[38] 2011 "Issues in the typographic representation of medieval primary sources", in Yuji KAWAGUCHI / Makoto MINEGISHI / Wolfgang VIERECK, Eds.: *Corpus-based Analysis and Diachronic Linguistics*, Amsterdam: John Benjamins Publishing Company (Tokyo University of Foreign Studies, 3), 153-73. [capítulo]

Issues in the Typographic Representation of Medieval Primary Sources

António EMILIANO

The transcription of *medieval primary sources* for linguistic and philological study entails complex problems and decisions regarding the typographic representation of medieval characters[1]. The linguistic study of medieval texts requires highly conservative transcriptions, in order that information about the medieval character sets be faithfully preserved in the editions: full access to the original character sets by researchers allows for grounded analyses of the graphemic systems and hence of the linguistic data present in manuscripts.

Many medieval characters, such as special lettershapes (with several case distinctions), abbreviation marks (combining characters) and signs (spacing characters) and punctuation signs have fallen into desuse since the arrival of printing (although the early printers up to the 18th century made use of several medieval characters and abbreviations). Traditional philologists have considered the diversity of medieval lettershapes as a problem and have dealt with it by simply suppressing that diversity. The fact is that different medieval scripts (like the Visigothic, Carolingian or Gothic families of scripts which were continuously used in the Iberian Peninsula for several centuries) had some very distinct lettershapes which could be regarded as separate characters when compared to the contemporary versions of the Roman alphabet we use in print.

1. Some general assumptions

I submit that the following assumptions regarding the nature and aims of editions of medieval primary sources cannot be circumvented:

1) 'transcription', 'transliteration' and 'edition' of a text are different tasks and steps of the philological work, each with its specific set of goals and procedures;
2) an edition will represent a text better (i.e. more faithfully) to the extent that it entails the least amount of transliteration operations;

---

[1] The concept of 'typographic representation' encompasses both physical and digital *media*, although this paper focuses on editions to be included in digital textual *corpora*.

3) an interpretive/diplomatic edition must always be based on a good transcription and a highly conservative transliteration—ideally, every interpretive/diplomatic edition should be preceded by a conservative/palaeographic edition;
4) no single edition of a medieval primary source will ever meet the needs of all potential users and audiences;
5) the edition of medieval texts cannot be considered and carried out solely in terms of printed editions, and the study of medieval texts requires the creation of not just electronic text archives but also of electronic *corpora*.

In order to explain fully the conceptual and practical consequences of these assumptions I will first consider the distinction between graphetics and graphemics and the meaning of such terms as 'letter', 'character' (and the correlated term 'character set') and 'glyph' (and the correlated term 'glyph set'). The following discussion will be confined to the Roman alphabet and its medieval and modern derivatives.

When one is faced with the analysis of an alphabet-based writing system the graphetic plane must crucially be distinguished and separated from the graphemic plane. On the former, a writing system is simply a code, a set of symbols (letters, diacritics, punctuation marks, auxiliary signs), or, to be more precise, a *character set*. Since graphetics is concerned with the description and the history of writing codes, the relation of characters to units of speech has no bearing on this plane of analysis. One can describe, analyse and discuss a historical set of characters and its development without ever referring to any language features or structures. In other words, alphabets *qua* character sets can be analysed without any reference to a particular language or spelling system. An alphabet is not an orthography and letters and characters are not graphemes. Both graphetic and graphemic factors should be considered when adopting any given strategy for transcribing medieval texts and for representing them typographically.

2. Letters, characters, glyphs and graphs

A letter, the basic unit of the medieval and modern versions of the Roman alphabet, is in reality a class of characters. For instance, what we call the letter A is in fact a set that comprises the characters capital A, small A, superscript small A, small capital A. In medieval writing the set was larger and encompassed other types of A. In our modern printed version of the Roman alphabet some letters have strikingly different shapes and graphic attributes, as shown in table 1:

Table 1. Letters and lettershapes

LETTER	*small*	*superscript*	*subscript*	*small capital*	*capital*
‹A›	a	^{a}	$_{a}$	A	A
‹G›	g	^{g}	$_{g}$	G	G
‹M›	m	^{m}	$_{m}$	M	M
‹Q›	q	^{q}	$_{q}$	Q	Q

Table 1 contains just four letters. However, the total count of lettershapes (characters) is twenty, because letters are not lettershapes: letters can be manifested by more than one lettershape or character which may be notoriously distinct but are considered to be formally equivalent, that is, the "same" letter. Small A and capital A are different characters: to a user of a script not related to the Roman alphabet they may be regarded at first sight as two letters. For users of the Roman alphabet they are the same letter and most users do not even notice how different they are as lettershapes.

A character is an abstract shape regardless of any graphic rendering. It is an ideal form, in the sense that the triangle or the circle are ideal forms with no concrete features or attributes beyond those which define them as distinct geometric entities. The triangle and the circle are two-dimensional entities, i.e. surfaces, regardless of any specific trait they may aquire in the real world: the triangle is simply a polygon with three vertices and three sides which are line segments; the circle is simply a sequence of points in a plane which are all equidistant from a given point (called the centre) which is located on the same plane. These are basic and archetypal geometric shapes: their definition is not bound or constrained by any aspects pertaining to the physical materiality of their *medium* (size, colour, texture, material).

In exactly the same way, the entity 'capital A' is an ideal or archetypal form, a basic writing symbol which can simply be defined as a shape made up of three line segments arranged in a distinctive manner. The differences between ‹A› (regular capital A), ‹**A**› (bold capital A), ‹*A*› (italic capital A) and ‹***A***› (bold italic capital A) are not contrastive, because they are subject to the 'style' of graphic presentation of the capital A character (which is essentially neither regular, nor italic, nor bold, nor bold italic—characters have no presentation style). Whilst capital A is a simple literal character (or just a 'simple literal'), Á, Â, Ã, etc. are composite characters: they are made up of two separate characters (a literal plus a diacritic). The association of the supralinear acute, circumflex and tilde combining characters to a literal results in new distinctive writing units.

