Modelling Linguistic Data at the Boundary of "Document"

Martin Klotz

Research Unit Emerging Grammars – project Pd
Humboldt-Universität zu Berlin, corpus linguistics

martin.klotz@hu-berlin.de

Basic concepts and general problem description

Units

- in linguistic corpora and beyond we require a concept of ***text***, which we may understand as a sequence of primary linguistic items (e. g. "words")
- a related concept is ***document***, which for the sake of this presentation shall be reduced to the technical representation or container of a text
- a technical definition of ***corpus*** can then refer to a container of (at least one) document(s)
- the division of a text into smaller meaningful and annotatable units will be referred to as ***segmentation***, with ***segments*** being those units, the same text can have multiple distinct segmentations
- a linguistic ***annotation*** marks a property through a key and value on such a segment, a group of segments, a document, or a corpus
- annotations can also be a label on an explicit relation between two or more items of such types
- these concepts and definitions are and need to be challenged (Krause 2019; Odebrecht 2018; Stede 2018; Krause et al. 2012; Zipser et al. 2010)

General problem

There are well-established workflows and tools to annotate, computationally model, and compile linguistic data within document boundaries. Achieving the same between and across documents is currently much more challenging. If we want to compile a corpus to be represented in ANNIS (Krause et al. 2016b), as an example, we face:

- document-oriented processing (Zipser et al. 2010), search, and analysis
- overlap-based mapping of annotations to annotated elements

General workaround

For document-based environments, a merging process can combine texts from multiple documents in a new document.

Examples

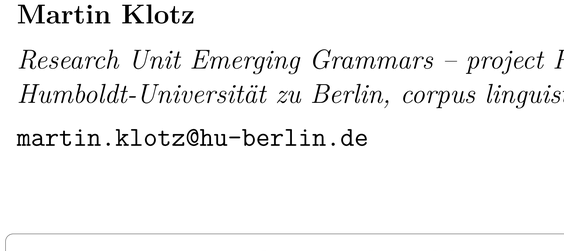

Figure 1: A prototype of a word-aligned parallel corpus of Old High German and Latin text, for previous versions see Donhauser et al. (2018).

Improving	the	Environment	,	Worker	Conditions	and	Communities	</h3>	<p>
VBG	DT	NN	,	NN	NNS	CC	NNS	--	--
Verbesserung	von	Umwelt	,	Arbeitsbedingungen	und	Lokaler	Gemeinden		
Ver|besser~ung	von	Um|welt	--	Arbeits#be|ding~ung	und	lokal	Gemeind~e		
FEM	--	FEM	--	FEM	--	--	FEM		
NN	APPR	NN	$,	NN	KON	ADJA	NN		
--	--	--	--	--	--	--	--		

Figure 2: The SMULTRON corpus contains aligned bilingual text (Volk et al. 2015), see also https://korpling.org/annis3/?id=dd10fe4-7c57-46b0-bed1-e2c83f86840b

Concrete problem

Problem description

This poster presents a solution to unify linguistically annotated documents in the **RUEG corpus** (Wiese et al. 2020). For a shared text segmentation, alternative text representations and annotations are distributed across multiple documents. This covers a special case of between-document annotation. Two types of documents **A** and **B** are given, featuring the following segmentations and annotations:

A segmentation of base text (`dipl`), a normalised segmented text (`norm`), morphological annotations mapped to normalised segments

B `dipl`, a segmentation into syllables (`syl`), prosodic annotations assigned to those syllables

As **output** we desire a single document (or corpus, respectively) holding all segmentations and annotations.

Challenge: Overlap-based mapping of morphological and prosodic annotations

The annotations of A and B are based on distinct segmentations of alternative texts to the same underlying base text. A mapping between the alternative texts' segments between A and B is unknown. Relying on overlap-based mappings of annotations to segments works due to the common base text, but introduces invalid mappings of annotations from one alternative text to the other (cf. figure 3).

`A:`	***morph***	m_1						m_2	m_3	
`A:`	**norm**	$n(d1, d2, d3)$						$n(d_4)_1$	$n(d_4)_2$	
`A∩B:`	**dipl**	d1		d2	d3			d4		
`B:`	**syl**	$s(d1)_1$	$s(d1)_2$	$s(d2)$	$s(d3)_1$	$s(d3)_2$	$s(d3)_3$	$s(d_4)_1$	$s(d_4)_2$	$s(d_4)_3$
`B:`	***pros***	…								

Figure 3: A visualisation of the two merged documents A and B in a single document. Mapping annotations by overlap leads to linguistically not motivated mappings of morphological annotations to syllables through the transitivity of the overlap relation.

