videnskabers Selskab. Trykt hos Carl C. Werner & Comp. [book]
Aasen, I. (1864): *Norsk Grammatik*. Kristiania. Mallings Bogtrykkeri. [book]
https://www.nb.no/items/6246e4cae7243d6ad5672e9ea66796d1
Aasen, I. (1873): *Norsk Ordbog. Med dansk Forklaring*. Christiania. Mallings Boghandel. [book]
https://www.nb.no/items/84ec2f60981ea9b9e23a8973d3ad030a
Aasen, I. (1925): *Norsk maalbunad*. Samanstilling av norske ord etter umgrip og tyding. Oslo. Det Norske Samlaget. [book]
Engh, J. (2014). IBMs leksikografiske prosjekt for norsk 1984–-1991. *Maal og Minne* 106 (1) : 67–101. (http://ojs.novus.no/index.php/MOM/article/view/225. (accessed: 25-03-2022)) [book chapter]
Grimm, J. (1819-1837): *Deutsche Grammatik*. Göttingen. Dieterichsche Buchhandlung. [book]
Hagen, K., Johannessen, J.B. and Nøklestad, A. (2000). A Constraint-Based Tagger for Norwegian. *17th Scandinavian Conference of Linguistics, Volume I,* no. 19. Odense Working Papers in Language and Communication. [journal contribution]
Hagen, K. & Nøklestad, A. (2010). Bruk av et norsk leksikon til tagging og andre språkteknologiske formål. *LexicoNordica, (17)* (https://tidsskrift.dk/lexn/article/view/18624 (last access: 25-03-2022)) [journal contribution]
Hovdenak, M. & al (1986, 3. ed. 2006): Nynorskordboka. definisjons- og rettskrivingsordbok. Oslo. Samlaget. [book]
Nøklestad, A. (2022) The Oslo Bergen Tagger, https://github.com/noklesta/The-Oslo-Bergen-Tagger (last access: 25-05-2022)
Nøklestad, A. (2022) The Compound analyzer software, https://github.com/textlab/mtag (last access: 25-05-2022)
Ore, C.-E. S. (2016): Gamle ordbøker og digitale utgaver. *Nordiske Studier i Leksikografi 13*. Rapport fra 13. Konference om Leksikografi i Norden København 19.-22. mai 2015. København. Nordisk forening for leksikografi. pp. 203-216. [conference proceedings contribution]
Roget, P. M. (1852): *Thesaurus of English Words and Phrases* Classified so as to Facilitate the Expression of Ideas and to Assist in Literary Composition. London. Bloomsbury.
Vinje, Aa. O. (1859): Det norske Landsmaals vigtigaste Bøygningsformer. Bilag til *Dølen* nr. 30, 1859.
Vinje, Aa. O. (1916–1921): *Skrifter i Samling I–V*. Oslo. J.W. Cappelens forlag [book]
Wangensteen, B. (1986, 3. ed. 2005): Bokmålsordboka. Definisjons- og rettskrivningsordbok. Oslo. Kunnskapsforlaget. [book]