2011 International Conference on Document Analysis and Recognition

The Four and a Half Challenges of Humanities Data

Marc Wilhelm Küster
University of Applied Sciences Worms
D-67549 Worms, Germany
kuester@fh-worms.de

***Abstract*—The lead medium of the humanities is text, but text with special characteristics that can be quite different from a normal monolingual article in most modern scripts. Text that can be derived from manuscripts, from retrodigitization of previous scholarly publications such as critical editions and dictionaries, from books printed centuries ago, applying conventions no longer in force today.**

The keynote identifies four major challenges for recognizing humanities data: Unusual characters, unusual layouts, unusual semantics and unusual segmentations. Each challenge is illustrated with concrete examples taken from a variety of times and places, starting with cuneiform tablets, an extract from a Greek manuscript, a page from a multilingual critical edition, a renaissance print, a lemma from a scholarly dictionary, and some more.

In addition, scholarly humanities data is typically marked up using domain-specific rich XML-based formats based on the TEI P5 guidelines. Any format that an OCR program produces must be sufficiently rich to permit for a mapping on TEI-compliant markup in order to be capable of reproducing the full richness of the original.

A closer view at the TextGrid virtual research environment for the humanities and its Text-Image Link Editor (TBLE) demonstrates how scholars currently tackle these tasks. It analyzes where automatization can facilitate their task and enable new dimensions of research.

I. Introduction

The lead medium of the humanities is text. Text, however, that can exhibit very special characteristics that are quite different from a normal monolingual article in, say, modern English or Simplified Chinese. Text that can be derived from manuscripts, from retrodigitization of previous scholarly publications such as critical editions and dictionaries, from books printed centuries ago, applying conventions no longer in force today. Text whose secondary characteristics[1] is often key to the message.

Scholars have to face the challenges of explicating those digitized texts for a digital age. A closer view at the TextGrid virtual research environment for the humanities and its Text-Image Link Editor (TBLE) demonstrates how scholars currently work with these tasks and where automatization can facilitate their task and enable new dimensions of research by permitting the markup of large corpora. This is all the more urgent as many humanities disciplines such as assyriology — dealing with all things cuneiform —, Persian and Byzantine studies have large libraries of manuscripts to analyze, but comparatively few scholars capable of handling them. Fully human analysis would in many cases still take literally centuries.

The paper identifies four major challenges for recognizing humanities data:

1) Unusual characters
2) Unusual layouts
3) Unusual semantics
4) Unusual segmentations

Each challenge is illustrated with concrete examples taken from a variety of times and places, starting with cuneiform tablets, an extract from a Greek manuscript, a page from a multilingual critical edition, a renaissance print, a lemma from a scholarly dictionary, and some more. Real-life specimen will typically pose a number of challenges at once, being thus positioned in a four dimensional problem space.

II. Handling Transcriptions Today: TBLE

The Text Image Link Editor[2] aka TBLE[3], primarily developed at the University of Applied Sciences Worms, is part of the TextGrid [13] ecosystem. TextGrid[4] is as a virtual research environment (VRE) for the humanities dealing with texts in a wide sense (philologies, epigraphy, linguistics, musicology, art history etc.), though the focus of the TBLE here are texts in the traditional sense.

The TBLE is primarily intended for linking the transcriptions of facsimiles or manuscripts with their digitized

[1] Secondary characteristics are characteristics of a text that do not have a direct counterpart in spoken language ([11], p. 55f), especially

- Shape of characters, for example in different fonts or the different forms of Chinese characters in China, Japan and Korea (including historical numerals such as the distinction between Arabic and Roman numerals)
- Employment of different font styles such as bold or italic typefaces or sizes of letters
- Use of (or refusal to use) specific ligatures or groups of ligatures
- Horizontal and vertical orientation of words and lines including script direction
- Employment of colours

[2] This section is heavily indepted to [12].

[3] Text-Bild-Link-Editor, German for Text Image Link Editor.

[4] The joint research project TextGrid is part of the D-Grid initiative, and is funded in its second phase by the German Federal Ministry of Education and Research (BMBF) for the period starting June 1, 2009 to May 31, 2012 (reference number: 01UG0901A).

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.206

sources, though it can also be used to build image annotations. Scholars can link segments of text with sections on the corresponding image. The information on the linking between manuscript fragments and the corresponding transcription is itself stored in TEI [3].

A. Architecture

The TBLE is integrated into TextGrid's user interface, the Eclipse-based TextGridLab, and hence also implemented as a group of Eclipse plugins. It exhibits the following subcomponents:

- Image View: shows the images and enables to select individual image sections to be linked
- Thumb View: is used for navigation. It displays a reduced version of the entire image and the active image detail (which is enlarged in the Image View) which can easily be moved and zoomed
- Toolkit: provides different functions for working on the Image View

In addition, it interfaces with the XML Editor component to type in the transcription with its markup.