A character can be presented in manifold variant forms which are called 'glyphs'. The glyph is no less abstract than the character but it is not simply an ideal shape: the *systematic glyph is a model or blueprint for the presentation of a character* in a writing *medium*. In other words the systematic glyph is an abstract entity that contains explicit information about the features of graphic (and typographic) rendering of a character. Table 2 shows the typographic representation of several glyphs and characters:

Table 2. Letters, characters and glyphs

	proportionally spaced sans serif typeface			*monospaced typeface*		
character name	*regular*	*italic*	*bold*	*regular*	*italic*	bold
capital A	A	*A*	**A**	A	*A*	**A**
small A	a	*a*	**a**	a	*a*	**a**
capital M	M	*M*	**M**	M	*M*	**M**
small M	m	*m*	**m**	m	*m*	**m**

Table 2 contains the typographic presentation of 2 letters, 4 characters and 24 glyphs. Each letter corresponds in this table to 2 characters and 12 glyphs.

Alphabets and character sets are closed (finite) sets but any character set can associated to an infinite number of glyph sets: a letter is a group of characters which are regarded as formally equivalent and each character can have an infinite number of systematic glyphs.

The actual materialization of writing in a specific *medium* is 'graphic or graphetic implementation' and any instance of writing *qua* implementation corresponds to a unique spatiotemporal event. In print (typographic implementation of writing), the 'implementational glyphs' or 'graphs' through which systematic glyphs are materialized are tendentially identical. In chirographic (manual) writing the differences between graphs which manifest the same glyph are greater: there are differences between different hands (scribes) and a single writer can adopt different presentation styles. Furthermore, a lettershape is never drawn exactly the same way by the same individual, even in the more elaborate calligraphic styles, due to the biologic/ physiologic nature of humans. This does not mean that we use, or that the medieval scribes used, an infinite number of glyphs: we use a finite set of characters, one or more finite sets of systematic glyphs (according to the adopted style of writing), and, yes, we do draw an infinite number of graphs for each character.

Letters, characters, glyphs and graphs are thus basic units in the graphetic analysis of writing: they belong to distinct logic planes, have different properties and attributes and should never be confused. These distinctions should be particularly important to philologists undertaking the edition of a text (or group of texts) or planning the annotation of a *corpus* of medieval primary sources. Careful consideration should be given to which type of entities one wishes to represent or annotate and for which purpose. It goes without saying that trying to encode graphs is tantamount to creating a typographic facsimile and simply makes no sense. However there is ample room for discussion regarding the amount of 'glyphic' information one should encode or leave out in a palaeographic edition of a medieval text.

3. Graphemes, allography and writing systems

The graphemic plane of writing (which is the object of graphemics) requires the consideration of the relation between writing units and linguistic units. The graphemic plane is, therefore, a (scripto)linguistic plane; in other words, graphemics is a subdivision of linguistics.

To make this clearer, consider the proposition:

A a is character of the Roman alphabet.

This proposition is of course true. No additional comments are needed.

Now consider a similar proposition:

A is a grapheme.

It is neither true nor false, it is simply meaningless. Because the grapheme (like the phoneme) is a linguistic relational concept, this proposition can only have truth or falsehood content in relation to a specific language. The definition of 'emic' units, contrary to 'etic' units, is a function of their status in a given symbolic system.

The proposition:

A is a grapheme of written Portuguese.

is true, whereas

A is a grapheme of written Arabic.

is false.

Graphemes are minimal units of a writing system, which is in turn the *minimal set of contrastive graphic elements*, which, in association with a set of concatenation and mapping rules, allows for the written representation of a linguistic system and hence makes written linguistic communication possible.

In a logographic system, mapping rules lay out the relation between lexemic units and graphemic units ('logograms' or 'grapholexemes') whereas in an alphabetic system (which is tendentially phonographic) the mapping rules are basically grapheme-phoneme correspondence rules[2].

In an alphabet-based system graphemes are mostly 'phonograms' (or 'graphophonemes'). A grapheme can be made up of one or more letters and it can be mapped to more than one phoneme.

The grapheme, like the phoneme, is an abstract unit, whose value is defined in terms of the relation between elements of the same type in a system. A phoneme can be actualized in speech by distinct phones (allophones) according to the phonological context. A grapheme can also have different allographemes or allograms, although the factors that govern allophony and allography are of distinct nature. In reality, two types of allography must be considered: 1) character-allography or 'deep allography' and 2) glyph-allography or 'shallow allography'. In deep allography the allographs are different characters (e.g. capitals vs. uncials vs. minuscules) whereas in shallow allography the allographs are variants of the *same* character, i.e. they correspond to similar glyphs whose occurrence is generally context-dependent (e.g. word-medial lettershapes vs. word-final lettershapes). Medieval texts present abundant examples of context-dependent allograms and careful thought should be given to the amount of allography that will be represented in an edition or encoded in a *corpus*.

4. Orthography vs. alphabet

An alphabet is a code, a set of symbols, created for the purpose of representing language. The letters have a basic or general phonographic value (which goes back to their Roman origin) but because the same Roman alphabet (with slight adaptations) is currently used in many orthographies associated to completely different languages the precise value of a letter can only be ascertaind after careful examination of a language and its orthography (i.e. after graphemic analysis). For instance, one can say that the letter P

---

[2] There is an obvious simplification in this description: there are no pure phonographic or logographic systems (the two categories phonography and logography overlap): alphabet-based systems are basically or originally phonographic but acquire over time logographic traits. Also, mature users of an alphabetic system process the written words holistically, i.e. they read logographically with no intermediate "letter-by-letter" sequential analysis of the written forms. There are graphemic elements in some alphabetic systems that represent syllables or parts of syllables and morphemes; in other cases there are graphemes with no phonemic, morphemic or lexemic status or content. For instance, the English verbal morpheme <ED> corresponds to three context-dependent phonetic possibilities; the two-letter sequence is a 'graphomorpheme' composed of two graphemes. Logographic systems can contain graphophonemes (grapholexemes/logograms used as graphophonemes) and contextual graphemes (determinatives) that do not correspond to any explicit linguistic element but define a linguistic category, like gender, animated/ nonanimated, type of object/entity, etc. Japanese writing is a mixed system: it contains logographic, syllabographic and phonographic subsystems.

represents a 'p-sound' (a voiceless bilabial stop); that is true in most instances, but in English *receipt* and European Portuguese *recepção* "reception" the letter P does not match any phonic unit; in English *pit* the P corresponds to an aspirated stop, but in *tip* it corresponds to an unaspirated unreleased stop. As for the letter C, it has so many strikingly different values throughout European orthographies (including Turkish), that it is impossible to state a 'basic value' for this letter regardless of a specific language.

