ONCO

Světlana Müllerová

David Šimeček

Ondřej Fúsik

LINDAT CLARIAH-CZ

MINISTRY OF EDUCATION, YOUTH AND SPORTS

FILOZOFICKÁ FAKULTA
Univerzita Karlova

What is a corpus?
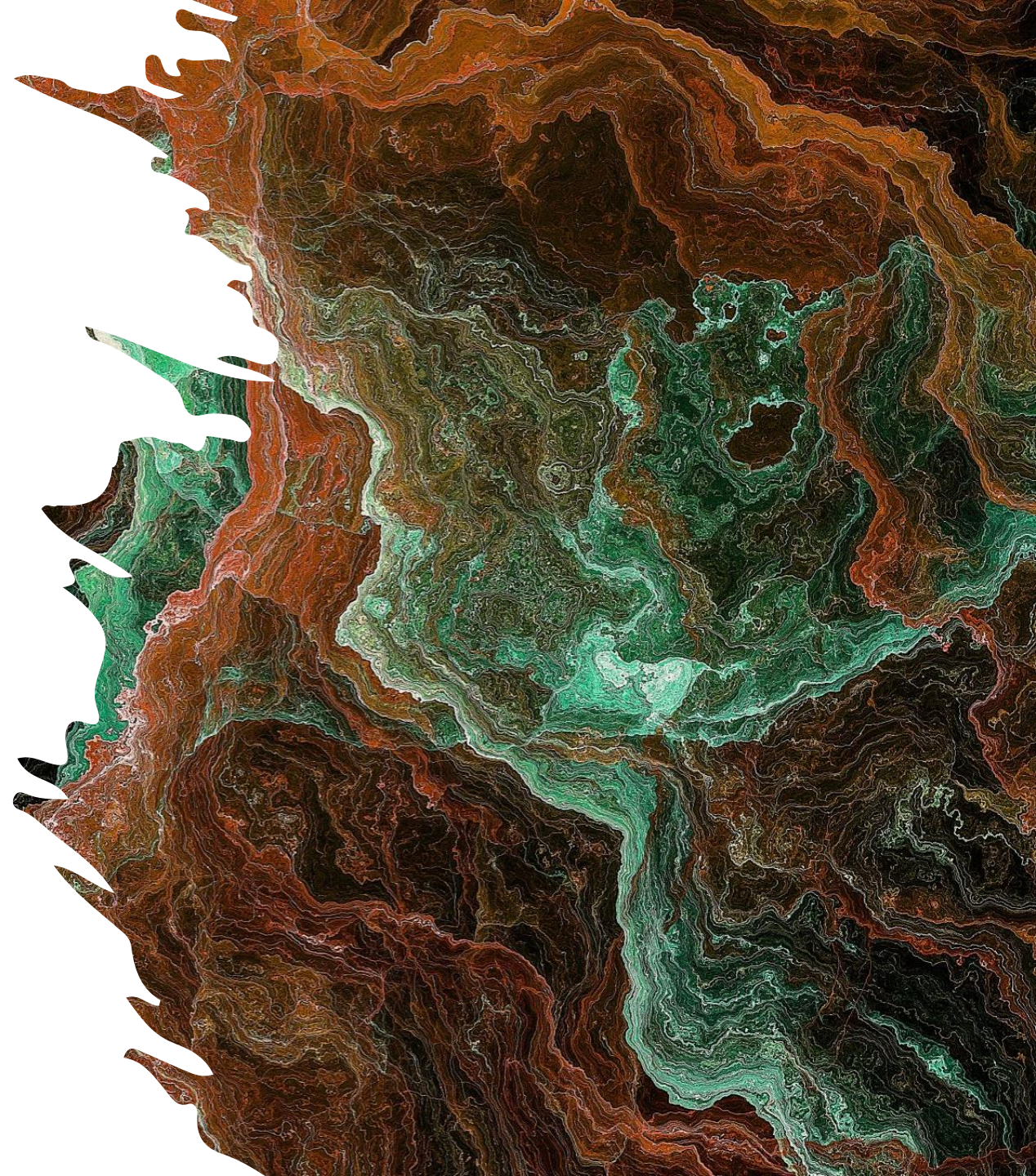
What is a corpus?

3.

a. A body or complete collection of writings or the like; the whole body of literature on any subject.

1728   E. CHAMBERS *Cycl.*   Corpus, is also used in Matters of Learning, for several Works of the same Nature, collected, join'd, and bound together... The Corpus of the Civil Law is compos'd of the Digest, Code, and Institutes... We have also a Corpus of the Greek Poets.

1865   J. B. MOZLEY *8 Lect. Miracles* i. 16   Bound up inseparably with the whole *corpus* of Christian tradition.

1876   W. E. GLADSTONE in *Contemp. Rev.* June 14   Assaults on the *corpus* of Scripture.

1886   *Athenæum* 14 Aug. 211/1   The completion of such a corpus of Oriental numismatics.

(Hide quotations)

b. The body of written or spoken material upon which a linguistic analysis is based.

1956   W. S. ALLEN in *Trans. Philol. Soc.* 128   The analysis here presented is based on the speech of a single informant..and in particular upon a corpus of material, of which a large proportion was narrative, derived from approximately 100 hours of listening.

1963   *Language* **39** 1   In the analysis of the data, the structural features of the corpora will first be described.

1964   E. PALMER tr. A. Martinet *Elements Gen. Linguistics* ii. 40   The theoretical objection one may make against the 'corpus' method is that two investigators operating on the same language but starting from different 'corpuses', may arrive at different descriptions of the same language.

1971   J. B. CARROLL et al. *Word Frequency Bk.* p. xxvii   How many types does one have to 'know' to know 95% of the tokens in the population of texts from which a corpus has been derived?

1983   G. LEECH et al. in *Trans. Philol. Soc.* 25   We hope that this will be judged..as an attempt to explore the possibilities and problems of corpus-based research by reference to first-hand experience, instead of by a general survey.

