Getting to grips with chips and Early Middle English text variants: sampling *Ancrene Riwle* and *Hali Meidenhad*

Manfred Markus
University of Innsbruck

1 Introduction

The method of compiling a machine-readable corpus obviously depends on the degree and type of complexity that the texts offer, and also on the purpose which the corpus is supposed to have.

Middle English texts, and Early Middle English prose texts in particular, are extremely complex indeed, and complex in bewilderingly different ways. As Margaret Laing (1994: 126) has demonstrated, Early Middle English word forms still show occasional traces of inherited inflexion, and there is, in general, much orthographic, phonological and morphological variation. Beyond the many idiosyncratic, yet still linguistically graspable features of texts, we have MS variants where one MS is really three times as long as, and entirely different from, the other, so that collating MS variants means comparing apples with pears.[1]

Margaret Laing and the Edinburgh team have taken the amorphous structure of Early Middle English texts as a reason to avoid 'analysis by hand', drawing the conclusion that all texts of a planned corpus have to be tagged before they are analysed. But by tagging in a certain way, for example for word types and word formation, we predetermine the analyses that can be applied on the texts later. We could call this 'wytiwyg' ('What you tag is what you get.').

As to the second point raised above (the main purpose of a corpus), we must be aware of the fact that the Edinburgh scholars have in mind a very special aim of research: to obtain complete lists of the spelling variants of words in line with regions so that an Early Middle English atlas can be made, in the wake of LALME (*Linguistic Atlas of Late Middle English*). This is a reasonable goal, but not everybody's. The language of the *Ancrene Riwle*/*Wisse* and of the Bodley MS (so-called language AB) is obviously less a regionalect than a literary style, so that language features such as metaphors, idiomatic phrases and alliteration are or could be of interest, features that the Early Middle English atlas planned in Edinburgh would not consider, of course.

It seems plausible then that corpus compilers who have more general target groups in mind should produce diplomatic texts and leave open the possibility of various types of analysis. If the main purpose of a planned corpus is to allow general comparison between different texts, for example, of a certain genre, the texts should, as I see it, still be kept naturally legible. The average corpus users will be less interested in graphemes, phonemes or morphemes than in syntagmemes or textemes.

In the following I will use some of the edited MS texts of language AB as a test case. So far, various computer specialists, like Peter Robinson and Raymond Hickey, have written programs for the analysis and collation of Middle English texts, but they obviously had ‘streamlined’ Chaucer manuscripts like Ellesmere and Hengwert in mind. The question remains: how can we come to grips with a complex literary tradition like that of *Ancrene Riwle*, as opposed, for example, to the much simpler tradition of the *Hali Meidenhad*? The answer given in this paper will be tentative and fragmentary, but it is hoped that the questions raised turn out to be necessary and fruitful ones for the analysis of Early Middle English texts in general.

2 Description of the versions available

Ancrene Riwle has come down to us in seventeen versions altogether, eleven English, four Latin and two French.[2] Most of these 17 versions are easily available in editions (cf Zettersten 1965) and thus invite collation. For practical reasons, I have scanned (and proofread) only the subset of five variants, all from the 13th century, namely:

1. Pepys 2498, Magdalene College Cambridge, ed. Arne Zettersten, EETS 274 (1976);
2. BL Cotton Nero A. xiv, ed. Mable Day, EETS 225 (1952);
3. BL Gonville and Caius College MS. 234/120, ed. R.M. Wilson, EETS OS 229 (1954);
4. Corpus Christi College Cambridge 402 (= *Ancrene Wisse*, c1230, an expanded version of *Ancrene Riwle* (c. 1200), ed. John R.R. Tolkien. EETS 249 (1962);
5. BL Cotton Titus D. xviii, ed. Frances M. Mack, EETS 252 (1963).

The text of *Hali Meidenhad* has survived in the English versions of two manuscripts of the early thirteenth century, namely

1. Oxford Bodleiana, Bodley 34
2. BL Titus D. xviii

While the number of manuscripts of *Hali Meidenhad* is, thus, very limited, the disagreement on what to make of them and, moreover, growing insights of critics have led to a few different editions. They include:

1. Ed. Frederick J. Furnivall 1922, *Hali Meidenhad*, EETS OS 18 (based on Cockayne's edition of 1866, OS 18, offering parallel diplomatic versions of the two manuscripts plus a modernized translation);
2. Ed. Bella Millett, *Hali Meiðhad* 1982, EETS 284 (a critical edition based on the two manuscripts).

