Die Normalisierung von Komposita in frühneuhochdeutschen Texten am Beispiel des RIDGES-Korpus

Laura Perlitz, Gohar Schnelle, Anke Lüdeling, Carolin Odebrecht, Humboldt-Universität zu Berlin

39. Jahrestagung der DGfS, Saarbrücken, 08.03.-10.03.2017

1. Wie lassen sich Komposita im Fnhd. variationistisch untersuchen?

- Motivation: Komposition ist ein produktives Wortbildungsmittel mit einer Vielzahl an Bildungsmöglichkeiten
- man möchte deshalb systematisch suchen und analysieren
- Herausforderung: eine Variable mit vielen unvorhersagbaren historischen Varianten (graphische, morphologische, phonologische Realisierung)

Variable	**Kompositum**		
Varianten	phonologisch	graphematisch	morphologisch
Beispiel	Repräsentationen durchgeführter und nicht durchgeführter Lautwandel	Getrennt- und Zusammen-schreibung	Existenz von Fugenelementen
Historische Belege	*wieh=rauch* *Weyhrauch*	*Kræuter Saltz* *Kræutersaltz*	*Kolfewer* *kolfeuwer*

2. Kompositum vs. Syntagma im Fnhd.

- zweifelsfreie Disambiguierung aufgrund fehlender graphematischer Kennzeichnung nicht immer möglich, v.a. für NLP-Methoden besteht darin eine Herausforderung[8], z.B.:
- NPn mit Adjektivattributen:
  *auff das Stro das **leinen tu*ſ*chlein*** vs. ***Leinentu*ſ*chlein***
- NPn mit vorangestellten Genitivattributen:
  Teuffels Bissz vs. *Teufelsbissz*

→ Definition im Gegenwartsdeutschen eher graphematisch (Zusammenschreibung[10])

3. RIDGES-Korpus

- **Version 5.0[4]:**
- 36 Kräuterkundetexte vom 15.-20. Jh.
  - 183.724 Token auf der diplomatischen Primärtextebene
  - Version 6.0 in Arbeit
  - entstanden im Rahmen von Bachelor- und Masterseminaren an der Humboldt-Universität zu Berlin
  - variationistischer Forschungsansatz:
    konstanter Faktor: Register; variabler Faktor: Zeit
  - multiple Segmentierung auf verschiedenen linguistischen Ebenen
  - ausführliche Annotationsrichtlinien[1]

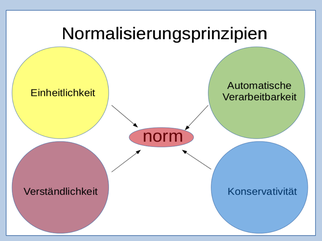
4. Normalisierung im RIDGES-Korpus

- **Primärtextebenen:**
  - **dipl:** diplomatische Transskription
  - **clean:** Normalisierung auf Zeichenebene
  - **norm:** Normalisierung auf verschiedenen linguistischen Ebenen

dipl	*den*	*zůuor*	*bereüchen*	*mit*	*wie=*	*rauch*
clean	*solt*	*zuuor*	*bereüchen*	*mit*	*wierauch*	
norm	*sollst*	*zuvor*	*beräuchern*	*mit*	*Weihrauch*	

- Normalisierung erfolgt manuell (bisher noch keine verlässlichen automatischen Verfahren zur Normalisierung frühneuhochdeutscher Texte entwickelt [2,8])
- nur auf den Ebenen Phonologie, Morphologie, Graphematik
- **nicht** lexikalisch und syntaktisch, da dies eine tiefere Interpretation erfordert

Diese Prinzipien werden bei der Normalisierung eingehalten:



5. Normalisierung fnhd. Komposita

	Phonologie	**Graphematik**	**Morphologie**
dipl	*wi**e**h=rauch*	*Kræuter Saltz*	*Kolfewer*
norm	*W**ei**hrauch*	*Kräutersalz*	*Kohl**en**feuer*

- phonologische und morphologische Normalisierung ermöglicht die automatische Verarbeitbarkeit, z.B. Lemmatisierung, POS-Tagging
- Zusammenschreibung für eindeutige Komposita
- Zweifelsfälle bleiben gemäß dem Konservativitätsprinzip getrennt geschrieben
- unbedingtes Miteinbeziehen der diplomatischen Transkription für eine Analyse des historischen Standes, da Normalisierungen immer Interpretationen der Annotatoren beinhalten