B. TBLE's underlying data model

TBLE stores transcriptions, segmentation information and links in a separate file (the "link editor linked file") using an extension of TEI P5. This file in turn references to both the image(s) and the corresponding transcription(s) or annotation(s) that are in separate files. It uses embedded overlay elements, expressed in the Scalable Vector Graphics (SVG) format, for image segmentation. The following list gives a very high-level overview over such an extended TEI document:

- `<teiHeader>`: metadata of the document
- `<facsimile>`: embedded SVG for descriptions of images and links
- `<body>`: link groups with link elements. Link elements represent the relationship between the image sections and the corresponding text segments

Fig. 1 illustrates the relationship between images and texts and links as they are mediated in the link editor linked file: the embedded SVG defines the segmentation chosen for the underlying bitmap image, in this case `2hbg.0.png` (1). The rectangle identified by `shape-1` (2) is linked in the `tei:linkGrp` to the identifier of an anchor set in the TEI transcription (3). This anchor in turn is exemplified either by the `xml:id` of an XML element in the transcription (case here, cf. 4) or by an explicit `anchor` element.

Segmentation is not limited to rectangles, but can take the form of other shapes, notably arbitrary polygons.

Ideally, each single word in the transcription can thus be traced to its source in the underlying digitized manuscript that in turn contains metadata identifying fully from which physical manuscript it was produced, by whom and when.

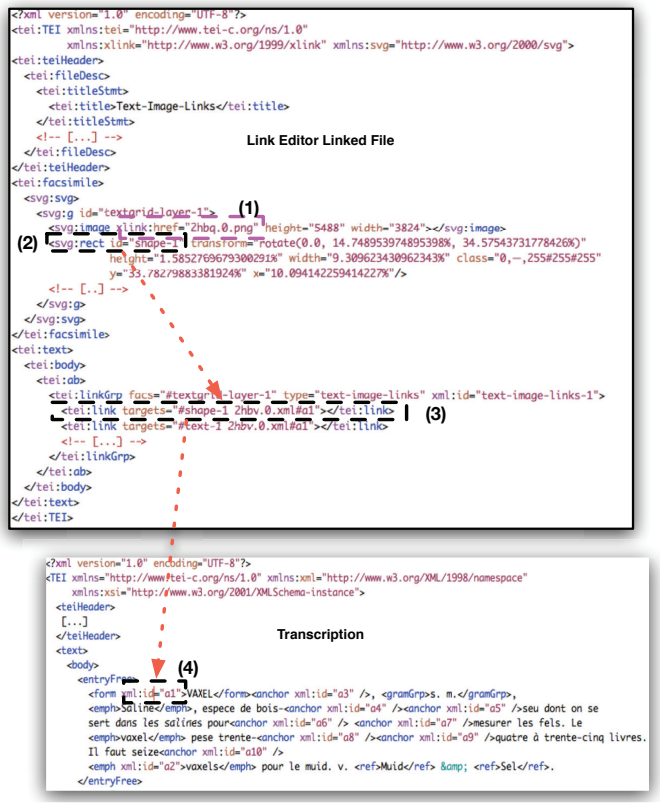


Figure 1. The TEI-5-based Link Editor Linked File

The schema thus strongly pushes the scholar to fully document the provenance chain from the physical manuscript up to the transcription and ultimately to the critical edition – a concept of built-in data provenance, based on the dominant standard in the research domain.

C. TBLE in action: Marking up a manuscript

The following example, taken from Ludwig Wittgenstein's "brown book" manuscript shows TBLE in action[5]. The editor has manually segmented the manuscript scan including the author's corrections and linked them with the corresponding transcription (cf. fig. 2).

Needless to say that not all authors have a handwriting that neat and modern, not all manuscripts are that well preserved and many use script forms that are far more remote from current usage than Wittgenstein's.

D. Automatization requirements

Working with the TBLE is a considerable advance over the traditional scholarly working methods for the prepara-

[5]Cf. [16] for a diplomatic presentation of Wittgenstein's manuscripts including Ms. 141).