A first comparison of the different versions in either case, *Ancrene Riwle* and *Hali Meidenhad*, makes clear that they confront us with different kinds of problems. The two manuscripts of the *Hali Meidenhad* differ only on minor points, such as spelling, punctuation or lexical meaning, so that they can easily be compared.

The *Ancrene* tradition, on the other hand, is so complex and some of the manuscripts, such as MS Cleopatra, cause so many different problems that a compiler seems well advised to leave the job of critical reading to previous editors and later corpus users. Thus, Dobson, in his edition of the Cleopatra MS(C), distinguishes four different hands and scribes, where A represents the original scribe, B the author improving on the quality of the text, D another later individual 'corrector' and E the most prominent of all other scribes leaving their traces on the manuscript at various stages. All in all, the text of this manuscript has countless annotations and alterations, both on the margin and between the lines, and, as Dobson convincingly claims, the various types of modifications can only be disentangled after 'years of patient study' (Dobson 1972: xii). The same kind of warning seems justified in view of the obvious disagreement between the different manuscripts of the *Ancrene Riwle*.

The conclusions to be drawn from these points of description are these: (1) The *Hali Meidenhad*, with its limited tradition of two manuscripts, invites collation of the two copy text editions, correction of their readings by recurrence to the manuscripts and better judgment due to easily producible indices, concordances, etc and perhaps on the basis of Millett's critical edition. (2) The *Ancrene*

Riwle, on the other hand, raises so many more global questions – of content, genealogy of subject matter, meaning, style, comparison of intention and so forth – that going back to the sources and tackling paleographical problems would overburden any corpus compiler and distract the corpus users.

In view of SGML specialists' claims of pedantic precision for the reproduction of manuscripts on the screen[3], it must be added that medieval manuscripts, and Early Middle English ones in particular, are simply subject to editorial interpretation and, moreover, often in bad physical shape, with erasures, etc. As a result, storing them as graphics for further special usage may be a better option than resorting to diplomatic computer texts which have been edited with incalculable care.

3 From image to code: from Prolector to HTML

If, then, compilers feel warned not to take over the job of the editors unnecessarily and prematurely, thus also disregarding editorial achievements of the last 150 years, they should try to make a given text available on its different levels of authenticity.

3.1 Manuscript

If a given text is available in MS form, even if only in facsimile (like *Hali Meidenhad*[4]), it is technically easy for it to be scanned and to be made available as a graphic database in a word processing program (such as WordPerfect). In Innsbruck we have been using a program called *Prolector* for scanning, with a resolution of 400 dots per inch, allowing moderate zooming so that single characters can better be seen on the screen than in the original MS. The tif-images can be imported without any problem by WordPerfect or WinWord so that they can be collated to corresponding text data bases, again with the help of zooming facilities.

I have not tried out another facility which certainly exists: the annotation of images. If manuscript words, lines or pages could be indexed or annotated, they would be better retrievable. This technique has been applied at the Medieval Institute at Krems near Vienna.[5]

3.2 Copy text editions

Diplomatic copy text editions of ME texts like *Hali Meidenhad* and *Ancrene Riwle* have to be treated 'diplomatically', indeed, by a corpus compiler. Since the corpus will have to omit the critical apparatus and the footnotes, and also both the sometimes very informative introductions and the editor's marginal comments, all the compiler can do is to convert the text to the screen as authentically as possible. Therefore, apart from characters and signs, an edition's layout and format should be preserved. These include lines, stanzas, paragraphs and pages. Initials, bolds and other formats of letters should, in principle, be marked and the policy of bracketing, which is very subjective in some editions, should be made clear.

But this only seems feasible up to a point. Thus, the Cleopatra text of *Ancrene Riwle*, reproducing the manuscript text as scribe A left it, uses claret brackets for interlined letters or words (at the point intended by A); erased letters and punctuation marks are printed within square brackets, in roman or italic, depending on whether the erased letter is legible or has to be conjectured. And subpuncted letters or words are printed as such, unless they are meant to be substituted by interlined letters or words.

We can use the different types of brackets, making their function clear in the header. But the distinction of different degrees of legibility and its marking by the roman or italic format of script goes well beyond what most editions have done so far and could, moreover, only be expressed in codes, since ASCII does not allow italicization.

