SEGMENTATION OF TEXTUAL ARTEFACTS IN DIGITAL HUMANITIES PROJECTS

CRISTINA VERTAN[1], WALTHER VON HAHN[2]
[1]Herder Institute for Eastern European History | [2]University of Hamburg
cristina.vertan@herder-institut.de | WvHahn@gmx.de | DOI: 10.18135/PG70.2024.17

Derived mainly from a European/American language technology background, digital humanities projects have used in a rather blind way for decades annotation as "golden tool". Less attention was paid to the type and scope of such annotations. This paper discusses segmentation decisions as an obligatory prerequisite of any annotation project. Also it introduces a number of parameters which should be the ground for these decisions. The paper tries to address a number of issues, some of them known from the research on rich morphology and less resourced languages, in which Gábor Prószéky plays an important role (Prószéky & Merényi, 2012).

1 Introduction

The tasks of representation, processing and analysis of textual artefacts is extending nowadays from mere language technology applications (e.g. machine translation or information extraction for marketing purposes) and corpus linguistics, to all fields of humanities. Methods, Models and Tools used once either for rather engineer applications or pure linguistic research are now one of the first (and most important) steps in the digitization and digital usage of textual objects in other disciplines as history, cultural heritage, ethnology, musicology or dedicated cultural area research (e.g. classical Ethiopic, old cultures on the American territory). This scope extension was done for almost two decades without a deep reflection on new data types, to which such methods and tools where exposed.

As a result, many projects in digital humanities either artificially limit themselves to some positive shallow representation or simply ignore text features (like vagueness and uncertainty of natural language) which in the hermeneutic (analogue) analysis plays a major role. (Dilthey, 1883). This gap between the physical digital analysis and the "traditional" (analogue) research on categories like Diltheys "Verstehen". They often led to misunderstandings and less enthusiasm for the new methods among humanities' researchers.

During the last years increased awareness about the potential of deep and text-type oriented representation in the humanities was gained together with the massive expansion of computational power, but looks like the representation of an increasing number of textual features. They can lead to better models for the artefacts.

In this paper we argue that not only the number of represented objects but their selection out of the text plays a central role. Although this may seem trivial we will show, focusing on the "segmentation" problem and its implication for annotation, that the contrary may be true. We claim that such multidimensional analysis of a project scope, available data and technical limitations has to be done right at the beginning of each DH-project, independently of the technical paradigm followed (deep-learning, ruled-based or mixed) and might avoid illusions about possible results.

The paper mainly concentrates on research projects and their practical success or failure. The contents of this paper may sound a bit trivial to non-experts in Digital Humanities, but the given examples might avoid illusions about possible results of projects. The central issue is how to prepare (represent) texts for digital processing of any kind.

2 Written artefacts, annotation and the segmentation problem

In the following sections we will use the term written artefacts for all sorts of objects containing language (in contrast to art objects or topographical pictures, audio or audio-visual objects). By language we understand written natural language but also other similar notation types like musical or phonetic alphabets.

Written artefacts in natural language can be represented though different script types: alphabetic (e.g. all languages from the Latin, Germanic, Slavic, Finno-Ugric families as well as Arabic, Jewish, Greek), syllabic (e.g. the Japanese Katakana or Hiragana), abugida (a mixture between alphabetic and syllabic as Amharic and classical Ethiopic (Ge'ez) or logographic (with symbols representing a concept like Maya glyphs, Egyptian hieroglyphs, Sumerian or Chinese).

Digitization of written artefacts is the prerequisite for any DH-Project and comprises the digital image of the text as well as any further transformation aiming at the enrichment with knowledge which may serve the research purpose (annotation). We distinguish among:

a) metadata-annotations (referring the entire object) and

b) content-annotation (on different parts of the text).

The current paper discusses the content-annotation, whilst one should keep in mind that the metadata granularity, i.e. the level at which the metadata are inserted (e.g. entire collection, book, page) faces also the segmentation problem.

2.1 Annotations-Levels and Processing Pipelines

In DH-projects we distinguish between layout, linguistic and domain/application annotations. They are created automatically or manually. In contrast to layout or domain annotations, linguistic annotations are often hidden in the visualization. They are mostly used in order to detect domain specific information, so their scope is strictly related to the aim of the research. Many projects grab automatically pre-compiled processing pipelines: tokenizer → lemmatization → part-of-speech tagger, followed in rare cases by a syntactic

annotation of noun or verbal phrases. Occasionally semantic roles or links to language specific subsets are annotated.