(Hide quotations)

What shall we do before we start to compile a corpus?
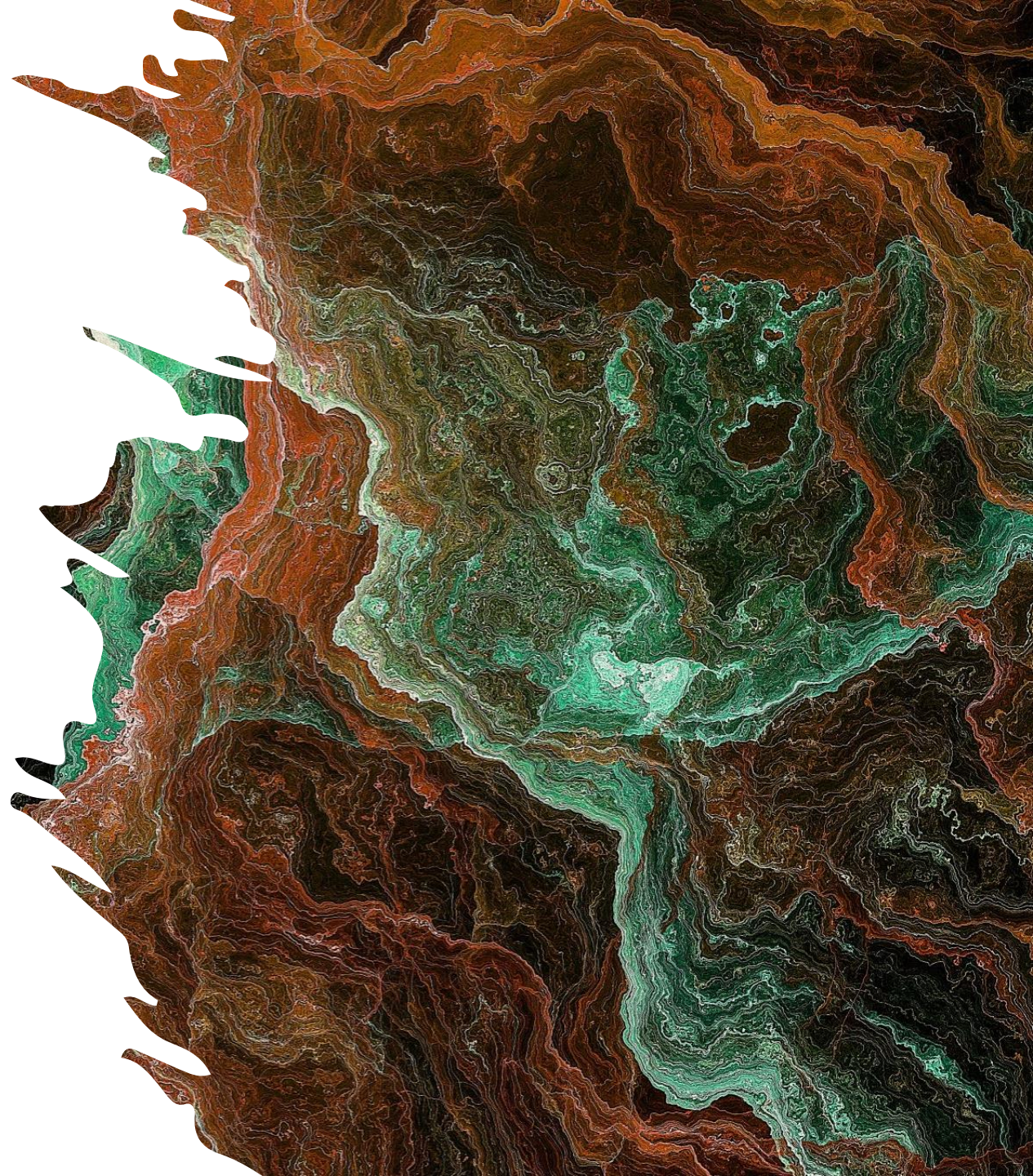
Task:
Google Old Norse Corpora

Preliminary Surveying

Old Norse corpora and Old Norse text databases

The Saga Corpus

- Through CLARIN ERIC infrastracture (75 historical corpora)

- Comprises OI (ON) narratives from 13th to 15th century: *Íslendingasögur, Sturlunga saga, Heimskringla, Landnámabók* (41 texts)

- concordancer *Korp* (concordance search tool of *Språkbanken* (Swedish Language Bank)

- tagged, lemmatised, normalized (> to Modern Icelandic)

Text	Words
Family Sagas	982.066
Sturlunga Saga	260.586
Heimskringla	231.502
Book of Settlement	37.120
Total	1.511.275

More about the corpus on Clarin:
https://clarin.is/en/resources/sagacorpus/

Saga corpus search here:
https://malheildir.arnastofnun.is/?mode=fornrit#?lang=en&stats_reduce=word
&isCaseInsensitive&searchBy=word&cqp=%5B%5D

ONP: Dictionary of Old Norse Prose

- University of Copenhagen, 65 000 headwords and 840 000 citations
- across genres, but poetry not included, also does not comprise Danish literature and ON literature composed in England
- texts from Norway up to c1370, Iceland up to c. 1540
- dictionary, but can sort occurrences based on the metadata (genres), lists forms, filters based on source, date (lists occurrences)
- also ONP reader – digital access to over 500 editions of ON texts with a parallel glossay generated based on the dictionary data
- orthography normalised, and selectively conservative – mixture of Icelandic and Norwegian

The Icelandic Saga Database (sagadb.org)

- includes only the Sagas of the Icelanders (Icelandic family sagas)
- normalized OI
- texts available in different translations, including English, German, or French
  → draws its resources from Heimskringla, Árni Magnússon Institute, etc.
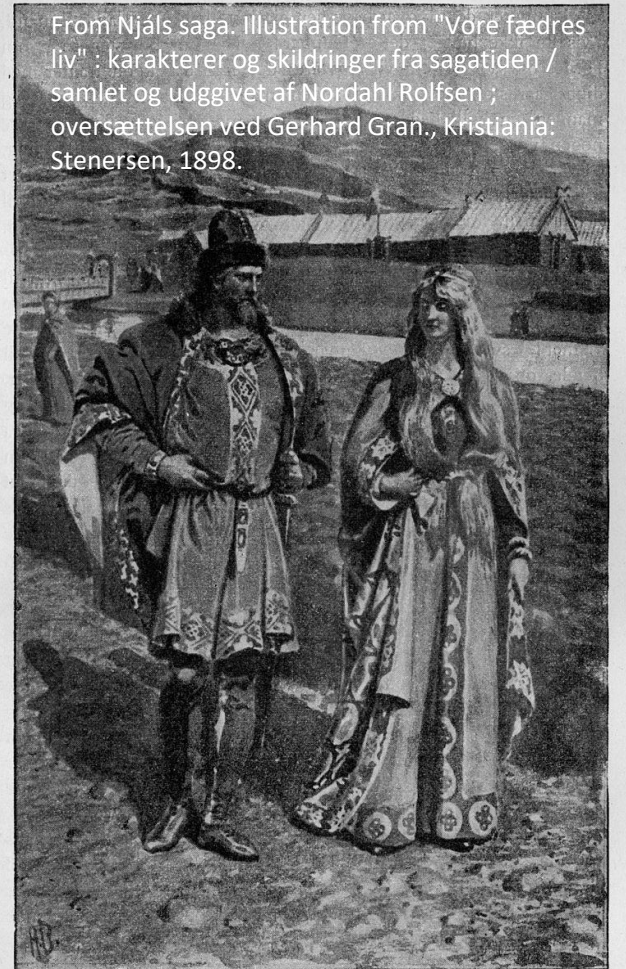
From Njáls saga. Illustration from "Vore fædres liv" : karakterer og skildringer fra sagatiden / samlet og udggivet af Nordahl Rolfsen ; oversættelsen ved Gerhard Gran., Kristiania: Stenersen, 1898.