But before we undertake manual marking of single or successive italics, used for reasons that are often only made transparent in the footnotes, with erasures and worm holes playing quite a role, we should consider that italics are difficult to identify, both by scanners and the human eye; that they are used in other editions for deviant purposes, mainly for emendation; that the two series of footnotes commenting on our text[6] will not be available to the corpus user anyway; and finally, that the degree of legibility, as stated by the editor and as marked by the italics, may appear in a different light when modern photostat technology for manuscript reading is applied.

In conclusion, the italicization seems negligible in this edition. After all, the letters at issue are still marked by square brackets and are thus eligible for further study. As to the subpuncted letters, they stand for the scribe's intended but unachieved correction and should be preserved. This can only be done by encoding in lower ASCII. Using rudimentary HTML here, I would suggest <s> ... </s> before and after the letters or words concerned (*s* for 'subpuncted'). By the same token, text segments crossed out by the scribe, and therefore underlined in the edition, have to be marked in some encoded way, for example by <d> ... </d> (for deleted).

4 Encoding characters and signs

The examples show that encoding is unavoidable. But where exactly the borderline is between mimesis and encoding and what the encoding system should be is difficult to say categorically, particularly in view of the characters and signs needed for a corpus. Of the various systems, one of the candidates is the upper ASCII – to the extent that the characters and signs are identical in dominant word processing systems, such as WORD5, WinWord6, WordPerfect6, and MacWrite. But as to texts, the import-export business is a risky one. By experiment one soon realizes that some keys have divergent functions, for example in the case of ALT 191 (which ICAMET uses for upper case yogh). But when transferred, these characters are automatically converted, given that they were saved in the ASCII mode before.

However, in view of the compatibility with other word processing programs, both of the present and the future, would it not be better for us to avoid such compromising makeshifts and relegate the whole question of special letters and signs, like that of layout and format, to the codes used by the Helsinki Corpus (<+t> for the lower case thorn etc), or, perhaps even more international, the allegedly neutral ground of SGML or its hypertext version HTML, the latter particularly in view of an INTERNET distribution?

My answer is that we really need at least two kinds of texts: the 'iconic' one – close to the MS, yet nicely legible – and the digitalized, ie encoded one for machine-readability and the 'internet-ional' community.[7]

There is not much point in discussing SGML codes in detail here – in spite of the great diligence of researchers in the field of TEI (Text Encoding Initiative) and in the different types and concerns of 'markup languages', particularly since the publication of Goldfarb's SGML 'bible' in 1990[8]. The codes are published in handbooks[9] and, for updates and corrections, in WWW too, of course. The task of changing a database in, for example, WORD5 into an HTML database can already be taken over by an editor or so-called converter programme (cf http://www.seas.upenn.edu/~mengwong/txt2html.html).[10]

For corpus compilers these are questions of secondary relevance. The first and foremost problem for them is to keep the text naturally legible and to define which parameters of the text can be expressed in ASCII, which others can be normalized and what simply has to be encoded. Many scholars take sides here from the very start, but the three methods are – at least as far as characters and signs are concerned – not mutually exclusive.

Relatively frequent OE characters – like ash and the runes – should not be encoded in a medieval text corpus from the very beginning; the codes would simply spoil the iconicity of the text. Upper ASCII makeshift characters[11] are preferable, even though the two ash-letters are the only ones which are correctly produced on the screen without a keyboard driver; in the other cases we get makeshift characters somewhat reminiscent of those we want. As mentioned, we can have these signs converted into the proper characters by a keyboard driver called ICAMET, written by Raymond Hickey.

Second, there is the possibility of normalizing. Several scribal habits have been reproduced in some editions, but considered negligible in many or most others, and are therefore suggested in corpora to be smoothed out of the texts: *wynn* is changed to <w>; long s to <s>, and vowels with tremas are converted to simple vowels. Moreover, colours and different sizes of letters can be left unmarked, as in almost all extant editions. So can italics, whose main function is to mark the editor's expansions of contractions.[12] Flourishes in contractions can be omitted, but the contractions as such are worth retaining for further study. The various Tironian notes used in the editions for *and* or Latin *et* can all be represented by the standard ampersand sign '&', even if they look like a '7' or the like.

