Multiple Tokenizations in a Diachronic Corpus
-
Corpus Demo Session Ridges Herbology

Thomas Krause, Anke Lüdeling, Carolin Odebrecht & Amir Zeldes

Corpus linguistic working group

Korpuslinguistik & Morphologie,
Humboldt-Universität zu Berlin

EALC Conference, 15th June 2012 Oslo

Outline

1. Project and Research Question

2. Linguistic Motivation
   Dealing with historical/diachronic texts
3. Examples leading to the Principle of Multiple Tokenizations
   Variance in historical/diachronic texts
4. Implementation
   Implementation and Visualization
5. Demo
   How does it work?

1. Project and Research Question

LAUDATIO (Long Term Access and Usage of Deeply Annotated Information)	SFB 632 Informationsstruktur D1 (Linguistic Database for Information Structure: Annotations and Retrieval)
Diachronic Corpus Ridges Herbology	Generic search tool for many kinds of corpora ANNIS
Texts from scientific register 1543-1870	Based on SQL Database
Multiple tokenizations for diachronic corpora allows the alignment of diplomatic transcripts with normalizations and a flexible application of further annotations on these layers.	

1. Project and Research Question

Important points for corpus linguistic research that we want to address in our projects:

- Access, usage and re-usage of primary data and annotations
- Open source character for data and software
- Transparency through detailed documentation

2. Linguistic Motivation

Dealing with historical/diachronic corpora…

...means preparing a maximally
flexible corpus architecture.

→This architecture needs to permit the addition of various texts and various annotations layers.

This architecture needs to capture transcriptions as well as normalizations, too.

→Above all, the architecture needs to be agnostic of all annotation layers, normalization and transcription guidelines.

3. Examples leading to the Principle of Multiple Tokenizations

Variability in historical/diachronic texts
…(see for example Claridge 2008).

→Orthography, separate spelling, special characters and special fonts occur in nearly every historical text.

→That is why we need normalizations to handle the variance.

→However, it is crucial to trace back the normalizations.

(1) *[…] gleich als wenn ſie aus vielen kleinen Blaͤt⸗lein* ***zuſammen geſetzet*** *waͤren […]*

’as if they were composed of many little leaves’

(Curioser Botanicus oder sonderbares Kräuterbuch, 1675)

(2) *[…] indem die krautartigſten Ge-wächſe bisweilen bloſs aus Mark , Fleiſch und Rinde* ***zuſammengeſetzt*** *ſind .*

’as the herbaceous plants occasionally are composed of only pith, flesh and bark’

(Grundriss der Kräuterkunde, 1792)

3. Examples leading to the Principle of Multiple Tokenizations

Our method of multiple tokenizations enables researchers to deal with all kinds of variation without loosing the retrieval for the transcripted data.

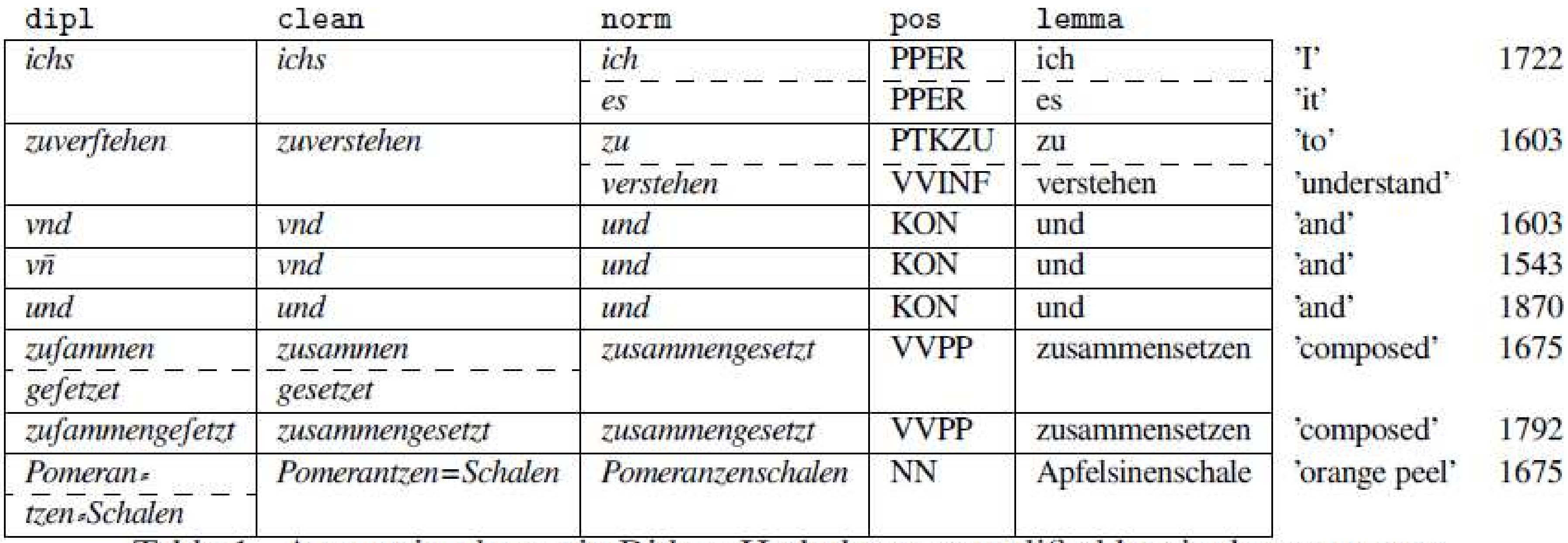
→We propose a step by step normalization whereby each step may get its own segmentation if necessary.

