Proposal for the encoding of material to be kept at the Computer Tape Bank in Copenhagen

by Andrea van Arkel

In the last decade some amount of computer aided research of Old Norse manuscript texts has been done. In the course of these projects several texts have been stored in computer readable form. In order to guarantee the continuing availability of those texts for scholarly research the Arnamagnæan Institute in Copenhagen has established a Computer Tape Bank, following a suggestion by Evelyn Firchow, Hans Fix and myself. This Computer Tape Bank (CTB) now contains copies of various Elucidarius manuscripts (Evelyn Firchow), Grágás Konungsbók and Járnsíða (Hans Fix) and Möðruvallabók (Andrea van Arkel). It is hoped that other texts will be added as texts are put into computer-readable form either for editing or other purposes. Also material on tape or diskette prepared for printing purposes can be converted for inclusion in the CTB. Only transcriptions that closely follow the manuscripts can be accepted however.

As no standards exist for the encoding of Old Norse texts, all of the text groups now available have their own coding and format. As long as there are only so few, this is inconvenient, but not yet disastrous. The accessibility of the CTB-texts will, however, be greatly enhanced when all texts stored are encoded and formatted in the same way. Only so can information be extracted in a simple way when one is interested in comparing and combining data from different manuscripts. Therefore I have drafted the following proposal for a coding standard for Old Norse texts, which I hope to discuss during the Saga Conference. Suggestions for changes and improvements can be brought up either in the discussion session or sent in before (preferably ready for copying). The CTB Committee will evaluate the suggestions and then write a coding standard.

It is hoped that the evolving coding standard will also encourage others to undertake work in this field, by giving them a norm to start from, thus freeing them from the time consuming development of a complete transcription and coding system.

The CTB standard applies only for the text to be stored in Copenhagen. Both the scholar who enters the text and the later user may prefer adapted versions, as the most practical code in each particular case depends on the locally available equipment (keyboards, visual displays, printers etc).

> It would, of course, not be a sound policy to leave keys on your keyboard unused if the signs they stand for in the CTB standard do not occur in your text at all, or may be just twice on a hundred pages, while encoding some frequent sign by a combination of two or three keystrokes.

As long as one recognizes the same categories, however, conversion to the CTB standard will be straightforward.

Work on the proposal and the discussions about it with Hans Fix and Gerard Walsh have brought up some points which are, I think, of wider interest than the coding of texts for computer work, but should be discussed in the context of edition standards in general. I hope they will lead to reconsideration of the current editing practice.

General requirements

Texts for the CTB should be texts which closely follow the manuscript, not normalized texts.
Texts should be submitted on magnetic tape, which should be 9 track, 1600 bpi and unlabelled. When more texts are submitted on one tape, each text should have its own file(s). Files should have fixed blocksize and fixed recordlength (preferably 120 characters). If possible all files should have the same blocksize and recordlength.
In order to check the datatransmission and the conversion to CTB standard it would be a good idea to add a small file (with a printout) containing the codes for all the different signs occurring in the text.

Types of transcription

Two types of transcription will 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 neck a will be distinguished; r and r rotunda will be kept apart). Both complete and partial graphetic transcriptions are possible (in the latter case only the allographs of some letters are kept apart). Transliterations are probably best handled as 2-line transcriptions where the upper line gives a literal transcription, the lower line number codes to identify the particular allograph:

```
        en for haN t(i) laxar
         1 1 2  12       2 12
```

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.

Text division

Each MS-line occupies one 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 chapternumbers etc.

Chapter headings

These 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 a pronouncedly different orthography from the main text. The start of the heading can be shown by @1, the end by @2 :

----- @1 capitulum @2 -----

Transition between chapters

At the borderline between two chapters the linear text order may be disturbed. For example the final words of a chapter can stand after the initial words and the chapter heading of the following chapter. To make the text accessible for computer work and searching, the natural word order should 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

Initials

Initials can be marked by a preceding ^. In paleographically oriented transcriptions the marking can be extended to indicate the size (in lines) of the initial and even the starting position : ^3 or ^3G (initial over 3 lines) or ^-1+3G (initial starting in the line above and extending over 3 lines). The latter approaches enable also an unambiguous encoding of one line (coloured) initials internally in a line: ^1G.

