Article Title: The Old Norse Computer Tape Bank at Copenhagen

Article Author: van Arkel-de Leeuw van Weenen, Andrea

Issue Date: December 1987

Publisher: William R. Veder, Slavisch Seminarium, Universiteit van Amsterdam, Postbus 19188, 1000 GD Amsterdam (Holland)

Citation: *Polata Knigopisnaia: an Information Bulletin Devoted to the Study of Early Slavic Books, Texts and Literatures* 17-18 (December 1987): 88-95.

The Old Norse Computer Tape Bank at Copenhagen

ANDREA VAN ARKEL - DE LEEUW VAN WEENEN
Vakgroep VTW, Postbus 9515, 2300 RA LEIDEN

The aim of the Old Norse Computer Tape Bank is to establish a collection of machine readable texts based directly on manuscripts, or on reliable diplomatic editions (of which, however, regrettably few exist). Through standardization of transcribing methods and methods of encoding transcriptions we hope in time to establish a database of Old Norse rather than a mere collection of texts.

Introduction. In the field of Old Norse philology,[1] especially textual editing, computers are not yet widely used. It seems almost a tradition that editing an Old Norse text not only takes up to 10 years for the scholarly work, but even longer to be printed. Innovations therefore are adopted only slowly by the Old Norse philological community. In the last decade, several texts have been rendered machine readable. In some cases, stylistic research was intended, and the texts entered for this purpose were entered as published in the normalized critical Íslenzk Fornrit editions, or even taken from the modern Icelandic reading versions. These need not concern us further, as they can not be used for other types of research like orthography or grammar, since all pecularities have been removed. A few texts, however, were entered in a manuscript transcription, either for textual editing purposes or for orthographical and linguistic research or for both, namely:

[1] The term Old Norse covers both Old Norwegian and Old Icelandic. Old Norwegian is replaced around 1350 by Middle Norwegian, Old Icelandic is commonly reckoned to continue until the end of the 16th century. It can be defended on linguistical grounds, however, that modern Icelandic is still an Old Germanic language (for example, the original three vowel system in unstressed syllables still survives). The vast majority of extant texts is Old Icelandic.

1. Barlaams saga og Josafats (Bergen)[2]
2. Grágás (Saarbrücken)[3]
3. Elucidarius (Minneapolis)[4]
4. Möðruvallabók (Eindhoven).[5]

Some of these projects were not institutionalized but financed by external means. Workers on the projects wanted to safeguard the continuing availability of the texts and the other materials developed during the projects (databases, indices, concordances). Therefore Hans Fix (Saarbrücken), Evelyn Firchow (Minneapolis) and I convinced the Arnamagnæan Institute at Copenhagen that its collecting activities should not be limited to manuscripts, photographs and microfilms of MSS, but extended to machine readable (narrow) transcriptions of MSS. A standing committee consisting of Evelyn Firchow, Hans Fix, Andrea van Arkel and Peter Springborg (as representative of the Arnamagnæan Institute) was installed to deal with the problems connected with this Computer Tape Bank.

So far, the texts of Grágás (GkS 1157 fol), Járnsída (AM 334 fol) and Möðruvallabók (AM 132 fol) have been deposited, as have two fragments of Elucidarius (AM 674 a 4° and AM 675 4°); the other

---

[2] Magnus Rindal, *Barlaams ok Josaphats Saga*, Norsk Historisk Kjeldeskrift-Institutt 1979; Magnus Rindal and Harald Solevåg, *Barlaams ok Josaphats Saga (Sth. Perg. fol no 6), KWIC-konkordanser og frekvensordliste*, Bergen; 1976. As this was the first major undertaking in computerized editing in Old Norse, the capacities of the computer were not maximally exploited. Abbreviations were expanded before entering, thus mixing up reading and interpretation.