Alphabets are not 'writing systems' proper, orthographies are. Letters and characters are the building blocks of orthographies but an orthography is much more than an inventory of symbols. An orthography is a normalized and codified writing system. It is a hard-learned protocol which is rigidly imposed either by social consensus or by law and the users have no say or choice in the way they spell. The concept of orthography entails the notions of norm, correction and error; it also precludes the possibility of particular or regional practices in a public context. Thus in written languages endowed with an orthography a deviant spelling is always a mistake, not a variant possibility.

The situation of the vernacular languages of Europe in the Middle Ages and the Renaissance was different from the present: not only different versions of the Roman alphabet were developed and extensively used but also individual and regional diversity and variation in writing were widespread. The concept of orthography was widely known and it was enforced in written Latin and Greek (by scholars) but not in the written vernaculars. Those languages did not have proper orthographies, although it has become common practice for many scholars to use 'orthography' as a neutral term for 'spelling system' when dealing with medieval and early modern printed texts.

5. Transcription vs. transliteration

The edition of a medieval text results from an editorial programme or agenda which presupposes an interpretation of textual data. An edition is always a process of mediation of a text. Through this mediation the text is stripped of its original mode of presentation, according to the editor's agenda. There is no such thing as a definitive or objective edition of a medieval text. Peter Robinson, an eminent example of a philologist of the computer age, noted wisely:

> *Interpretation is fundamental to transcription.* It cannot be eliminated, and must be accomodated. [...] Transcription of a primary textual source cannot be regarded as an act of substitution, but a series of acts of translation from one semiotic system (that of the primary source) to another semiotic system (that of the computer). Like all acts of translation, it must be seen as fundamentally incomplete and fundamentally interpretive. (Robinson 1994: 9; my emphasis)

The type of edition that linguists, language historians and philologists require is one that presents a high degree of faithfulness regarding textual data, graphemic data and graphetic data. Any other kind of edition will always require the direct examination of the manuscript or of a good facsimile.

Taking for granted that there is no such thing as 'philological truth', one can still argue the case for 'philological truthness' or, better, 'faithfulness'. But how does one measure or gauge the degree of faithfulness of an edition? In my view the issue of philological faithfulness is contingent to the adoption of a set of clearcut principles and procedures which take into account *the crucial distinction between transcription and transliteration.*

There has always been a (serious) problem with the correct understanding of what 'transcription' is in medieval philology. In most editions the 'transcription' of the text(s) is generally preceded by a list of transcriptional criteria or procedures adopted by the editor(s). Most editors fail to realize that their criteria for transcription are in fact criteria for transliteration.

Transcription is the representation of a text by means of the original character set and the systematic glyph set: transcribing a text requires that both the original character and glyph sets are represented faithfully or in an unambiguous way. To me this is the measure of 'truthness' of an edition. Thus an egyptologist will use hieroglyphic, hieratic or demotic characters to transcribe an Egyptian text, a sanskritologist will use the *devanagari* script to transcribe Vedic or Classical Sanskrit texts and a Norse philologist will use some form of *fuþark* to transcribe runic texts, whether they are transcribing their specimens manually or using a computer. Their need for faithful transcriptions will not exempt those scholars from making transliterations using the Roman alphabet in order that their texts be more accessible to nonspecialists. Likewise the medieval philologist should use glyphs that unmistakeably match the special medieval glyphs that are no longer in use in present times in order to transcribe the original character set of medieval texts. Ensuing transliterations should be grounded on faithful transcriptions.

We should bear in mind that transcription is not depiction of a text in its original *medium*, a facsimile. Facsimiles are no doubt useful and each and every Portuguese medievalist would be grateful if they could have direct access to good quality digital reproductions of the thousands of medieval manuscripts kept at the Lisbon National Archive. Facsimiles can be invaluable aides to research. But facsimiles are not editions; a *corpus* of digital images of manuscripts is just a *corpus* of images not a textual *corpus*. Archives and *corpora* of images are not queriable objects like textual *corpora*. Philologists work with texts, not images: when a philologist transcribes the text their goal is to capture or represent the text (not an image of the manuscript). A highly conservative palaeographic edition will never be a facsimile nor should it ever purport to be one.

Transliteration implies substituting a different character set for the original one. To transliterate a text is to represent it by means of a character set and a glyph set that are structurally and formally different from those present in the manuscripts. If transliteration were always performed *homothetically* there should be no real problem, but the fact remains that most editions of medieval texts involve criteria for transliteration that deliberately mutilate and disfigure both the graphemic and graphetic reality of the texts: they do not just transliterate, instead they 'adapt', 'normalize', 'modernize' and whatnot. Most editions of medieval texts reduce the amount of graphemic and graphetic information present in the texts because of the limitations of the current typographic version of the Roman alphabet.

Most editors in fact transliterate their texts when they state that they are transcribing them. This common misunderstanding stems, in my view, from the fact that traditional philologists and historians (who are responsible for many editions of medieval sources which are unfortunately useless for certain types of linguistic analyses) fail to recognize that medieval character sets *are* different from their modern counterparts. Most editors do not seem to realize that medieval scripts and medieval scribal practices present a reality that is completely different from our contemporary printed versions of the Roman alphabet: the fact that a medieval Portuguese text makes use of the Roman alphabet does not mean that the structure of the script and of the spelling system thereof are the same as those of a contemporary Portuguese text. Replacing medieval characters (both literals and nonliterals) by modern print is transliteration not transcription.

