Tekst, context en debat – Quine's *Word & Object*

Opdracht I/2 (Zotero/SalVe)

General

For this assignment you will help make all the published work of Willard van Orman Quine digitally and quantitatively analysable. You have been given one or more texts in pdf format to turn into digital text files (txt and TEI) suitable for processing by SalVe and other tools. This assignment consists of four elements: (a) cleaning all the texts that are assigned to you, (b) preserving their structure by converting them to the TEI format, (c) split it in smaller parts (so they can be separately analysed in SalVe) and (d) document any problems you run into in a report form.

TEI (Text Encoding Initiative), is an XML-based standard format used for encoding the structure of digital texts. Therefore, it is popular with digital humanities scholars, because it provides a standard way to tell computer tools that something is a paragraph, a quote or a footnote. We will be working with TEI Lite, a simplified version that is widely used.

There is a lot of information in this document but we strongly recommend you to read it all through before starting to work, that will likely save you time and confusion later.

You have successfully completed the assignment when you have: (1) cleaned and structured the text(s) you were assigned thoroughly and precise *and* splitted it in the way we asked you to (2) documented the problems you ran into and solve it where possible (3) handed in the files at the end of the assigned dates (so at last 21:00 on 23 November).

More in detail: to complete the assignment you must hand in the following files:

(1a) For every text (see below 'Starting with ABBYY' and 'Cleaning / splitting plan) :

- [*Year_Title_Author*_metadata] in txt format
- [*Year_Title_Author*] in TEI format
- [*Year_Title_Author*] (when there was nothing to split), or (when you had to split) a series of files in the form [*Year_Title*_sec*digit_Author*] or [*Year_Title*_ch*digit_Author*] all in txt format

(1b)For some text:

- In some cases: [*Year_Title*_introduction_*Author*] or [*Year_Title*_preface_*Author*] in txt format or a file we mentioned in the text specific instructions (emailed to you).

You will receive information on what text(s) you have to clean and split *and* specific instructions for that text by email by Lisa and Yvette.

(2) For every student:

- [*Year_Title_Author_YourName*_Report] in doc or docx format (see below in 'Report form/question gdoc'). The Report Form can be downloaded here.

Be sure to save the files in the right format, specified above. Put these files in this folder https://drive.google.com/open?id=0BxuO3YCtqjLxeTdkMTBIRm93NUE by making a separate folder inside this one for each work you cleaned. Have separate report forms for each work and put them in the folder as well.

Report form/Question gdoc

One important thing to keep in mind is that a badly cleaned text will corrupt the data, lead us to false conclusions and jeopardize the research. The cleaned texts will be checked but if you have corrected something that shouldn't have been corrected or deleted too much it would in some cases be hard to make it right again. Bottom line: we'd rather receive no text than one that is badly cleaned.
While cleaning, structuring and splitting, there might be things you are unsure about what action to take. For every doubt you have or problem you run into we need you to make a note in your Report Form. Download the Report Form and fill it in for each decision you took you were not fully certain of, including the questions and problems may have contacted us about. Your report form file should be named [*Year_Author_Title_Your*Name_Report].
When the work you got was too big to finish completely put in the report form exactly what you did and did not do. Keep in mind that we rather want less well cleaned texts then more badly cleaned ones.

How do you deal with a problem?
- Check if the solution really isn't mentioned in the instructions in this document.
- Try looking in the question gdoc. In this gdoc we collect the questions students ask us about the assignment along with the solutions. If someone ran into a similar problem you could find your answer there.
- Ask us; that's what we're there for. Don't forget to still make a note in the Report Form (unless we say it's not necessary).

Starting with ABBYY

You will use an OCR (Optical Character Recognition) program, ABBYY, which is also a helpful cleaning interface. Two general guidelines: (1) we need the main text (including the title(s) of the text/ chapters/ paragraphs) and the footnotes in the final documents, nothing more, and (2) while cleaning always keep in mind SalVe's functions and the way they are performed. If you have questions concerning the latter, please contact us.

Find and open ABBYY on the UvA computer in All Programs/ABBYY Finereader 12/ABBYY Finereader 12. If you get a licence agreement request, select English and click accept.

You'll first need to check and change some settings. These settings are lost when you logout (keep that in mind if you don't want to do the assignment in one go). Open the option menu: Tools/Options. You will get a window with different tabs.

Document tab

- Click 'Edit Languages' and make sure English, German, Greek, French and Latin are selected as document languages.
- Change the color mode to black and white.

Scan/Open tab

- Make sure 'Automatically process pages as they are added' is selected

Read tab

- Make sure 'Thorough reading' is selected in reading mode

Save tab

- In subtab DOCX/ODT/RTF deselect 'Keep headers footers and page numbers', 'Keep line numbers' and 'Keep pictures'. Select 'Highlight low-confidence characters'.
- In subtab TXT select 'Use blank line as paragraph separator' and deselect 'Keep headers and footers'

View tab

- Select 'Highlight low-confidence characters and non-dictionary words'

Advanced tab

- Select 'Stop at unknown compounds', 'Correct spaces before and after punctuation marks', and 'Participate in the Customer Experience Improvement Program'.