Finally, some missing characters simply have to be encoded. Of the alphabet, this concerns the two eth-characters (for upper case and lower case), but they are rare, even in Early Middle English editions. It seems advisable to use <+d/+D> in these few cases, in line with the Helsinki Corpus. A second point concerns accent signs. Since accents on vowels may have a prosodic function, they should not be lost. The accents are no problem with <a,e,i,o,u>, but <y> does not allow an accent on a normal keyboard, so that we have to code the accent, for example by adding it after the <y>[13]. Tildes or dashes above a letter often have an abbreviating function and therefore may both be substituted by a tilde[14] after the letter concerned (with Alt 126). Generally, superscripts, ie characters or signs above other letters, can be inserted in the lines where they belong and marked by surrounding signs of equality; if the superscribed text has more than three words, it is given a separate line. Raised characters, which often stand for contractions, can also be suggested to be indicated by signs of equality.[15]

Bolds and *initials* may be indicated by <b> before single letters and by surrounding <b> ... </b> in the case of several letters or words.[16]

The production of punctuation marks on the screen is almost no problem. Comma, full stop, colon, question mark and exclamation mark are part of the standard keyboard. Some editors have observed the difference between a full stop (syntactic function) and a middle dot (for marking speech breaks); it should by all means be preserved, for example marked by two successive full stops for the prosodic dot.[17]

Another frequent text marker in Middle English texts is the paragraph sign within the lines. In order to keep it apart from the usual return sign needed for breaking the lines, the in-line paragraph signs can be represented by ALT 20. While the produced sign '¶' looks like the RETURN sign on the screen, it is kept apart from it on the computer, for example in search routines, and does not have the effect of the RETURN key of breaking the line.

Another specifically medieval text marker is the inverted semicolon, the so-called *punctus elevatus*, which often has a clearly prosodic function ('raise your voice')[18]. This can be displayed by two normal semicolons. Hyphens can generally be used in line with the source text. But when a word is broken due to the change of line (*hus*- + *bondman*), the second element can be taken to the end of the first line, marked by an underline (_). This marker and method can also be used when a word is broken by change of line without a hyphen, or when the break of line is marked otherwise (for example by a short vertical stroke).

Long vertical strokes or slashes (| or /) are used in many editions to mark the change of folio number, with the number added in the margin. But generally there are all kinds of markers. For the sake of uniformity it seems advisable always to use the same marker, for example square brackets, for this purpose; in ICAMET, they are given within the lines, not marginally at their ends.[19]

Many more things have to be encoded in an early medieval text, for example the editor's emendations and comments and the compiler's comments; in Innsbruck we take over the editor's markers for the former purpose and use curly brackets for the latter. The comments of both editors and compilers (including headlines) are marked by the 'bybing' strokes (|; produced by ALT 124), interpreted by the WORDCRUNCHER program as command signs. On the other hand, the texts have many accessories which have not consistently been marked by editors and can or even have to remain unmarked in the corpus, such as deviant fonts and words or texts in foreign languages.[20]

5 Collation of texts

Collating text variants can, in principle, take place on different levels, from differences of spelling to discongruities of macrostructure. But normally we have words and their smaller units in mind. Accordingly, well-known collating programs such as COLLATE or LEXA suggest lines as the basis of collation. While the latest update of COLLATE, COLLATE2, allows the definition of bigger blocks of text, up to 32,768 words in length per block, Peter Robinson (1994:

34), the author, has to admit that COLLATE2 works less efficiently on longer blocks. For prose, he suggests the paragraph as the collatable unit that is most likely to be sensible.

What exactly is worth correlating in two or more texts depends of course on the special case. The two manuscripts of the *Hali Meidenhad* in Furnivall's edition, and his modernized version added on the bottom of the pages, would allow tags or anchor spots pagewise. So, in addition to Furnivall's different pagination for the three texts at issue, they would be given correlated page markers, like <p.1>, <p.2>, etc. Beyond this correlation of pages, one could of course also consider correlating the lines. Hickey's LEXA and Robinson's COLLATE do this automatically, but the problem is that, even in the case of fairly similar manuscripts like Bodley and Titus, the lines do not fully correlate. And Millett's pagination deviates entirely from Furnivall's, one of the main achievements in her edition being to have the text restructured in terms of paragraphs. Since these – unlike Furnivall's pages – are content-based units, the option of transferring them to the Furnivall texts seems to be a better way of correlating the various text versions. The anchors based on Millett's paragraphs could also be implemented in Furnivall's modernized version. But all of this must, of course, be done manually.

