DFG-Projekt „Entwicklung der satzinternen Großschreibung im Deutschen“
Leitung:
Prof. Dr. Klaus-Michael Köpcke (Universität Münster), Prof. Dr. Renata Szczepaniak (Universität Hamburg)
Mitarbeiter:
Fabian Barteld, Marc Schutzeichel
Annotationsteam (Uni Hamburg):
Annemarie Bischoff, Lisa Dücker, Eleonore Schmitt, Annika Vieregge
Annotationsteam (Uni Münster):
Julia Hübner, Johanna Legrum, Katja Politt, Nicholas Wieling

Julia Hübner
Katja Politt
Marc Schutzeichel

Hexenverhörprotokolle: Erstellung eines historischen Korpus

Die Großschreibung in den Hexenverhörprotokollen

In dem DFG-Projekt „Die Entwicklung der **s**atz**i**nternen **G**roß**s**chreibung im Deutschen“ (SIGS) werden linguistische Fragestellungen aus dem Bereich der **historischen Graphematik** und der **kognitiven Linguistik** an die Textsorte „**Hexenverhörprotokolle**“ herangetragen. Die Grundannahme ist, dass sich die satzinterne Großschreibung in linguistisch beschreibbaren Stufen vollzogen hat. Am Ende dieses Prozesses steht dann eine konsequent syntaktisch motivierte Großschreibung.

Die für das Projekt relevanten kognitiv-semantischen (Referentialität und Belebtheit) und morphosyntaktischen Faktoren (PoS, Numerus) wurden bereits in einem Korpus aus Hexenverhörprotokollen durch Annotation abfragbar gemacht. Um den spezifischen Eigenschaften eines historischen Korpus gerecht zu werden, wird hier eine doppelte Tokenisierung, d.h. Unterscheidung zwischen graphischen und syntaktischen Token, durchgeführt. Das Korpus ist in ANNIS (ZELDES et al. 2009) überführt und so für die Untersuchung zur Verfügung.

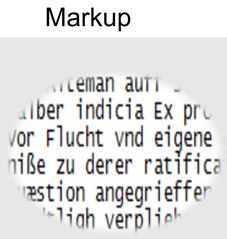
Das so aufbereitete Kernkorpus besteht aus 18 Protokollen vergleichbarer Länge, die gleichmäßig in Zeit (1588 - 1630) und Raum verteilt sind.

- 56 Protokolle aus ganz Deutschland (Macha et al. 2005)
- Kernkorpus: 18 Protokolle
- durchschnittliche Länge: ca. 1.400 Token
- Tokenanzahl gesamt: ca. 26.000

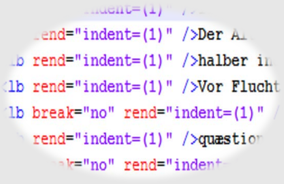
Ort	Region	Zeit
Jever	NW	1593
Meldorf	NW	1618
Alme	NW	1630
Perleberg	NO	1588
Güstrow	NO	1615
Stralsund	NO	1630
Hamm	MW	1592
Gaugrehweiler	MW	1610
Lemberg	MW	1630
Georgenthal	MO	1597
Rosenburg	MO	1618
Ostrau	MO	1628
Riedlingen	SW	1596
Günzburg	SW	1613
Baden-Baden	SW	1628
München	SO	1600
Schweinfurt	SO	1616
Bamberg	SO	1628

Datenaufbereitung

1. digitalisierte Editionstexte als .txt mit einfachem Markup



2. TEI XML-Version der digitalisierten Editionstexte



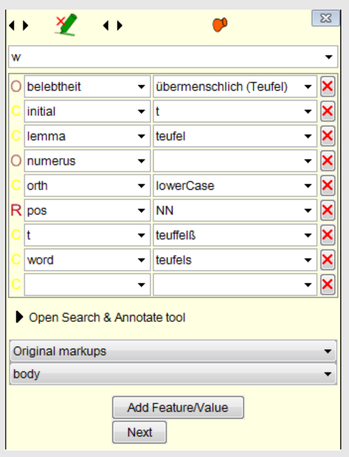
3. automatische Vorannotation (CAB)
   - normalisierte Wortform
   - Lemma
   - PoS
   - Sprache (de, lat)