[3] H.Beck et al., *Projekt eines Grágás-Wörterbuches*, skandinavistik 4, 1974, 67ff.; M.Bonner - H.Fix, *Projekt "Untersuchungen zu altisländischen Rechtstexten"*, Computers and the Humanities 12,1978.

[4] Kaaren Grimstad, *Editing the OI 'Elucidarius' with the Aid of the Computer*, Amsterdamer Beiträge zur Älteren Germanistik 1986, 91ff. Evelyn S.Firchow, *Editing Medieval Manuscripts with the Help of the Computer: The Case of the Old Icelandic 'Elucidarius'* in *Sprachen und Computer*, Festschrift zum 75. Geburtstag von Hans Eggers, 9.Juli 1982, eds. Hans Fix, Annely Rothkegel and Erwin Stegentritt (Dudweiler 1982), 172-186.

[5] Andrea van Arkel – de Leeuw van Weenen, *Möðruvallabók, AM 132 fol, I. Lemmatized Index and concordance*. Leiden, 1987.

versions of the text of Elucidarius will follow upon its publication.

As the projects producing these machine readable texts were completely independent and as no standards for transcription existed, these transcriptions differ widely, both in the level of transcription, the actual format and the method of encoding. For some texts a graphemic transcription is used, with an occasional phonemic decision (*r* and *r rotunda* are not distinguished), while for others a graphetic transcription is used. Two types of graphetic encoding are employed, both making use of a base sign with an additional sign. In one type, the graphetic markers are tagged behind the word;[6] in the other, they are placed on the line below. Furthermore, the transcriptions differ in their encoding of additional characters in the Old Norse alphabet like *æ* or *thorn* and of the abbreviation signs which abound in ON manuscripts.

This state of affairs was felt to be rather impractical. The committee therefore decided to advocate some standardization. At the Sixth International Saga Conference in Helsingør in 1985, I presented a paper in the Workshop "Computer Aided Editing" concerned with the establishing of standards for material to be kept at the Computer Tape Bank at Copenhagen. As no changes were proposed either by the participants of the conference or by the readers of the published version, these standards are now operative. This means that the texts already in Copenhagen will be changed to this format and that new texts will only be accepted for storing in this format. It is not implied that these standards have to be used in future projects, as in any particular case, depending on hardware, software and manuscript, a different solution may be better.

However, we hope that the availability of the standards will lead to the entering of more texts in the computer, by giving aspiring MS editors a norm to start from, freeing them from the time-consuming task of developing their own complete transcription and coding system. With the availability of more texts, it might then be possible to build up a database system for investigating those texts.

Paleography.

The earliest extant Old Norse manuscripts date from the 12th century. Old Icelandic original manuscripts were produced until the late 16th

[6] Evelyn S. Firchow, Kaaren Grimstad and Stephen Gilmour, *The Old Icelandic 'Elucidarius': A diplomatic edition with the help of the Computer*, ALLC Bulletin 6/3 (1978) 292-301 and 7/1 (1979) 60-65.

century and as copies even until the early 20th century. Paleographically therefore, the relevant MSS are far from uniform. A uniform encoding system can not be made. The best option seemed to be to choose for a two-tiered system of transliteration: One line with graphemes and the next with codes for paleographic variants.[7]

Ex. for
1 2

In this way the odd (main) lines will comply to a uniform standard whereas the even ones can only be standard within a certain script type. For grammatical purposes it suffices to use odd lines only. Even lines can also be used to indicate the uncertainty of certain readings:[8]

n
*

Selection of texts.

Only texts which follow the MS closely can be submitted to the Computer Tape Bank. This does not imply that texts have to be transliterated (see below); also transcriptions proper (without paleographic variants) can be submitted. Effectively this means that only the odd lines of a complete transcription are given.

No Old Norse texts are exempted.

Transcription.

