The Árni Magnússon Institute for Icelandic Studies

A new resource of Icelandic sagas
Digitizing normalized scholarly editions and enhancing textual data

Ellert Þór Jóhannsson, Finnur Ágúst Ingimundarson, Þórður Ingi Guðjónsson

The Árni Magnússon Institute for Icelandic Studies
E-mail: etj@hi.is, fai@hi.is, thingi@hi.is

Abstract
This poster accounts for a new project which aims to digitize and make available normalized scholarly editions of Old Icelandic texts. The project is based on the print publication of the Icelandic family sagas by the Old Icelandic Text Society, which consists of 14 volumes published in the period 1933–1991. The texts are digitized and subsequently processed as a tagged and lemmatized language corpus. The material is further enhanced by providing a link to a leading dictionary resource (ONP) which in addition to lexical information about individual words contains detailed information about source manuscripts. Another enhancement involves an inflectional database providing detailed information about morphological forms and inflectional patterns of each word.

1. Introduction

The project involves digitizing Old Icelandic text editions and further enhancing the resulting data. This includes:

- Digital text editions and creating a lemmatized corpus
- Linking lemmas to a lexicographic resource
- Developing comprehensive morphological database.

The aim is to make the material accessible online on a user-friendly platform. The initiative seeks to provide scholars, researchers, and language enthusiasts with a dynamic resource for exploring and understanding Old Norse vocabulary and linguistic structures in a detailed and straight forward manner.

2. Background

Icelandic family sagas represent the peak of Icelandic medieval writing culture. Until now these sagas have only been freely available online in translations or in normalized form compatible with Modern Icelandic language standards. The mould of Modern Icelandic orthography as well as morphology, eliminates various nuances present in the original Old Norse texts.

Íslenzka fornritafélagið (Old Icelandic Text Society = OTS) was founded in 1928 and has ever since worked on publishing Icelandic medieval texts in print as part of the *Íslenzk fornrit* series, with detailed introduction and textual commentary. The editions of the society use their own normalization standard, which reflects the state of the language around the year 1200. The text editions were produced before the digital era and have only been available in print. The current project aims at facilitating access to the texts and creating a versatile online resource for the study of Old Norse in a standardized form.

Figure 1
The website of the Old Icelandic Text Society: *www.fornrit.is*

3. Methods and enhancements

- **Digital texts and a lemmatized corpus:**

A digital version of the texts is generated through an innovative and iterative process. First a handful of digitally available Old Icelandic saga texts were manually corrected in accordance with OTS standards. Subsequently these texts were aligned with data from the Saga Corpus (SC). SC contains Modern Icelandic transcriptions of the family saga texts that have been tagged and lemmatized with high-level of accuracy (Rögnvaldsson et al. 2011).

ch	p	s	modern	lemma	pos	old
78	12	1	vil	vilja	sfg1en	vil
78	12	1	ég	ég	fp1en	ek
78	12	1	að	að	c	at
78	12	1	við	ég	fp1fn	vit
78	12	1	förum	fara	svg1fn	farim
78	12	1	eina	einn	tfveo	eina
78	12	1	leið	leið	nveo	leið
78	12	1	bæði	bæði	fohfn	bæði

Figure 2
A table showing alignment of Modern Icelandic word forms with their corresponding Old Icelandic forms in a sentence from Egils saga

This approach is based on the fact that tags for Modern Icelandic almost always can be applied to Old Icelandic, i.e. the grammatical structure is the same, with only a few minor exceptions that are easy to account for.

Aligning the information from the SC data results in digital Old Icelandic texts that contain grammatical information from Modern Icelandic transcriptions. The aligned and labelled data is then used to re-generate Old Icelandic versions of the remaining family saga texts with the missing vocabulary added manually and gradually, text by text. With this iterative approach, the missing vocabulary is reduced with each step as the lemmatized corpus expands.

- **Linking lemmas to** *A Dictionary of Old Norse Prose* (ONP):