Characters

In any single manuscript occur more different signs than can be represented by single characters in a computer. The total of manuscripts even from a single period shows even more different signs. The signs have therefore 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 initial). Within one class there may arise confusion 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 the ae-ligature æ representation as one character is to be chosen as the sign stands for a phoneme that has not a more unambiguous single character representation; it is also very frequent. The case of the aa-ligature is not so easy; the phoneme can be simpler represented by á. Encoding as a ligature has the advantage that the position of accents can be better described: [aa]' for á , [a'a] for á , [aa'] for á and [a'a'] for á . The abundance of spellings for ǫ (or ö): o, ǫ, ǫ̊, ø, ꝛ, etc., the difficulties of predicting all possibilities and the difficulties of assigning the accents over the aa-, av- and ao-ligatures makes it preferable to store these as combinations of characters with diacritics respectively as ligatures (with diacritics). An exception should probably be made for the ae-ligature in those texts which distinguish æ and œ. The thorn is so frequent that we will allow it a standard representation. The choice has fallen upon w (and W). The few instances where real w or W occurs can be represented as ligatures of v and v: [vv] .

A character has both a particular shape and a size. In the manuscripts both shape and size can be used to emphasize a letter. As we are used to do this by capitalizing, both capital letters, majuscules and enlarged minuscules are transcribed and

printed as upper case letters. So and N are both transcribed as N, n as n, while remains a small capital. It will be less ambiguous to transcribe primarely the shape and add a marker for size. Not only will the shape be clear from the transcript, but in the many instances where one hesitates whether the letter is enlarged or not, the decision will not be so crucial. Allographs, as mentioned above can be handled by numbers on alternating lines.

manuscript	usual transcription	proposal
n	n	n
ɳ	N	*n
N	N	*N
ᴎ	N	N
ſ		s
s		S
S		*S

A second problem is posed by the pairs i and j, u and v. Very likely in both cases the two letters are just graphical variants in the same way as r and r rotunda. There is not a distribution like i for vowel and j for consonant. The fact that these allographs have been faithfully transcribed seems not so much to lie in their phonemic status or intrinsic importance, but rather in the accidental occurrence of the two in our orthographic system and on our typewriters. Banning one of each pair would free space on the keyboard for abbreviation marks and the like, just as the shape/size transcription introduced above will free C and O.

Ligatures

As most ligatures are accidental, apart from some frequent ones like av, almost any combination of characters can be involved. Ligatures usually consist of two parts, but can also consist of three parts. Moreover, each part can be accompanied by diacritics. Therefore a bracket structure is required to indicate both beginning and end of the ligature: ar [ar], a'a' [a'a']

Accents

Accents should be placed after their base sign.

NB. When using devices allowing the superimposing of accents (keyboards supporting 'dead' keys), it will be handier to have them in front of their base characters. After is chosen for the CTB to correspond to the rules for superscript signs, which belong logically after the base sign. When converting from before to after and vice versa, beware of combinations of accents and superscripts.

As accents can occur over any character a single code representation for accented characters is out of the question.

Superscript signs

Superscript signs like ~ are to be transcribed by single signs without special marking for superscript (these signs occur only as superscripts).

Superscript characters

Of these there are too many to allot special codes to each individually. Two systems are possible:

1. a marker to indicate that the following character is superscript: $\overset{aN}{m}$ could be represented by m↑a↑N ;
2. a bracket structure indicating beginning and end of the superscripts: $\overset{aN}{m}$ will be written m(aN).

The latter possibility is chosen as it has the possibility of coding supersuperscripts and is easier to handle by programs.

Some superscript entries are corrections by the scribe and indicated in the line by a correction mark. Here the superscript passage should not be enclosed in round brackets but by \.../.

Corrections

Superscript corrections should be inserted at their proper place (often indicated by a comma in the line) and placed in \../.

Marginal corrections should in the same way be placed within /...\. Letters that are crossed or dotted out can be preceded by a Ø. As crossed out words are usually few, repeating the code for each character is feasible. Anyway, most transcriptions will probable skip the deletions completely.

Subscripts

Subscript diacritics and abbreviation marks are fewer than their superscript counterparts. One might be tempted to introduce a subscript marker and where possible use this marker plus the corresponding superscript. Even though this would be theoretically nice, it would be tedious in practice. As the subscripted marks are few, they can get their own representation easily.