Two types of transcription can be distinguished, a *literal* one or a transcription proper (all different types of *a* or *r* will be transcribed as *a* or *r* and a *graphetic* one or transliteration (different types of *a* like long necked *a* will be distinguished; *r* and *r* rotunda will be kept apart). Transliterations are handled as 2-line transcriptions where the upper line gives a literal transcription, the lower line (number) codes to identify the particular allographs:

---

[7] This system was used by Hans Fix in his transliteration of the Konungsbók text of Grágás. See Hans Fix, Grágás, *Graphemische Untersuchungen zur Handschrift GKS 1157 Fol.*, Frankfurt;1979

[8] This should not be taken too far. Many ON manuscripts are such that quite a few identifications depend more on the context (word or phrase) than on the actual shape of the character. As long as the word is unambiguous, one is usually not even aware of ambiguities on the character level.

en for haN t(i) laxar
2 1 1 2

In this way, any transliteration can be used as a literal transcription by skipping the even records, while any transcription can be supplemented to a partial or complete graphetic transcription, depending on how much graphetic detail will be incorporated.

Format.
Each MS-line occupies one double record and is preceded by a reference to enable quick searching. The reference consists of a 3-digit folio number, *r* or *v* (for recto or verso), a column indication (*a*, *b*, *c*) and a 2-digit line number. So 011ra07 stands for line 7 in the first column of page 11r. The text itself starts on position 11. Positions 8, 9 and 10 remain free; they can be used for chapter numbers etc. Fixed record length is used. For the treatment of exceptionally long records see under layout.

Chapter headings have to be distinguished from the main text, as they are often by a different scribe and/or in a different colour. They can show an orthography pronouncedly different from the main text. The start of the heading is shown by @1, the end by @2:
... @1 capitulum @2 ...

Transition between chapters.
At the borderline between two chapters, the linear texts order may be disturbed. For example, the final words of a chapter can stand after the initial words and the chapter heading of the next chapter. To make the text accessible for computer work, the natural text order has to be restored.
== can be used to indicate a break of this type within a word:
fram==an,
==+ for a break between words: haN==+kom heim
Markings might be added to indicate the original line.

Initials.
Initials are marked by a preceding initial marker ^. In paleographically oriented transcriptions the marking can be extended to indicate the size (in lines) and the place of the initial: ^G or ^3G (initial over 3 lines) or ^-1+3G (initial G starting in the line above and extending over 3 lines). By the latter approach an unambiguous encoding of (coloured) one line initials is also possible: ^1G.

Representation of characters.
In a single manuscript more different signs occur than can be represented by single characters on a computer. The total number of manuscripts even from a single period contains even more different signs. The signs therefore have to be classified so that various classes can be represented by a combination of a character and a class symbol (as above with the ¯ for an initial). Within one class, confusion may arise as to whether a sign has to be represented as a single character or as a combination of a "basic" character and a diacritic. One tends to try for a phonemic solution, but this is not always possible. For æ a representation as a single character is chosen. Other ligatures are encoded by putting the elements of the ligature between square brackets. In the case of vowel phonemes this enables us to transcribe the position of the accents accurately.

The *thorn* is so frequent that a standard representation (*w*) is given. In the few instances where *w* or *W* do actually occur in the manuscript they can be represented as ligatures: [vv] or [VV]. The abundance of spellings for ǫ (or ö): o, ǫ, ó, ø, ǿ, etc and the difficulty of predicting all variants make it preferable to treat these as combinations of a base sign (o) and diacritics.

A character has both a particular shape and size. In the manuscripts, both size and shape (capitals, majuscules) can be used to emphasize a letter. Modern editions tend to transcribe both capitals, majuscules and enlarged minuscules as upper case letters. It is less ambiguous to transcribe primarely the shape and add a marker (*) for size. This makes the transcription less ambiguous, as the shape of the letter is always clear. In the many instances where one hesitates about whether the letter is enlarged or not, the decision is not so crucial, as the user will realize that size is continuous and therefore many decisions about size may be arbitrary. The pairs *u/v* and *i/j* are handled as graphical variants. No loss of information is involved as the nature of the variant is registered on the even (variant) lines of the transcription.