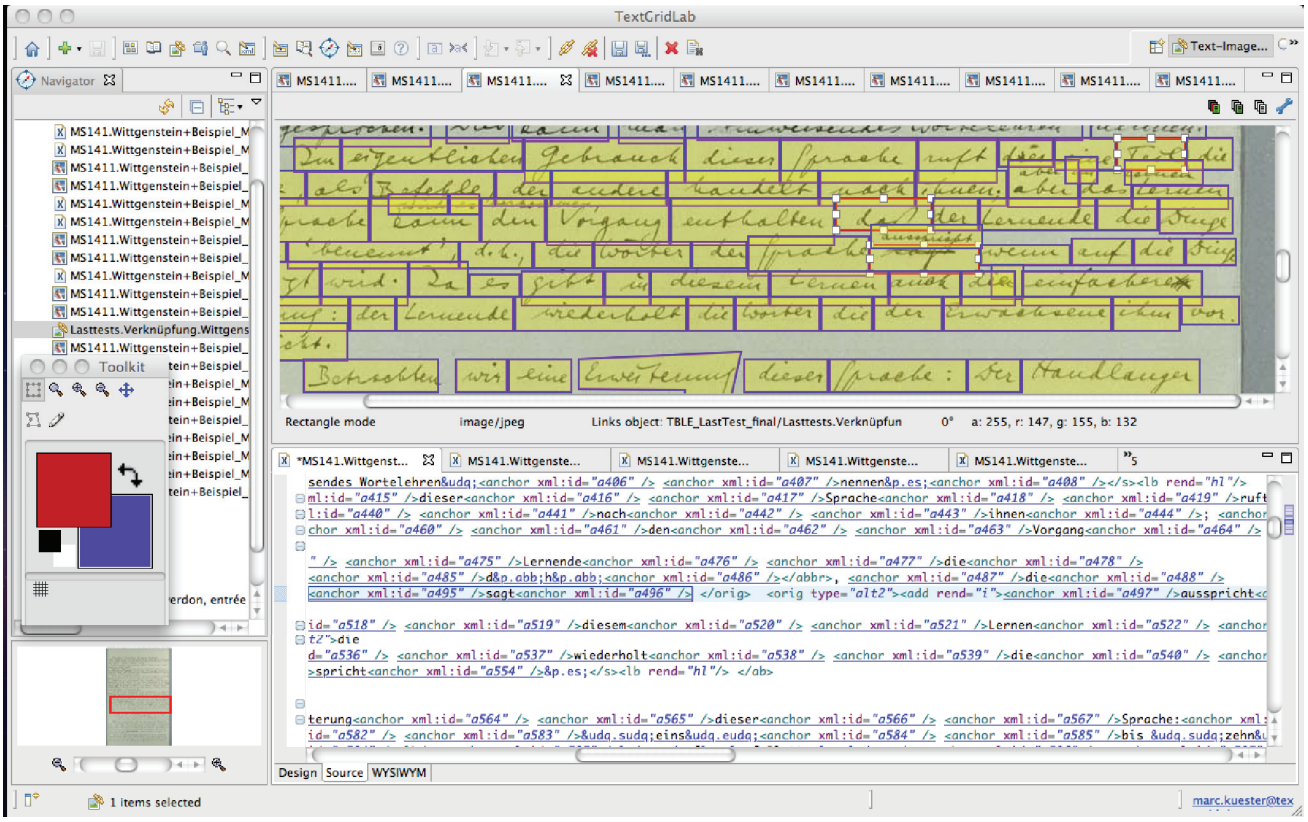


Figure 2. TBLE in action: editing a manuscript of Ludwig Wittgenstein (MS 141)

tion of critical editions. In particular it helps to improve traceability of the research by making all decisons verifiable (provenance). Nevertheless, the work remains intensely manual. Scholars have to segmentize the manuscript by drawing polygones, type the proposed transliteration and link both texts and polygones. While possible these steps are both slow and not very rewarding intellectually — in other words, ideal candidates for an at least partial automization. It would be great if TBLE itself could segment the manuscript automatically on a word and character level and already propose a first reading that the scholar then just has to validate or adapt those proposals.

This is however easier said then done as humanities texts both in manuscript and print pose their own set of challenges. Even Wittgenstein's simple, well-preserved and well-layouted manuscript page in modern German hides a number of challenges that defy current systems:

- Segmenting the text at least by words, identifying strikethroughs, overwrites and corrections
- Recognizing the individual characters in Wittgenstein's neat and modern handwriting
- Explicating the semantics of the author's actions (e.g. correction)

Other types of humanist data pose bigger challenges still which this keynote looks at in more detail.

III. Unusual Characters

As scripts evolve, characters change form, are newly invented or lost. In order to speed up writing or for aesthetic reasons specific ligatures or abbreviations become popular or go out of fashion, both in print and in handwriting. What we perceive as *the* script — and what we therefore typically train OCR engines on — is really just the last snapshot in a continuous development. Historical texts and hence much of humanities data represent other, equally valid snapshots that follow different rules.

A particular striking example is a script in plain development. If we compare three incarnations of the so-called professions list Lú-A, this becomes quite obvious. This cuneiform list is one of the oldest written cuneiform document we know of, like many of the oldest examples actually a sort of dictionary. It originates from the so called Uruk IV period (ca. 3350-3200 BC) and forms a kind of inventory of the social structure of Sumerian society, proceeding from the more prestigious ranks downwards. Over the centuries this list was reproduced again and again, reflecting the then current choices of glyphs. Two of the three versions presented here are both archaic, the first

with CDLI[6] number P000006 being the Uruk IV version, the second one, CDLI number P000161, being its Uruk III counterpart (ca. 3200-3000BC). If these two already differ markedly, the third, CDLI number about 500 years younger (ca. 2600-2500 BC) seems to have little similarity, though its contents are essentially the same. The style of CDLI P010078 already resembles more closely the "classical" cuneiform style as it would continue to be used for over two and a half millennia, albeit with significant variations.