In the case of *Ancrene Riwle* (we are better off having Morton's early ordering system. All later editions refer to it, and in view of this) the M-tags should be preserved in all computerized text versions. On that basis, we can easily find corresponding passages by a simple search routine, even in common text processing programs. In COLLATE 2 we can define 'M' as the basic tag and thus get a list of all the deviations.

Collating texts in their linear structure can, however, only lead to eclectic results. We are not yet able to decide whether the spellings of, for example, 'David' or 'salmwrihte' in the Titus MS vs 'Davið' and 'psalmwruhte' in the Bodley MS of *Hali Meidenhad* are at all typical of the text.

6 In need of index lists

In order to compare a word, for example its spelling, on the syntagmatic axis, ie in relation to other occurrences elsewhere in the text, we need an index list of the text concerned. Nowadays such index lists can easily be produced, for example in WORDCRUNCHER. By analysing and comparing the entries of these lists with each other, we have a better idea of what a special text variant is worth.

The individual lists can help us to discover inconsistencies of spelling and thus pave the way for regularization. However, before we wish to regularize the text of *Ancrene Riwle*, we want to find out about the norms of language AB, which the Bodley MS ('B' for Bodley) shares with *Ancrene Wisse* of the Corpus Christi MS. The next step of analysis, therefore, can be a grouping of different AB-texts and a comparison of the different index lists produced by WORDCRUNCHER. The four groups that can be distinguished in the present case are the works of the Katherine Group, which *Hali Meidenhad* (Bodley) belongs to, as group one, the four prose meditations of the so-called wooing group[21] (group 2), the prose allegory *Sawles Warde*, and, of course, *Ancrene Wisse*.

The general index list of the works of all these groups allows for a group- and work-specific analysis of different words as long as all the words are marked before they are merged in the general index list. This marking could easily be done by giving an origin code to each single word, for example according to the following tree of origin (Figure 1):

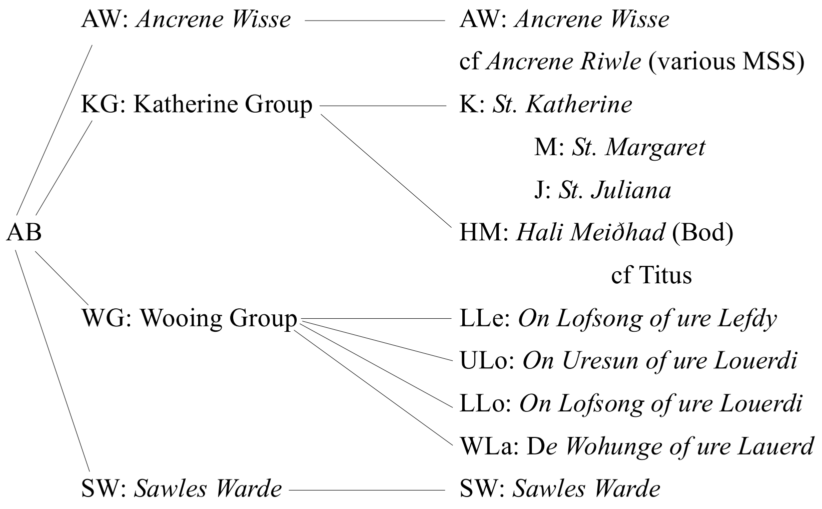

Figure 1: Grouping and marking of AB-texts in a general index

On the basis of the list that can thus be produced, it is possible to isolate what has been called language AB in its purest form, namely based on the *Ancrene Wisse* and the Bodley MS alone, from the other ('satellite') texts.

Since language AB has been a field of research for quite some time, work specific dictionaries and also more general vocabulary lists already exist, for example those compiled by Zettersten (1965) and Millett (1982). But our machine-readable index lists allow for more flexibility. WORDCRUNCHER can correlate the word lists with the concordance lines and with wider contexts, and if we take the word lists over into text processing programs, which is easy, we can rearrange the lists according to new criteria, for example according to etymological origin (cf Zettersten 1965: 275-283) or according to alliterative phrases (particularly in the Katherine group). Millett (1982: lv), in the excellent introduction to her edition (*Hali Meiðhad)*, also mentions, among many other features of Latin rhetorics in the work, the figures of *traductio* and *similer cadens*; the identity of morphological stem (as in *forleoseð*, *leosen, leoseð*) and the identity of suffixes (as in *biwinneð, biwiteð, forleoseð*, etc) allow for formal questions to be raised and answered with the help of the computer.

