

Beyond the Edition: On the Linguistic Annotation of Vernacular Texts

Odd Einar Haugen

1. Annotation as an editorial enterprise

Preparing an edition of a classical or medieval text is more often than not a long-term project for an editor. The aim of the exercise is the *constitutio textus*, the bringing together of divergent and often contradictory sources ideally into an unmitigated whole. The edition provides a point of entry to the text, making the text accessible in its variety, although at the same time it incorporates the editor's critical understanding of the sources and his or her recommended routes through them. The publication of the edition is not necessarily the end of the editorial work, but beyond this point, the editor will have the company of other contributors in the form of scholars who draw attention to specific aspects of the text by commenting upon it, by annotating it in various ways, or by contextualising it more generally.

The focus of this chapter will be on annotation, and specifically on the linguistic annotation of medieval texts. Although the examples given here are drawn from the Nordic vernaculars, the principles are in the main the same for other language families, and there are in fact several annotation projects that span a wide variety of texts that are diverse with respect to their language, their provenance, their dating and their contents. After an initial discussion of the representation of manuscript text, looking in particular at vernacular texts, the chapter will move on to two central types of linguistic annotation, those of morphology and of syntax. While many projects concerning medieval texts have been annotating morphology, there are far fewer that have included syntactic structures in their annotation.

This lecture was given at the *Ars edendi* concluding conference 18 August, 2016.

How to cite this book chapter:

Haugen, O. E. 2020. Beyond the Edition: On the Linguistic Annotation of Vernacular Texts. In: Kihlman, E. and Searby, D. (eds.). *Ars Edendi Lecture Series. vol. V*. Pp. 67–88. Stockholm: Stockholm University Press. DOI: <https://doi.org/10.16993/bbd.e>. License: CC-BY

In a digital workflow, annotation is so interwoven with the encoding of a text that it may be useful to see annotation as an integral part of the whole editorial process. The annotation establishes a link to a number of external resources such as dictionaries and grammars, and it makes the text accessible for a broader user base. As I will argue below, morphological and syntactic annotation is not a strictly linguistic endeavour. It also means that literary and historical investigations of a much wider scope can be conducted.

2. Multi-level representation of a text

Some works are only preserved in a single manuscript, a situation that is probably more common in medieval than in classical literature, and this is certainly the case in Nordic vernacular literature. The majority of the Eddic poems, for example, have been handed down to us in a single manuscript, and the same goes for many other of the most prominent early Nordic works.¹ A work preserved in a single manuscript makes life simple for the editor, but whenever a work has been preserved in more than one manuscript, the editor has to decide whether the various manuscripts should be transcribed *in extenso* or whether some of them, perhaps all of them except the *Leithandschrift*, should be reflected solely through variants in the apparatus of the edition.

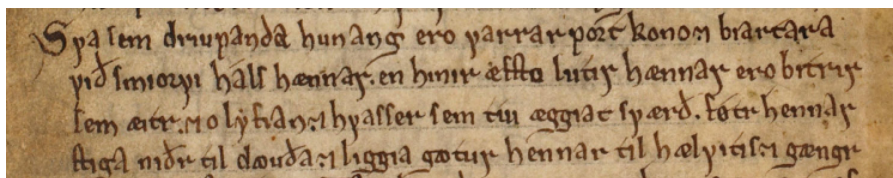


Figure 1. An admonition in the Old Norwegian Homily Book. Copenhagen, The Arnamagnæan Collection, AM 619 4°, f. 8v, l. 23–26.² Copyright: The Arnamagnæan Collection, License: CC BY-NC-ND.

¹ The major Eddic manuscript is Reykjavík, The Árni Magnússon Collection, GKS 2365 4° (dated ca. 1270), often referred to simply as *Codex Regius*. Another example is the translation from Old French of the *lais* by Marie de France, *Strengleikar*, in Uppsala, University Library, De la Gardie collection, DG 4–7 fol (dated ca. 1270). This is the only manuscript of *Strengleikar*, apart from a fragment, Copenhagen, The Arnamagnæan Collection, AM 666 b 4°, which was once part of the same codex.

² The text of the Homily Book was edited by Gustav Indrebø, *Gamal norsk homiliebok* (Oslo: Kjeldeskriftfondet, 1931); the passage quoted here is on p. 17, l. 15–19 in his edition. The text of the Homily Book is also available with full colour facsimiles in *The Medieval Nordic Text Archive* at <http://clarino.uib.no/menota/catalogue>.

As for the actual representation of the source, the text may be represented with various degrees of fidelity, ranging from a close reproduction of its graphical form to an extensive regularisation of its orthography. In the encoding of vernacular sources, I suggest that three focal levels can be identified along this axis: a facsimile level, a diplomatic level and a normalised level.³ This multi-level rendering can be illustrated with a short extract from the Old Norwegian Homily Book, with my rather literal translation into English at the end:

Sva fem driupanda [< driupande] hunang ero varrar port kono. 7 biartara
við smiorvi hálfi hænna. en hinir æfto lutir hænna ero bitrir fem æitr. 7 o
lyfian. ok hvasser fem tíu æggjat sværð. fótr henna stiga niðr til dauða. 7
liggia gætur hennar til hælýtif.

Sva sem driupanda hunang ero varrar port kono. ok biartare [< biartara]
við smiorvi hálfs hennar. en hinir æfsto lutir hennar ero bitrir sem æitr. ok
o lyfian. ok hvasser sem tuí [< tíu] æggjat sværð. fótr hennar stiga niðr til
dauða. ok liggja gætur hennar til hælvtis.

Svá sem drjúpanda hunang eru varrar portkonu, ok bjartari viðsmjórvi hals
hennar. En hinir öfstu hlutir hennar eru bitrir sem eit ok úlyfjan, ok hvassir
sem tvíeggjat sverð. Fótr hennar stíga niðr til dauða, ok liggja gætur hennar
til hælvtis.

Like dripping honey are the lips of a harlot, and her throat is brighter than
oil. But her end is bitter as wormwood and poison, and sharp as a two-
edged sword. Her feet lead down to death, and her road to hell.

This admonition, which ultimately is a faithful rendering of Proverbs 5.3–5, was quoted in Ch. 18 ‘De castitate’ of Alcuin’s *De virtutibus and vitiis*, written between 801 and 804.⁴ Alcuin’s immensely popular text was translated into Old Norwegian around 1200 and became part of the Homily Book.

The three levels exemplified above can be seen as three perspectives on the same passage in the text. They differ in what might be termed granularity, i.e. the degree to which they adhere to the source.