Even without knowing much about cuneiform texts, it is clear that the early glyph versions differ significantly amongst each other and from the "final", more or less codified version (cf. also [14] and [9]).

Much the same observations I could have made with examples from other scripts, be they Chinese, Egyptian, Semitic scripts, Greek or Latin. Furthermore, even synchroneously glyphs are largely shaped by the medium they are formed in and by the degree of formality. Chinese knew early on formal and informal styles of writing (cf. [1], p. 198) (seal script, clerical script, not to speak of the very early script forms), as did Egyptian hyroglyphics where most everyday texts were not composed in the formal hyroglyphs that we have all seen on Egyptian monuments, but in the so called hieratic and later Demiotic scripts far quicker to use on papyrus and other media.

Another character challenge is less obvious, but from a scholarly perspective actually more relevant still. Historical texts in dead scripts are today typically not presented as such in print, but are in parallel or even exclusively rendered in a modern script — in other words, in transliteration. However, since modern scripts do not normally have the necessary character repertoire, this poses significant problems for OCR and until very recently also for typesetting. Let us have another look at Lú-A, this time in transliteration:

> 1 NAMEŠDA, 2 NAM_2 KAB, 3 NAM_2 DI, 4 NAM_2 NA_2, 5 NAM_2 URU_{a1}, 6 NAM_2 ERIN, 7 GAL_a ŠUBUR[7]

Each syllable here stands conventially for one cuneiform character and must be interpreted as a unit.

While the cuneiform transliterations are not that complicated, given that all the Latin characters are encoded in the Universal Character Set (UCS)[10] aka Unicode[6] and are reasonably frequently used elsewhere, this is not true for many other transliteration and transcription schemes. An example for this is the transcription system Teuthonista, widely used for well over a century for transcribing German dialects in many scholarly publications, including e.g. the the *Wörterbuch der bairischen Mundarten in Osterreich* (Dictionary of Bavarian dialects in Austria):

[6]The Cuneiform Digital Library Initiative (CDLI) http://cdli.ucla.edu is the leading initiative build a digital library of cuneiform texts and is becoming a standard. Using the CDLI number a user can directly access to the cited documents.

[7][8], p. 14f.

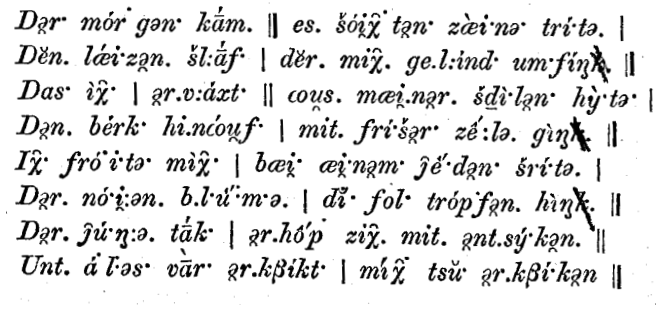


Figure 3. Extract from [2], cited after the proposal ISO/IEC JTC1/SC2/WG2/N4031

In order to support Teuthonista in the UCS no less than 26 new combining diacritics and 55 new letters must (and most likely will) be encoded. For fuller automatization, OCR tools must be trained to recognize those letters and diacritics — and the Teuthonista system is only one transcription system in scholarly use out of many.

Similar examples could be presented for the many abbreviations (abbreviaturae) and sigles used in European medieval manuscripts of which Cappelli [5], in spite of its age still the standard inventory of Latin and Italian abbreviations, identifies some 14.000 (cf. also fig. 5). Doubtlessly similar inventories could be dressed for other languages, scripts and periods, adding large quantities of glyphs that are used in humanist data.

Cappelli's inventory of abbreviaturae is a case in point where the classical Unicode character / glyph model starts to break down. Unicode traditionally distinguishes clearly between (abstract) characters and glyphs:

> The Unicode Standard draws a distinction between characters and glyphs. Characters are the abstract representations of the smallest components of written language that have semantic value. They represent primarily, but not exclusively, the letters, punctuation, and other signs that constitute natural language text and technical notation. [...] Glyphs represent the shapes that characters can have when they are rendered or displayed. In contrast to characters, glyphs appear on the screen or paper as particular representations of one or more characters [6], p. 11f.

