2025-09-04 - HTR transcription conventions

!!! The following rules have to be tested in order to complete them and to know if they can reasonably be applied or not. Possibilities for changes towards a better time-result balance will be checked. !!!

We propose a flexible, double-tracked annotation system which can be reused as well for HTR and NLP tasks as by historians and diplomatists in the sense of a classical transcription.

It is based on the differentiation between “visual abbreviations”, where the abbreviation is indicated by a specific sign that indicates how the abbreviation has to be resolved (e.g. “pro”, “pr(a)e”, “per”, and many others) and “other abbreviations”, where the full extension is not indicated by the abbreviation (for instance “AD” for “Anno Domini” or any first letter standing alone for the name of a person)

The second assumption is that we keep as near to the original as possible, given the fact that normalizations/adaptations to modern norms can or will soon be changeable automatically, whereas the way it is in the charter cannot be reconstructed this way.

The character set is restricted to the 26 letters of the Latin alphabet plus the numbers from 0 to 9, which are only used as numbers and not as replacement for resembling abbreviations (like “7” for the “et” or “9-” for “con-”). This means that different forms of a letter are ignored (for instance the “long s”).

The text is transcribed line per line, as in the original. Therefore, line breaks do not need to be indicated with a particular sign.

- Abbreviations: For abbreviated words, a special representation is used containing two qualities of transcription, which in the best case is fully adopted to obtain gold standard transcriptions for model training, but can also be used partially.
  The structure is as follows: [what it means|what I see as a letter of the alphabet], for instance: [gratia|gra]
  Ideally, for the full version of the transcription, two persons should work on this, one doing the abbreviated part and one the expanded. If only one person transcribes, the two parts should be made separately.
  Depending on the aim of the transcription/the motivation of the transcriber/the time reserved for the transcription, only one of the two sides of the vertical bar can be used (the other side can always be added later). Still, the long form is always preferred.
  Only the full word: Keep the bracket on the left to indicate that originally, the word was abbreviated: [gratia
  Only the abbreviated word: Keep the bracket on the right to indicate that the word is abbreviated: gra]

- Special characters that are not part of the alphabet are replaced by the percent sign “%”, e.g.: [conditione|%ditione], [et|%]

- Missing space between words: Add the space -> helpful for NLP : add an underscore (_). In general, take the spaces as they are in the original.

- Superscripts are treated as normal letters: for instance an “o” upon a “w” is transcribed like this: hausfrawon (in this case, the superscript “o” has a phonetic value). Superscripts that stand for an abbreviation are also transcribed, for instance: “[Millesimo|Mo] cc lvi” for “Mo cc lvi”
  For (Middle High/Early Modern) German: Two dots/accents over a vowel are rendered as an “e”, excl. two dots on the “y” which are used to recognize this specific letter.

- Symbol for illegible “$”. Can be used in the brackets too. one dollar sign for every letter you think there is: Certio$a (the correct word being “certiora”)

- Punctuation: Keep as near as possible to the original if this is reasonably feasible.. We are aware that this might be difficult for many charters, especially if you work on a scan and not the original. If it is not possible to apply the punctuation like it is in the original, it is preferable to ignore it completely.

- Capital letters: Keep it like the original. Here again, normalization can be done automatically, reconstructing the original not.

- If a word is split up over two lines, keep the separation and do not use a hyphen: “[con|%] (here comes the line break) [ventus|vent]” for conventus.

- No “ß”, write “ss” or “sz” instead, depending on what the ligature looks like.

- Diacritics:
  Discussion: Especially Middle High German and Early Modern German use many different diacritics: These can be superscript letters, diagonal dots, hooks, etc. (see for example https://www.monasterium.net/mom/AT-DASP/Urkunden/1378_V_01/charter - a rather extreme example). It is a common practice for the translation of medieval German texts to transcribe the diagonal dots as superscript “e”, dashes and hooks are usually transcribed using an accent (´). As we want to reduce the number of signs, we do not use the accent. Concerning the diagonal points, the transcriber has to decide whether they are considered as an “e” or not and then apply the rules accordingly to that. Using the Umlaut may be a solution, but a) using it for medieval texts looks weird and b) we do not have the “Umlaut” for “e” and “i” on our keyboard.
  The following rules apply:

- If you ignore the diacritics, use the closing bracket to show that this word is (possibly) abbreviated (in the sense of a missing diacritic/superscript letter which is not an abbreviation)
- If you do not ignore them:
  - Diphtongue: A clearly recognizable superscript letter is transcribed as such and added after the principal letter: e over o becomes oe ("ho rent" -> "hoerent" (without brackets if you consider the superscript "e" as a full letter, or with the opening bracket if you consider the "e" as an expanded abbreviation) - you have to specify this in your personal transcription rules

  - If you do consider diagonal dots/hooks etc as "e" -> write them like a superscript "e" e.g. "ú" -> "ue" and marked with opening brackets to show that the word has been expanded "hórent" -> "[hoerent

  - If you do not consider diagonal dots/hooks etc. as "e" -> see the first point "ignoring diacritics") -> "hórent" -> "horent]

As it is particularly difficult to set up rules that cover all aspects of a transcription, the described rules shall be applied to a small quantity of charters which will be shared as examples for the users. See for instance: https://app.transkribus.eu/share/99b9823c453d10a5200dccf0b5d65d9f (example for a 13th century Austrian charter in Latin)

Objection/additions:

- If you transcribe a text in this way in for instance Transkribus, it will not be possible to train a model in Transkribus based directly on this transcription: The transcription must be reworked first before training.

- How do we indicate deletions, additions, any form of damage (and the assumed text if we can amend it), change of the script (Kurrent to Latin in Early Modern German texts for instance) -> reference in the critical apparatus?

- How do we signal missing (omitted) words added by the transcriber?

- How to transcribe the dots standing for the unknown/unspecified forename of a person - simply two dots as they are?

Further transcription rules

- Arabic numbers are transcribed as such, same for the Roman numbers: "[Millesimo|Mo] cc lvi" for "Mo cc lvi"

- "u" and "v" as well as "i" and "j" are transcribed according to their phonetic value.

- If they cannot be clearly distinguished, “c” and “t” are transcribed uniformly throughout German texts according to their phonetic value, and in Latin words according to the classical spelling. “J” is normalized to “I” in Latin words, including names.

- If they cannot be clearly distinguished, “cz” and “tz” are transcribed uniformly throughout the text. As a rule, ‘cz’ is assumed to be used until 1550 and “tz” after 1550.