→Doing so, researchers are free to choose on which normalization layer a tool or the manual annotation should be applied.

→It is possible to investigate direct and indirect precedence, e.g. particle verb constructions, orthography, e.g. special characters, and graphical information, e.g. line breaks, independently.

3. Examples leading to the Principle of Multiple Tokenizations

Examples:



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
vn̄	*vnd*	*und*	KON	und	’and’	1543
und	*und*	*und*	KON	und	’and’	1870
zuſammen	*zusammen*	*zusammengesetzt*	VVPP	zusammensetzen	’composed’	1675
geſetzet	*gesetzet*					
zuſammengeſetzt	*zusammengesetzt*	*zusammengesetzt*	VVPP	zusammensetzen	’composed’	1792
Pomeran⸗	*Pomerantzen=Schalen*	*Pomeranzenschalen*	NN	Apfelsinenschale	’orange peel’	1675
tzen⸗Schalen						

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

Now, we need a way to implement and visualize the data…

4. Implementation

Implementation and Visualization…

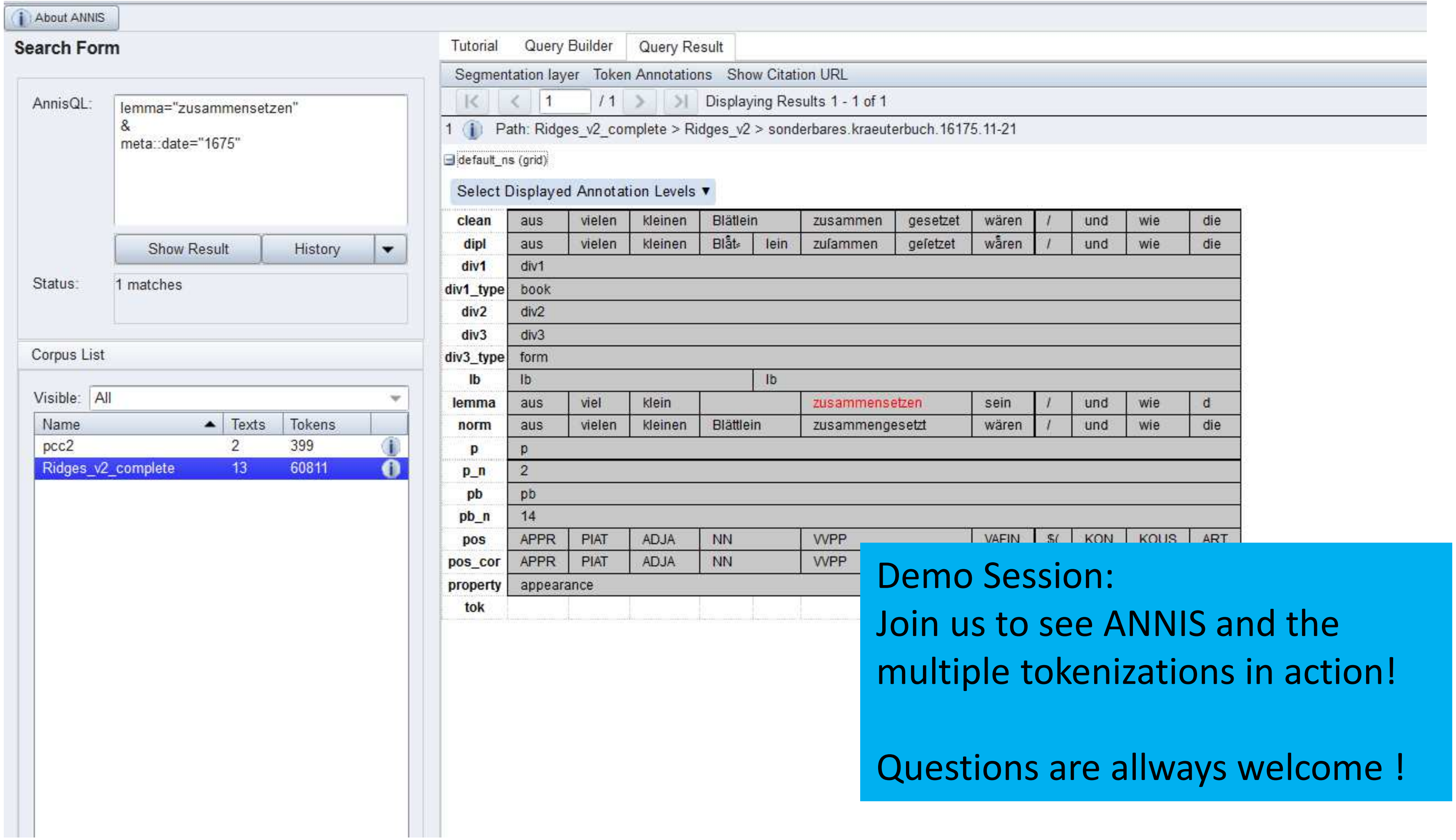
→Our corpus search tool ANNIS uses a relational database (Zeldes et al. 2009).