The problem of this type of approach is, that such pipelines are applied without a deep definition of the annotated segments and the appropriateness of this employed segmentation to the particular research problem, respectively underlying language. In the next section we will argue that the decision on text-segmentation is a prerequisite for any successful DH-project dealing with textual artefacts. We will discuss different levels of segmentation and the possible pitfalls by annotations, keeping always an eye on the software economy.

3 Segmentation Levels

In this section we will discuss possible segmentation levels in written artefacts languages with alphabetic scripts and give some hints about challenges when dealing with other language types as mentioned in section 2.

However before deciding on a proper segmentation, the first decision to be taken is the alphabet which will be the basis for the annotation. It often happens in DH projects that languages based on non-Latin scripts are transliterated in Latin script. However, not all (historical) scripts have standardized transliterations (e.g. for Ottoman Turkish there exist several transliteration schools). Thus, one should decide which text version (original script or one of the transliterations) will be the basis for the annotation. Secondly, one should decide if there is a one to one linear correspondence between transliteration and original script alphabet. In case of Semitic languages e.g. transliteration may contain the vocalization, while the original script is might be missing it. The most complicated case is an annotation, where the text must have a right-to-left direction and the annotation (of separable verb forms, e.g.) have to be annotated left-to-right.

In order to illustrate different levels of annotation-complexity we present in Figure 1 three cases. Case 1: Annotation in a traditional basic language technology project (Part-of-Speech-Annotation /PoS in a German text). The focus here is on the annotation of the supposedly most relevant PoS (nouns, verbs, adverbs, adjectives). The second case shows a more complex annotation model in the annotation of classical Ethiopic (Vertan, 2018a). Here the annotation is done on several layers, there is a more fine-grained segmentation, and the original script and transliteration are both involved. The third case shows an even more complex situation in which annotation also involves segments at the text level, references to external sources (ontologies) (von Hahn & Vertan, 2019).

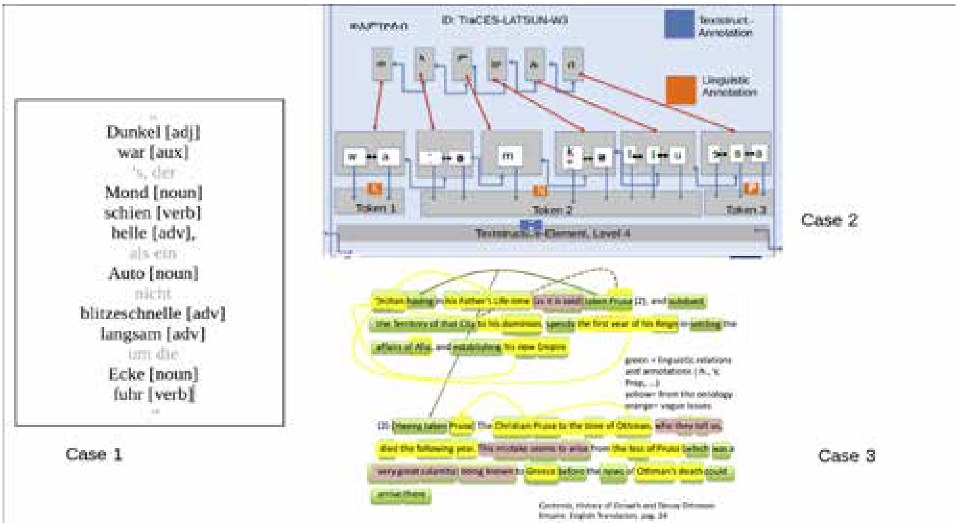

Figure 1 Different segmentation levels

From the above examples one can see that different levels of segmentation may be considered. Our claim is, that before starting to design the annotation model one should go through the segmentation scale presented in Table 1 across a multidimensional parameter scale presented in Table 2.

Segments	Targets, blind spots
Characters: **letters (a, ț, í, ڂ, Ĥ, ኈ,) separators (፡, empty character), single numbers (1, 6, ፸,)**	For annotation of transcription doubts (in Edition projects), or when the letter itself is a meaning carrier; for the detection of abbreviations or segment borders (take into account, that non-European languages may have other separators as the blank) and that one letter may be encoded with 2 Unicodes (like the case of Ĥ in Amharic
Punctuation marks (blank, , ;, ?, -, ፧, ፨)	Often neglected and just erased from the source text, punctuation marks deliver information on the level of discourse as well as syntax (relative clause). They must be separated from the meaning segments (words, letters).
Letter groups (sch)	E.g. for Semitic languages in which one letter may be transliterated by a group of characters in the Latin alphabet. (consonant-vowel)
Morphemes (ver-, -ul,)	To mark postfixed articles (like in Bulgarian or Romanian) or prefixed particles like in German
Tokens (Dunkel, war's, der, ቀንከም)	Group of characters between two separators. In the Amharic example ቀንከም we have a combination of conjunction (ቀ=and) and pronoun (ንከም = together). In the German example "war's" is an abridged form of 2 words (war es =it was)