A noteworthy case is the way most philologists and historians handle medieval abbreviations and punctuation.

Abbreviations were are an important element of medieval writing and because our present form of the Roman alphabet does not contain the special signs that were used in abbreviations most editors feel that they must alter the texts by substituting modern characters for the old abbreviatory signs or marks. The procedure most commonly used is to expand abbreviations (i.e. to replace brachygraphs by sequences of letters) according to the interpretation and the intuition of the editor. As for punctuation many editors simply ignore the original system of punctuation and insert punctuation in their editions according to the principles of their own language. Furthermore they separate words and text units—such as titles, paragraphs and verses—according to modern practice, 'normalize' capitalization and simply ignore intermediate letter cases such as uncials, enlarged minuscules or small capitals. All these editorial procedures are taken for granted as part of transcription.

If an edition of a medieval text is intended for use by an audience of nonspecialists all the aforementioned operations of transliteration are of course

legitimate because they insure that the text be received by a contemporary audience of nonphilologists. People who just want to know and enjoy the works of the past as part of their education, self-improvement or entertainment are not willing to tackle with a foreign or alien writing system (nor should they be forced to be!). They need, nay, they demand a fully accessible rendition of the text.

Scholarly editions are a whole different business: they are made for a scholarly community and for scholarly purposes in compliance to strict requirements of accuracy and faithfulness.

I wrote above «an edition will better represent a text to the extent that it entails the least amount of transliteration operations». A practical consequence of this statement which I regard as a crucial guideline is the need for special typefaces as a means of ensuring i) that the original character sets be faithfully preserved in editions and ii) that the original character set be unambiguously recognized by a (human) reader. Another consequence is that all special medieval characters and systematic glyphs should be recognized and encoded by the Unicode Consortium and ISO and should be included in the Universal Character Set (UCS). An interim solution, which several philologists and projects have adopted, is of course the use of provisional codepoints in the Private Use Area of a Unicode-compliant font.

6. The typographic representation of medieval primary sources

There are three basic strategies in the typographic representation of medieval texts:

1) indirect (deferred) representation
2) direct (straightforward) representation
3) normalization

All modern scholarly editions can be adequately labelled according to this simple scheme. Most fall under the heading of 'normalization' due to their extensive use of transliteration procedures. Only the first two editorial strategies allow for the representation of medieval character sets. I present below in an *addendum* some examples of these different approaches to typographic representation by means of a short excerpt from a Latin-Portuguese 10th-century charter written in cursive Visigothic script.

Strategy 1 (indirect representation) is especially suited for electronic processing because it requires the use of a text-encoding application, such as the *Text Encoding Initiative* (TEI), based on a markup language like XML (previously SGML). Editions produced according to this strategy thus make use of annotation and entities (cf. *addendum*—'indirect representation'): word

abbreviations can be encoded with the TEI core elements <abbr> and <expan> (among others available) whereas abbreviatory signs and modifications of literals for abbreviatory purposes can be encoded as entities (cf. Robinson 1994, Parkinson & Emiliano 1999 and 2003 and Chapter 11, Representation of Primary Sources, of TEI P5[3]). All special medieval characters, both literal and

---

[3] Chapter 11 of TEI P5 presents and discusses briefly the encoding of medieval abbreviations; several approaches, using different elements are proposed. They all seem, however, to rely on the (implicit) assumption that the ideal edition of a medieval text is a normalized edition: representation of medieval characters seems to be peripheral. The simple fact that abbreviation encoding is discussed in a section called "Altered, Corrected, and Erroneous Texts" (11.3) downplays the role of abbreviations as an important feature of medieval character sets. Abbreviations were not simply scribal devices used to shorten words and to speed up writing: they formed a rich subsystem of characters and were an integral and fundamental part of a scribe's graphetic and graphemic competence. The proposed use for the <abbr> and <expan> elements is not grounded in good and sound palaeographic doctrine: «The content of the abbr element should usually include the whole of the abbreviated word, while the expan element should include the whole of its expansion.» (11.3.2 Abbreviation and Expansion) This proposal does not take into account that there were distinct types of abbreviations, namely, lexical abbreviations, whereby a whole word was abbreviated by contraction and/or suspension and the simultaneous use of a generic brachygraph (usually an overline of variable length), abbreviation signs (modified letters or special spacing characters) and abbreviation marks (combining supra- and infralinear characters). Version 5 of the TEI Guidelines contains the additional elements <g> (=glyph!), <am> (=abbreviation mark) and <ex> (=editorial expansion) and uses the term 'brevigraph' to refer to special abbreviatory characters. This term is an unwarranted admixture of Latin and Greek formatives; 'brachygraph' or 'brachygram', of Greek origin, should be preferred since the words 'brachygraphy' and 'brachygraphic' already exist in the English language. As for the <am> and <ex> elements they seem to be redundant, at least in some cases, with respect to <abbr> and <expan>. To make this point clear: TEI illustrates the use of these elements with a sequence from an English medieval text that contains the word *eu(er)y* (the letters enclosed by round brackets are the expansion of a combining supralinear brachygraph). TEI proposes the following encoding solutions: **<abbr>**eu**<am><g ref="#b-er"/></am>**y**</abbr>** and **<expan>**eu**<ex>**er**</ex>**y**</expan>**, among other possibilities. If the <abbr> and <expan> elements are defined and used properly (i.e. associated strictly to the abbreviated letter-sequences, not to the whole word as a matter of principle), the use of <am> and <ex> in abbreviation signs and marks is redundant; like this: eu**<abbr><am><g ref="#b-er"/></am></abbr>**y and eu**<expan><ex>**er**</ex></expan>**y. The first example, with the element <am>, is odd from a philological and a palaeographic point of view: the encoded text is not really text but a <g> (=glyph) element whose attribute is a character reference (!). The Unicode block Latin Extended-D (range A720-A7FF) contains several medieval characters that were recently accepted into the UCS (cf. Everson et al. 1996a, 1996b). The designations proposed and accepted for several abbreviatory signs include the word *letter*; e.g. the abbreviatory sign small P with stroke (which stood for *per*, also *par* in Medieval Portuguese) was deliberately encoded with the name LATIN SMALL LETTER P WITH STROKE THROUGH DESCENDER (U+A751). Other abbreviatory (spacing) signs included in Latin Extended-D that bore no relation to an existing literal were also named *letters*. This stresses the fact that abbreviatory signs had the same importance as literal characters in medieval scripts and that they are not simply a problem to be solved by editorial policy. They possessed full "character-ness" and were not glyphs as the TEI <g> element might suggest to the uncautious reader of the TEI Guidelines.

