Articles and notes

Encoding Text and Encoding Texts: Some Reflections on Theory and Practice*

Ralph M. Cleminson,
Winchester

Summary

Both traditional and digital editing essentially consist of the inputting of text. When the source is a manuscript, this is still a very labour-intensive process, unlikely to be repeated. Scrupulous accuracy in the transference of information is therefore needed, continuing the tradition of diplomatic editions. The editing of texts within kleine Philologien differs from the classical mainstream, because of the different histories both of the traditions and of the texts themselves. For Slavonic, this reflects in particular the predominance of linguistics in early textual studies, and the national variants of the Slavonic language. Digital methods offer a new approach to these problems, principally through the more complete capture of information and greater flexibility in its presentation. Users of digital technology should ensure that their editions are enhanced rather than limited by it.

In the twenty years since the conference at Blagoevgrad that marked the beginning of the digital age in mediæval Slavonic studies, and in the fifty years of the *Monumenta Linguæ Slavicæ Dialecti Veteris*, a series closely associated with Freiburg, much has been done in the transference of data from manuscripts to another medium—which is essentially the operation with which both these initiatives are concerned. Fundamental to both the digital analysis and the editorial process is the *inputting of text*, which in recent years has come to mean almost without exception inputting of text into the computer, so that one may expect, as a by-product of even the most traditional edition, an electronic text which might be used for further research.[1]

* This paper was originally delivered at the Interdisciplinary Conference 'eHumanities: Nutzen für die historischen Philologien', Freiburg, 8–10 October 2015 (see the conference report in this issue p. 132).

1 This is dependent on the recognition that 'a computer is not just a better typewriter' (Birnbaum 1995, 19–28), so that it is incumbent upon inputters of digital text to ensure that it is created in, or converted to, a format which will allow, and continue to allow, multiple use of the material.

This cannot yet be done automatically from the manuscripts: the advances in optical character recognition that are progressively making the inputting of a greater and greater variety of printed material easier and easier cannot yet be applied to handwritten material,[2] let alone mediæval Slavonic material. It follows that the inputting, or transcription, of a manuscript still represents a considerable investment of the time and effort of highly qualified specialists. For this reason it is unlikely to be done more than once, and indeed, digital editions have in practice tended to be made not directly from the manuscripts, but from already extant print editions: an excellent example is the Corpus Cyrillo-Methodianum Helsingiense,[3] which explicitly states that 'The e-text should be considered to be a tertiary source as it is not based on the manuscript itself', but on the printed edition. This of course means that any errors or other peculiarities in the printed edition will be perpetuated in the digital text,[4] which in the case of the CCMH is a very minor problem, given the extremely high quality of the editions used, but still gives us cause to reflect on the principles of the inputting, and in particular on the decisions to be made about what information is to be preserved—given that, effectively, it is being encoded once and for all, and that the decisions made at this point will affect all subsequent use.

Fundamentally, then, the essential requirement is to maintain scrupulous accuracy at the lowest level, and, equally, not to add anything (such as punctuation) that is not clearly identified as an editorial addition and easily removable. This seems, so far, to have been taken for granted (it is notable that the existing literature on digital texts is almost exclusively concerned with how, not what information is to be encoded), and it may well be that the existing tradition of preparing diplomatic editions provides a completely adequate methodological basis for the operation. In that case no new standard for transcription is required, though if, to borrow Manfred Thaller's terminology, the computer will introduce not only 'changing modes of study', but 'changing modes of thought',[5] it will be necessary to ensure that the old best practice is carried over into the new mentality.

If the digital encoding, editing and presentation of text is can thus continue established traditions in a relatively straightforward manner, this is not the case with the encoding, editing and presentation of texts. For Slavists, this

2 Though efforts are being made in this direction: see, for example, <http://transcriptorium.eu/>, last accessed 15 February 2016.

3 <http://www.helsinki.fi/slaavilaiset/ccmh/>, last accessed 15 February 2016.