³ Recently discussed in Odd Einar Haugen, ‘Levels of Granularity: Balancing Literary and Linguistic Interests in the Editing of Medieval Vernacular Texts,’ in *Philology Matters! Essays on the Art of Reading Slowly*, ed. by Harry Lönnroth (Leiden: Brill, 2017), pp. 118–135.

⁴ Cf. Paul E. Szmarach, ‘A Preliminary Handlist of Manuscripts Containing Alcuin’s *Liber de virtutibus et vitiis*’, *Manuscripta*, 25 (1981), 131–40. Alcuin’s text is published in *Patrologia Latina* vol. 101, cols 613–638. A new edition of the text remains a desideratum.

On the facsimile level, all characters, including diacritical marks and abbreviation signs, are copied in their position along the base line. On the diplomatic level, a smaller number of characters are used, so that allographs, such as the round and the straight ‘r’, are represented by a single character, ‘r’, and abbreviations are expanded. On the normalised level, the orthography is regularised according to the standard grammars and dictionaries of the language. The latter level is in many respects unique for Old Norse (i.e. Old Icelandic and Old Norwegian) texts; there are no comparable standard orthographies for most other European vernaculars. Even for medieval European vernaculars without a standard orthography, however, a certain degree of regularisation is not uncommon, such as the introduction of punctuation (often according to modern rules), the capitalisation of proper names and the first word after a full stop, and perhaps the ironing out of minor spelling variations.

It should be underlined that the three levels illustrated here are not variants of the text, since variants would be quoted on the same level of granularity, and there are no variants unless there are at least two manuscript witnesses of the same passage. The levels exemplified here are representations of a single source. They are alternative ways of seeing and representing a specific manuscript rather than a work as such.

As can be seen from the square brackets in the Old Norwegian transcription above, there are several comments and corrections that the editor might wish to add to the text. The first is visible on the facsimile level and points to the fact that the scribe had obviously corrected an ‘e’ to an ‘a’ in the third word of the first line. On the diplomatic level, there is another comment regarding the form of the comparative *biartare* ‘lighter, more luminous’, which is spelt *biartara* in the manuscript. However, this form is wrong according to Old Norse grammars, which specify the plural form of the comparative as *bjartari*, or, in the case of a manuscript with vowel harmony, as *bjartare*.⁵ At some point, the editor might want to correct the text here, although not necessarily on the diplomatic level. In fact, this is an early reflection of the merger of the endings of comparatives that took place in Old Norwegian

⁵ Vowel harmony was a characteristic trait of Old Norwegian and meant that the height of an unstressed vowel in the ending of a word was controlled by the height of the stressed root vowel. For example, in a word with a high root vowel, such as *lif* n. ‘life’, the unstressed vowels would also be high, *lifi* and *lifum*, while in a word with a non-high root vowel, such as *lof* n. ‘praise’, the unstressed vowels would also be non-high, *lofe* and *lofom*.

during the thirteenth century, and should probably be transcribed as such. Finally, the form *tíu* in the manuscript has been analysed as a metathesis of *tuí* ‘double’, part of the adjective *tvíeggjaðr* ‘two-edged’. This correction seems obvious and also helpful for the users of the edition.

One of the strengths of the multi-level type of editing exemplified here is that it allows uncorrected and corrected text to live side by side, and it can also draw a useful distinction between what may be termed scribal interventions, such as the correction *driupande* > *driupanda*, and editorial interventions, such as *tíu* > *tuí*. At the facsimile level, the text is rendered “as is”, in an uncorrected state, apart from corrections made by the scribe himself, while on the diplomatic level, and even more so on the normalised level, editorial interventions are allowable. Furthermore, this division into levels has ramifications for the annotation of the text, especially, as we shall see below, with respect to linguistic annotation.

For classical scholars, the focus on minute variation may seem odd. Why immerse oneself in the accidentals of a text, when there are substantives to behold?⁶ The answer lies in the fact that vernacular texts are important sources for the history of the early stages of a written language. When the provenance and dating of a source have been established, the orthography has its story to tell. Often, it is a conflated story, since the orthography of a manuscript has to be understood in the context of the orthography of the exemplar, the linguistic norm of the scribe, and, sometimes, even the intended audience.⁷ In order to use the orthographical representation of any vernacular manuscript, these influences need to be identified and isolated. In some cases, this can be done with a high degree of certainty. In other cases, the language of the source appears inconsistent, which usually is understood as a conflict of norms, between the copy and the exemplar, or between the orthography of the manuscript and the internalised orthographical norms of the scribe. A normalisation of the orthography will remove these traces of norm conflict and lessen the source value of the

⁶ These terms were introduced in the much-quoted essay by Walter Wilson Greg, ‘The Rationale of Copy-Text’, *Studies in Bibliography*, 3 (1950–51): 19–37. Although referring to the editing of printed works, the distinction is equally relevant for classical and medieval works, in which the textual variation may be as plentiful and as complex as in printed works.

⁷ It has been argued that a number of manuscripts produced in Iceland in the late fourteenth century were written by Icelandic scribes in Old Norwegian orthography since these manuscripts were intended for export to Norway, cf. Stefán Karlsson, “Íslandsk bogeksport til Norge i middelalderen,” *Maal og Minne* 1979: 1–17.

manuscript. This might be fine for some scholars, but certainly not for those who are trying to extract the linguistic norm from it. In the model discussed above, the facsimile level and to some extent the diplomatic level offer the necessary level of granularity here, while the normalised level moves the text closer to a representation of the work and away from the source itself.

3. Encoding procedures

Most editorial projects nowadays will situate their work in an open, digital environment, encoding their texts in an interchangeable format. In recent years, this has become more or less equivalent with XML, *Extensible Markup Language*, an open and stable format for a variety of texts.⁸ This format was chosen for the archive that the author of this chapter heads, the Medieval Nordic Text Archive.⁹ A great number of other text archives also use XML, and many follow the guidelines set up by the Text Encoding Initiative (TEI).¹⁰

It has to be admitted that texts encoded in XML look forbidding as they stand. While the raw XML is not intended for most users of an archive, the editor must nevertheless understand and be comfortable with the basics of it. The upside is that when a text has been encoded in an open and to a great extent self-documenting format like XML, it will be accessible to a wide range of users hopefully over a very long period of time. An XML file is a straightforward text file, as simple as they come, and it will be readable as long as basic, unformatted text files can be read.

The *Guidelines* published by the Text Encoding Initiative, now in its fifth version, specify the encoding of a wide variety of sources, prose as well as poetry.¹¹ However, in our experience handwritten medieval sources require a number of additional specifications. *The Menota*

⁸ There are numerous introductions to XML, e.g. "What is XML and why should humanists care" by David J. Birnbaum, <http://dh.obdurodon.org/what-is-xml.xhtml>. Another introduction is Ch. 2, "Text encoding using XML" in v. 3.0 of *The Menota handbook*, <http://www.menota.org/handbook.xml>.

