MULTIPLE TOKENIZATIONS IN A DIACHRONIC CORPUS

THOMAS KRAUSE, ANKE LÜDELING, CAROLIN ODEBRECHT AND AMIR ZELDES
HUMBOLDT-UNIVERSITÄT ZU BERLIN

This paper deals with the construction of a maximally flexible corpus architecture for building and analyzing diachronic corpora. Historical data poses many challenges with regard to representation and analysis, and diachronic corpora are even more varied and unsystematic (Claridge, 2008). Since historical and diachronic corpora are so difficult and expensive to build, it is crucial that they be stored in an architecture that permits the addition of new texts and annotation layers at any point in time. In this paper we focus on two issues of corpus construction - multiple normalizations and multiple tokenizations in a multi-layer architecture. We exemplify our methodological issues using a diachronic corpus of German scientific texts (Ridges Herbology[1]). The corpus contains excerpts from texts about herbs from 12 different sources written between 1543 and 1870.

Multiple normalizations. In most corpus architectures it is necessary to specify one token layer that is the basis for all annotation layers[2]. This unique token layer uses a single concept of 'token' (this can in principle be anything but for modern European language corpora often a token is something like an orthographic word, or a character sequence between spaces, Schmid, 2008). This is problematic for corpora where not all texts follow the same regularities. Diachronic corpora are interesting in this respect because spelling has changed over time, especially with regard to the crucial features often used for tokenization (punctuation and insertion of spaces): Applying the same tokenization method across diachronic data (e.g. tokenization by spaces and punctuation marks) will not always result in the expected consistant output. Consider for example German particle verb constructions in (1) and (2). The same lemma is spelled in two different ways which would lead to different tokenizations:

(1) *[...] gleich als wenn ſie aus vielen kleinen Blåt-lein* ***zuſammen gefetzet*** *wären [...]*
'as if they would be composed of many little leaves'
(Curioser Botanicus oder sonderbares Kräuterbuch, 1675)

(2) *[...] indem die krautartigſten Ge-wächſe bisweilen blofs aus Mark , Fleiſch und Rinde* ***zuſammengefetzt*** *ſind .*
'as the herbaceous plants occasionally are composed of only pith, substance and bark'
(Grundriss der Kräuterkunde, 1792)

Some historical or diachronic corpora therefore add an annotation layer with a normalized variant of each token. But normalization - which involves abstraction and categorization - can be achieved in several ways, and each normalization layer (and each combination of normalization layers) can help answering different research questions. Consider the examples in Table 1, which show three different layers in the Ridges corpus. The column called `dipl` represents tokenization and character representation following the original spelling in the text, words separated by line breaks are separated here. In the second column (`clean`), only slight changes are made: characters like long s are mapped onto modern characters and clear compounds that were separated by line-breaks are combined here. This facilitates search. The combination of these two layers can be used to analyze the development of character use but other spelling changes cannot be found. The third column (`norm`) is a hyperlemmatization layer where each word is mapped onto its Modern German (MG) equivalent. This layer can be used to search for developments across all texts. In this layer we also use the MG spacing which leads to the fact that sometimes two (or more) original tokens are represented as one token here or that one token in the original text has to be represented by several tokens in the normalization layer. Note that all three layers have different token counts, which has interesting consequences for quantitative studies. Each

[1] **R**egister **i**n **D**iachronic **Ge**rman **S**cience. The corpus is freely available at `http://korpling.german.hu-berlin.de/ridges/index_en.html`

[2] These issues are sometimes discussed under the heading 'primary data' and 'secondary data' (Himmelmann, 2006). ISO uses the notion of equated 'primary data objects' (Ide & Romary, 2002)

dipl	clean	norm	pos	lemma		
ichs	*ichs*	*ich*	PPER	ich	'I'	1722
		es	PPER	es	'it'	
zuverſtehen	*zuverstehen*	*zu*	PTKZU	zu	'to'	1603
		verstehen	VVINF	verstehen	'understand'	
vnd	*vnd*	*und*	KON	und	'and'	1603
vñ	*vnd*	*und*	KON	und	'and'	1543
und	*und*	*und*	KON	und	'and'	1870
zuſammen gefetzet	*zusammen gesetzet*	*zusammengesetzt*	VVPP	zusammensetzen	'composed'	1675
zuſammengefetzt	*zusammengesetzt*	*zusammengesetzt*	VVPP	zusammensetzen	'composed'	1792
Pomeran= tzen=Schalen	*Pomerantzen=Schalen*	*Pomeranzenschalen*	NN	Apfelsinenschale	'orange peel'	1675

Table 1. Annotation layers in Ridges Herbology exemplified by single occurences.

normalization layer can have its own annotations. The use of an automatic pos-tagger is more suitable for the `norm`-layer than for other layers.