4. Nachbearbeitung mit SIGS-Tool (VB.NET)
   - Säuberung
   - Initial, Großschreibung

5. Säuberung der CAB-Annotation, Anpassung an Projekt-Tagset

6. Entwicklung von Annotationsrichtlinien, manuelle Korrektur der Annotation in GATE, Fortlaufende Anpassung der Richtlinien (MAMA-Zyklus; s. Pustejovsky/Stubbs 2012)

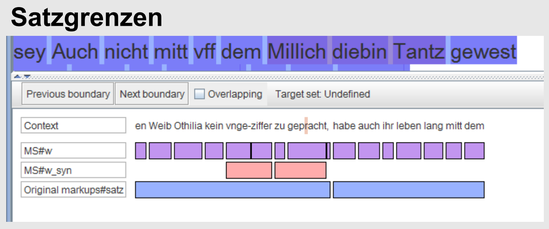
avffm	teuffelß	dantz	Text
avffm	teufel	tanz	**Lemma**
APPRART	NN	NN	**PoS**
	übermenschlich (Teufel)	abstrakt	**Belebtheit**
		sg	**Numerus**



Tokengrenzen

avffm		teuffelß	dantz	Text
avffm		teuffelß	dantz	**graphisches Token**
avff	m	teuffelßdantz		**syntaktisches Token**

Satzgrenzen



Export und Auswertung in ANNIS

Weitere Annotationsebenen, z.B. in Synpathy

Annotation

Neben der technischen Umsetzung der Annotationen ist es eine wichtige Aufgabe, die sprachlichen Besonderheiten der Textsorte zu erfassen. Existierende Annotationsrichtlinien, z.B. STTS (SCHILLER et al. 1999), HiTS (DIPPER et al. 2013) für die PoS-Annotation, decken nicht alle Aspekte ab und müssen daher angepasst bzw. erweitert werden.

Als Beispiel für Erweiterungen sei hier die Unterscheidung von Propria (NE, z.B. *Geörg*) und Appellativa (NN, z.B. *garten*) genannt. Da Berufsbezeichnungen (häufige Quelle für Eigennamen) nicht klar in eine der beiden Kategorien einordenbar sind, werden sie als getrennte Gruppe (Berufsbezeichnungen, NB, z.B. *schneider*) annotiert.

Ein Bereich, in der die Entwicklung eigener Annotationsrichtlinien notwendig war, ist die Einteilung des Textes in syntaktische Einheiten, da sowohl Interpunktion als auch Großschreibung keine zuverlässigen Indikatoren darstellen. Als Grundkategorie in der Annotation dient der Minimalsatz.

Besonderes Augenmerk liegt bei der Annotation auf dem transparenten Umgang mit strukturellen Ambiguitäten wie sie im Falle der Unterscheidung zwischen Komposition und freier syntaktischer Fügung mit Genitivattribut (*ins teuffels nahmen*) vorliegen können.

Fazit

Die Aufbereitung der Hexenverhörprotokolle als Korpus stellte das Projekt vor Herausforderungen: 1) Die für die Untersuchung der Entwicklung der Großschreibung relevanten Faktoren machten eine technisch komplexe Mehrebenenannotation notwendig. 2) Aufgrund der Besonderheiten der historischen Texte mussten die für Standardkorpora entwickelten Annotationsrichtlinien in vielen Fällen modifiziert werden.

Literatur

- DIPPER, S. et al. (2013): „HiTS: ein Tagset für historische Sprachstufen des Deutschen“. In JLCL 28, 85-137.
- MACHA, J. et al. (Hgg.) (2005): Deutsche Kanzleisprache in Hexenverhörprotokollen der Frühen Neuzeit. 2 Bde. Berlin/New York: de Gruyter.
- PUSTEJOVSKY, J. & STUBBS, A. 2012. Natural Language Annotation for Machine Learning. O'Reilly.
- SCHILLER, A., S. TEUFEL, C. STÖCKERT, Ch. THIELEN (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset).
- ZELDES, A., J. RITZ, A. LÜDELING, und Ch. CHIARCOS (2009): „ANNIS: A search tool for multi-layer annotated corpora“. In Proceedings of corpus linguistics, Bd. 9.

- ANNIS: http://annis-tools.org
- CAB: http://www.deutschestextarchiv.de/doku/software#cab
- GATE: http://www.gate.ac.uk
- Synpathy: http://www.mpi.nl/tools/synpathy.html

DFG
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG
wissen.leben
WWU Münster