Open your pdf file and wait a little until all pages are processed. The yellow box at the right upper corner disappears. If there happen to be 10 or more words wrongly OCR-ed, first try to do something with the settings or the contrast of the pdf image. You can assess ABBYY's manual here, or contact us.

Cleaning and splitting plan

1. *Metadata*

Once ABBYY has processed your text you can start with cutting out all metadata (publishing information, index, acknowledgements, bibliography, references, dedication, etc.). Paste the metadata into a separate file with the same name as your original document, except for the addition of '_metadata'. So your file will be named [*Year_Title*_metadata_*Author*].

2. *Redundant elements*

- *Structural ones.* Recall, we need the main text (analysis-relevant), including the title(s) of the text/chapters/paragraphs, and the footnotes in the document, nothing more. Remove headers, footers[1] and page numbers, if still needed. You can easily delete the green blocks in the left pdf window. Dragging the green blocks, e.g. to exclude a page number, does only work if you let ABBYY re-analyze the relevant page. Removing the page number manually is faster in this case.
- *Rare ones.* Remove all other pieces of text that can influence SalVe's similarity value calculation in a bad way. All analysis-irrelevant text, e.g. a closing of a paper like 'Harvard University, Cambridge, Massachusetts U.S.A.', JSTOR Term & Conditions of Use, Quine's signature, or descriptions below images. Use your common sense and your knowledge of SalVe's way of functioning as a guide. When you are in doubt it is very important that

[1] A header or footer can be the name of the author, the title of the book, chapter or section, or a combination of those, appearing respectively on the top or bottom of the pages.

you ask us, report your doubts to us (if your question is not answered in the question gdoc), and make a note in your report form (depending on the weight of the problem).

4. *The real cleaning: checking words*
Check all blue marked or red underlined words. When you are in doubt about the needed improvement, click on the word and the relevant passage in the pdf file appears. Realise that you do not have to make the text fully perfect. For example, an OCR-error such as 'as Carnap5 pointed out' (the 5 is a footnote reference) needs to be corrected in Carnap 5, because Carnap5, not being part of SalVe's dictionary, will not be included in the analyses. But OCR-errors concerning punctuation such as 'as. Carnap 5 pointed out' do not matter, because SalVe does not read punctuation. Again, keep in mind SalVe's way of functioning while cleaning.

6. *Footnote transfer and marking the splits (change to keep)*
Move all footnote text to the end of the section or chapter (depending on the way your text needs to be split). Mark the end of the section or chapter at once as a splitting point, by putting [split] in the text on the spot the split must be made, so later on you can easily search in your .txt document for the places the document must be split.

7. *Deleting the rest*
The titles 'Bibliography', 'Notes' or 'Reference' are not titles of the text/chapters/paragraphs so they should be deleted. Everything that is not written by Quine should be removed, e.g. prefaces written by others. An introduction or preface that is written by Quine, is always a separate section, put in a separate file, named [*Year_Title*_introduction_*Author*] or [*Year_Title*_preface_*Author*].

8. *Providing readability*
Rewrite the section or chapter title in capitals and separate it from the text with an enter. Subsequently, check whether accompanying order indicating signs are correctly represented. For example, sometimes an 'I', meant to indicate the beginning of a first section, is OCR-ed as an 'i'. Change that to an 'I' again.[2]

9. *Dealing with the symbols*
Mathematical symbols or Greek letters are often not OCR-ed well. They might turn into the wrong symbol or a blank spot. There are two cases:

- In text symbols: change the symbols into *s*
- Formulas: delete the whole formula and change it to <formula></formula> (an empty formula element in TEI)

10. *Hyphens*
Move on to the last cleaning step: correct end-of-page hyphens if needed. ABBYY corrects all redundant hyphens except if they appear at the last word of a page.

[2] The goal of this step is to provide basic readability in the overview page that SalVe shows when presenting the results of the analyses.

11. *Save and finish*

Save the document as a Word document (.doc) with the name [*Year_Title_Author*]. From this Word document, you will make the TEI version and the TXT version. To make the TXT version, first open the doc file and save it as .txt in Word. Then open the txt file, search for your [split] points, cut out the text per section or chapter and put it in a separate file.

How to name your files depends on the way the splits are made:

- per chapter → [*Year_Title*_ch*digit_Author*]
- per chapter section → [*Year_Title*_ch*digit.digit_Author*] (ch5.2 = section 2 of chapter 5)
- per section --> [*Year_Title*_sec*digit_Author*] (when each section/paragraph has its own number, like in W&O)

To illustrate: [1970.2nd1986_Philosophy of logic_ch4.3_Quine] contains the third section of the fourth chapter of Philosophy of Logic. [1960_Word and object_sec45_Quine] contains paragraph 45 of Word and Object.

When nothing in the email is mentioned about splitting the text, you do not have to split the main text in smaller sections (but you still have to split the metadata from the main text).

Do not forget to delete all [split] indications.

12. *Last check*

Do a last check on your separate txt files. Are all sections or chapters separate files now? Is the section title separated from the rest of the text with an enter? Are there a lot of redundant enters anywhere in between the text?