4 See the detailed discussion in respect of the *Codex Suprasliensis* in Cleminson 2012, 329–342.

5 The expression is taken from his contribution to the Freiburg conference (Thaller 2015).

may be particularly problematic, because of the history of the discipline, and it is the subject of ongoing and still unresolved debate. The typical complaint is that the earliest editions were all made by linguists and are thus virtually useless from a text-critical perspective (though it is only fair to say that if the first editors of Slavonic texts had been textual critics, the linguists would have equally just cause for complaint). The Slavonic tradition, as one of the 'kleinen Philologien', is very different from the classical tradition which is the primary point of reference for European academic culture. Whereas Renaissance scholars used language to establish text, the founders of Slavonic philology—roughly, in the period from Dobrovský to Jagić—used text to establish language. That is to say, Renaissance philologists assumed that the classical authors had correctly observed the grammatical and prosodic rules of the classical tongues, and thus applied these rules, which were known, to emend and establish the texts of their writings. The pioneer Slavonic philologists, by contrast, were dealing with a language which did not have an established grammar, and one of the primary tasks which they set themselves was to extrapolate that grammar from the texts which they were editing.

As a result, our idea of an edition, as Slavists, is very different from that of the classical philologists. To the criticism that no one would edit a Greek text the way we edit Slavonic texts, one might reply that no one would edit a Slavonic text the way one edits Greek texts—or one might have so replied if William Veder had not recently done that very thing.[6] His edition of the *Scete Patericon* does indeed aim to present a reconstruction of the cyrillic *textus receptus* and its glagolitic protograph, purged of any of the accidentals of manuscript transmission. Now it is perfectly possible, highly probable indeed, that many of Veder's emendations restore what Methodius wrote – but we shall never be certain which of them do so. It is however certain that the text as a whole is not identical with the Methodian original: it is a modern construct (as its very regularity proves!). It is another textual variant, not *the text*.

The argument in favour of such an edition is that Veder's edition of the *Scete Patericon* bears the same relation to Methodius as a modern edition of Sophocles does to the text as originally written. This, moreover, is true, provided that one considers only the two end-points of the process, and disregards everything that comes in between. Classical texts, from Homer onwards, are cultural artefacts which are made up of their origins, the activities of Alexandrian and humanist scholars, the *textus receptus*, and modern criticism, and a modern edition embodies the whole of that tradition, which also

6 Veder 2012a, Veder and van Tak 2012, Veder 2012b. The actual edition is the third of the three volumes (Veder 2012b). For a detailed critical discussion, see Krys'ko 2014.

includes the standardised orthography and grammar in which the text is presented, and which were embodied in that tradition at a relatively early stage. These editorial principles are not, however, applied to early Modern Greek works (roughly, the period from the Liberation to the Second World War),[7] nor to neo-Latin.[8] Byzantinists have also begun of late to take manuscript features into account when editing texts.[9] The vernaculars likewise have their own traditions: the works of Shakespeare, which have an uninterrupted publication history from the early seventeenth century to the present day, are printed in standard modern English spelling, but modern editions of the works of his contemporaries, which do not, preserve the original Elizabethan spelling, with standardisation, if any, normally confined to the regularisation of *i* and *j*, *u* and *v*.

From this we may extrapolate the principle that how a text is edited, in respect of any norms and standardisations, is a statement about the nature of the text and its tradition, transmission, and history; and such a statement may be true or false. Such a principle may be simple to enunciate, but it is by no means simple to apply, particularly for Slavists, who are faced with a very specific form of interaction between text and language in the tradition with which they are dealing. No editor, after all, in either Alexandria or Oxford, would present us with an Attic Sappho, but the manuscript tradition really does confront us with a Serbian Clement of Ohrid and a Russian Gregory Camblak.

The problem can be avoided if a text lends itself to a Bédieriste treatment, but by no means every text does, and in such a case we are fain to do our best and accept whatever opprobrium proceeds from offended national susceptibilities. If there were any simple solution, it would no doubt have been discovered at some time during the last two hundred years; however, the electronic age does at last offer some mitigation of the quandary. One can, for example, switch back and forth between witnesses within an electronic edition; one can have parallel texts; one can, indeed, in principle, preserve within the edition all the information provided by the entire tradition. This is not to be understood as a new path to the New Philology. By no means: while one may willingly concede that the text is the text in the totality of its realisations (or rather, in the totality of its *extant* realisations – the data are incomplete), one may decline to follow the New Philology to its logical conclusions, which seem to preclude the possibility of any editorial activity whatsoever. (It may

7 Ricks 2009 argues cogently against orthographic standardisation of such texts.

8 See, for example, Rammlinger 2006.

9 Giannouli 2014.

provide a suitable framework for a literary or sociological approach to texts, but not for textual criticism.)