Gunnar møder Hallgerd paa thingssletten.

Snerpa.is

- across genres: old poetry, prose, other texts, Bible

- not restricted to ON period → up to (and including) Modern Icelandic texts

- normalized Modern Icelandic spelling throughout

Heimskringla.no

- ON texts – Eddas, Icelandic sagas, scaldic poetry and others

- includes not only ON, but also Icelandic, Danish, Norwegian, Swedish and Faroese versions of (some of) the texts

- some texts not complete

- some normalized ON, some not

The Library of Old Norse
in *Bibliotheca Polyglotta*

- selected ON texts (sagas of the Icelanders, kings' sagas, Poetic Edda, Snorra Edda, Stjórn…)

- includes translations (incl. English), diplomatic editions as well as normalized (*Íslenzk fornrit* series)

- Basic searching for phrases available

- Parallel full text view with other editions or sentence view

Old Norse Web text corpus

- excerpted from the internet – 'based on material from 2012', listed sources for individual examples given (Heimskringla.no)
- not lemmatized, normalized
- 1 138 843 tokens
- provides frequencies, context (words with similar context or lists context words within the same sentence) – sentences defined?
- In Leipzig University Corpora Collection – 'collects and processes documents available from the internet' = corpus-based dictionaries in more than 250 languages

Medieval Nordic Text Archive (menota.org)

Haugen, Odd Einar (2001) *Medieval Nordic Text Archive*. University of Bergen. Available here: https://www.menota.org/
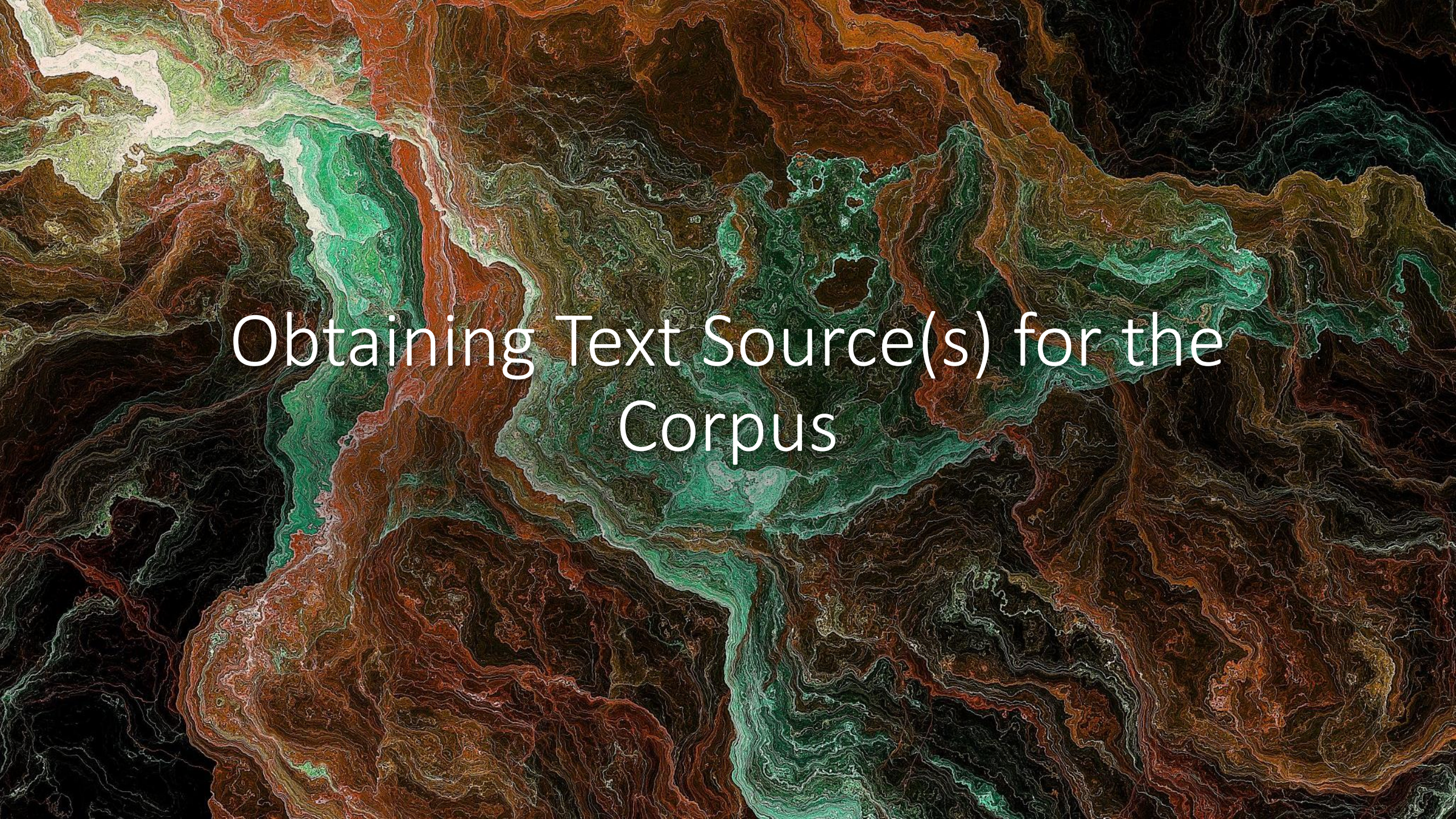
= 'A network of Nordic archives, libraries and research departments working with medieval texts and manuscript facsimiles'

- Nearly 50 texts (approx. 1.8 million words), some lemmatized, some syntactic annotation and accessible through *INESS* (Norwegian Infrastructure for the Exploration of Syntax and Semantics)

- Old Icelandic, Old Norwegian, Old Swedish, Old Danish (and Latin)

- Most texts normalized, some close transcriptions (non-normalized), also diplomatic editions

- Also aims to provide translations of the texts

- Links to dictionaries, digital archives and libraries

→A corpus interface: query, concordance, collocations, word list (acc. to the latest report a corpus of 765,000 words)