nonliteral, can be indirectly (but unambiguously) represented by XML entities listed in the *corpus'* Document Type Definition (DTD). "Special" characters are transcribed as entity references in the body of the edition and no attempt is made to represent them directly ('glyphically'). This procedure has been advocated by the *Text Encoding Initiative* (since its inception) and the *Digital Scriptorium* among others.

Strategy 2 (direct representation) implies that each and every medieval character (abbreviation marks and signs included) is explicitly and unambiguously represented in the edition.

There are two ways to achieve direct typographic representation:

1) anisomorphic direct representation
2) isomorphic direct representation

The terms 'isomorphic' and 'anisomorphic' in this context refer to the presence or absence of a direct and absolute match between characters in the edition and characters in the manuscript.

Anisomorphic representation can in turn be implemented in two very distinct ways:

1) creation of a set of typographic conventions that map sequences of characters of the Basic Latin character set to medieval characters;
2) creation of a set of SGML/XML entities and use of entity references in the body of the edition; the entities are mapped to UCS or Private Use Area character codepoints for display purposes.

In Solution 1 each medieval character that has no direct match in the Basic Latin set is mapped to a combination of characters. It is a cumbersome way of transcribing and representing a medieval text (an edition made according to this strategy is not easy to read) but it is effective (cf. Parkinson 1983). With the widespread use of markup languages and the development of TEI and similar XML applications this strategy is clearly outdated.

Solution 2 has been successfully adopted and used by the *Medieval Nordic Text Archive—MENOTA*[4], a project of outstanding quality and scholarship. The actual text file of their editions is in XML format and contains entity references representing "special" medieval letters and signs. The editions are

---

[4] *Medieval Nordic Text Archive (Arkiv for nordiske middelaldertekster)*—Forskergruppe for tekstteknologi / Avdeling for kultur, språk og informasjonsteknologi (AKSIS) / Universitetet i Bergen, Norge, http://www.hit.uib.no/menota/; cf. *The Menota Hanbook 2.0.*

meant to be read with a web browser. The philologists who prepare the editions work solely with entity references but the end-users of the editions see only glyphs by means of a Unicode-compliant font, not the entity references (cf. *addendum*—'anisomorphic direct representation with XML/TEI entities'). The MENOTA project uses the MUFI guidelines and the MUFI font (*q.v. infra*).

In my view, a important downside of this solution is the fact that the base text contains entity references for special medieval characters, while the browseable version containing special glyphs is meant for display purposes only[5]. Any type of search or data extraction operation (such as generating wordlists and concordances) will have to be performed directly on the base text and the search parameters will have to refer to entities not to medieval characters. The upside is the fact that any change in the codepoints of medieval characters (for instance, when a character that was provisionally located in the Private Use Area is officially recognized and accepted into the UCS) requires a single correction in the *corpus*' DTD instead of multiple corrections in the body of the edition. But this is an upside for the encoders/curators not for the researchers which are the *corpus*' end-users. I strongly believe that *corpus* building and *corpus* structure should always be user-oriented not encoder- or curator-oriented.

Isomorphic representation is achieved by designing, and using in the body of the edition, a Unicode-compliant computer typeface containing medieval characters and glyphs (cf. *addendum*—'isomorphic direct representation').

The last two solutions for direct representation—entity-based (anisomorphic) and character-based (isomorphic), respectively—are not mutually exclusive and the use of the latter does not preclude the philologist to adopt the former at any given point in time: substituting one solution for the other is just a matter of automatically replacing entity references for characters and *vice versa*.

For methodological reasons I strongly prefer solution 2, which dispenses with entity references in character representation (Emiliano 2002, 2004a, 2004b). Thus I have collaborated with the *Medieval Unicode Font Initiative* (MUFI)[6] in the development of an inventory of medieval characters, a Unicode-compliant 'medieval' font and two medievalist proposals submitted to the Unicode consortium[7].

---

[5] Actually there is not a separate version of the editions with medieval characters: what a browser does is to interpret the entity references as Unicode or Private Use Area codepoints and to render them as medieval glyphs for display purposes only.

[6] *Medieval Unicode Font Initiative (MUFI)*—Forskergruppe for tekstteknologi / Avdeling for kultur, språk og informasjonsteknologi (AKSIS) / Universitetet i Bergen, Norge, http://www.hit.uib.no/mufi/; cf. MUFI Character Recommendation 3.0.

The main advantage in adopting isomorphic typographic representation of medieval characters in a *corpus* is to make life easier for end-users, i.e. the researchers that actually use the editions a raw material for their work. After all it is for their use that *corpora* are created in the first place: *corpora* are a means to an end not an end in themselves; the end in this case is research and knowledge. Isomorphic representation, which relies on the creation of a Unicode-compliant font, does not require that entity references be converted into glyphs: this means that the edition can be actually read with any text editor. Also, searches and data extraction are simple to perform because the base text contains all the required special characters. A final noteworthy upside of isomorphic direct representation concerns the work of the *corpus*' encoders and curators: this type of representation allows for the immediate visualization and verification of an edition in progress. Proof-reading and correction of the edition are also greatly simplified. This upside should not be underestimated when one is dealing with a large *corpus* and with entities whose names are long sequences of letters or just codepoints and thus easily subject to mistakes.