⁹ Cf. <http://www.menota.org>.

¹⁰ Among high-profile archives are the British National Corpus, Deutsches Textarchiv, Digitales Wörterbuch der deutschen Sprache, and the Perseus Digital Library of Latin and Greek texts. There are a host of other archives, of early and modern texts, and the list grows continuously. Many of the archives using the TEI guidelines, among them the Menota project, are listed by TEI at <http://www.tei-c.org/Activities/Projects/>.

¹¹ TEI: P5 *Guidelines*, <https://tei-c.org/guidelines/p5/>.

Handbook is intended to supplement the TEI Guidelines explaining and exemplifying how XML encoding as recommended by the Text Encoding Initiative can be used for the specific purpose of encoding vernacular medieval documents.¹²

For a linguistic annotation of a text, there are two categories that need to be clearly marked up: sentences and words. In XML, the <s> element groups each sentence, and the <w> element each word. To this pair of elements, a <pc> element should be added for punctuation characters, such as full stop, comma, colon and the like. Each element states its opening, e.g. <s>, and its end, e.g. </s>. So, this is how a sentence is contained in the <s> element, and the words in the <w> element and punctuation in the <pc> element:

```
<s>
  <w>This</w>
  <w>is<w>
  <w>a</w>
  <w>sentence</w>
  <pc>.</pc>
</s>
```

A multilevel edition can be encoded as parallel readings for each word, using elements such as <facs>, <dipl> and <norm> for the three levels exemplified above.¹³ The very first word would receive the following encoding:

```
<w>
  <facs>Sȳa</facs>
  <dipl>Sva</dipl>
  <norm>Svǎ</norm>
</w>
```

On the <facs> level, the usage of an Insular form of the ‘v’ is recorded. On the <dipl> level, this character is merged with the ordinary ‘v’, and on the <norm> level, an accent is added to the vowel to indicate that

¹² *The Menota Handbook* was published in v. 1.0 on 20 May 2003 and in v. 3.0 on 12 December 2019. See <http://www.menota.org/handbook.xml> for an overview of versions.

¹³ Since the elements <facs>, <dipl> and <norm> are not part of the standard TEI repertoire, they have been prefixed with ‘me:’ (for Menota-specific elements) in the XML encoding, but this prefix has been left out in the examples here for the sake of conspicuity.

it is phonemically long. In order to make the multi-level structure explicit, the element <choice> states that the contents of this element are alternatives:

```
<w>
  <choice>
    <fac>Sȳa</fac>
    <dipl>Sva</dipl>
    <norm>Svá</norm>
  </choice>
</w>
```

It should be underlined that the encoding examples given here are not meant to be typed, character by character, by a transcriber. They are the representations of transcriptions that would be done by various input methods.¹⁴

The actual encoding of a text on several levels is no guarantee that it can easily be displayed in a manner that is accessible for any non-technical reader. In the Menota archive, the display is based on the Corpuscle application.¹⁵ As shown in Fig. 2, this application allows a text to be displayed at up to three parallel levels, including a photographic facsimile. The Corpuscle application is used for several other archives, some of which are rather close in scope and format to Menota, such as the Georgian National Corpus, covering Old, Middle and Modern Georgian.¹⁶

While the characters on the normalised level and largely on the diplomatic level can be rendered by almost any font, many of the characters on the facsimile level require specialised fonts. Until a few years ago, this meant that users of the archive had to install a font containing the necessary characters. Such fonts have been offered by the Medieval Unicode Font Initiative since 2004, and several of these fonts can be

¹⁴ *The Menota Handbook* exemplifies this in its Tutorial, which was introduced in v. 3.0 of the handbook, https://www.menota.org/HB3_T1.xml.

¹⁵ The Corpuscle application is a corpus management system for annotated texts developed by Paul Meurer at the University Library, Bergen, cf. <http://clarino.uib.no/korpuskel>.

¹⁶ The Georgian National Corpus is available at <http://clarino.uib.no/gnc/page>. For other corpora, see Paul Meurer, 'Corpuscle – a new corpus management platform for annotated corpora', in *Exploring Newspaper Language: Using the Web to Create and Investigate a large corpus of modern Norwegian*, ed. by Gisle Andersen (Amsterdam: John Benjamins, 2012), pp. 29–50.

NRA 7 – A fragment of Landslog Magnúss Hákonarsonar v. 1.1

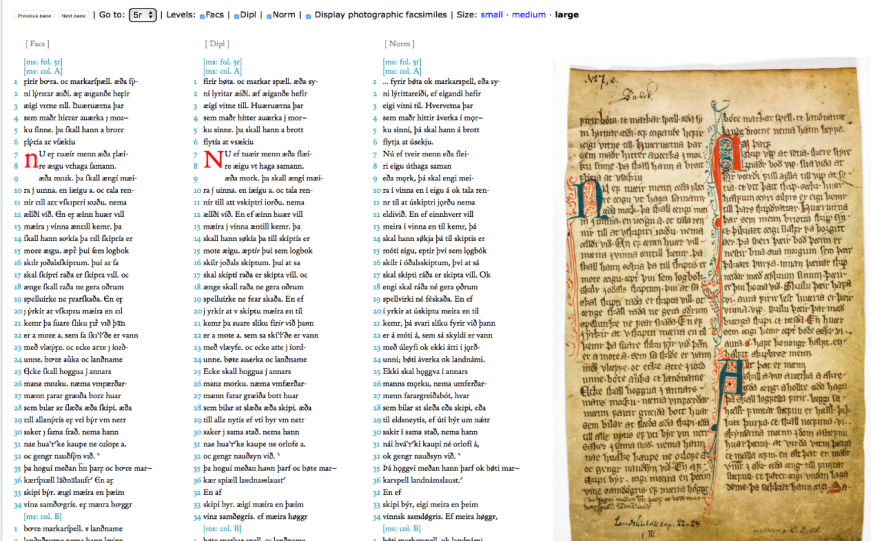


Figure 2. Display of a multilevel transcription of an Old Norwegian law fragment. Oslo, National Archives, Norse Fragment 7, f. 5r, col. A (ca 1320–1350). The three columns render the text on a facsimile, diplomatic and normalised level, based on a single transcription in XML. A photographic image of the fragment complements the display. Copyright: Odd Einar Haugen, License: CC BY-NC-ND.

downloaded, installed and used free of charge.¹⁷ In spite of the availability of free fonts, Menota could not offer a “plug and play” solution, and some users were surely put off by missing characters, boxes or question marks in the web display. It was a great step forward when the Web Open Font Format (WOFF) was officially launched in 2012.¹⁸ This means that any recent browser can display all necessary characters on the fly, irrespective of which fonts happen to be installed on the user’s computer.