Tokenization. If it is useful to have several layers of normalization (there could be many more than the ones represented here) it is necessary to have a corpus architecture that can deal with this. By definition a token is the smallest unit of a corpus. Therefore, we call the original tokens 'segments' when they are combined into one corpus. The corpus architecture itself is agnostic to the chosen storage format as long as it is convertable to the meta model SALT (Zipser & Romary, 2010). Most standoff formats are able to represent multiple segmentations either as an explicit concept or by using more general span like structures (PAULA: Chiarcos, Ritz, & Stede, 2009, GraF: Ide & Suderman, 2007, MaF: ISO 24615:2010, 2008, SynAF: ISO/DIS 24611, 2010). Our implementation proposal is based on a concept of several segmentation paths, where each original token (e.g. `dipl`, `clean`) is part of exactly one path. Thus each original token has a precedence order which is only meaningful together with its specific segmentation path. None of the original token layers is minimal and thus we construct an artificial token layer as the most granular level. Every start and end of an original token in any of the segmentation layers leads to a split in the artificial token layer.

Search & Analysis. The multiple segmentations aligned on the artificial token layer enable us to deal with the examples (1) and (2) of particle verb constructions in several ways. Research questions concerning direct precedence can be investigated with the help of `norm` where all those constructions are normalized as one token (*zusammengesetzt*). If all occurrences of a certain particle verb construction should be found in a diachronic corpus, it is also useful to search on the basis of the `norm`-layer or on the lemma layer which is built on the basis of `norm` (*zusammengesetzt* / *zusammensetzen*): For investigating variants of particle verb constructions, the results can be matched either with the elements in `dipl`, if one is interested in the exact spellings, or with `clean` if it is necessary to match the special characters (*zusammen gesetzet*, *zusammengesetzt*). When using automatic taggers on `norm`, the adverb *zusammen* will not be separately tagged as ADV but as a component of the verb which might be useful to make a distinction between adverbial and particle verb constructions. These kinds of searches can be done in the corpus search tool ANNIS (Zeldes, Ritz, Lüdeling, & Chiarcos, 2009). As next step, we plan to support the explicit specification of a segmentation layer when using the precedence search operator or defining a search context. Coverage searches are possible using the artificial token layer.
Several independently tokenized but still aligned layers do not force a choice between conflicting interpretations of a diachronic text. Instead new annotations can be based on any tokenization whithout restrictions on possible research questions.

References

Chiarcos, C., Ritz, J., & Stede, M. (2009). By all these lovely tokens...: merging conflicting tokenizations. In *Proceedings of the third linguistic annotation workshop* (pp. 35–43).

Claridge, C. (2008). Corpus linguistics. In A. Lüdeling & M. Kytö (Eds.), (pp. 242–259). Berlin: Mouton de Gruyter.

Himmelmann, N. P. (2006). Essentials of language documentation. In J. Gippert, N. P. Himmelmann, & U. Mosel (Eds.), (pp. 1–30). Mouton de Gruyter.

Ide, N., & Romary, L. (2002, May). Standards for Language Resources. In *Third International Conference on Language Resources and Evaluation - LREC 2002* (p. 9 p). Las Palmas, Spain, France: none. Available from `http://hal.inria.fr/inria-00100771` (Colloque avec actes et comité de lecture. internationale.)

Ide, N., & Suderman, K. (2007). GrAF: a graph-based format for linguistic annotations. In *Proceedings of the linguistic annotation workshop* (pp. 1–8). Stroudsburg, PA, USA: Association for Computational Linguistics. Available from `http://dl.acm.org/citation.cfm?id=1642059.1642060`

Language resource management – morpho-syntactic annotation framework (Norm No. ISO 24615:2010). (2008). ISO, Geneva, Switzerland.

Language resource management – syntactic annotation framework (SynAF) (Norm No. ISO/DIS 24611). (2010). ISO, Geneva, Switzerland.

Schmid, H. (2008). Tokenizing and part-of-speech tagging. In A. Lüdeling & M. Kytö (Eds.), (pp. 527–551). Berlin: Mouton de Gruyter.

Zeldes, A., Ritz, J., Lüdeling, A., & Chiarcos, C. (2009). ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of corpus linguistics* (pp. 20–23).

Zipser, F., & Romary, L. (2010). A model oriented approach to the mapping of annotation formats using standards. In *Workshop on Language Resource and Language Technology Standards, LREC 2010*. La Valette, Malta. Available from `http://hal.inria.fr/inria-00527799`