Of course, in the case of the *Hali Meidenhad*, the alphabetical index list is fairly limited, due to the short length of the text. But generally speaking, index lists based on different manuscripts are not always easily comparable, since the different spellings of words may give them an unexpected position in the alphabet. And even spelling variants in one and the same manuscript cause an analyst to be uncertain about what words are used in a text. Logically, we can only identify spelling and other variants (morphological, for example) if we know what normal forms they deviate from. In other words, what we need are profile lists which make clear what the main variants in a special text or group of texts are in comparison to normal forms. According to Laing (1994), the Edinburgh program BASEMAKER is the very program which can create such profiles. But it presupposes manually tagged texts, the acquisition of which is – to say the least – a time-consuming business.

Hickey's LEXA programme also has a routine for this purpose; I am, however, not so sure whether it brings the complexity of Early Middle English texts under control.

7 Words and phrases

Big corpora are liable to challenge many questions that do not concern the smaller units of language structure (spelling, phonemes, morphemes, etc), but the text's 'macrostructure', as reflected in less frequent and transphrastic features. The questions that could be tackled range from collocation via idioms and stylistic features to cohesive markers and characteristics of usage. Below, I illustrate briefly the role of such features in *Hali Meidenhad*.

This work has beeen praised in conventional research for the liveliness of its images and idioms. Due to the conservativism of language AB and idioms in particular, many idioms that we find when reading the text show less resemblance to modern English than to Old English and – due to the common West-Germanic ancestry – German. Since these old-fashioned idioms are difficult to trace, it would be useful to have them marked in a corpus for further analysis.

Here is a small collection of such idioms with the modern English meaning and a German formal equivalent (Table 1):

Table 1: Idioms in *Hali Meidenhad*, compared to ModE and German

idiom	ModE	cf German pattern
a hwile	*for a time*	*eine Weile*
a(ne) dale	*partly*	*einen Teil*
moni an	*many people*	*manch einer*
te olde feond	*the devil*	*der 'altböse' Feind*
do awei	*(to) put away*	*wegtun*
awei warpe	*to throw away*	*wegwerfen*
brekest ti wedlak to	*(you) are unfaithful to*	*(du) brichst die Ehe mit*
on ealre erst	*first of all*	*zu allererst*
beon efterwart	*(to) pursue*	*hinterhersein*
on *ende*	***in*** *the end*	***am*** *Ende*
on *Englisch*	***in*** *English*	***auf*** *Englisch*
on Englische ledene	*in* ***the*** *Engl. lang.*	*in englischer Sprache*
wiþouten euenunge	*incomparably*	*ohnegleichen*
turne to god	*to turn out well*	*zum Guten wenden*
imaket hal	*healed*	*heil gemacht*
as þu turnest þin hond	*in a moment*	*im Handumdrehen*
þe alre measte	*the most*	*das allermeiste*
on ealre earst	*first of all*	*zu allererst*
to him halden	*(to) stick to him*	*zu ihm halten*

Apart from such idioms, there are many other phrases in the text which we can more easily recognize because they have ModEnglish analogues. These are phrases of emphasis, such as *pine ouer pine*, *sorhe up-o sorhe*, *blisse up-o blisse*, and *crune up-o crune*, or other phrases like *wullen ha, nullen ha* ('willy, nilly') or prepositional verbs like *to warp ut*, *to schawe forþ, nim ȝeme* ('take heed') or *to beseon on* ('to look at'). Another most striking type is that of alliterative formulae: *fleschliche fulthen* ('carnal filth'), *here of helle* ('army of hell'), *i body ant i breoste, liues lauerd* ('life's lord'), etc.

All these phrases, no matter what their structure or function, are lost for machine legibility unless marked as phrases. I am not saying that marking such phrases is the task of the compiler who initiates a corpus. But if idiomatic and formulaic texts like the *Hali Meidenhad* are parsed and tagged, they should also profitably be marked for phrases. This could be done in a twofold way: for a phrase to be retrievable in a sequential text, its beginning and end must be marked; to find it in WORDCRUNCHER, the gaps between the individual words of the phrase must be bridged. So, what we need is a semi-automatic makro which converts *wullen ha, nullen ha* into something like <phrase>wullen_ha_,_nullen_ha</phrase>.