Having established this simple (although admittedly verbose) model of representing sentences and words on more than one level in XML, the next question is how to enhance the text with additional information.

¹⁷ The Medieval Unicode Font Initiative was established in 2001 and has published several recommendations for font usage as well as several free Unicode fonts that have a good selection of medieval characters. See <http://www.mufi.info>.

¹⁸ The Web Open Font Format was developed in 2009 and made a recommendation by the World Wide Web Consortium (W3C) in December 2012, cf. <https://www.w3.org/TR/2012/REC-WOFF-20121213/>.

4. Annotating a text

There is a plethora of textual features that can be identified and annotated: motives and themes, allusions, rhetorical devices, names of persons and places, stylistic features, metrical properties, allusions to or readings from other texts, and so on. From a linguistic point of view, phonological, morphological, syntactic and lexical features are all relevant, but, in my experience, morphology and syntax are particularly well suited for annotation.

4.1 Morphological annotation

A morphological annotation will as a minimum specify the lemma, i.e. headword, of each running word in the text, and usually also the grammatical form. In the Menota XML, this information is added to each word by way of attributes to the <w> element. In the example of the adverb *svá* (as the entry is spelt in an Old Norse dictionary), the lemma attribute is simply *svá*, as shown in a slightly simplified encoding:

```
<w lemma="svá">
  <fac>Sya</fac>
  <dipl>Sva</dipl>
  <norm>Svá</norm>
</w>
```

Adverbs like *svá* are not inflected, so the only grammatical information in addition to the lemma will be its word class (part of speech). In the Menota project, the *msa* attribute (the full form being *me:msa*) specifies the morpho-syntactic analysis of the word. An adverb will receive the value *xAV*, in which *x* signifies word class and *AV* adverb:

```
<w lemma="svá" msa="xAV">
  <fac>Sya</fac>
  <dipl>Sva</dipl>
  <norm>Svá</norm>
</w>
```

Words with more grammatical categories, such as nouns, adjectives and verbs, have a longer list of values for the *msa* attribute, but the principle is the same, so that, for example, the noun *varrar* ‘lips’ will, in addition to its word class, be annotated for gender, case, number and species (the latter category has two values, indefinite as in *varrar* ‘lips’ or definite as in *varrarnar* ‘the lips’):

fótr	hennar	ftiga	niðr	til	dauða
common noun	pers. pron.	verb	adv.	prep.	common noun
nom., pl., m., strong	gen.	ind., pres., act., 3rd p., pl.	pos.	non-infl.	gen., sg., m., strong
<i>fótr</i>	<i>hon</i>	<i>stiga</i>	<i>niðr</i>	<i>til</i>	<i>dauði</i>
ok	liggia	gætur	hennar	til	hælvitíf
conj.	verb	common noun	pers. pron.	prep.	common noun
non-infl.	ind., pres., act., 3rd p., pl.	nom., pl., f., strong	gen.	non-infl.	gen., sg., n., strong
<i>ok</i>	<i>liggia</i>	<i>gata</i>	<i>hon</i>	<i>til</i>	<i>helvíti</i>

Figure 3. A morphological analysis of the last sentence in Fig. 1. The first line has the diplomatic reading of each word, the second line the word class, the third line the grammatical categories, and the fourth line the head words in a standard, Old Norse dictionary. Copyright: Odd Einar Haugen, License: CC BY-NC-ND.

```
<w lemma="vqrr" msa="xNC gF cN nP sI">
  <fac>varrar</fac>
  <dipl>varrar</dipl>
  <norm>varrar</norm>
</w>
```

The *msa* attribute contains one or more name tokens, each specifying a grammatical category; in this case *xNC* for “noun common”, *gF* for “gender feminine”, *cN* for “case nominative”, *nP* for “number plural” and *sI* for “species indefinite”.¹⁹

Assuming that each word of a text has received morphological annotation, the information can be displayed in various ways, including tabular displays such as the one in Fig. 3. The actual encoding and the display is more or less self-explanatory and draws on a long tradition of traditional dictionary archives based on the venerable index card.²⁰

It should be pointed out that the usefulness of a linguistic annotation is dependent on the variability of the orthography of the texts. For a literature in which the language of the texts is highly regularised, such as in most corpora of modern texts, a morphological analysis can to a

¹⁹ In v. 3.0 of *The Menota Handbook*, levels are discussed in Ch. 4 and linguistic annotation in Ch. 11.

²⁰ One of many archives of this type is the *Gammelnorsk Ordboksverk* [Old Norwegian Dictionary Project], which worked with index cards from the 1950s until the early 1980s. Many of these archives have been converted into digital databases, but still present part of their material in facsimiles of the original index card collections, such as the *Ordbog over det norrøne prosasprog* [Dictionary of Old Norse Prose] and *Gammeldansk Ordbog* [Old Danish Dictionary], both in Copenhagen.

high degree be done by semi-automatic analysis. For medieval vernacular texts in which the orthography is variable not only between sources but even within sources, a linguistic annotation really comes into its own. By the same token, the annotation is more time-consuming and less suited for automatisisation. The inflection of many Old Norse words illustrates this point, e.g. the verb *verða* ‘become’, which has these (amongst other) forms:

verða	verð	verðum	varð	urðum	yrða	yrðim	orðit
	verðr	verðuð	vart	urðuð	yrðir	yrðið	
	verðr	verða	varð	urðu	yrði	yrði	

The initial *v-* is dropped in several forms, and the root vowel shifts between *e*, *a*, *u*, *y* and *o* due to a combination of Ablaut and Umlaut. This degree of variation within a paradigm is known from many other languages, but what really complicates the analysis here is the fact that each form could be spelt in more than one way, sometimes in a frustratingly high number of ways. For example, the regularised form *urðu*, 3rd person preterite indicative of *verða*, might (at least in theory) be spelt *urðu*, *urðv*, *vrðu*, *vrðv*, *urþu*, *urþv*, *vrþu*, *vrþv*, *urþu*, *urþv*, *vrþu*, *vrþv*, *urþo*, *urðo*, *vrþo*, *vrðo*, and more. On the normalised level, there would only be the form *urðu*, but on the diplomatic, and especially the facsimile level, there will be many more forms.