Ligatures are mostly accidental and triggered by lack of space. Almost any combination of characters can be involved. Ligatures can consist of 2 or 3 base characters and any part can be accompanied by diacritics. They are encoded by putting the elements between square brackets: [ar] for ar.

Accents are placed *after* their base sign. Taken in itself accents could

also be encoded *before* the base signs. There are, however, a number of superscript signs which logically belong *after* their base sign, like the tittle for an *er*-abbreviation. To transcribe v̄a for *vera* as *erva* would be inconvenient to say the least. Therefore these superscript signs are placed after their base signs, and for reasons of consistency the accents are placed there as well.

Superscript signs other than accents are of two types:
special abbreviation marks which occur only as superscripts. These are represented by a single code character.
alphabetic characters. In principle every alphabetic character can also occur as superscript. Therefore no individual codes can be allotted. Instead one has to use a marking for superscript. Of the two available possibilities (a marking to indicate that only the following character is superscript, or a bracket structure to indicate beginning and end of the superscript) the bracket structure is chosen. The main reason for this choice was the more natural way in which superscripts with superscripts could be handled. Thus v̊ is transcribed as v(ı), v̊ as v(ı$) or v(í).

Scribal corrections.
These are indicated in the usual way with ` ... ´ for intralinear insertions and ´ ... ` for marginal insertions. Crossed out or expunged characters are preceded by a 0 (zero).

Subscripted diacritics and abbreviation marks are only few. They get a single character representation like the superscript special signs.

Layout.
Each (pair of) record(s) should contain one manuscript line, and conversely each MS line should be placed in its own (pair of) record(s). In this way no end-of-line markers have to be introduced in the text. The reference field in each record contains page and line in the manuscript, so that markers to indicate new pages are not needed. If in an exceptional case, a very long MS line occurs (for example caused by a lengthy marginal insertion), the line can be split over two records, where the second record has a continuation mark (+) in position 8.

Editorial additions.
The editor may wish to introduce certain marks to make the text more understandable (like end-of-sentence markers), or to make computer processing of certain features possible (e.g. poetry markers). As it

is not possible to predict what kind of markers might be needed, no standards have been set. The accompanying information to each text will state what markers of this type are used. These markers can be suppressed whenever uniform marking is necessary.

Future development.

In the first place we hope and aim for a rapid extension of the number of available texts, now that the first publications show the advantages of computer aided editing and, most of all, the linguistic insights to be gained from auxiliary material like (automatically produced) concordances, frequency lists, retrograde wordlists and the like and especially from the tagged versions (which can be produced semi-automatically). In our situation, it is not possible to dictate which type of text or which period should be covered first.[9] Texts will be made computer readable either for editing purposes, in which case factors like existence, availibility and quality of previous editions count more than the requirements with regard to period or style of the Tape Bank, or as base material for linguistic analysis where mostly shorter texts will be chosen and the choice of text/period will depend on the linguistic phenomena one wants to study.

Secondly, we hope to extend this project from a mere collecting of machine readable texts to a full-fledged data base (text material plus query language) consisting of tagged texts, i.e. texts where each word is tagged with lemma, word class and relevant grammatical information (case, person). Searching can then be done not merely on word forms (e.g. manni) or character patterns (e.g words containing the character *x*, words starting with *c/q/k*, containing an *a* and ending in *n*, words containing *nt*), but also on lemma (all forms of a lemma), wordclass plus pattern (all verbal forms ending in *z* or grammatical classification (neutral dative singular) or combinations of the above (neutral dative singular ending in *e/i*. On this subject see for example: Hans Fix und Maria Bonner, *Ein Computerwörterbuch – Spielerei oder Hilfsmittel*, skandinavistik (1981) 107-113.

[9] If possible, we would give preference to untranslated Icelandic prose texts (not in the so-called learned style), and from amongst those we would opt for large texts evenly spaced through the complete Old Icelandic period.