This works excellently for modern scripts where we have a clear concept of what constitutes a character, e.g. the LATIN CAPITAL LETTER A in fig. 4, but much less for many historical texts. As fig. 5 illustrates, it is more often than not a matter of interpretation if a given abbreviatura is a character — a sign constituting natural language text, seen e.g. as a logogram —, or "just" a glyph. This ambiguity is incidentally also reflected by different preferences in scholarly critical editions, some opting to faithfully reproduce

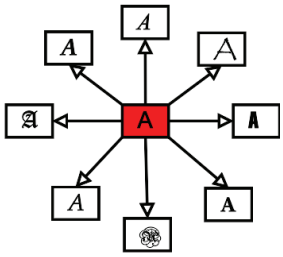


Figure 4. Glyph variants of the character LATIN CAPITAL LETTER A

the abbreviaturae of the source text, considering them an essential part of the original text, others to render them fully spelt out.

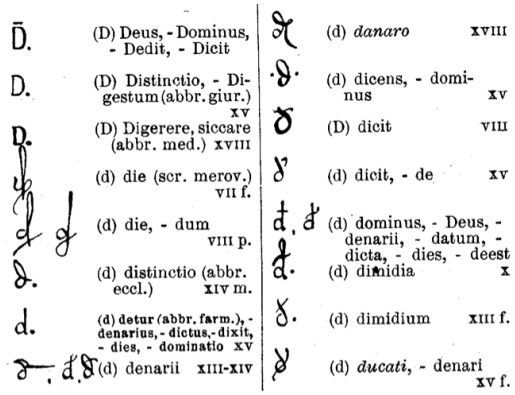


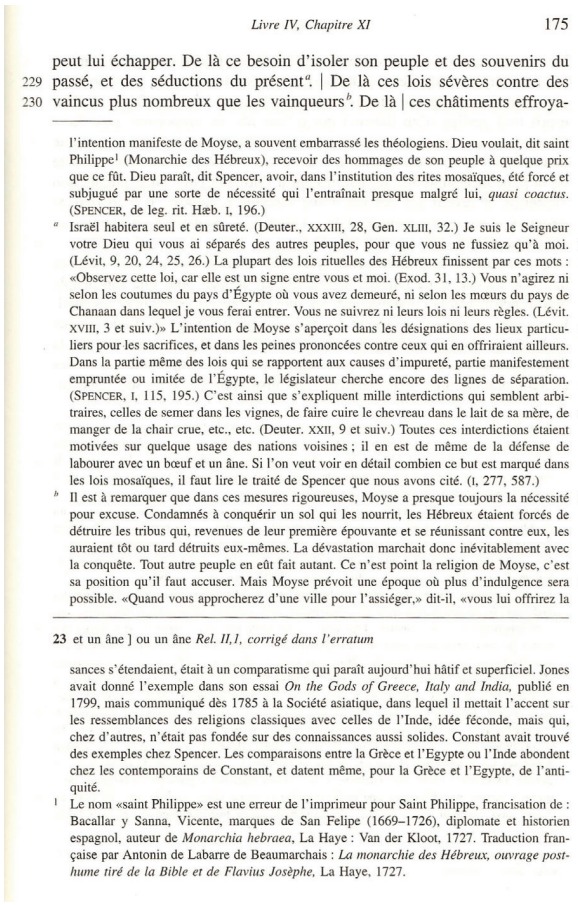
Figure 5. Abbreviaturae Deus to ducati, as listed by [5], p. 86

Either way, the correct handling of abbreviaturae and other historical signs is a formidable challenge for automatization, one that has to be able to handle the dual nature of these entities somewhere in between characters and glyphs.

IV. Unusual layout

Critical editions are a key product of philological research. They often attempt to establish an (ideally) error-free "base text" while fully describing alternative existing witnesses of the text (old manuscripts, early prints,...) and fully covering the genesis of the text (authorial or scribal additions, deletions, comments, etc.). They are supposed to note ideally all, but at least all significant variants between witnesses down to the word level in one or more critical apparatuses, indicating which of the used sources is the witness for which reading. In addition critical editions typically also to explain the historical background, unclear names, words, potentially corrupt passages. At the end of the day and up to now the typical end result is a high-quality print publication.

Critical editions in print use specific layout conventions, as fig. 6 demonstrates. The edition, whose layout is rather simple, has in addition to the main text Benjamin Constant's original footnotes, and the editors' documentation of an



Livre IV, Chapitre XI 175

peut lui échapper. De là ce besoin d'isoler son peuple et des souvenirs du
229 passé, et des séductions du présent[a]. | De là ces lois sévères contre des
230 vaincus plus nombreux que les vainqueurs[b]. De là | ces châtiments effroya-

l'intention manifeste de Moyse, a souvent embarrassé les théologiens. Dieu voulait, dit saint Philippe[1] (Monarchie des Hébreux), recevoir des hommages de son peuple à quelque prix que ce fût. Dieu paraît, dit Spencer, avoir, dans l'institution des rites mosaïques, été forcé et subjugué par une sorte de nécessité qui l'entraînait presque malgré lui, *quasi coactus*. (Spencer, de leg. rit. Hæb. I, 196.)