Finally, strategy 3 (normalization) can be implemented in a variety of ways, according to the aims and agenda of the editor. Normalization basically relinquishes any attempt at direct representation and results in extensive editorial intervention (cf. *addendum* for an example of what can be called "deep" normalization; the general outlook of the edition is modern, with insertion of modern capitalization, punctuation, word separation and generalized expansion of abbreviations). Scholarly editions of this type, usually called 'diplomatic editions', are extremely useful to extract historical information, textual information and linguistic data concerning lexis and syntax. These editions can also provide the basis for wordlists (*indices verborum*) and glossaries.

Since normalization is in fact 'non-representation' (or 're-representation', so to speak), it should be done ideally by philologists with a good palaeographic and diplomatic background. Normalized editions should rely on *bona fide* transcriptions.

7. To sum up

The most important requirement that must be met in the typographic representation of medieval texts is, in my view, accuracy of transcription (with clarity and consistency in the definition of editorial criteria). The actual strategy and the precise tactic or expedient adopted will ultimately depend i) on the type of study one intends to carry out and ii) on the nature of the *corpus* one wants to build.

---

[7] See Emiliano, A. and S. Pedro, S. 2003, Everson et al. 2006a, 2006b, Everson et al. 2007, Unicode Consortium 2009—Latin Extended D.

Digital typography has come a long way in the last decade. There are virtually no limits to what one can represent typographically in web-accessible plain text. As a philologist I wonder every day at how far we have come in terms of possibilities. But the edition of medieval texts in the computer age is not about what *can* be represented in print or in a computer screen: it is about what *should be* represented. Both over-representation and under-representation can be instances of bad philology: the former eludes the importance of characters and systematic glyphs as basic transcriptional units, the latter simply ignores the reality of medieval texts and thus disfigures them. *In medio virtus*.

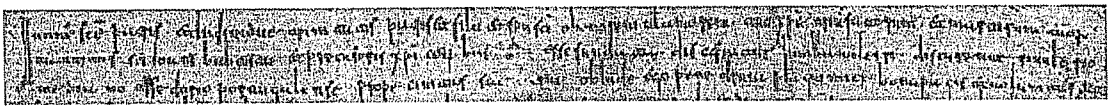
This paper is respectfully and gratefully dedicated to Michael Everson.

Addendum

Latin-Portuguese 10th-century charter (a.D. 977)—Excerpt of a deed of gift to a monastery

Date:	0977/04/22
Place:	Municipality of Santa Maria da Feira (Northern Portugal)
Archive ref.:	Instituto dos Arquivos Nacionais / Torre do Tombo, Sé de Coimbra, maço 1, n.°5 (reference code PT-TT-CSC/1/5)
Type:	Private document (deed of gift); original ms. in cursive Visigothic script
Editions:	Herculano, A. (ed.) 1867-73, vol. I, doc. CXX, 75; Santos, M.J. 1994: 323-4
Scribe:	*Inuenando*
Subject:	*Penedruia* donates several real estate items to the Monastery of *S. João-de-Ver*.

Facsimile (detail)[8]



Indirect representation

Editorial conventions

abbr	XML/TEI element=abbreviation
add	XML/TEI element=scribal addtition

[8] Detail extracted from matrix No. PT-TT-CSC-1-5_1_m0001.tif of the project *Origins of the Portuguese Language: Digitization, Edition and Linguistic Analysis of Charters from the 9th and 10th centuries* (ORIGENS DO PORTUGUÊS: DIGITALIZAÇÃO, EDIÇÃO E ESTUDO LINGUÍSTICO DE DOCUMENTOS DOS SÉCULOS IX-X, POCI/LIN/58815/2004, funded by Fundação para a Ciência e a Tecnologia).

expan	attribute of XML/TEI element <abbr>=expansion
l	XML/TEI element=line
n	attribute of XML/TEI element <l>=line number
place	attribute of XML/TEI element <add>=place of scribal addition
punc	XML/TEI element=punctuation
#a775	Unicode character reference (U+A775—LATIN SMALL LETTER RUM)
_	underscore sign=word juncture
=	equals sign=word disjuncture[9]
bold	bold text is "visible text" (regular text is annotation)

```
<text>
<body>
<l n="1"> Jn <abbr expan="nomine">nne</abbr> <abbr
expan="sancte">sce</abbr> patris <punc>.</punc> et_ Indiuidue trinitatis
patris et filii et_ <abbr expan="spiritu">spu</abbr> <abbr
expan="sancti">sci</abbr> ob_ onorem <abbr expan="uel">ul_</abbr> amorem
tuo <abbr expan="christe">xpe</abbr> triu<abbr expan="m">~</abbr>fatorum
<punc>.</punc> et martirum tuorum </l>
<l n="2"> natiuitas <abbr expan="sancti">sci</abbr> Ioanes babtista et
precusoris <abbr expan="christi">xpi</abbr> co<abbr
expan="rum">&#a775;</abbr> baselica esse fundata dis=cernitur In_ uilla
ualeiri <punc>.</punc> discurrente riuulo rio </l>
<l n="3"> media=no t<add place="supralinear">e</add>rredorio portugalense
prope ciuitas <abbr expan="sancta">sca</abbr> maria <punc>.</punc> obinde
ego penedruia plagui mici bona_ pacis et uolumtas ut </l>
...
</body>
</text>
```

Anisomorphic direct representation with XML/TEI entities

List of entities used

CHARACTER	*ENTITY*	*NAME*	*UCS#*
ɑ	aopen	LATIN SMALL LETTER OPEN A	—
ȡ	dins	LATIN SMALL LETTER INSULAR D	U+A77A
ſ	slong	LATIN SMALL LETTER LONG S	U+017F
ɇ	evis	LATIN SMALL LETTER VISIGOTHIC E	—
ꝿ	gvis	LATIN SMALL LETTER VISIGOTHIC G	—
⎮	itall	LATIN SMALL LETTER TALL I	—

[9] No elements are provided in the TEI Guidelines to handle word juncture and word disjuncture in a straightforward way. The only elements that could be used in theory to that effect are <seg> (= segment) and <join>. However their suggested use encompasses many different situations, none of which corresponds to the common medieval problem of word separation. Cf. TEI Character Encoding Workgroup 2004 (CE W 12) and Bański & Przepiórkowski 2009 for short discussions of related issues.