An added difficulty is that while Old Norwegian and Old Icelandic (i.e. Old Norse) have a normalised orthography, this is not the case for Old Swedish or Old Danish. The Old Norse normalised orthography was established in the late nineteenth century and is used with minor variation in standard grammars, dictionaries and in many editions.²¹ No similar norm exists for Old Swedish and Old Danish, even if Old Swedish texts in particular might be suited for normalisation.²² The Old Danish language is less conducive to normalisation, partly due to the fact that it evolved so quickly in the Middle Ages and partly due to the sparsity of sources up to ca 1300.²³ A similar lack of orthographic

²¹ Ivar Berg, ‘Om normalisert norrønt’, *Arkiv för nordisk filologi*, 129 (2014), 21–54.

²² Henrik Williams, ‘Normalizing Old Swedish texts: Why Not?’, in *Beyond the Piraean Lion: East Norse Studies from Venice*, ed. by Jonathan Adams and Massimiliano Bampi (Odense: Syddansk universitetsforlag, 2017), pp. 51–58.

²³ For Old Swedish and Old Danish, see the discussion in Odd Einar Haugen, ‘Normalisering av vest- og østnordiske middelaldertekster’, in *A Copenhagen Miscellany: Studies in East Nordic Philology*, ed. by Simon Skovgaard Boeck and Seán D. Vrieland (Copenhagen: Syddansk Universitetsforlag, 2019), pp. 161–181.

norm applies to other European vernaculars, and as a consequence, morphological annotation is a desideratum across the board for the European vernaculars.

In spite of these difficulties, morphological annotation of the type discussed above is a fairly simple undertaking and it is not the object of many linguistic controversies. After all, the aim of the annotation is to link texts to existing resources such as grammars and dictionaries, and, consequentially, the grammatical categories will be traditional. There is some variation in matters such as the lemma orthography and the grammatical categories, especially word classes, but at least within Medieval Nordic philology, these problems are relatively small.²⁴

4.2 Syntactic annotation

Syntactic annotation is more of a challenge than morphological annotation. Competing syntactic models have evolved over the years, and there is a varying degree of compatibility between them. For a language of comparatively free word order, such as the Medieval Nordic languages, it seems that dependency analysis is a suitable and fairly simple syntactic model.²⁵ In dependency analysis, each word is described by its function and hierarchical position within the sentence. This is typically displayed in a tree with labels for each word specifying its function, as shown in Fig. 4. It is a characteristic and perhaps unexpected trait of dependency analysis that words rather than phrases are assigned syntactic functions. There are some non-intuitive consequences of this, for example that conjunctions are analysed as heads (as in Fig. 4) and for this reason have full sentences as their dependents. However, the internal hierarchy of the coordinated sentences, each having a predicate as its head, is not affected by the fact that the conjunction has been elevated, as it were, to the position of a head.

While morphological annotation easily can be incorporated in the XML discussed above, syntactic annotation is better carried out in a separate module. The PROIEL project developed at the University of

²⁴ *The Menota handbook* v. 3.0, Ch. 10 and 11.

²⁵ For a general introduction to dependency analysis, see Igor A. Mel'čuk, *Dependency Syntax: Theory and Practice* (Albany: State University of New York Press, 1988). An implementation for Old Norwegian is offered by Odd Einar Haugen and Fartein Th. Øverland, *Guidelines for Morphological and Syntactic Annotation of Old Norwegian Texts* (Bergen Language and Linguistic Studies, vol. 4:2, Bergen, 2014).

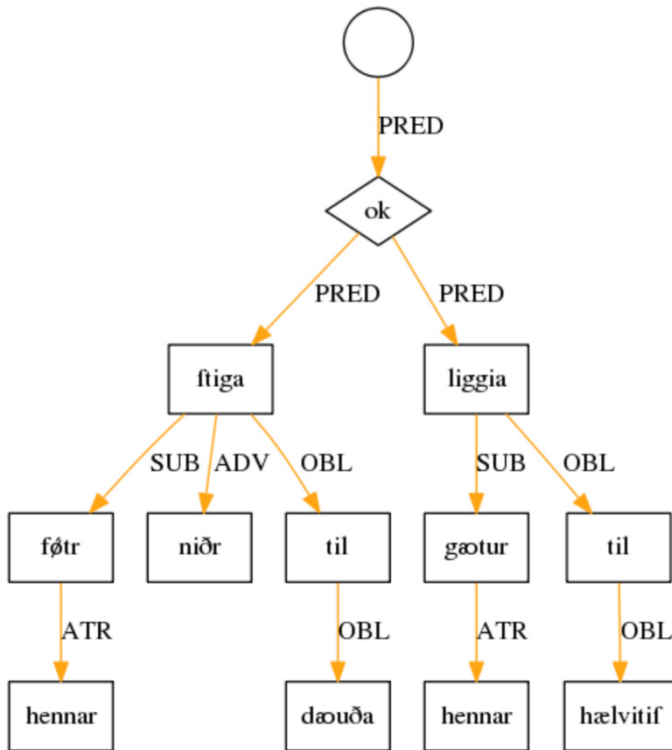


Figure 4. A dependency analysis of the last, coordinated sentence in Fig. 1. The conjunction *ok* ‘and’ functions as the head of the two coordinated sentences, while each of these has a predicate as its head, *stiga* ‘go’ and *liggia* ‘lie’. The subject of the first sentence is *fótr* ‘feet’, with *hennar* ‘her’ as an attribute, and the subject of the second is *gætur* ‘roads’ also with *hennar* ‘her’ as an attribute. The preposition *til* ‘to’ in both sentences is heading an oblique specifying the goal of the predicate, dominating *dauða* ‘death’ and *hælvitis* ‘hell’ respectively, and finally, the adverb *niðr* ‘down’ specifies the direction of the predicate *stiga* ‘go’. Copyright: Odd Einar Haugen, License: CC BY-NC-ND.

Oslo offers exactly this type of annotation environment.²⁶ PROIEL initially undertook a syntactic analysis of the New Testament in five old Indo-European languages, the original Greek and early translations into Latin, Gothic, Armenian and Old Church Slavonic. Through cooperation with other projects, the annotated corpus in the PROIEL format has later been extended to many more languages, among them

²⁶ The original project was funded by the Norwegian Research Council in the period 2008–2012 and was led by Dag Trygve Truslew Haug. The core group also included Hanne Martine Eckhoff and Marius Larsen Jøhndal.