→ Project *Menotec* transcribed texts published in *MENOTA;* dependency model for syntactic annotation, modified *PROIEL* scheme
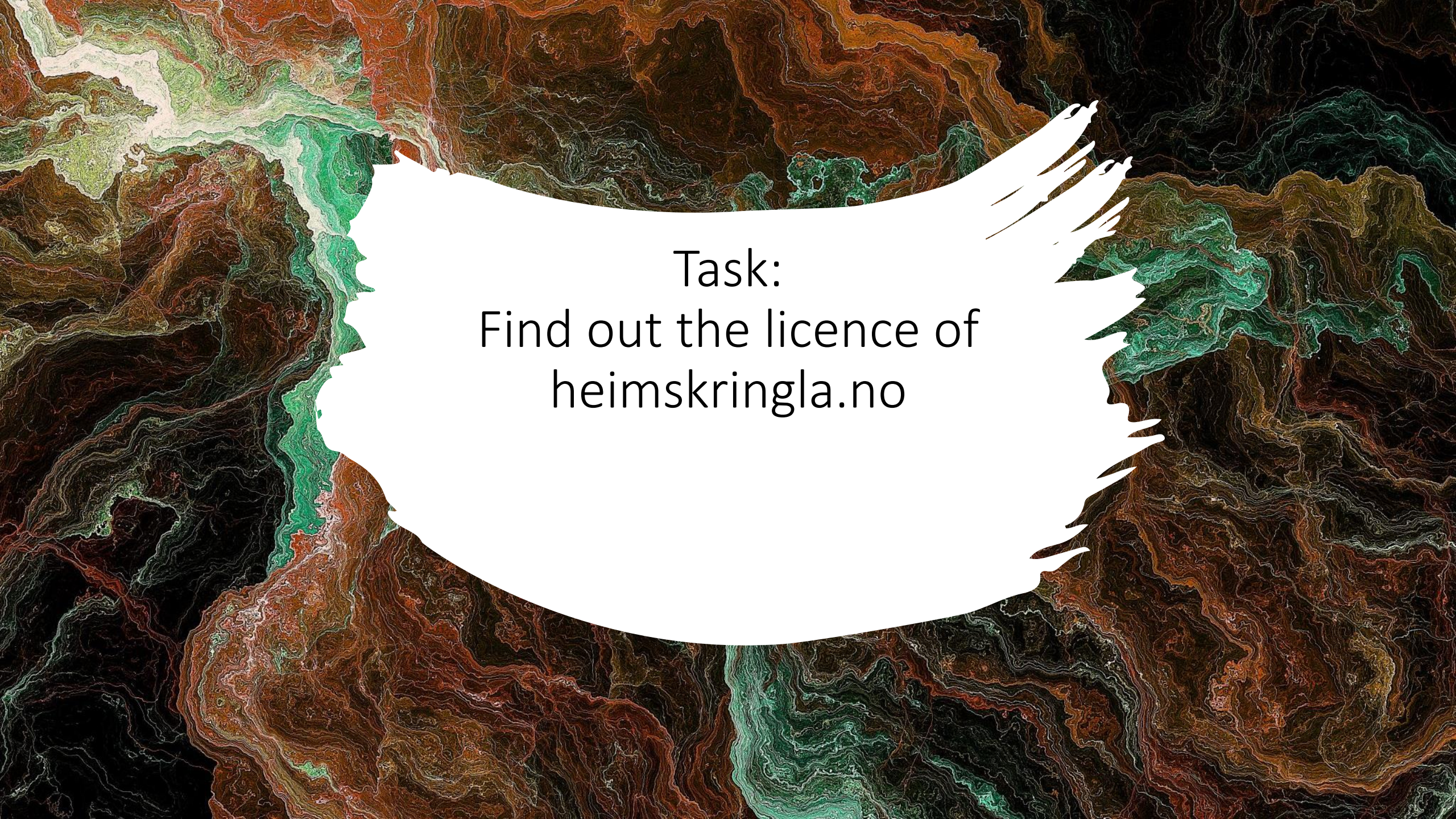
Overview

Saga corpus	ONP	Sagadb	ON Library	Snerpa	Heimskringla	ON web	MENOTA
13-15th ce (ON/OI)	N (up to 14th ce), Ice (up to 16th)	Up to 14th ce (OI) + translations	13th ce – 14th(?) (ON/OI)	Up to Modern Icelandic	ON up to modern Scandinavian languages	Internet excerpted data	up to the Reformation (1200-1700)
prose (41 texts)	across genres, but no poetry	Íslendingasögur	across genres	across genres	across genres	? across genres	across genres
tag, lemma	metadata sort	--	parallel view, phrase search	--	--	-- (context listing)	lemma, syntactic annotation
norm (Ice)	norm (+ variants)	norm (OI, Ice)	norm (+ dipl)	norm (Ice)	norm/non-norm	?norm (ON)	norm/non-norm (+ dipl)

→ Mostly databases without corpora managers, interface, functions or consisting only of normalised texts (OI, Modern Icelandic) with the exception of *MENOTA* (so far not complete)