Words (Common or proper names, Verbs, etc.)	Usually the main segment in DH and language technology projects
Compounds (Donaudampfschifffahrtsgesellschaft, Alu|mini|umher|stell|ung)	Must be fragmented into its (sets of) parts and thus produces ambiguity. The second example has 12 readings!
Multiword lexemes (The United States of America)	Often these segments should be isolated together as meaning carriers. One should decide for every singular case if only the idiom as such or also its components need to be segmented for the projects' result.
Idioms (Solving that math problem turned out to be a piece of cake for her)	Very important for semantic representation as the meaning of the idiom differs from the simple combination of the meaning of its parts.
Syntactic categories	It is important to define the category and the grammar (constituent vs. dependency)
Concepts [IDEA]	Concepts in the sense of the "triangle of reference" can have word equivalents per language and can be suitable for synonym analysis or language comparisons. However, concept hierarchies (ontologies) cannot be built with words.
Propositions (logical combination of group of words)	Whereas propositions do not comprise questions or relative clauses, negations with their scope are standard versions of (declarative) propositions. Propositions can be orders, modal utterances, declarations, e.g.
Text segments beyond sentence level	It is primarily used for summarization or discourse analysis. Textual segmentation may comprise different (possibly hierarchical segmentation levels). In critical edition projects one may face the problem of overlapping text segments (different editions with different segmentation layers). Syntactically ambiguous sentences may require overlapping annotations too.

Table 1 Segmentation Layers

One particular aspect, which may affect any of the above mentioned levels is the (partial) destruction of one or more segments. This has to be represented and the representation form is essential for any further processing. In texts based on hieroglyphs for example it is crucial not only to mark a specific area as destroyed but also to distinguish about the position and the type of the destruction. This is something to be kept in mind by any project working with historical /archeological material.

It is apparent that there is no "natural" segmentation of texts across any of the levels in Table 1. Thus the choice of the segmentation levels depends on a series of parameters which we summarize in Table 2.

Segmentation parameter	Consideration
Project outcome	What has to be visualized; which algorithms have to be run in the background?
Available technical capacities	A large collection of texts in which each letter is annotated (i.e. represents a segment) may require very fast storage and processing capacity.
Available software and/or representation language	Does the representation allow for reasoning? Representing and annotating uncertainty at all levels in the text may lead to a very slow performance of the used reasoner; moreover it may exceed the logic implemented within the available reasoner. If e.g. the number of uncertain places at the level of single characters are limited one may omit this segmentation level.
Language-type	Is a transliteration level needed? Is the text an agglutinative language (e.g. Hungarian)? Is it a rich compound language (like German). Which is the written orientation of the language and of the mark-up?
Text-type	Inscription, novel, legal text, legend.

Table 2 Parameters for the choice of segmentation layers

3 CONCLUSION

In this paper we present a possible workflow for DH or language technology projects dealing with annotation of textual artefacts. We argue that the first step in such projects is to set the segmentation levels according to a set of parameters. Although written from a perspective of projects in the European language setting, the paper tries to go beyond these borders and address issues in other language families. Possible representation formats which are flexible enough to model this multidimensional problem are presented in (Vertan, 2018b).

References

Dilthey, W. (1883). *Einleitung in die Geisteswissenschaften (Großdruck): Versuch einer Grundlegung für das Studium der Gesellschaft und ihrer Geschichte* (p. 357). Leipzig: Teubner.

von Hahn, W., & Vertan, C. (2019). Modelling linguistic vagueness and uncertainity in historical texts. In *Proceedings of the Workshop on Language Technology for Digital Historical Archives in conjunction with RANLP-2019* (pp. 34–38). Varna: INCOMA Ltd. http://doi.org/10.26615/978-954-452-059-5_007

Prószéky, G., & Merényi, Cs. (2012). Language Technology Methods Inspired by an Agglutinative, Free-Phrase-Order Language. In *Multilingual Processing in Easternand Southern EU Languages* (pp. 182–206). Cambridge: Cambridge University Press.

Vertan, C. (2018a). Data Modelling for Historical Corpus Annotation. In Burghardt, M., & Müller-Birn, C. (Eds.), *INF-DH 2018 – Workshopband, 25. Sept. 2018, Berlin.* https://doi.org/10.18420/infdh2018-17

Vertan, C. (2018b). Supporting hermeneutic interpretation of historical documents by computational methods. In Piotrowski, M. (Ed.), *Proceedings of the Workshop on Computational Methods in the Humanities 2018* (pp. 77–82).