ꝛ	rlleg	LATIN SMALL LETTER R WITH LONG LEG	U+027C
ꭓ	tvis	LATIN SMALL LETTER VISIGOTHIC T	–
ꭗ	xvis	LATIN SMALL LETTER VISIGOTHIC X	–
◌̄	cmbmacron	COMBINING MACRON	U+0304

Editorial conventions

abbr	XML/TEI element=lexical abbreviation
add	XML/TEI element=scribal addtition
l	XML/TEI element=line
n	attribute of XML/TEI element <line>=line number
place	attribute of XML/TEI element <add>=place of scribal addition
punc	XML/TEI element=punctuation
&text;	XML entity reference
_	underscore sign=word juncture
=	equals sign=word disjuncture
bold	bold text is "visible text" (regular text is annotation)

```
<text>
<body>
<l n="1"> Jn <abbr>nne&cmbmacron;</abbr>
<abbr>&slong;ce&cmbmacron;</abbr> p&aopen;&tvis;&rlleg;i&slong;.
&evis;&tvis;_ &itall;ndiui&dins;ue
&tvis;&rlleg;ini&tvis;&aopen;&tvis;i&slong; p&aopen;&tvis;&rlleg;i&slong;
&evis;&tvis; filii &evis;&tvis;_ <abbr>&slong;pu&cmbmacron;</abbr>
<abbr>&slong;ci&cmbmacron;</abbr> ob_ ono&rlleg;em <abbr>u&lstr;_</abbr>
&aopen;mo&rlleg;em &tvis;uo <abbr>&xvis;pe&cmbmacron;</abbr>
&tvis;&rlleg;iu&cmbmacron;f&aopen;&tvis;o&rlleg;um <punc>.</punc>
&evis;&tvis; m&aopen;&rlleg;&tvis;i&rlleg;um &tvis;uo&rlleg;um </l>

<l n="2"> n&aopen;&tvis;iui&tvis;&aopen;&slong;
<abbr>&slong;ci&cmbmacron;</abbr> &itall;o&aopen;ne&slong;
b&aopen;b&tvis;i&slong;&tvis;&aopen; &evis;&tvis;
p&rlleg;ecu&slong;o&rlleg;i&slong; <abbr>&xvis;pi&cmbmacron;</abbr>
co&rndvbar; b&aopen;&slong;elic&aopen; e&slong;&slong;e
fund&aopen;&tvis;&aopen; di&slong;c=&evis;&rlleg;ni&tvis;u&rlleg;
&itall;n_ uill&aopen; u&aopen;lei&rlleg;i .
di&slong;cu&rlleg;&rlleg;en&tvis;e &rlleg;iuulo &rlleg;io </l>

<l n="3"> medi&aopen;=no &tvis;<add
place="supralinear">e</add>&rlleg;&rlleg;edo&rlleg;io
po&rlleg;&tvis;u&gvis;&aopen;len&slong;e p&rlleg;ope
ciui&tvis;&aopen;&slong; <abbr>&slong;c&aopen;&cmbmacron;</abbr>
m&aopen;&rlleg;i&aopen; . obinde &evis;&gvis;o pened&rlleg;ui&aopen;
pl&aopen;&gvis;ui mici bon&aopen;_ p&aopen;ci&slong; &evis;&tvis;
uolum&tvis;&aopen;&slong; u&tvis; </l>

...
</body>
</text>
```

Isomorphic direct representation
Editorial conventions

abbr	XML/TEI element=lexical abbreviation
add	XML/TEI element=scribal addititon
l	XML/TEI element=line
n	attribute of XML/TEI element <line>=line number
place	attribute of XML/TEI element <add>=place of scribal addition
_	underscore sign=word juncture
=	equals sign=word disjuncture

```
<text>
<body>
<l n="1"> ]n_ <abbr>nnē</abbr> <abbr>ſcē</abbr> puaꝓıſ . ꝯa_ ]ndıuıꝺue
    aꝓınıauaıſ puaꝓıſ ꝯa ſı]ıı ꝯa_ <abbr>ſpū</abbr> <abbr>ſcī</abbr>
    ob_onoꝓem <abbr>uſ</abbr>_ umoꝓem auo <abbr>ψpē</abbr> aꝓıūfuaoꝓum .
    ꝯa maꝓaıꝓum auoꝓum </l>
<l n="2"> nuaıuıauſ <abbr>ſcī</abbr> ]ouneſ bubaıſau ꝯa pꝛecuſoꝓıſ
    <abbr>ψpī</abbr> coꝫ buſelıcu eſſe funduau dıſ=cꝯnıaur ]n_ uılla
    uuleıꝓı . dıſcurꝛenae ꝓıuulo ꝓıo </l>
<l n="3"> medıu=no a<add place="supralinear">e</add>ꝛꝛedoꝓıo poꝓauguleníe
    ꝓrope cıuıauſ <abbr>ſcū</abbr> muꝓıu . obınde ꝯo penedꝓuıu pluguı
    mıcı bonu_ pucıſ ꝯa uo]umauſ ua </l>
...
</body>
</text>
```

"Deep" normalization
Editorial conventions

Normalized capitalization, punctuation and word separation (according to contemporary practice in print).
Abbreviations are expanded.
The text is divided into numbered paragraphs (a procedure which is not unfortunately current practice in diplomatic editions).

lb	XML/TEI element=line break
n	attribute of XML/TEI element <p>=paragraph number
p	XML/TEI element=paragraph
bold	bold text is "visible text" (regular text is annotation)

```
<text>
<body>
<p n="1"> In nomine Sancte Patris et indiuidue Trinitatis Patris et Filii
et Spiritu Sancti, ob onorem uel amorem tuo Christe triumfatorum, et
martirum tuorum, <lb/> natiuitas Sancti Ioanes Babtista et precusoris
Christi, corum baselica esse fundata discernitur in uilla Valeiri,
discurrente riuulo rio <lb/> Mediano, terredorio portugalense, prope
ciuitas Sancta Maria. </p>
<p n="2"> Obinde ego Penedruia plagui mici bona pacis et uolumtas ut
[...]</p>
...
</body>
</text>
```