Obtaining Text Source(s) for the Corpus

Task:
Find out the licence of
heimskringla.no

Licence

- Texts taken from Heimskringla.no
- No licence provided on the webpages.
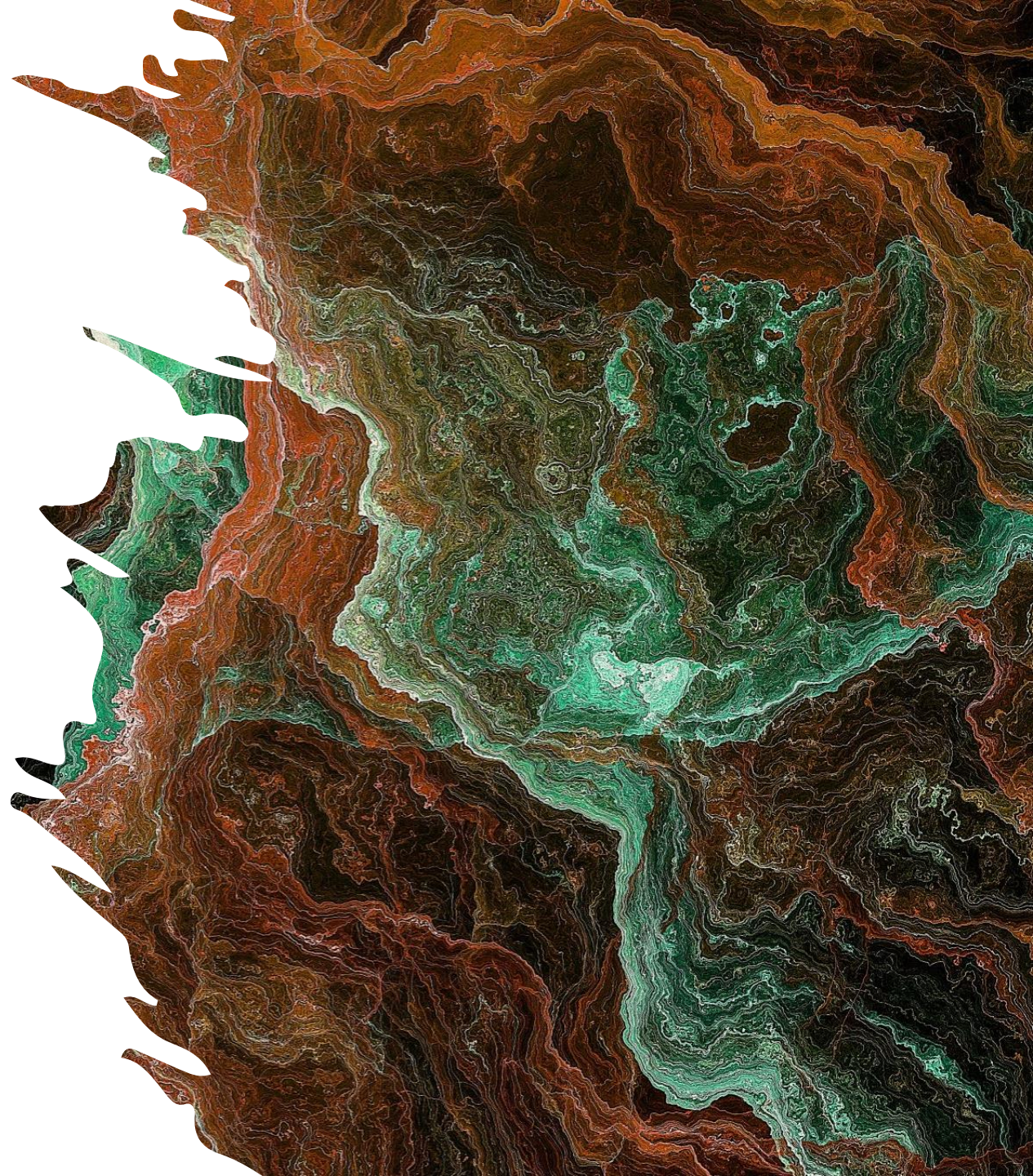- The contact with the authors:

„Hei David

Absolutt. Det er hele formålet med Heimskringla.no at tekstene skal kunne brukes.

Mvh. Jon Julius Sandal"

1. What do we want to study in the corpus?

2. How actually are we going to define the data ?

Text-related Issues

Non-normalised and normalised texts, similarities between Scandinavian (Modern) languages

Similarities between modern Scandinavian languages

- ***Snorre Sturelsons, Edda samt Skalda, Skáldskaparmál, 55:***

Einn maðr er nefndr Ægir eða Hlér. Hann bjó í eyju þeiri, er nú er kölluð Hlésey. Hann var mjök fjölkunnigr.

En Mand hed Ægir eller Hler. Han boede paa den Ø, som nu hedder Hlesø (Læssø); han var meget troldkyndig.

En man är nämnd Äger eller Hler: han bodde på ön Lessö och var mycket klok.

- Source: http://heimskringla.no/wiki/Sk%C3%A1ldskaparm%C3%A1l

Non-normalised vs normalised text

	heimskringla
Edda	norm
Snorra Edda	norm
Fagrskinna	non-norm
Fornmanna sogur	norm
Flateyjarbok	norm/non-norm
Fornaldarsogur	norm
homiliubok	non-norm
Hauksbok	non-norm
heilagra manna sogur	non-norm
heimskringla	norm (/non-norm)
islandske annaler	non-norm

- Many spelling conventions

- Variant letter forms

- Unique letters and signs for ON (Norrønt)

- Degree of normalisation may vary > 19th Century scholars > a standardized normalization of ON > based on *Fyrsta málfræðiritgerðin (Codex Wormianus)*

Source: https://en.wikipedia.org/wiki/Old_Norse_orthography

ON definition and perception

Established scholarly terminology:

- *Altnordisch* in German studies since Grimm
- D: *oldnordisk* – older than *gammelnordisk*
- I and S: *Icelandic* without discriminating further
- N: *oldnorsk/ gamalnorsk* and *norrønt* specifically for OWN
- preceded by *runedansk, runesvenska...*

ON speaker perception:

- 'one language': *dǫnsk tunga* - in Iceland and Norway retained throughout Middle Ages
- 12th century: Norway-Iceland - name *norrønt mál* – 'Norwegian'

> conscious choice, distinguishing W branch from the E

- 3rd Grammatical Treatise from the mid13th ce – distinguishes btw 'Danish usage' and *'norrønt mál'*

(Ottosson in Bandle, 2002: 787-789)

Task:
Define Old Norse

Old Norse vs Old Icelandic

A Historical Development Crash-course

Germanic Language, North-West Germanic

- branches of Germanic after its division in a100 – differences in divisions – traditionally tripartite N, W, E (19th century division)

- currently added midstep subgrouping: North-West Germanic

– shared development and even after splitting possibly features shared, even at the time when the WG itself began splitting (Ringe and Taylor, 2014: 10)

- term 'North Sea Germanic', subsuming Old English, Old Frisian, Old Norse due to their close affinity by Hans Frede Nielsen