Structuring plan

Next, you will convert the Word document version to TEI, and check the resulting structure. The basic idea is to have enough structure to be able to separate the running text from other textual elements such as footnotes, headings, formulas and bibliographical information. We also want to be able to process elements written by others, such as quotes, separately. We would also like to know what elements were highlighted, as these could be considered differently when processing the text. Lastly, we want to be able to link the scanned images of pages to the digital text.

Word is very good at markup, layout and other visible elements of structure, but not so good at semantic aspects of structure, such as quotes. Therefore, it is likely that the converted TEI documents has correctly marked paragraphs, but no marked quotes, for example. Below, we will go through each type of structure (and TEI element) that should be checked. TEI is a huge format and there are many others, so it is almost impossible to encode every element of structure. We have tried to list only the most useful ones, on the basis of a selection made by the Deutsches Tekstarchiv (DTA): DTA Basisformat level 2. An overview of it can be found here: http://www.deutschestextarchiv.de/doku/basisformat/uebersichtText.html

Go through the converted TEI document and check the following structural elements:

1. Title-related elements

All front matter, before the main text, should be enclosed in <front> tags. All back matter, after the main text, should be enclosed in <back> tags. Within either, there may be a title page, which is encoded with <titlePage>. Parts of the title should be encoded with <titlePart>.
The byline, containing for example the name of the author(s), should be encoded with <byline>, also if it occurs at the end of a text.
Copyright statements or other statements authorizing the publication of a work should be encoded with <imprimatur> tags.
Subdivisions of the front/back matter that are important, but not covered by another tag, should be enclosed in a element. For example, the table of contents or a preface.

2. Paragraphs and text

In TEI, paragraphs are enclosed in tags: paragraph . The converter probably got most of these correctly. If not, make sure all paragraphs are enclosed in tags. Furthermore, the entirety of the text (excluding back matter and front matter) should be enclosed in a <body> tag. Subdivisions of text that are important, but not covered by another tag, should be enclosed in a element. In the body of a text, these would be sections or chapters. For example, or .
Text that is somehow highlighted (i.e. in italics or bold) should be enclosed in <hi> tags.
If a text contains text in a foreign language (compared to the language of the rest of the text), enclose this in <foreign> tags. This is not normally part of DTA level 2 but it is important for us.
If the text contains verses with numbered lines, or otherwise numbered lines that are important, they should be encoded with the <l> tag. If such lines form a group, such as a poem, this group of lines should be enclosed with the <lg> tag.
If a text contains of multiple columns on one page (besides margin notes and such), an empty <cb> tag is placed at the beginning of a new column on the multi-column page.
If there is an unusually large space in the running text (not a normal space), this should be marked with an empty <space> element.
If a significant portion of the text is somehow missing (i.e. lost due to OCR mishaps, or simply not attested), put a <gap> element there.

3. Pages and page numbers

At the beginning of each page, there should be an empty <pb> element, with the number of the page as an attribute: <pb n=”1”> is the start of page 1

4. Headings

Any type of heading can be enclosed in <head> tags. For example, the title of a section or a table, or the heading of a list.
Epigraphs, whether in sections or on the title page, are encoded with <epigraph>.
Sometimes, you may encounter something that summarizes a subdivision of the text (i.e. a chapter). This summary can be prose, or a list of things. These should be encoded with <argument> tags.

5. Footnotes and margin notes

Any type of note should be enclosed in <note> tags. This also includes annotations. To specify that it is a footnote or a margin note, use *place* properties such as *bottom*: <note place="bottom">, <note place="margin">. A full overview of possible places is here: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-att.placement.html.
Different types of note, such as author notes or editor notes, can be distinguished with *type* properties. In the XML file, the note should be placed within the text at its point of attachment.

6. Lists
Lists should be placed within <list> tags, with each item in the list being encoded with an <item> tag.

7. Quotes
Quotations without an author reference should be enclosed in <quote> tags. Quotations with an author reference are encoded with a <cit> tag. This <cit> element should then contain both the quote in a <quote> tag and the bibliographic reference, as explained in the next section.

8. References
Bibliographic citations (of any form) should be encoded with the tag <bibl>, such as an entry in a bibliography. Other types of references, such as a (Bloem, 2018) in the running text, should be encoded with <ref>. It is not necessary to precisely specify the target of that reference as that is a lot of work.

9. Tables
If there are tables that are important and/or have been preserved well enough by the OCR, they should be encoded within <table> elements and consist of <row> elements that contain one <cell> per column. Example: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-table.html

10. Formulas and figures
Use empty formula element as marker: <formula/>. If a graphical element or something else which could be deemed a figure was present, put empty <figure/> tags.

11. Names
(to be decided)

12. Abbreviations
(to be decided)

13. End of line splits
(to be decided)

Last remark
If you have suggestions for the cleaning/splitting plan or if you happen to discover a great, yet unmentioned option in ABBYY or strategy for cleaning, please let us know and include this in your report form!

Contact info

Lisa Dondorp: lisadondorp@ymail.com
Yvette Oortwijn: yvette.oortwijn@gmail.com
Jelke Bloem: j.bloem@uva.nl