The lemmatized texts can be linked to ONP on a headword-level, integrating lexicographic and manuscript information from an authoritative resource.
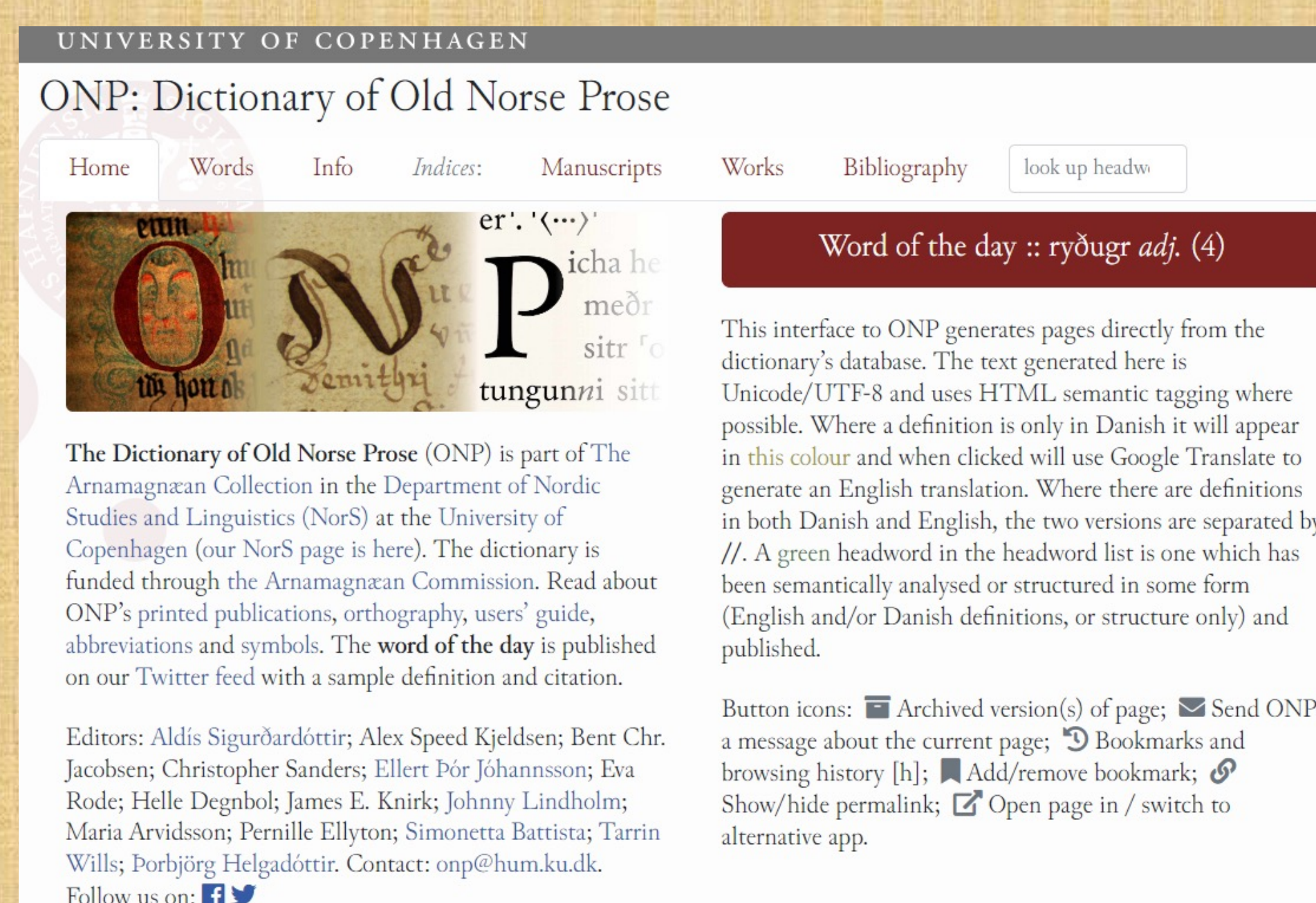