On the contrary, the very existence of a critical apparatus indicates a concept of the text that goes beyond any particular realisation of it, and it is a fundamental error of the neophilologist approach to state that in traditional criticism 'la variante est du non-textuel'.[10] Quite the reverse: in an edition of the New Testament, for example, the inclusion of a variant in the apparatus is a positive assertion by the editors that that variant requires consideration. One fears that at bottom the New Philology represents a postmodernist rejection of any form of judgment, of the notion that one variant may be 'better' than another (though the people who wrote the manuscripts certainly believed that it might be), which is fundamentally opposed to the very concept of textual criticism, which depends upon κρίσις, on *judicium*. In other words, we do not simply gather information: we have to do something with it.

Digital editing does to an extent relieve us from some of the harder choices that an editor in other media has to make, and may go some way, for example, towards resolving such conflicts of interest as that mentioned above between the linguist and the textual critic. Within a manuscript one may find quite substantial passages in which that which is of interest to the textual critic is irrelevant to the linguist and *vice versa*; in such a case, given a complete encoding, each could generate a secondary encoding stripping out all unnecessary information, which, given that the initial encoding remains, would not entail the loss of information which in the pre-digital age was inherent in the choice between one or other type of edition. Similarly, the machine can handle much greater quantities of material than the unaided scholar, and, if correctly programmed, does not introduce errors. This in itself is a great advance.[11] Nevertheless, as already noted, at present it still requires considerable effort to input the data, automatic collation, for example, requiring complete digital encodings of every witness, which it may not be practical to provide.

This may change, and as the technology now available has made many operations quicker, cheaper and easier than they were in the past, so we may hope that further advances may assist with tasks which are difficult or impractical now. Indeed, one of the main difficulties faced by scholars at the moment (particularly if they are not engaged full-time with digital text) is keeping up with developments, the more so that digital humanities have become such a wide and complex field in which it is not always easy to be aware of events outside one's own immediate sphere of activity. This rapid and ever-expanding

10 Cerquiglini 1989, 111.

11 'No one ever checks anybody else's collations (or his own for that matter) without finding mistakes in them' (West 1973, 63).

progress is both empowering and disempowering for the individual scholar. Until quite recently one would work at one's digital application until it finally did what one wanted, which probably absorbed considerable time and effort, but, on the other hand, one fully understood the process. Now between input and output one is much more likely to have something developed by someone else for their own purposes, which for most of us is likely to mean something far more powerful than we would have been able to build by ourselves, but of which we are no longer entirely in control—a 'black box'.[12]

Our black boxes are not yet quite so black. Even though there are now computers that a five-year-old child can operate, one needs to have considerable philological experience to make CollateX do anything useful; but still, one does not need to know how to build CollateX. The use of such a black box may be seen as a sort of vicarious collaboration in one's project by the creators of the device, who thus become (in Latour's terminology) not 'actors', in the sense of direct participants, but 'actants', in the sense of having an indirect input through the medium of the machine. This account of interactions is very different from that put forward by Latour in his later work and in the actor-network theory currently very influential in sociology, which recognises both human and non-human actors and applies a principle of 'generalised symmetry' which treats them all in the same way. Leaving aside any alarm we might feel at an approach that equates us with non-human or even post-human agents, this concept of the 'agency of things' appears methodologically unsuited to textual criticism, and even empirically false, in that it ignores the realities of scholarly activity, of what we actually do. We all know that one of the features of collaborative research is that we argue with our colleagues. ('Was this written by two scribes or one? Is this a fifteenth- or sixteenth-century manuscript? Is this character ъ or ь?') One can even have some sort of intellectual interaction with the actants behind a construct: for example, in the course of using the TEI one may gain an insight into why its authors have structured it as they have, which is not necessarily obvious at the outset. One cannot argue with the machine: it does what it does. Interaction with it is not in reality symmetrical, and any attempt to treat it as such is methodologically barren.