Such phrase tagging could give a convincing empirical basis to conventional research presented by Oakden as early as 1935 and to the many stylistic observations that can be gleaned from the introductions of *EETS* editions.

8 Summary and outlook

In this article, some of the problems involved in compiling a corpus of medieval texts have been tackled, but not fully solved. Definite solutions are unlikely, since compilers of machine-readable texts are very dependent on the speedy development of computer technology. It is against the background of these unstable conditions that this article offers the author's tentative views on the encoding and collating of Early Middle English texts.

I hope to have shown that the strategy of compiling an Early Middle English corpus is conditioned by some variables:

a) the target texts;
b) the target groups of users, who either want to go back to the original/manuscript or want to have the text prepared for them;
c) the different dimensions of the texts (characters, signs, format, etc);
d) the comparability of the texts involved, with different features to be tagged, depending on the individual case.

In view of these imponderables, we may finally come back to two basic questions, that of normalization/regularization, on the one hand, and the problem of the graphical reproduction of manuscript pages, on the other.

It stands to reason that a multi-language corpus[22], driving at and allowing comparisons between similar texts of different languages, dialects and manuscripts, needs normal word forms of Middle English as a basis of comparison, or perhaps not of Middle English as a whole, but of a subtype of it, such as language AB. There are different degrees of normalization, and no doubt different types, from the expansion of contractions via the regularization of idiosyncratic spellings within one work to the change of some or all words into Middle or Modern English for the sake of better understanding and comparability. The subject is apt to divide scholars – all the more reason for it to be given closer attention in the near future.[23]

Another problem, not dealt with in detail here, is that of the graphical reproduction of manuscript pages on the screen. Present-day scanner technology allows the digitalization of manuscripts or of facsimile pages, but in everyday work with the computer, problems may arise in view of the available storage space, the speed of the processor or in view of the resolution mastered by the scanner and the graphic card. In Innsbruck, we have recently experimented a little with different scanner programs such as RECOGNITA and PROLECTOR. The latter allows a resolution of 400 dots. It is, I trust, only a question of a few more years before we may all indulge in the brilliance of 20" monitors with a resolution and a range of colours that equal the present TV standard.

Notes

1. For a listing of the specific problems that medieval texts offer for the corpus compiler, cf Markus (1994).
2. The figures include fragmentary texts and adaptations.
3. Cf Gaylord's (1995: 59) word of caution suggesting that if a MS has five forms of <o>, it may well be important to encode them in the TEI scheme.
4. Cf Ker (1960).
5. For further information, cf Kühnel (1992).
6. There are 'special' ones recording scribe B's modifications and 'general' ones recording 'everything else' (xviii).
7. For further details of a multi-version strategy see the handbook of ICAMET (Markus 1999).

8. Cf the series of articles in *Computer and the Humanities* in 1995.
9. Cf for example Russ Jones, and Adrian Nye (1995). *HTML und das World Wide Web*. Bonn: O'Reilly; Morris (1995).
10. The difference is that the 'editor' does the whole job, including formatting and layout, whereas converters and filters do part of the job.
11. ALT 145, 232, 168; 146, 233, 191 for æ, þ, ʒ, Æ, Þ, Ʒ.
12. The function of italics is not always clear without a careful reading of the preface (cf for example Dobson 1972: xv).
13. *Accent aigu* and *accent grave* are no problem, but the *accent circonflex* is used as a control character in WORD5 and should therefore not be used. For the extremely rare cases when it might occur, I suggest coding by name after *y*: y<circonflex>.
14. The difference seems negligible and often difficult to make out in the MSS.
15. Cf the same policy in Kytö (1993: 27).
16. Initials can be of different sizes and (in the manuscripts) colours (mainly red and/or blue), but for the sake of simplification such different types and sizes of initials have not been marked as such in the present corpus.
17. In higher ASCII, ALT 249 would be available for the mid-line dot.
18. Cf Ker in Tolkien (1962, xiii).
19. When a word is interrupted by the change of folios, we have marked the position in the word by an asterisk and referred to the folio in brackets after the word.
20. For further details concerning encodings of Middle English texts cf *Manual of ICAMET* (Markus 1999).
21. D*e Wohunge of ure Lauerd*, etc (see Figure 1 above).
22. This article was originally written as a vanguard contribution meant to help prepare the compilation of a corpus of medieval saints' legends in different languages at the University of Liverpool (Antoinette Renouf et al). Unfortunately, the project has, mainly due to copyright problems, not materialized.
23. The present author dealt with the problem at the 17th ICAME Conference in Stockholm in May 1996 (cf Markus 1997).