Figure 3
Screenshot from ONP Online

This linking enables users to cross-reference each words with their definitions, contextualizing the language within a linguistic framework and the broader context of other medieval text genres.
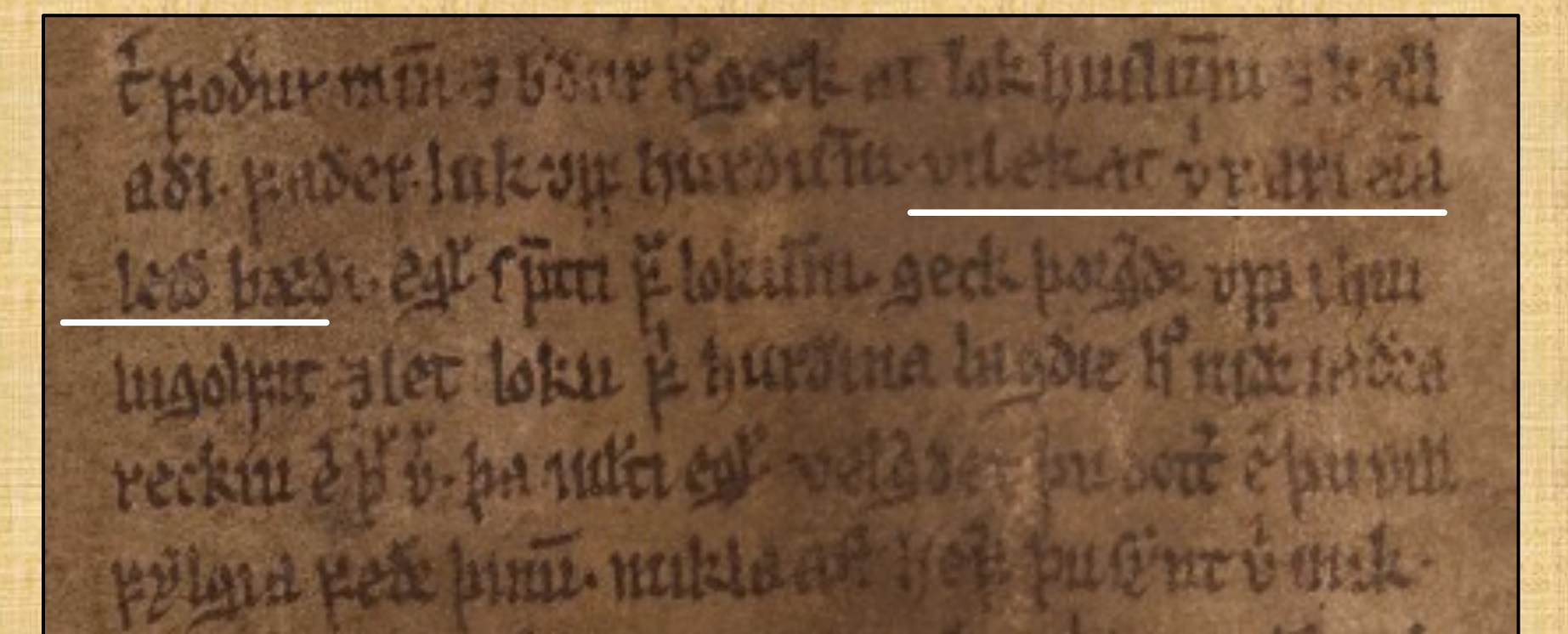

Figure 4
The manuscript source of the Egils saga text from *Figure 2* available from the online manuscript catalogue *handrit.is* through ONP.

- **Developing a Morphological Database:**

The database will showcase the inflectional forms for each lemma, allowing users to explore paradigmatic variations (inspired by the work on BÍN/DIM cf. Bjarnadóttir et al. 2019).

Traditional handbooks of Old Norse Morphology (e.g. Noreen 1923 & Iversen 1990) give only representative paradigmatic examples of selected words. Since the proposed database will contain all attested morphological patterns it provides users with comprehensive inflectional information about all occurring word forms in the saga texts.

	sg.	pl.
nom.	maðr	menn
acc.	mann	menn
dat.	manni	mǫnnum
gen.	manns	manna
nom.	maðrinn	menninir
acc.	manninn	mennina
dat.	manninum	mǫnnunum
gen.	mannsins	mannanna

Figure 5
Morphological database displaying all forms of the noun *maðr* 'man'

4. Conclusion

The result will be a useful web-resource for learning and teaching Old Norse. The platform will offer users new and innovative ways to explore original texts in a standardized form closer to the linguistic state of their composition and thereby gain deeper insight into Old Norse language and literature.

Bibliography

BÍN = *Beygingarlýsing íslensks nútímamáls*. Kristín Bjarnadóttir (ed.). Reykjavík: Stofnun Árna Magnússonar í íslenskum fræðum. <https://bin.arnastofnun.is/> (April 2024)

Bjarnadóttir, Kristín, Kristín Ingibjörg Hlynsdóttir og Steinþór Steingrímsson. 2019. DIM: The Database of Icelandic Morphology. *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pp. 146–154. (NoDaLiDa 2019, Turku, Finnland).

DIM = *Database of Icelandic Morphology* <bin.arnastofnun.is/DMII/LTdata/> (April 2024).

Iversen, Ragnvald. 1990. *Norrøn grammatikk*. 7. utg. revidert ved E.F. Halvorsen. Oslo: Tano.

Noreen, Adolf. 1923. *Altnordische Grammatik I: Altisländische und altnorwegische Grammatik (Laut- und Flexionslehre) unter Berücksichtigung des Urnordischen*. 4th edn.. Halle: Niemeyer.

ONP Online = *Ordbog over det norrøne prosasprog/A Dictionary of Old Norse Prose*. <onp.ku.dk> (April 2024).

Rögnvaldsson, Eiríkur, and Sigrún Helgadóttir. 2011. Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change. In Sporleder, Caroline, Antal P.J. van den Bosch og Kalliopi A. Zervanou (eds.): *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*. pp. 63–76. Springer, Berlin.

SC = *Saga Corpus* <https://clarin.is/en/resources/sagacorpus/> (May 2024).