Old Icelandic, Old Norwegian and Old Swedish.²⁷ While the PROIEL project was originally designed for the study of information structure, all texts were annotated for morphology and on the basis of this also for syntax. The result is a deep annotation with considerable linguistic information about each text, organised in the form of a treebank.²⁸ The original PROIEL treebank project and the subsequent projects now form what may be called a treebank family of early attested Indo-European languages, ranging from classical to medieval, and in some cases modern, stages in their development.²⁹ In total, approx. 1,6 million words have been annotated manually at a high level of accuracy. For Old Norwegian and Old Swedish, there are so far no other treebanks than those in the PROIEL family.³⁰

The dependency model illustrated here is in some respects close to lexical-functional grammar (LFG), but it contrasts with especially phrase structure models. However, as can be seen in the examples here, the functional categories used in dependency analysis are by and large familiar, such as predicate, subject and object, although some categories, especially the obliques and the external objects, are less familiar. Even so, there seems to be a sufficiently high degree of recognition between central syntactic models. The major criterion in such cases is, I believe, that an insight is only a fruitful insight if it can be transferred from one model to another; if not, it may be an insight solely into the model, not

²⁷ These projects include the ISWOC project for Old English and several Romance languages, the TOROT project for Russian, the Menotec project for Old Norwegian, and the MAPIR project for Old Swedish. PROIEL has also added many Greek and Latin texts to the original New Testament texts. The exact number of languages depends on the classification; five major language families are represented, i.e. Armenian, Germanic, Greek, Romance and Slavic, and all but Armenian have several branches. The total of languages (or linguistic stages) covered by PROIEL is 18.

²⁸ *Treebank* is a term that reflects the fact that a syntactic analysis typically takes the form of a tree, and that a collection of such trees make up a bank. The term gained usage in the 1990s after the Penn Treebank Project was established in 1989. Cf. Mitchell P. Marcus, Beatrice Santorini and Mary Ann Marcinkiewicz, 'Building a Large Annotated Corpus of English: The Penn Treebank', *Computational Linguistics*, 19/2 (1993): 313–30.

²⁹ Hanne Martine Eckhoff, Kristin Bech, Gerlof Bouma, Kristine Gunn Eide, Dag Trygve Truslew Haug, Odd Einar Haugen and Marius Larsen Jøhndal, 'The PROIEL treebank family: A standard for early attestations of Indo-European languages', *Language Resources and Evaluation* (2017): 1–37, <https://doi.org/10.1007/s10579-017-9388-5>.

³⁰ The Old Norwegian texts are part of the Menotec collection accessible through the INESS portal at <http://clarino.uib.no/iness>, and they are in the process of being published on the *Syntacticus* website at <http://www.syntacticus.org>.

what the model purports to explain. A case in point is the fact that a dependency tree (with some modifications) can successfully be converted to an LFG representation, and the other way round; in other words, dependency and LFG models are able to express similar analytic insights.³¹

5. The appeal of annotation

A text annotated for morphology and syntax is indeed a boon for the linguist and the language historian. For many other scholars, for example of a literary or historical inclination, the annotation is one of several resources for a better understanding of opaque or ambiguous passages in a text. The annotated text is a close cousin of the commentary; while the latter can go into extensive detail and list a number of interpretations, the annotated text makes a decision and is usually unambiguous, unless the categories themselves have been designed to be polyvalent.

Old Norse poetry offers a host of complex and enigmatic passages. In the Eddic poem *Völuspá* ‘The Prophecy of the Seeress’, stanza 2.4–6 is still unresolved. As Fig. 5 from *Codex Regius* shows, the poem is written in continuous lines, and the script is somewhat difficult to read here.

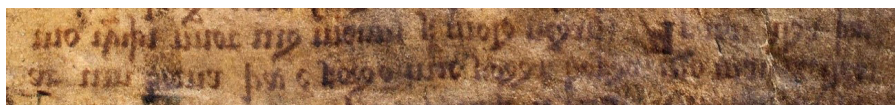


Figure 5. Stanza 2.5–8 of *Völuspá* in Reykjavík, The Árni Magnússon Collection, GKS 2365 4º, f. 11, l. 4–5. The words “nio man ec heima” can be read (with some difficulty) at the end of the first line in the photograph (i.e. l. 4 in the manuscript), while the remaining part follows on the next line, “nio iðiþi [?] miot uið mæran fyr mold neðan”. Copyright: The Arnamagnæan Collection, License: CC BY-NC-ND.

Below, the stanza is quoted in the edition by Gustav Neckel and Hans Kuhn³² and the translation by Henry Adams Bellows.³³

³¹ Dag [Trygve Truslew] Haug, ‘From dependency structures to LFG representations’, in *Proceedings of LFG12*, ed. by Miriam Butt and Tracy Holloway King (Stanford: CSLI Publications, 2011), pp. 271–291, and Paul Meurer, ‘From LFG structures to dependency relations’, in *The very model of a modern linguist – in honor of Helge Dyvik*, ed. by Victoria Rosén and Koenraad De Smedt (Bergen: Bergen Language and Linguistic Studies, vol. 8, 2017), pp. 183–201.

³² Gustav Neckel and Hans Kuhn, eds, *Edda: Die Lieder des Codex Regius nebst verwandten Denkmälern*, Vol. 1: *Text*. 5th ed. (Heidelberg: Winter, 1983).

³³ Henry Adams Bellows, transl., *The poetic Edda*, Scandinavian classics, vol. 21–22 (Princeton: Princeton University Press, 1936).

1 Ec man iotna 2 ár um borna,
 3 þá er forðom mic 4 fædda hofðo;
 5 nío man ec heima, 6 nío íviði,
 7 miotvið mæran 8 fyr mold neðan.

1 I remember yet 2 the giants of yore,
 3 Who gave me bread 4 in the days gone by;
 5 Nine worlds I knew, 6 the nine in the tree
 7 With mighty roots 8 beneath the mold.

A much-debated question concerns the reading and understanding of the phrase “nío iviði” in stanza 2.6 of the poem. This has been taken as normalised *níu í viði* ‘nine in the tree’ in many editions, since there commonly was no word division between a preposition and its complement, and *viði* is a *bona fide* accusative of *viðr* m. ‘wood, tree’. However, after studying the manuscript closely, some philologists conclude that there is an almost invisible abbreviation character after “viði”, in the form of an *ur* sign. If this is correct, the reading of the word becomes “iviði-ur”, normalised *íviðjur*, meaning ‘giantesses’. This reading happens to be supported by another manuscript, *Hauksbók* in Copenhagen, The Arnamagnæan Collection, AM 544 4^o, f. 20r, l. 3, so even if the reading in GKS 2365 4^o is pending, *íviðjur* ‘giantesses’ has been adopted in the latest edition of the Eddic poems.³⁴

In the annotated version of this poem, the latter interpretation has been selected, as can be seen from the morphological analysis in Fig. 6.³⁵

níu	man	ek	heima	níu	íviðjur
card. num.	verb	pers. pron.	common noun	card. num.	common noun
non-infl.	ind., pres., act., 1st p., sg.	nom., sg.	acc., pl., m., strong	non-infl.	acc., pl., m., weak
<i>níu</i>	<i>muna</i>	<i>ek</i>	<i>heimr</i>	<i>níu</i>	<i>íviðja</i>
mjötvið	mæran	fyr	mold	neðan	
common noun	adj.	prep.	common noun	adv.	
acc., sg., m., strong	pos., acc., sg., m., strong	non-infl.	acc., sg., f., strong	pos.	
<i>mjotviðr</i>	<i>mærr</i>	<i>fyrir</i>	<i>mold</i>	<i>neðan</i>	

Figure 6. The morphological annotation of *Völuspá* stanza 2.5–8. Copyright: Odd Einar Haugen, License: CC BY-NC-ND.