Layout

Each line should contain one manuscript line. In this way no end-of-line markers have to be introduced in the text. Also the reference numbers in each line give page and line in the MS, so that markers to indicate new pages are not needed.

Editorial additions

The editor may wish to bring in certain marks of his own to make the text more understandable or to make computer processing of certain features possible. I have added to the Möðruvallabók text, for example, end-of-sentence markers (|) ,markers to indicate beginning and end of direct speech (' "), commata , markers around dittographs (< >), and around verses (<< >>). It is, of course, not easy to predict what kind of markers may be needed for different text types, or to prescibe a standard for them. On the other hand it would be a waste to erase this type of information in order to get uniform texts. Probably it will be best if the editor uses his own discretion and states in the accompanying information very clearly what markers of this type occur. The user who is not interested in them can then always remove them from his private copy.

Codelist

Letters

MS	transcription	EBCDIC	MS	transcription	EBCDIC
a	a	81	A	A	C1
b	b	82	B	B	C2
c	c	83	C	C	83
d	d	84	D	D	C4
ð	{	C0			
e	e	85	E	E	C5
f	f	86	F	F	C6
g	g	87	G	G	C7
h	h	88	H	H	C8
i	i	89	I	I	C9
j	i j	89 91	J	I J	C9 D1
k	k	92	K	K	D2
l	l	93	L	L	D3
m	m	94	M	M	D4
n	n	95	N	N	D5
o	o	96	O	O	96
p	p	97	P	P	D7
q	q	98	Q	Q	D8
r	r	99	R	R	D9
ſ	s	A2	S	S	E2
s	S	E2			
t	t	A3	T	T	E3
u	v u	A5 A4	U	V U	A5 A4
v	v	A5	V	V	A5
w	[vv]		W	[vv]	
x	x	A7	X	X	A7
y	y	A8	Y	Y	E8
z	z	A9			

ø	Ø	D6			
þ	w	A6		W	E6
æ	æ (æ)	7C	Æ	Æ (')	79
œ	œ	51			
A	*A		a	*a	
B	*B		b	*b	
C	*c		c	*c	
D	*D		ð	*d	
E	*E		e	*e	
F	*F		f	*f	
G	*G		g	*g	
H	*H		h	*h	
I	*I		ı	*i	
J	{*I *J		J	{*i *j	
K	*K		k	*k	
L	*L		l	*l	
M	*M		m	*m	
N	*N		n	*n	
O	*o		o	*o	
P	*P		p	*p	
Q	*Q		q	*q	
R	*R		r	*r	
S	*S		s	*S	
T	*T		t	*t	
U	{*V *U		u	{*v *u	
V	*v		v	*v	
W	[*v*v]		w	[*v*v]	
Y	*y		y	*y	
Þ	*W		þ	*w	
Æ	*Æ (*')	5C 79	æ	*æ (*æ)	
Œ	*Œ	5C 52			
Ø	*O	5C D6			

Special signs

℔	4	F4
ℨ	6	50
;	;	5E

Ligatures

æ	æ (@)	7C	Æ	*Æ (*`)	5C 79
œ	œ	51	Œ	*Œ	5C 52
ꝛ	}	D0			
ꝓ	Z	E9			
ꝑ	X	E7			

all other ligatures are encoded with [ ]

Accents and abbreviation marks

´ accent	´ ($)	5B	˝	˝ ($$)	5B 5B
˛	˛ (#)	7B	°	(°) ((#))	
˙	˙ (:)	7A			
˜	˜ (|)	4F			
˘	˘ (¯)	A1			
ꝰ	9	F9			
˚	8	F8			
ˀ	ˀ (%)	6C			

Numerals

0	0	F0
1	1	F1
2	2	F2
3	3	F3
4	4	F4
5	5	F5
6	6	F6
7	7	F7
8	8	F8
9	9	F9

Coding signs

(	(	4D
)	)	5D
[	[	4A
]	]	5A
^	^	5F
*	*	5C
\	\	E0
/	/	61

Punctuation marks

.	.	4B
,	,	6B
- (hyphen)-		60
(space)	(space)	40

Pre-editing signs

<	4C	begin dittograph
>	6E	end dittograph
|	6A	end of sentence
'	7D	begin of direct speech
"	7F	end of direct speech
+	4E	divides pseudo-compounds
=	7E	joins elements of compounds
?	6F	stands for unreadable character
Ø	F0	before expunged letter