- Old Norse even closer affinity with Old Northumbrian (Bibire, 2001: 91)
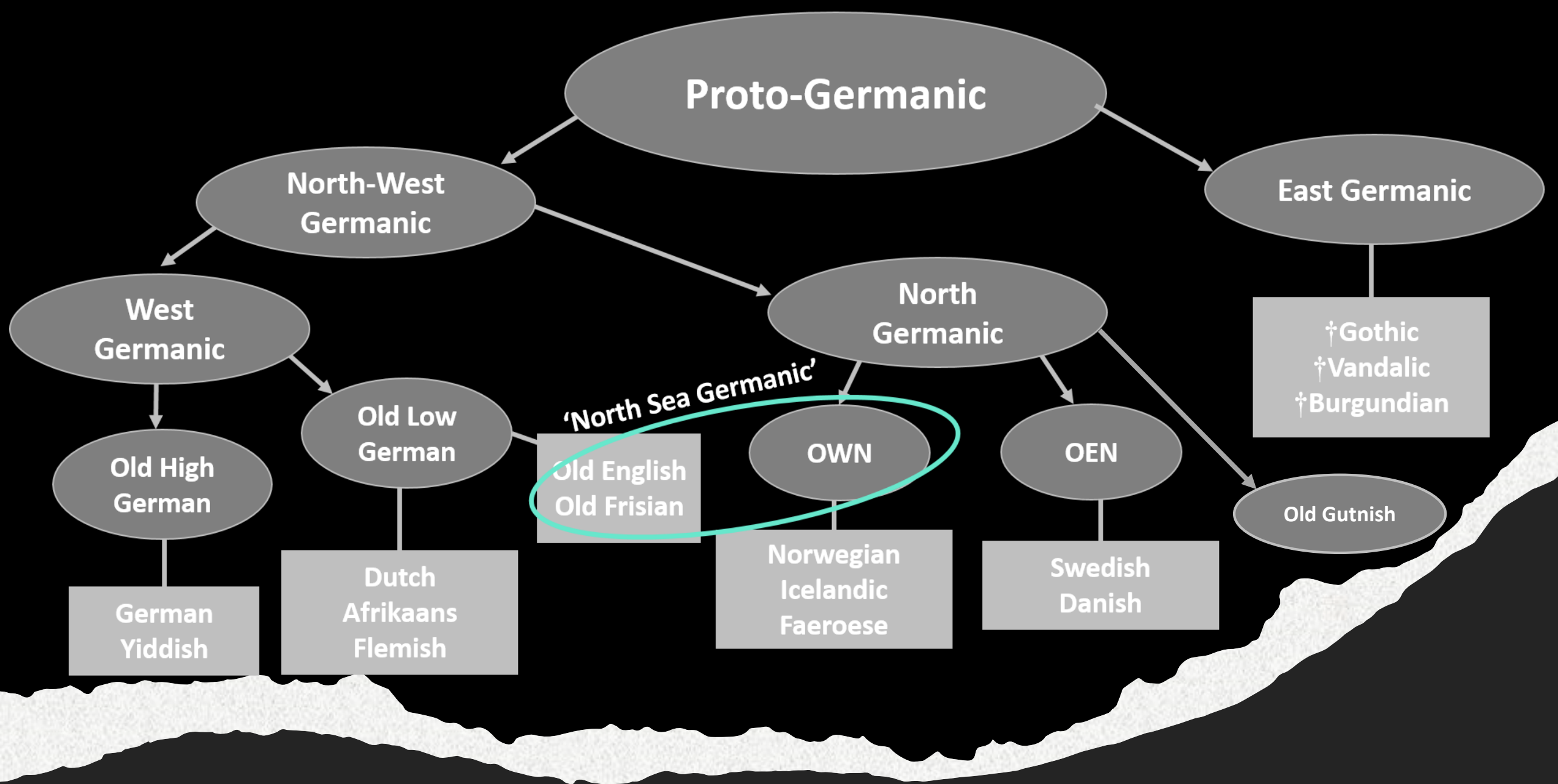
Langauge tree based on development suggested in Brinton and Arnovick, 2011: 128-129; N-W Gmc grouping based on Ringe and Taylor, 2014: 10; North Sea Gmc based on Bibire, 2001: 91.

Old West Norse (OWN) vs Old East Norse (OEN)

- **OEN** = Old East Norse after a1000 comprises old Danish and Swedish – Danish least conservative of all (Gordon and Taylor, 1957: 320)
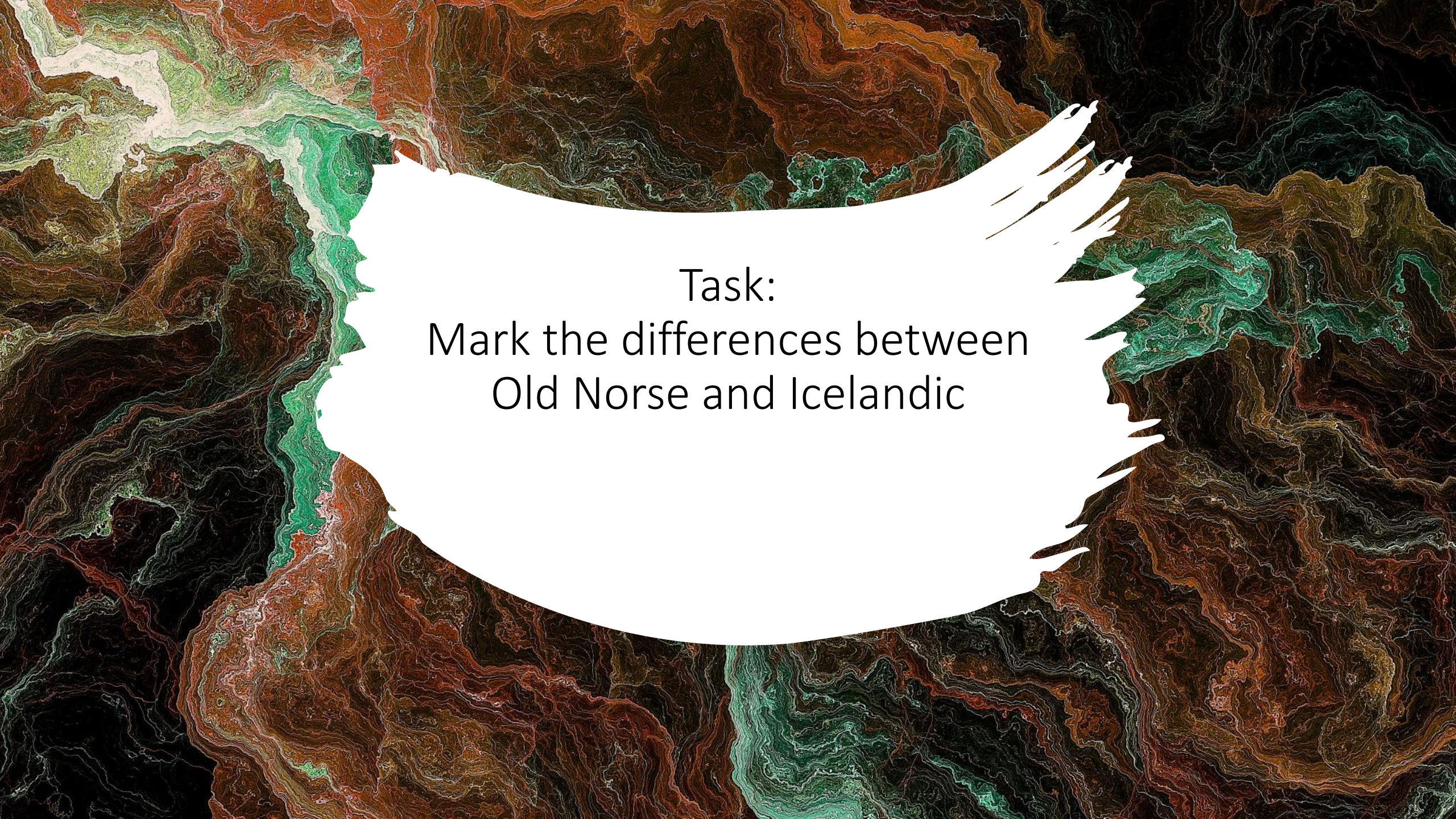- **OWN** = Old West Norse, after a1000 comprises Old Norwegian and Old Icelandic (G and T, 1957: 265)