As for the syntax of the half-stanza, the analysis in Fig. 7 juxtaposes one full sentence, *níu man ek heima*, with two elliptical sentences, in

³⁴ *Eddukvæði*, ed. by Jónas Kristjánsson and Vésteinn Ólason, vol. 1, p. 291 (*Konungsbók* [= Codex Regius]) and p. 308 (*Hauksbók*) (Reykjavík: Hið íslenska fornritafélag, 2014).

³⁵ See PROIEL at <http://foni.uio.no:3000>, under “Old Norse”. This text may eventually be moved to the *Syntacticus* website, <http://syntacticus.org>.

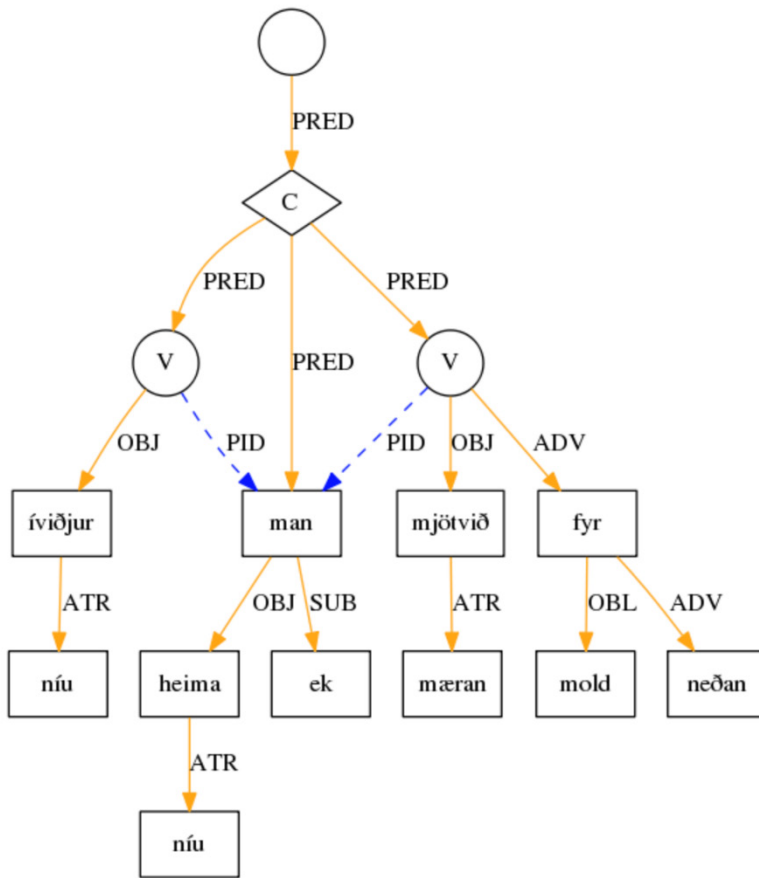


Figure 7. The syntactic annotation of *Völuspá* stanza 2.5–8. The main predicate is *man* (of *muna*) ‘remember’, which is present in st. 2.5 and must be seen as covert in the two other sentences, 2.6 and 2.7–8, represented by an encircled “V”, and slashed lines pointing to *man* in order to indicate identity with the overt predicate. Copyright: Odd Einar Haugen, License: CC BY-NC-ND.

which there is no overt predicate or subject, *nú* [*man ek*] *íviðjur* and *mjötvíð* [*man ek*] *mærán fyr mold neðan*.

In this analysis, *íviðjur* ‘giantesses’ is explicitly analysed as the object in *nú* [*man ek*] *íviðjur* ‘nine [I remember] giantesses’, and *nú* ‘nine’ as an attribute. It is an analysis that makes rather heavy assumptions about ellipted words, but, assuming that *íviðjur* is a noun rather than a preposition and its complement, the present analysis seems to be the best one.

The point of this example is not that a linguistic annotation offers the definitive answer to an enigmatic reading. That would be presumptuous. Rather, what it offers is a clear and consistent analysis of each

word in each sentence of the poem in question, not skipping over any difficult passages, and as such it has a high degree of explanatory potential. It offers a point of reference for any interpretation of the text.

6. Costs and benefits

Nobody would deny that annotation enhances an edition and makes it accessible for a wider audience. What is a matter of discussion is the cost of annotation compared to the benefits it offers. For a literature where many texts are still awaiting an edition (or a sufficiently good edition), the priority would be to edit the remaining texts, unless these texts by common consent were regarded as being of too little value even for the most avid scholar of the period. This is not the case for the field this author is most familiar with, Old Norse literature. The great majority of Old Norse works have been edited, many several times, and there are few works that are not available in a decent edition. Most works are preserved in more than one manuscript, but the really large manuscript traditions known from e.g. classical scholarship are few and far between. In Old Norwegian literature, the 15 preserved manuscripts of *Barlaams saga ok Jósafats* count as a rather large tradition, and *Konungs skuggsjá* with 60 preserved manuscripts (the majority being younger Icelandic ones) is one of the largest manuscript traditions, only surpassed by the law code of Magnús Hákonarson. A notable trait is the fact that almost all of the earlier manuscripts of these work are fragmentary, so that the text of each work has to be pieced together from several textual witnesses. And, as stated above, these manuscripts typically have different orthographies, representative of their time and locale.

In comparatively small textual traditions it makes sense and is in many cases feasible to transcribe each of the manuscripts. Any critical edition can only offer a glimpse of the textual variation through its apparatus, and while for many scholars it is fine to have an apparatus that only contains the substantive variants, for other users, the variation in accidentals is as interesting. Even in a fairly small vernacular manuscript tradition, there are simply too many variants for a workable apparatus, so the only way to record them is to edit each manuscript as an individual witness to the work. These transcriptions will preferably be digital ones published in text archives and searchable within these archives.