Fallbeispiel

dipl	wer	der	beyfufz	wurczeln	v̊ber	die	thor	des	hauſes	legt	oder	hencket
clean	wer	der	beyfusz	wurczeln	über	die	thor	des	hauses	legt	oder	hencket
norm	wer	der	Beifuß	Wurzeln	über	die	Tor	des	Hauses	legt	oder	hängt
lemma	wer	die	Beifuß	Wurzel	über	die	Tor	die	Haus	legen	oder	hängen
pos	PWS	ART	NN	NN	APPR	ART	NN	ART	NN	VVFIN	KON	VVFIN

Trefferreferenzlink: https://korpling.org/annis3/?id=e62923a0-96ff-46a2-aaa2-59c434ab09c7

Dieser Beleg ist auch im Kontext nicht disambiguierbar, daher greift das Prinzip der Konservativität.

6. Bisherige Forschung:

Perlitz, L. (2014): Konkurrenz zwischen Wortbildung und Syntax – historische Entwicklung von Benennung[7]

- **Forschungsfrage:**
  - "Haben Syntax und Wortbildung bei der Entwicklung von Fachtermini in der deutschen Wissenschaftssprache miteinander konkurriert?" (Fokus auf Komposita und NPn mit Genitivattribut)
- basierend auf RIDGES 4.1
- Erstellung zusätzlicher Annotationen
  - graphematische Annotationen (Zusammen-, Getrennt- und Bindestrichschreibung)
  - Annotation des Status potentieller Komposita auf der Skala "eindeutig – wahrscheinlich – zweifelhaft"

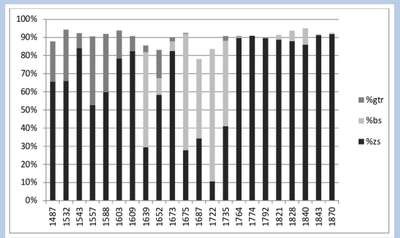


Abbildung: Anteil der Getrennt- ('gtr') , Zusammen- ('zs') und Bindestrichschreibung ('bs') an allen Komposita pro Erscheinungsjahr in % (abzüglich Komposita mit Zeilenumbruch).

In dieser Abbildung wird deutlich, dass in einem bestimmten Zeitraum aufgrund fehlender graphematische Kennzeichnung Komposita und konkurrierende Syntagmen nicht eindeutig disambiguierbar sind. Dies geht einher mit dem Vorkommen vorangestellter Genitivattribute im selben Zeitraum.

Referenzen:
[1] Belz, M.; Odebrecht, C.; Perlitz, L.; Schnelle, G. & Voigt, V. (2016): Annotationsrichtlinien zu Ridges Herbology Version 5.0, Humboldt-Universität zu Berlin. https://hu.berlin/ridges_annotationsrichtlinien_v5. ● [2] Bollmann, M. (2012): Automatic normalization for linguistic annotation of historical language data. Masterarbeit, Ruhr-Universität Bochum. http://publikationen.ub.uni-frankfurt.de/frontdoor/index/index/docId/31076. ● [3] Labov, William (2006): The social stratification of English in New York City. 2. Auflage. Cambridge; New York: Cambridge University Press. [4] Lüdeling, A.; Odebrecht, C.; Zeldes, A.: RIDGES-Herbology (Version 5.0), Humboldt-Universität zu Berlin. https://korpling.org/ridges. ● [5] Odebrecht, C., Belz, M., Zeldes, A., Lüdeling, A., Krause, T. (erscheint): RIDGES Herbology - Designing a Diachronic Multi-Layer Corpus. In: Language Resources and Evaluation. Malden, Oxford: Blackwell. ● [6] Pavlov, V. M. (1983): Zur Ausbildung der Norm der deutschen Literatursprache (1470-1730). Berlin: Akademie-Verlag. ● [7] Perlitz, L. (2014): Konkurrenz zwischen Wortbildung und Syntax - historische Entwicklung von Benennung. Bachelorarbeit, Humboldt-Universität zu Berlin. http://edoc.hu-berlin.de/master/perlitz-laura-2014-08-08/PDF/perlitz.pdf. ● [8] Piotrowski, Michael (2012): Synthesis Lectures on Human Language Technologies: Natural Language Processing for Historical Texts. Morgan & Claypool. ● [9] Wegera, K.-P.; Prell, H.-P. (2000): Wortbildung des Frühneuhochdeutschen. In: Sprachgeschichte 2.2. Berlin, New York: de Gruyter, 1594-1605.● [10] Wöllstein A., Eisenberg P. (2016): Duden - Die Grammatik: unentbehrlich für richtiges Deutsch. Berlin: Dudenverlag:722-723.