[a] Israël habitera seul et en sûreté. (Deuter., xxxiii, 28, Gen. xlii, 32.) Je suis le Seigneur votre Dieu qui vous ai séparés des autres peuples, pour que vous ne fussiez qu'à moi. (Lévit, 9, 20, 24, 25, 26.) La plupart des lois rituelles des Hébreux finissent par ces mots : «Observez cette loi, car elle est un signe entre vous et moi. (Exod. 31, 13.) Vous n'agirez ni selon les coutumes du pays d'Égypte où vous avez demeuré, ni selon les mœurs du pays de Chanaan dans lequel je vous ferai entrer. Vous ne suivrez ni leurs lois ni leurs règles. (Lévit. xviii, 3 et suiv.)» L'intention de Moyse s'aperçoit dans les désignations des lieux particuliers pour les sacrifices, et dans les peines prononcées contre ceux qui en offriraient ailleurs. Dans la partie même des lois qui se rapportent aux causes d'impureté, partie manifestement empruntée ou imitée de l'Égypte, le législateur cherche encore des lignes de séparation. (Spencer, I, 115, 195.) C'est ainsi que s'expliquent mille interdictions qui semblent arbitraires, celles de semer dans les vignes, de faire cuire le chevreau dans le lait de sa mère, de manger de la chair crue, etc., etc. (Deuter. xxii, 9 et suiv.) Toutes ces interdictions étaient motivées sur quelque usage des nations voisines ; il en est de même de la défense de labourer avec un bœuf et un âne. Si l'on veut voir en détail combien ce but est marqué dans les lois mosaïques, il faut lire le traité de Spencer que nous avons cité. (I, 277, 587.)

[b] Il est à remarquer que dans ces mesures rigoureuses, Moyse a presque toujours la nécessité pour excuse. Condamnés à conquérir un sol qui les nourrit, les Hébreux étaient forcés de détruire les tribus qui, revenues de leur première épouvante et se réunissant contre eux, les auraient tôt ou tard détruits eux-mêmes. La dévastation marchait donc inévitablement avec la conquête. Tout autre peuple en eût fait autant. Ce n'est point la religion de Moyse, c'est sa position qu'il faut accuser. Mais Moyse prévoit une époque où plus d'indulgence sera possible. «Quand vous approcherez d'une ville pour l'assiéger,» dit-il, «vous lui offrirez la

23 et un âne] ou un âne *Rel. II,1, corrigé dans l'erratum*

sances s'étendaient, était à un comparatisme qui paraît aujourd'hui hâtif et superficiel. Jones avait donné l'exemple dans son essai *On the Gods of Greece, Italy and India,* publié en 1799, mais communiqué dès 1785 à la Société asiatique, dans lequel il mettait l'accent sur les ressemblances des religions classiques avec celles de l'Inde, idée féconde, mais qui, chez d'autres, n'était pas fondée sur des connaissances aussi solides. Constant avait trouvé des exemples chez Spencer. Les comparaisons entre la Grèce et l'Egypte ou l'Inde abondent chez les contemporains de Constant, et datent même, pour la Grèce et l'Egypte, de l'antiquité.

[1] Le nom «saint Philippe» est une erreur de l'imprimeur pour Saint Philippe, francisation de : Bacallar y Sanna, Vicente, marques de San Felipe (1669–1726), diplomate et historien espagnol, auteur de *Monarchia hebraea*, La Haye : Van der Kloot, 1727. Traduction française par Antonin de Labarre de Beaumarchais : *La monarchie des Hébreux, ouvrage posthume tiré de la Bible et de Flavius Josèphe,* La Haye, 1727.

Figure 6. Critical Edition of Benjamin Constant's *De la Réligion* [7], p. 175

alternative reading ("et un âne / ou un âne" in line 23). In addition, the editor presents his comments and explanations in a critical apparatus which constitutes a fourth flow on the page. More complex layouts might still involve more critical apparatuses and / or annotations in margins.

Since most critical editions currently in use predate the digital era, a lot of effort is spent in retrodigitizing them. In addition to the character issues the key challenge is to reconstitute the links between the various apparatus and the passages in the original author's text to which they refer. It is, incidentally, also not easy to rebuild the original print layout from the XML encoding, as few standard typesetting tools can automatically handle more than two flows on a page with all the required cross-references and synchronizations between them.

In addition to unusual print conventions, humanities text can appear on other media than paper. We have already seen cuneiform text on clay tablets, but texts can appear on basically all other types of objects, most frequently on vases, on gravestones or on buildings. Fig. 7 shows an Etruscan syllabary on an ink-well from the 7$^{\text{th}}$ century BC, the so

called Calamaio of Cerverteri.