The implication of this is that while we are, always have been and always will be limited by what our tools cannot do, we should not allow ourselves to be limited by what they can do. The course of research must be determined by the problems which we as scholars believe need to be solved, and not by the

12 'The word *black box* is used by cyberneticians whenever a piece of machinery or a set of commands is too complex. In its place they draw a little box about which they need to know nothing but its input and output' (Latour 1987, 2–3).

digital applications which we happen to have at our disposal. As far as digital text is concerned, we are still living in the Neolithic Age, and it is not surprising if we do not yet fully grasp the potential of digital technology for our discipline, nor that that potential is far from being fully realised. This should not discourage us: even palæolithic texts[13] are still extremely valuable, and we may be confident that our own productions, however primitive they may seem in a few years' time, will be of lasting use and be susceptible to modes of study which at present are impractical or have not yet been imagined. The essential prerequisite is that the initial transcription should be accurate, and the primary encoding as comprehensive as possible in its informational content. This is a law which we have inherited from pre-digital scholarship, and shall no doubt bequeath to whatever follows the digital age; but so long as we observe it, we have absolute freedom in what we do with the material.

References

Birnbaum, D.J. 1995. 'How Slavic Philologists Should Use Computers', in *Computer Processing of Medieval Slavic Manuscripts: Proceedings [of the] First International Conference, 24–28 July, Blagoevgrad, Bulgaria* (Sofia: Akademično izdatelstvo Marin Drinov, 1995), 19–28.

Cerquiglini, B. 1989. *Éloge de la variante* (Paris: Seuil, 1989).

Cleminson, R. 2012. 'The Codex Suprasliensis and the Principles of Digital Editing', in A. Miltenova, ed., *Преоткриване: Супрасълски сборник, старобългарски паметник от X век* (Sofia: Institut za literatura, Iztok-Zapad, 2012), 329–342.

Giannouli, A. 2014. 'Byzantine Punctuation and Orthography. Between Normalisation and Respect of the Manuscripts. Introductory Remarks', *Comparative Oriental Manuscript Studies Newsletter*, 8 (July, 2014), 18–22.

Krys'ko, V.B. 2014. 'Об издании Скитского патерика', *Voprosy jazykoznanija*, 5 (2014), 122–148.

Latour, B. 1987. *Science in Action: How to Follow Scientists and Engineers through Society* (Cambridge, Mass.: Harvard University Press, 1987).

Rammlinger, J. 2006. 'The Orthography of Machine-readable Neolatin Texts: A Plaidoyer for Minimal Intervention', <http://www.neulatein.de/orthography/>, version of 5 October 2006, last accessed 16 February 2016.

Ricks, D. 2009. 'Orthographic Standardisation of the Modern Greek Classics: Gain and Loss', in A. Georgakopoulou and M. Silk, eds., *Standard Languages and*

13 If one may without disrespect so characterise digital texts like the CCMH which use only ASCII.

Language Standards: Greek, Past and Present (Farnham: Ashgate, 2009), 131–147.

Thaller, M. 2015. ‘Digital Humanities, Big Data und die ‘kleinen’ Philologien’, paper presented at the Interdisciplinary Conference *eHumanities: Nutzen für die historischen Philologien*, Freiburg, 8–10 October 2015.

Veder, W.R. 2012a. *The Scete Paterikon: Introduction, Maps and Indices*, Pegasus Oost-Europese Studies, 12 (Amsterdam: Pegasus, 2012).

— 2012b. *Скитский патерик: Славянский перевод в принятом тексте и в реконструкции глаголического архетипа*, ed. W.R.Veder, Pegasus, Oost-Europese Studies, 14 (Amsterdam: Pegasus, 2012).

— and J.G. Tak 2012. *Patericon Sceticum: Greek Text. Latin Translation of the 6th c. English Translation of the Slavonic* Textus Receptus, ed., tr. by J.G. van der Tak and W.R. Veder, Pegasus Oost-Europese Studies, 13 (Amsterdam: Pegasus, 2012).

West, M. 1973. *Textual Criticism and Editorial Technique, applicable to Greek and Latin Texts* (Stuttgart: Teubner, 1973).