Many editors would be happy with an edition of the text as it appears in a single, typically best, manuscript. Probably the majority of texts published in an archive like the Medieval Nordic Text Archive will remain at this level. However, some editors would like to add

annotation in order to open up the text, and other scholars would like to contribute by annotating editions by previous editors. Incrementally, more and more texts will be annotated. On the whole, this process is likely to be self-regulatory, although canonical texts are more prone to receive annotation than other texts. GKS 2365 4^o, the major source for the Eddic poems, is a prime example of this type of text. The question of cost will ultimately be decided on the background of the canonicity and thus the general interest in the texts to be annotated.

Bibliography

- Alcuin of York, *De virtutibus et vitiis liber, ad Widonem comitem*, in *Patrologia Latina*, ed. by Jacques-Paul Migne, vol. 101, cols 613–638 (Paris: Garnier freres, 1863).
- Bellows, Henry Adams, transl., *The poetic Edda* (Princeton: Princeton University Press, 1936).
- Berg, Ivar, ‘Om normalisert norrønt’, *Arkiv för nordisk filologi*, 129 (2014), 21–54.
- Eckhoff, Hanne Martine, Kristin Bech, Gerlof Bouma, Kristine Gunn Eide, Dag Trygve Truslew Haug, Odd Einar Haugen and Marius Larsen Jøhndal, ‘The PROIEL treebank family: A standard for early attestations of Indo-European languages’, *Language Resources and Evaluation* (2017): 1–37, <https://doi.org/10.1007/s10579-017-9388-5>.
- Eddukvæði*, ed. by Jónas Kristjánsson and Vésteinn Ólason, vol. 1 (Reykjavík: Hið íslenska fornritafélag, 2014).
- Gamal norsk homilieboek*, ed. by Gustav Indrebø (Oslo: Kjeldeskriftfondet, 1931).
- Greg, Walter Wilson, ‘The Rationale of Copy-Text’, *Studies in Bibliography*, 3 (1950–51): 19–37.
- Haug, Dag [Trygve Truslew], ‘From dependency structures to LFG representations’, in *Proceedings of LFG12*, ed. by Miriam Butt and Tracy Holloway King (Stanford: CSLI Publications, 2011), pp. 271–291. <http://web.stanford.edu/group/cslipublications/cslipublications/LFG17/papers/lfg12haug.pdf>.
- Haugen, Odd Einar and Fartein Th. Øverland, *Guidelines for Morphological and Syntactic Annotation of Old Norwegian Texts* (Bergen Language and Linguistic Studies, vol. 4:2, Bergen, 2014), <https://bells.uib.no/index.php/bells/issue/view/158>.
- Haugen, Odd Einar, ‘Levels of Granularity: Balancing Literary and Linguistic Interests in the Editing of Medieval Vernacular Texts,’ in *Philology Matters!*

- Essays on the Art of Reading Slowly*, ed. by Harry Lönnroth (Leiden: Brill, 2017), pp. 118–135.
- Haugen, Odd Einar, ‘Normalisering av vest- og østnordiske middelaldertekster’, in *A Copenhagen Miscellany: Studies in East Nordic Philology*, ed. by Simon Skovgaard Boeck and Seán D. Vrieland (Copenhagen: Syddansk Universitetsforlag, 2019), pp. 161–181.
- Karlsson, Stefán, ‘Islandsk bogeksport til Norge i middelalderen’, *Maal og Minne*, 1979: 1–17.
- Marcus, Mitchell P., Beatrice Santorini and Mary Ann Marcinkiewicz, ‘Building a Large Annotated Corpus of English: The Penn Treebank’, *Computational Linguistics*, 19/2 (1993): 313–330.
- Mel’čuk, Igor A., *Dependency Syntax: Theory and Practice* (Albany: State University of New York Press, 1988).
- The Menota handbook: Guidelines for the electronic encoding of Medieval Nordic primary sources*, gen. ed. Odd Einar Haugen, version 3.0. (Bergen: Medieval Nordic Text Archive, 2019), http://www.menota.org/HB3_index.xml.
- Meurer, Paul, ‘Corpuscle – a new corpus management platform for annotated corpora’, in *Exploring Newspaper Language: Using the Web to Create and Investigate a large corpus of modern Norwegian*, ed. by Gisle Andersen (Amsterdam: John Benjamins, 2012), pp. 29–50.
- Meurer, Paul, ‘From LFG structures to dependency relations’, in *The very model of a modern linguist – in honor of Helge Dyvik*, ed. by Victoria Rosén and Koenraad De Smedt (Bergen: Bergen Language and Linguistic Studies, vol. 8, 2017), pp. 183–201, <https://bells.uib.no/index.php/bells/issue/view/249>.
- Neckel, Gustav, and Hans Kuhn, eds, *Edda: Die Lieder des Codex Regius nebst verwandten Denkmälern*, Vol. 1: *Text*. 5th ed. (Heidelberg: Winter, 1983).
- Szmarach, Paul E., ‘A Preliminary Handlist of Manuscripts Containing Alcuin’s *Liber de virtutibus et vitiis*’, *Manuscripta*, 25 (1981), 131–140.
- Williams, Henrik, ‘Normalizing Old Swedish texts: Why not?’, in *Beyond the Piraeus Lion: East Norse Studies from Venice*, ed. by Jonathan Adams and Massimiliano Bampi (Odense: Syddansk universitetsforlag, 2017), pp. 51–58.

Manuscripts

Copenhagen, The Arnamagnæan Collection/Den Arnamagnæanske Samling

AM 619 4°

AM 666 b 4°

AM 544 4°

Oslo, National Archives

Norse Fragment 7

Reykjavík, The Arnamagnæan Collection/Handritasafn Árna Magnússonar
GKS 2365 4^o

Uppsala, University Library

De la Gardie collection, DG 4–7 fol

Web resources (last accessed 3 April 2020)

The *Corpuscle* application: <http://clarino.uib.no/korpuskel>

Gammeldansk Ordbog: <http://gammeldanskordbog.dk>

The INESS portal: <http://clarino.uib.no/iness>

The Medieval Nordic Text Archive (website): <http://www.menota.org>

—— (catalogue): <http://clarino.uib.no/menota/catalogue>

The Medieval Unicode Font Initiative: <http://www.mufi.info>

Ordbog over det norrøne prosasprog: <http://onp.ku.dk>

The PROIEL Treebank: <https://proiel.github.io>

Projects using the Text Encoding Initiative: <http://www.tei-c.org/Activities/Projects/>

The Syntacticus web site: <http://www.syntacticus.org>

The Text Encoding Initiative, P5 Guidelines: <https://tei-c.org/guidelines/p5/>

The Web Open Font Format: <https://www.w3.org/TR/2012/REC-WOFF-20121213>