Figure 7. Calamaio of Cerverteri (author's own photograph, cf. also [11], p. 250)

The media, of course, influences and occasionally distorts the layout of the text, a challenge that OCR tools must face also in other fields of application.

V. Unusual semantics

Print dictionaries have to visually encode dense structural information on each lemma in as little space as possible. A typical entry will present the lemma itself, give grammatical information, declensions, etymology, multiple senses, cross-references to other lemmata and, especially in the case of scholarly dictionaries, pointers to other publications. In order to achieve this, dictionaries often make use of minuscule variations in font size, font face and style. This can be seen in fig. 8 for the very simple lemmata "Aalfang", "Aalflöße", "Aalfrau" and "Aalförmig", printed in this case in a Fraktur font.

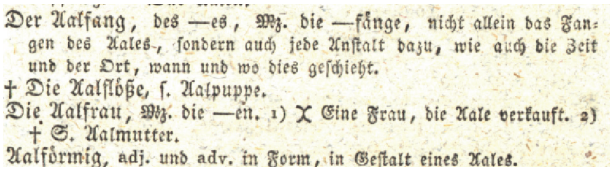

Der Aalfang, des —es, Mz. die —fänge, nicht allein das Fangen des Aales, sondern auch jede Anſtalt dazu, wie auch die Zeit und der Ort, wann und wo dies geſchieht.
† Die Aalflöße, ſ. Aalpuppe.
Die Aalfrau, Mz. die —en. 1) ✕ Eine Frau, die Aale verkauft. 2) † S. Aalmutter.
Aalförmig, adj. und adv. in Form, in Geſtalt eines Aales.

Figure 8. Lemma "Aalfang" from Campe's Dictionary of the German Language (1807-1812)[4]

[15] analyzes many of the problems of marking up such an semantically charged text, using the Campe dictionary as example. The only way to handle this challenge at present is to combine a precise capture of text and typography — in and of itself a research challenge in view of the intricacies of Fraktur fonts — with a manual analysis of the grammar the dictionary uses to structure its lemmata. Building on this [15] presents a custom-programmed Prolog parser for the Campe dictionary to actually perform the semantic markup.

At present, the author knows of no better solutions, but it is clear that hand-crafting parsers for each single dictionary to be retrodigitized is not an approach that scales to large retrodigitization programmes for semantically rich data.

VI. Unusual segmentation

We have above seen cases for the segmentation between lines e. g. in manuscripts or on unusual media. The complexity of inner-word segmentations can be higher still and differs heavily from script to script. Notoriously complex in this case are the challenges of ligatures in Arabic manuscripts and early prints, of which the small extract from the divan of the famous fourteenth century Persian poet Hafiz / Hafez of Shiraz is a simple example (cf. fig. 9).

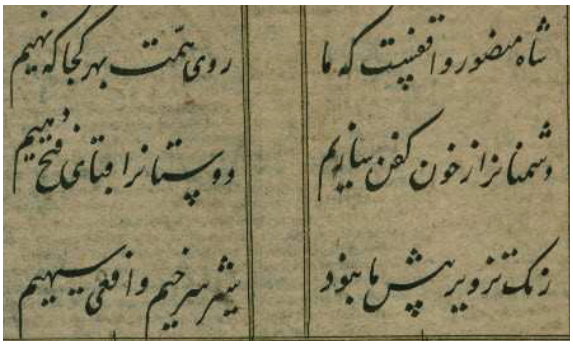


Figure 9. Hafiz: Divan, W.629: Collection of Poems (divan), Walters Art Museum, p. 122, manuscript dated 1552 AD

In addition to a number of ligatures we also see in this example clearly the traditional way of grouping in the Arabic script. Individual words form a complex unit, graphically set apart from other words on the same line of text. Fig. 10 overemphasizes this underlying design principle:

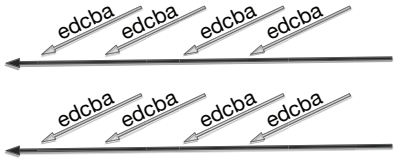


Figure 10. Traditional writing style of texts in Arabic script, cf. [11], p. 58

Within the overall limits of the script the individual scribe had a considerable degree of freedom to choose or not to choose certain ligatures that can comprise whole words. To segment these ligatures into individual characters while at the same time also documenting the scribe's choice of ligature is much like the handling of medieval abbreviaturae a major challenge for OCR on humanities data.

Much the same that has been demonstrated here on a simple example of a Persian manuscript in Arabic scripts is equally applicable to many other scripts including most middle Persian scripts such as Sogdian and Pahlavi scripts

— but even something as simple as handwriting in the Latin script can follow similar patterns.