Some differences between OWN and OEN – features:

- Old Danish as opposed to ON: intervocalic k, *p, t* > *g, b, d* (*lágedóm* x OI *læknisdómr*); *g* dropped next to *i* in unaccented syllables *æi* (x *eigi* OI)
- Old Swedish marked by careful distinction of length degrees and stress in weak syllables:
    - *komá, livá* (final inf. ending of short root syllable verbs long)
- also 'vowel balance' – ending vowels varied according to stress and length of the preceding syllable: *a, i, u* > *æ, e, o* with weaker accent (*móþer* x *faþir*, cp. OI *móðir, faðir*) (G and T, 1957: 319-326)

Is it still Old Icelandic? Some notes on delimitation

Mostly period 1100-c1350

- OI used for unattested ON forms, until the 13th century very similar (G and T, 1957: 319)
- Surviving materials issue – late records – OWN, OEN differing development (Danish especially rapid – even the earliest records show morphological 'simplification')
- OI as a 'faithful enough representative': (beginnings)
  - for OD and OS mid13th ce oldest manuscripts preserved
  - for OI early 12th ce, ON fragments from mid12th ce
  - Cultural date – 'end of the Viking period, 1066' (Ottosson in Bandle, 2002: 787-789)
- End of the period:
  - Norwegian – breakdown of scribal norms as a result of Great Plague in 1349-50, S and D as well as strong Low German influence
  - Swedish – emergence of unifying written norm
  - Danish – rapid morphological changes underway early (1100 *middeldansk*), *olddanks/runedansk* before 1100 > division *yngre* and *ældre middeldansk* a1350 (Skautrup 1944, Brøndum-Nielsen, 1928-73) (O in B, 2002: 788)

Task:
Mark the differences between
Old Norse and Icelandic

	Old Icelandic (ON)	Modern Icelandic
merger æ < œ, æ	hœla, bæði	hæla, bæði
merger ö < ø, ǫ	sǫgu (saga), ǫl, døkkr	sögu, öl, dökkur
epenthetic u (r > ur)	maðr, konungr	maður, konungur
Final ss, rr simplification > s, r	íss, herr	ís, her
Final unstressed t, k to ð, g	þat, ok, kallat	það, og, kallað

(ON)
Hét karlmaðrinn Askr, en konan Embla, ok ólst þaðan af mannkindin, sú er byggðin var gefinn undir Miðgarði. Þar næst gerðu þeir sér borg í miðjum heimi, er kölluð er Ásgarðr. Þat köllum vér Trója. (taken from Heimskringla.no)

(I)
Hét karlmaðurinn Askur en konan Embla, og ólust þaðan af mannkindin, þeim er byggðin var gefinn undir Miðgarði. Þar næst gerðu þeir sér borg í miðjum heimi er kallað er Ásgarður. Það köllum vér Trója. (taken from snerpa.is)

The man was called Ash and the woman Elm, and from them humankind was created, who were given the dwelling under Midgard. There in the middle they then build for themselves a town that is called Asgard. That we call Troy.

1) How do we build our corpus?

2) What is going to be our methodology?

3) Do we have a tagger?
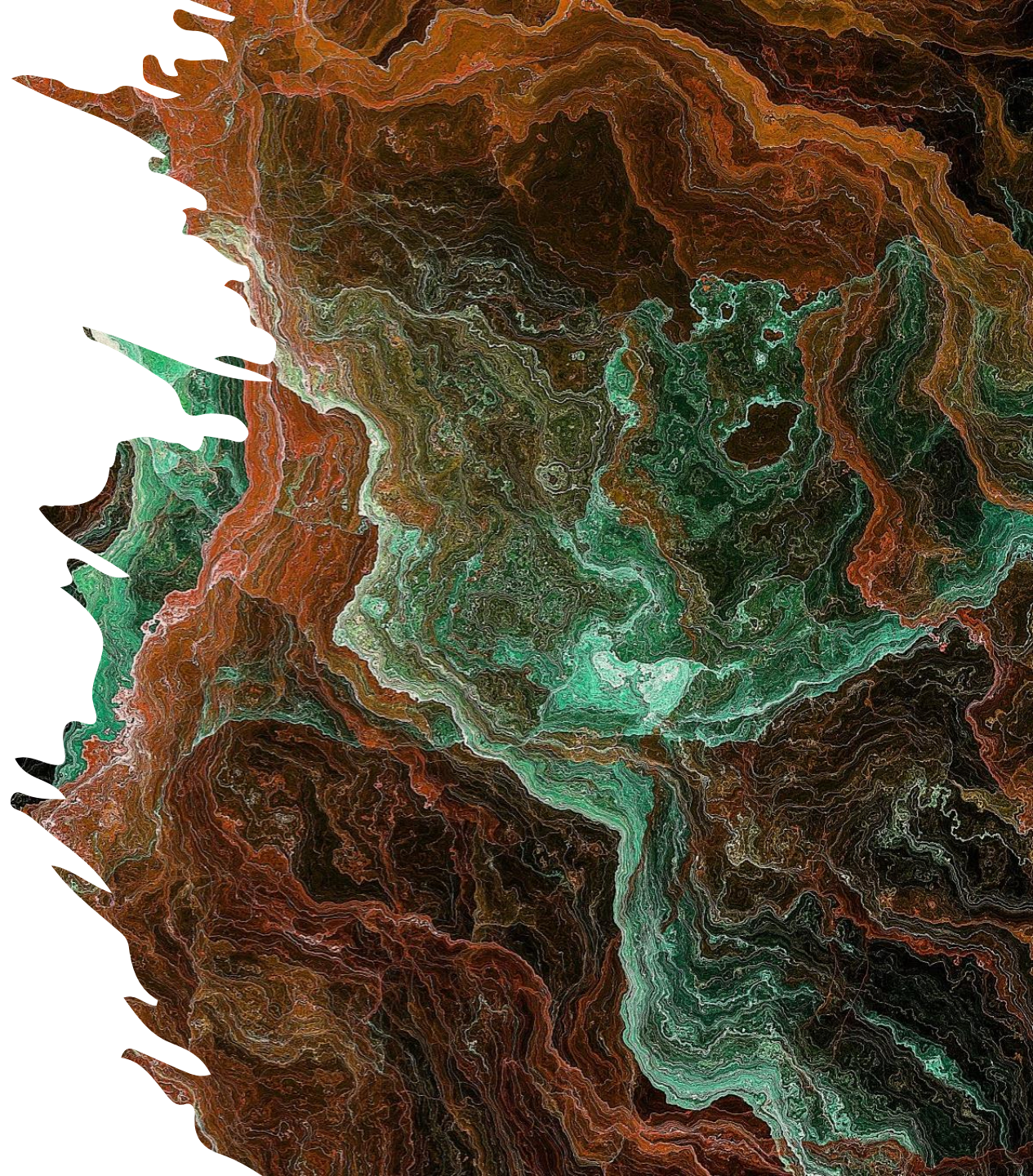
Task:
1) Google some Old Norse
taggers

Task:
1) Google some Old Norse taggers
2) Icelandic tagger – can we use them?

