

Instructions for transcription, I90822

General: The text pages to be transcribed are delivered on a portable hard drive named "texts_for_transcription". Each page is available in two file formats and the drive contains two corresponding folders: "1_export_images.jpg" and "2_export segmentation.pdf". The jpg-files should be used as the base for transcription and the pdf-files indicate the segmentation scheme for each page (where the text of each page is segmented in white frames and numbered). The text in each such frame will be transcribed in a separate file and named with the file name of the page (to be found in the top of each page and in the actual file name of the page) followed by the number of the segment (separated by an underscore) and ending by the .txt suffix. The naming of the files follows the principle:

page filename_segment number.txt

Line breaks and hyphens should be kept as in the original. Sometimes the frames that indicate the segmentation do not match the margins of the columns on the text page (e.g. if the text column is skewed due to printing conditions). In such case it should be possible to estimate the intended section of the text column, even if parts of letters are not enclosed by the white frame. Blank lines in the original are to be represented in the text by two consecutive line breaks. Indentations can be omitted, line centered text does not need to be marked as such. Images, flourishes and other decorative elements as well as lines, borders, arrows and other graphic elements that form part of the page layout can be omitted.

Font and style marking: Text in Fraktur should be enclosed in <fr></fr>, text in Antiqua in <aq></aq>, text set in Schwabacher in <sw></sw>. Italic text is tagged with <i></i>, small capitals with <sc></sc>. Note that font changes are common for single words or short stretches of text, for purposes of highlighting. When several tags apply to the same parts of the text, the order of tags does not matter.

Text set at the size of the page body text does not need size marking, but stretches of (relatively) larger text should be enclosed within <big></big> and smaller text within <small></small>.

Decorative initials and those spanning several lines should be marked with <init></init> tags.

Characters and ligatures: In general, characters must be transcribed as they appear in the document. Case must be preserved. Special characters that are common enough to have a standard Unicode representation will be transcribed as such, for instance 'ß' (the sz ligature), 'þ' (thorn), '¶' (pilcrow), '§' (section), 'ʒ' (Tironian et sign, Unicode U+204A), '@' (at), '#' (hashtag), '%' (percent sign), '‰' (per mille sign), '©' (copyright sign), '*' (asterisk). The character 'ſ' (long s) should also be transcribed as such and *not* be converted to s. Slashes should be transcribed as / (slash), even when functioning as commas, the double hyphen common in Fraktur is to be rendered as = (equals sign). Other accents on letters (mostly in antiqua) should just be typed as their normal Unicode representation.

Obvious ligatures, that is, those where the shape of the letters is changed for the sake of the ligature should be transcribed as their composing characters separated by an underscore. For Fraktur and Swabacher this is notably the **tz** ligature, which will be marked as **t_z**, for Antiqua these are predominantly **c_t**, **s_t**, **o_e**, **a_e** and **t_z**. Only the cases where the composing letters look different are to be marked as ligatures, e.g., 'Œ' becomes **O_E** but 'OE' is just **OE**. Other cases that might be considered ligatures (e.g., 'ch', 'ck', and several combinations starting with 's' or 'f' in Fraktur) need not be marked as a ligature.

The character **ṃ** (m with doubling sign) should be coded **m_m**, its capital counterpart **M_M**.

As in German Blackletter, the 'ä' and 'ö' are written with a tiny 'e' above the 'a' or 'o': $\overset{e}{a}$, $\overset{e}{o}$. They can be transcribed as just ä or ö. However, in Swedish there is the additional character 'å', which both in Blackletter and in Antiqua is marked with a small ring above the 'a'. Care must be taken to distinguish 'å' from 'ä', as they may be very similar in Blackletter.

Fractions should be written out using a slash, that is, $\frac{3}{8}$ and $\frac{3}{8}$ become **3/8**.