References

Bański, P. and A. Przepiórkowski. 2009. "Stand-off TEI Annotation: the Case of the National Corpus of Polish". *Proceedings of the Third Linguistic Annotation Workshop—LAW III (Suntec, Singapore, 6-7 August 2009)*, ACL-IJCNLP. 64-7. http://140.116.245.248/ACL-IJCNLP-2009/LAW-III/index.html, http://www.aclweb.org/anthology/W/W09/W09-3011.pdf [30/6/2010]

Digital Scriptorium. 2007. "Technical Information". Columbia University of Libraries, https://www1.columbia.edu/sec/cu/libraries/bts/digital_scriptorium/technical/index.html[30/06/2010]

Emiliano, A. 2002. "Problemas de transliteração na edição de textos medievais". *Revista Galega de Filoloxía*, 3. 29-64.

Emiliano, A. 2004a. "Tarefas da Filologia Portuguesa face à documentação antiga de Portugal". *Actas do XIX Encontro Nacional da Associação Portuguesa de Linguística (Faculdade de Letras da Universidade de Lisboa, 1-3 de Outubro de 2003)*. Lisboa: Colibri, APL. 58-68.

Emiliano, A. 2004b. "A edição e interpretação da documentação antiga de Portugal: problemas e perspectivas da Filologia Portuguesa face ao estudo das origens da escrita em Português". *Aemilianense. Revista Internacional sobre la génesis y los orígenes históricos de las lenguas romances* 1. *(Proceedings of I Congreso Internacional sobre «Las Lenguas Romances en su Origen», Fundación San Millán de la Cogolla, Logroño, Spain, Monastery of San Millán de la Cogolla, 16-20 December 2003)*. 33-63.

Emiliano, A. and S. Pedro. 2003. *The Portuguese Medieval Font Project and the Medieval Unicode Font Initiative*. http://www.hit.uib.no/mufi/portuguese/TM_Unicode.pdf, http://www.fcsh.unl.pt/philologia/tm_unicode.pdf [30/06/2010]

Everson, M., Baker, Emiliano, Grammel, Haugen, Luft, Pedro, Schumacher and Stötzner (ed). 2006a. *Proposal to add medievalist characters to the UCS*. Universal Multiple-Octet Coded Character Set, International Organization for Standardization, Organisation internationale de normalisation, Международная организация по стандартизации, ISO/IEC JTC1/SC2/WG2 N3027, L2/06-027, 2006-01-30 (Working Group Document), http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3027.pdf [30/06/2010].

Everson, M., Baker, Emiliano, Grammel, Haugen, Luft, Pedro, Schumacher and Stötzner (ed). 2006b. *Response to UTC/US contribution N3037R, "Feedback on N3027 Proposal to add medievalist characters"*. Universal Multiple-Octet Coded Character Set, International Organization for Standardization, Organisation internationale de normalisation, Международная организация по стандартизации, ISO/IEC JTC1/SC2/WG2 N3077, L2/06-116, 2006-03-31 (Expert Contribution for consideration by JTC1/SC2/WG2 and UTC), http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3077.pdf [30/06/2010].

Everson, M., Baker, Dohnicht, Emiliano, Haugen, Pedro, Perry, Pournader (ed). 2007. *Proposal to add Medievalist and Iranianist punctuation characters to the UCS*. Universal Multiple-Octet Coded Character Set, International Organization for Standardization, Organisation internationale de normalisation, Международная организация по стандартизации, Working Group Document for consideration by JTC1/SC2/WG2 and UTC, http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3193.pdf [30/06/2010].

Herculano, A. (ed). 1867-73. *Portugaliae Monumenta Historica a Saeculo Octavo post Christum usque ad Quintum Decimum—Diplomata et Chartae*, vol. I. Lisboa: Real Academia das Sciencias.

Medieval Nordic Text Archive. 2008. *The Menota Hanbook 2.0*. http://www.menota.org/guidelines/index.page [30/06/2010].

Medieval Unicode Font Initiative. 2009. *MUFI Character Recommendation 3.0*. http://www.mufi.info/specs/ [30/06/2010].

Parkinson, S. 1983. "Um arquivo computorizado de textos medievais portugueses". *Boletim de Filologia* 28. 241-252.

Parkinson, S. and A. Emiliano. 1999. *Encoding Medieval Portuguese and Latin Texts for Computer Analysis: development of TEI conformant tagging guidelines for linguistic study of electronic corpora of Medieval texts*. ms., Oxford and Lisbon, Final Report of Project Ref. B-38/98 of the Anglo-Portuguese Joint Research Programme Treaty of Windsor 1998, http://www.fcsh.unl.pt/philologia/WindsorReport1999.pdf [30/06/2010].

Parkinson, S. and A. Emiliano. 2002. "Encoding medieval abbreviations for computer analysis (from Medieval Latin-Portuguese and Portuguese nonliterary sources)". *Literary and Linguistic Computing* 17. 345-360.

Robinson, P. 1994. *The transcription of primary textual sources using SGML.* Oxford: Office for Humanities Communication Publications, Oxford University Computing Services.

Santos, M.J. 1994. *Da Visigótica à Carolina. A escrita em Portugal de 882 a 1172.* Lisboa: Fundação Calouste Gulbenkian, Junta Nacional de Investigação Científica e Tecnológica.

TEI Character Encoding Workgroup. 2004. "The 'end of word' problem in Sanskrit: report of the workgroup". CE W 12: Report from Sanskrit Workgroup, 2004, http://www.tei-c.org/Activities/Workgroups/CE/cew12.pdf [30/6/2010]

TEI Consortium (eds). 2007. *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* [TEI P5 1.0], http://www.tei-c.org/Guidelines/P5/ [30/06/2010].

Unicode Consortium. 2009. "Latin Extended D, Unicode 5.2 Character Code Chart". *The Unicode Standard, Version 5.2.* http://www.unicode.org/charts/PDF/UA720.pdf , [30/06/2010]