Methodology

Tagger, Tagging, Success Rate

Method

- Heimskringla.no

- > tagger (Mälföng; IceNLP tagger; open source) > tokenization + lemmatization

- > source view

- > encoding the text itself + the text information (ON text database)

Tagger IceNLP

- Tagger is for Icelandic; encode in inconvenient format > thanks to Ondřej Tichý, PhD > transformation to XML compatible format

- Tag set example:

Hann fpken (Hann) átti sfg3eþ (eiga) son nk eo (sonur) þann fakeo (sá) , , (,) er sfg3en (v era) Eirekr nken-
s (Eirekr) hét sfg3eþ (heita) . . (.)

Tagset for the electronic files of the IFD corpus

Category	Analytical symbol - information
Word class	n-noun
Gender	k-masculine, v-feminine, h-neuter, x-gender unspecified
Number	e-singular, f-plural
Case	n-nominative, o-accusative, þ-dative, e-genitive
Article	g-with suffixed definite article
Proper noun	m-name of a person, ö-place name, s-other proper noun

The Primary Analysis

Primary Analysis (IceNLP):

Repeated words correctly tagged:
Á 4x, *Hann* 6x, *Þeir* 13x (but occasionally wrong lemma attribution),
Var 3x, *Átti* 2x, *þann* 4x, Hét 2x
- 38 mistags

- David Šimeček, PhD has conducted a primary analyses by the IceNLP and ABLTagger:
- analysed a chapter of *Eireks saga*, normalized ON
- IceNLP showing different <u>types of issues:</u>
  - Mixed lemma (group 1)
  - Morphosyntactic (homoforms) (group 2)
  - Repeated problems with singular lemmas (group 3)

Including the 3rd group:
+ 3 mistags (group 3)
-25 words
i.e. 267 words, 42 mistags
Including the group 3 and 1:
- 245 words, 53 mistags
Including the group 3, 2 and 1, without the issues in the second group:
- 231 slov, 53 chyb
- Success rate 77%
If the 2nd group mistakes are included:

- 234 words, 68 mistags

- <u>Success rate 71%</u>

Tagger ABL

- Due to low success rate of IceNLP, ABL tagger suggested (also made for Modern Icelandic, but should have a higher success rate nonethelss

- Suggested by Steinþórr Steingrímsson and Starkaður Barkarson, *Stofnun Árna Magnússonar í íslenskum fræðum*

- Tag set example:

{"lemma":"\u00fer\u00e1ndr","tag":"nken","word":"\u00deR\u00c1NDR"},{"lemma":"vera","tag":"sfg3en","word":"er"},{"lemma":"nefndr","tag":"lkensf","word":"nefndr"},{"lemma":"konungr","tag":"nken","word":"konungr"},{"lemma":"s\u00e1","tag":"faken","word":"s\u00e1"},{"lemma":",","tag":",","word":","},{"lemma":"er","tag":"c","word":"er"},{"lemma":"fyrstr","tag":"nken","word":"fyrstr"},{"lemma":"r\u00e1\u00f0a","tag":"sfg3e\u00fe","word":"r\u00e9\u00f0"},{"lemma":"fyrir","tag":"a\u00fe","word":"fyrir"},{"lemma":"\u00der\u00e1ndheimur","tag":"nke\u00fe-s","word":"\u00der\u00e1ndheimi"},{"lemma":".","tag":".","word":"."}

→ Uses Unicode for special characters
→ Same tagset as employed by the IceNLP
→ Tags also sentences (need to be separated by an empty line)

→ Based on the analysis, higher success rate than IceNLP

Tagging Issues

Lemma Mixing, Morphosyntactic Properties, Specific Lemma Issues

Different lemmas mixed together

Lemma	*þá*	*sá*	*er*	*at*	*sé*
POS	ADV, V, Pron	Pron, V	Pron, Conj, V	Prep, Conj, inf. marker	V (subj. vera)
tagged as	ADV	V, Pron	V	foreign word, N (n)	V (sjá)

Same lemma different tags

Lemma	*váru*	*konungr*
POS	V. *vera*	m N
tagged as	f N (*vára*)	ADJ n/f

Morphosyntactic issues (homonymy):

eitt – acc, n. sg. -- nom, n. sg.
skip – acc n. sg -- nom n. pl.
þeim – dat pl. m. -- dat pl. n., dat sg. m.
væri - subj pl. -- subj sg.
ætlaði - ind sg. -- konj pl.
mörg - acc pl. n. -- nom pl. n

Solutions

- Manual tagging of specific issues

- Some issues of 'mis-tagging' may be remedied – some occurrences may be used to form precise identifications (as in the case of complex conjunctions with particles *er, at*)

- Regular but very flexible word order – individual word classes in different environments – difficult to encode a „rule" – tag manually, but inflections in some cases may help(?)

→ teach tagger of TEITOK to do the rest of the work

The next step: TEI TOK

http://lindat.mff.cuni.cz/services/teitok/index.php

Sources and References

- Bibire, P. (2001) 'North Sea Language Contacts in the Early Middle Ages: English and Norse', in *The North Sea World in the Middle Ages: Studies in the Cultural History of North-Western Europe*, ed. by T. R. Liszka and L. E. M. Walker. Dublin: Four Courts Press, pp. 88-107.
- Brinton, J. L. and Arnovick, K. L. (2011) *The English Language: A Linguistic History, the Second Edition*. Oxford University Press.
- Brøndum-Nielsen, Johs (1928-1973) Gammeldansk Grammatik i sproghistorisk Fremstilling 1-8. København.
- Gordon, E. V. and A. R. Taylor (1957) *An Introduction to Old Norse: Second Edition*. Oxford: Clarendon Press.
- Ottosson, Kjartan (2002) 'Old Nordic: A definition and delimitation of the period'. In: Bandle, O. (2002) *The Nordic Languages: An International Handbook of the History of North Germanic Languages, Volume 1*. Berlin: Walter de Gruyter.
- Ringe, Don and Taylor, Ann (2014) The Development of Old English: A Linguistic History of English, Volume 2. Oxford University Press.
- Skautrup, Peter (1944) Det danske sprogs historie 1. København.