A solution for the RUEG corpus

Parallel corpus approach: Instead of transferring all annotations to a common base segmentation, each segment from the common diplomatic segmentation in A is aligned with its corresponding segment from B; cf. figures 4 and 5.

`A:`	***morph***	m_1						m_2	m_3	
`A:`	**norm**	$n(d1, d2, d3)$						$n(d_4)_1$	$n(d_4)_2$	
`A:`	**dipl**	d1		d2	d3			d4		
		↑		↑	↑			↑		
`B:`	**dipl**	d1		d2	d3			d4		
`B:`	**syl**	$s(d1)_1$	$s(d1)_2$	$s(d2)$	$s(d3)_1$	$s(d3)_2$	$s(d3)_3$	$s(d_4)_1$	$s(d_4)_2$	$s(d_4)_3$
`B:`	***pros***	…								

Figure 4: Duplicating the base text and aligning the segments via alignment relations blocks the undesired transitive mapping of annotations to other segments.

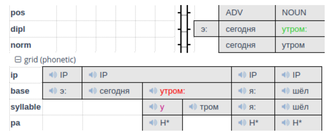

Figure 5: Merging prosodic and morphological annotations in one text allows to search for one linguistic feature in contexts defined by the other.

Summary

- the presented approach unifies documents without ill-representing their annotations
- the illustrated solution overcomes issues originating from overlap-based mappings of annotations to linguistic items
- this is a test case of ideas and concepts for a potential solution to current problems in modelling and representation, and
- a reliable solution for the RUEG corpus; it is transferable to similar, but not generally related problems

Limitations

- A general solution for obtaining between-document annotations and avoiding conflicts of overlap-based mappings of linguistic items and annotations is not provided by the presented approach (cf. Krause 2019; Krause et al. 2016a)

References

Donhauser, Karin, Jost Gippert, and Rosemarie (2018) Lühr (2018). *Deutsch Diachron Digital - Referenzkorpus Altdeutsch Version 1.1 (Version 1.1)*. Humboldt-Universität zu Berlin. URL: https://doi.org/10.34644/laudatio-dev-WiWkDnMB7CArCQ9CyBEw.

Krause, Thomas (2019). "ANNIS: A graph-based query system for deeply annotated text corpora". Doctoral Dissertation. Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät. DOI: 10.18452/19659. URL: https://doi.org/10.18452/19659.

Krause, Thomas, Ulf Leser, and Anke Lüdeling (2016a). "graphANNIS: A Fast Query Engine for Deeply Annotated Linguistic Corpora". In: *Journal for Language Technology and Computational Linguistics (JLCL)* 31.1, pp. 1–25.

Krause, Thomas and Amir Zeldes (2016b). "ANNIS3: A new architecture for generic corpus query and visualization". In: *Digital Scholarship in the Humanities* 31.1, pp. 118–139. ISSN: 2055-7671. DOI: 10.1093/llc/fqu057.

Krause, Thomas et al. (2012). "Multiple Tokenization in a Diachronic Corpus". In: *Exploring Ancient Languages through Corpora Conference (EALC)*. Universitetet i Oslo. URL: http://www.hf.uio.no/ifikk/english/research/projects/proiel/ealc/abstracts/Krause_et_al.pdf.

Odebrecht, Carolin (2018). "MKM – ein Metamodell für Korpusmetadaten". PhD thesis. Humboldt-Universität zu Berlin, Sprach- und literaturwissenschaftliche Fakultät. DOI: http://dx.doi.org/10.18452/19407.

Stede, Manfred (2018). *Korpusgestützte Textanalyse : Grundzüge der Ebenen-orientierten Textlinguistik / Manfred Stede*. ger. Tübingen.

Volk, Martin et al. (2015). *SMULTRON (version 4.0) — The Stockholm MULtilingual parallel TReebank*. An English-French-German-Quechua-Spanish-Swedish parallel treebank with sub-sentential alignments. URL: http://www.cl.uzh.ch/research/parallelcorpora/paralleltreebanks_en.html.

Wiese, Heike et al. (2020). *RUEG Corpus*. Version 0.3.0. Zenodo. DOI: 10.5281/zenodo.3765218. URL: https://doi.org/10.5281/zenodo.3765218.

Zipser, Florian and Laurent Romary (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. URL: https://hal.inria.fr/inria-00527799.