References

Dobson, E.J. (ed). 1972. *The English text of the Ancrene Riwle*, ed. from B.M. Cotton MS. Cleopatra C.vi. EETS OS 267.

Furnivall, Frederick James (ed). 1922. *Hali Meidenhad: An alliterative homily of the thirteenth century: From MS. Bodley 34, Oxford, and Cotton MS. Titus D. xviii*. EETS OS 18.

Gaylord, Harry E. 1995. Character representation. C&H 29: 51-73.

Goldfarb, Charles F. 1990. *The SGML handbook,* ed. Yuri Rubinsky. Oxford: Clarendon Press.

Hickey, Raymond. 1994. *Lexa 6.0. Update documentation*. Bergen: The Norwegian Computing Centre for the Humanities.

Jones, Russ, and Adrian Nye. 1995. *HTML und das World Wide Web*. Bonn: O'Reilly.

Ker, Neil Ripley. Introd. 1960. *Facsimile of MS. Bodley 34: St. Katherine, St. Margaret, St. Juliana, Hali Meiðhad, Sawles Warde*. EETS 247.

Kühnel, Harry. 1992. Zwanzig Jahre Institut für Realienkunde des Mittelalters und der frühen Neuzeit: Ein Resumee. In G. Jaritz (ed), *Zwanzig Jahre Institut für Realienkunde des Mittelalters und der frühen Neuzeit der Österreichischen Akademie der Wissenschaften*. Krems b. Wien: *Medium Aevum Quotidianum* 25:9–11.

Kytö, Merja. 1993. *Manual to the Diachronic Part of the Helsinki Corpus of English Texts*, 2nd ed. Helsinki: University of Helsinki, Department of English.

Kytö, Merja, Matti Rissanen and Susan Wright (eds). 1994. *Corpora across the centuries. Proceedings of the First International Colloquium on English Diachronic Corpora. St Catherine's College Cambridge, 25–27 March 1993*. Amsterdam etc: Rodopi.

Laing, Margaret. 1994. The linguistic of medieval vernacular texts: Two projects at Edinburgh. In M. Kytö, M. Rissanen and S. Wright (eds). *Corpora across the centuries*, 121–141.

Markus, Manfred. 1994. The concept of ICAMET (Innsbruck Computer Archive of Middle English Texts). In M. Kytö, M. Rissanen and S. Wright (eds). *Corpora across the centuries*, 41–52.

Markus, Manfred. 1997. Normalization of Middle English prose in practice. In M. Ljung (ed). *Corpus-based studies in English: Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17)*, 211-226. Amsterdam-Atlanta, GA: Rodopi.

Markus, Manfred. 1999. *Manual of ICAMET (Innsbruck Computer Archive of Machine-Readable English Texts)*. Innsbrucker Beiträge zur Kulturwissenschaft, Anglistische Reihe 7. Innsbruck: Institut für Anglistik.

Millett, Bella (ed). 1982. *Hali Meiðhad*. MS. Bodley 34. EETS OS 284.

Morris, Mary E. S. 1995. *HTML for fun and profit*. Mountain View, CA: Sun Soft Press.

Morton, James (ed). 1853. *The Ancrene Riwle*, Camden Society, 57.

Oakden, J.P. (with the assistance from Elizabeth R. Innes). 1935. *Alliterative poetry in Middle English. A survey of the traditions*. Manchester: Manchester University Press.

Robinson, Peter. 1994. *Collate 2: A user guide*. Oxford University Computing Services. Oxford, 13 Banbury Road.

Tolkien, John R.R. 1962. *The English text of the Ancrene Riwle*, ed. from MS. Corpus Christi College Cambridge 402, intr. by N.R. Ker, EETS 249 (1962, for 1960).

Zettersten, Arne. 1965. *Studies in the dialect and vocabulary of the Ancrene Riwle*. Lund Studies in English 34. Lund.