→An implementation was needed for

- Extend automatic generation from AQL (ANNIS Query Language) to SQL
- SALT Data Model and Pepper Converter Framework (Zipser & Romary 2010)
- Converter for extracting the segmentation from EXMARaLDA
- Modification of the search engine interface.

5. Demo

Multiple Tokenization in ANNIS



Search Form

AnnisQL:

```
lemma="zusammensetzen"
&
meta::date="1675"
```

Status: 1 matches

Corpus List

Visible: All

Name	Texts	Tokens
pcc2	2	399
Ridges_v2_complete	13	60811

1 / 1 Displaying Results 1 - 1 of 1

1 Path: Ridges_v2_complete > Ridges_v2 > sonderbares.kraeuterbuch.16175.11-21

default_ns (grid)

clean	aus	vielen	kleinen	Blätlein		zusammen	gesetzet	wären	/	und	wie	die
dipl	aus	vielen	kleinen	Blåt⸗	lein	zuſammen	geſetzet	wåren	/	und	wie	die
div1	div1											
div1_type	book											
div2	div2											
div3	div3											
div3_type	form											
lb	lb				lb							
lemma	aus	viel	klein			zusammensetzen		sein	/	und	wie	d
norm	aus	vielen	kleinen	Blättlein		zusammengesetzt		wären	/	und	wie	die
p	p											
p_n	2											
pb	pb											
pb_n	14											
pos	APPR	PIAT	ADJA	NN		VVPP		VAFIN	$(	KON	KOUS	ART
pos_cor	APPR	PIAT	ADJA	NN		VVPP						
property	appearance											
tok												

Demo Session:
Join us to see ANNIS and the multiple tokenizations in action!

Questions are allways welcome !

References

Claridge, C. (2008). Corpus linguistics. In A. Lüdeling & M. Kytö (Eds.). *Corpus Linguistics. An International Handbook*. Vol 1. (Reihe Handbücher zur Sprach- und Kommunikationswissenschaft). Berlin: Mouton de Gruyter (pp. 242–259). .

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.

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).

LAUDATIO: http://www.laudatio-repository.org/

SFB 632 D1: http://www.sfb632.uni-potsdam.de/~d1/

Ridges Herbology: http://korpling.german.hu-berlin.de/ridges/download_en.html

EXMARaLDA: http://www.exmaralda.org/index.html