VII. And some more

In this paper we have seen four major challenges for humanities data:

1) Unusual characters: Humanities data can not only be heavily multilingual and multiscript, they are often also printed in old font faces, rare glyph forms, idiosyncratic transcription systems or even characters at present not yet standardized in the UCS. This not to speak of the challenges of segmenting manuscripts in historic handwriting.
2) Unusual layouts: (Not only) printed text can exhibit multiple flows including types of flows that do not exist or a little used in normal printing such as critical apparatusses, text in margins etc. It can contrast the original and related passages or translations (synoptic editions). Other conventions may differ as well. Especially historical Asiatic texts can be written in writing directions that are not typically found in modern text in those scripts, e.g. from top to bottom.
3) Unusual semantics: Minuscule font variations in size or font face can encode very specific semantics which must be preserved. Especially dictionaries often employ small typographic variations to encode different parts of a lemma and its explanations.
4) Unusual segmentations: Segmenting printed text and linking transcriptions against their underlying image base is a well-understood task. However, this is not necessarily the case for text that uses a high number of ligatures such as historical Arabic even in its printed form. Segmenting manuscripts with corrections and overwritten passages pose bigger challenges still, often even for well-trained scholars themselves.

However, meeting these challenges is only a means to an end, namely to preserve to the maximum degree the semantic richness of the underlying source that is implicit in its typographic choices. These choices are then typically explicated by marking up the text using domain-specific rich XML-based formats, today normally based on the TEI P5 guidelines. Any format that an OCR program produces must therefore be sufficiently rich to permit for a mapping to TEI-compliant markup in order to be capable of reproducing the full richness of the original. This can include the explication of personal names and other named entities, of precise links and references inside and outside of the text, of authorial corrections, and much more — we have seen examples of this in the TBLE data model. To extract this rich markup from the image, to reconstitute the semantics of typography, that is the ultimate half of the humanities data challenge.

References

[1] William G. Boltz. Early chinese writing. In Peter T. Daniels and William Bright, editors, *The world's writing systems*, pages 191–199. Oxford University Press, New York, Oxford, 1996.

[2] Otto Bremer. *Deutsche Phonetik*. Leipzig, 1893.

[3] Lou Burnard and Syd Bauman. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative, 2007.

[4] Joachim Heinrich Campe. *Wörterbuch der deutschen Sprache*. Braunschweig, 1807.

[5] Adriano Cappelli. *Lexicon Abbreviaturarum. Dizionario di Abbreviature Latine ed Italiane*. Hoepli, Mailand, 3 edition, 1929.

[6] Unicode Consortium. *The Unicode Standard, Version 6.0.0*. The Unicode Consortium, Mountain View, CA, 2011.

[7] Benjamin Constant, author, and Kurt Kloocke, editor. *De la religion, considérée dans sa source, ses formes et ses développements. Tome 2.*, volume 18 of *Oeuvres complètes de Benjamin Constant*. de Gruyter, Berlin, 1999.

[8] R. K Englund, H. J Nissen, P. Damerow, and Deutsches Archäologisches Institut. Abteilung Baghdad. *Die lexikalischen Listen der archaischen Texte aus Uruk:*. Archaische Texte aus Uruk. Gebr. Mann, 1993.

[9] Stephen D. Houston, editor. *The First Writing. Script Invention as History and Process*. Cambridge University Press, Cambridge, 2004.

[10] ISO/IEC. Information technology – universal coded character set (UCS). Technical Report 10646:2011, ISO/IEC, 2011.

[11] Marc Wilhelm Küster. *Geordnetes Weltbild*. Niemeyer, Tübingen, 2006.

[12] Marc Wilhelm Küster, Christoph Ludwig, Yahya Al-Hajj, and Thomas Selig. TextGrid provenance tools for digital humanities ecosystems. In *Digital Ecosystems and Technologies, 2011. DEST 2011. 5th IEEE International Conference on*. IEEE/IES, 2011.

[13] Marc Wilhelm Küster, Christoph Ludwig, and Andreas Aschenbrenner. TextGrid as a digital ecosystem. In Elizabeth Chang, editor, *DEST 2007*, 2007.

[14] Hans J. Nissen, Peter Damerow, and Robert K. Englund. *Frühe Schrift und Techniken der Wirtschaftsverwaltung im alten Vorderen Orient*. verlag franzbecker, Bad Salzdetfurth, 2 edition, 1991.

[15] Christian Schneiker, Dietmar Seipel, and Werner Wegstein. Schema and variation: digitizing printed dictionaries. In *Proceedings of the Third Linguistic Annotation Workshop*, ACL-IJCNLP '09, page 82–89, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[16] Ludwig Wittgenstein, author, and Alois Pichler, editor. *Wittgenstein Source - Bergen Text Edition (BTE) | Diplomatic presentation*. Wittgenstein Archives at the University of Bergen, Uni Digital, Bergen, 2009.