UPB-Annotate: Ein maßgeschneidertes Toolkit für historische Texte

Seemann, Nina
nina.seemann@upb.de
Universität Paderborn, Deutschland

Merten, Marie-Luis
mlmerten@mail.upb.de
Universität Paderborn, Deutschland

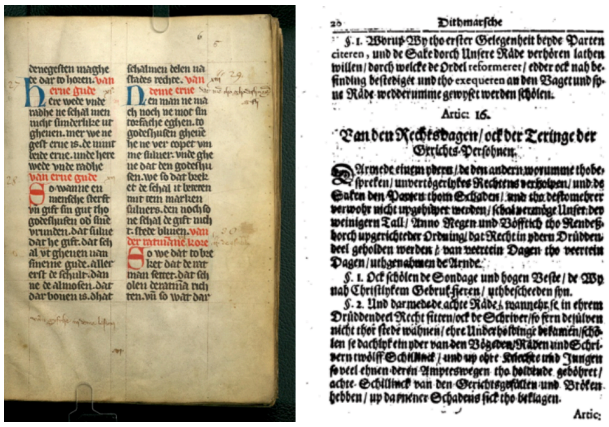
Einleitung

Das interdisziplinäre Projekt InterGramm geht dem literaten Ausbau des Mittelniederdeutschen bis hin zum Schriftsprachenwechsel zum Frühneuhochdeutschen nach. Im Zuge von Sprachausbauprozessen wandeln sich (häufig) bereits existierende Konstruktionen und neue literate Konstruktionen entstehen (Traugott / Trousdale 2013). Unser Ziel ist es, diese Wandelphänomene durch die Analyse historischer Rechtstexte und Arzneibücher zu erfassen. Insbesondere interessiert uns die Entwicklung komplexer Satztypen (subordinierende Konstruktionen), attributiver Techniken, die in komplexen Nominalphrasen resultieren, und textstrukturierender Elemente etc. Voraussetzung zur Erfassung dieser Elemente ist es, unser Korpus mit *Part-of-Speech*-Tags sowie mit Konstruktion-Tags (eine Art semantisch-syntaktische Annotation auf Phrasenebene) zu versehen. Naheliegenderweise steht für die linguistische Annotation von Daten bereits eine Vielzahl an Tools zur Verfügung, z.B. WebAnno (Yimam et al. 2014) oder CorA (Bollmann et al. 2014). Jedoch benötigen wir im Hinblick auf unsere graphematisch stark variierenden Texte und die herausfordernde Grammatikanalyse ein maßgeschneidertes Toolkit, das das Layout der historischen Texte 1:1 abbildet, das Editieren von Primärdaten erlaubt sowie morphologische und konstruktionale Annotationen unterstützt. Das im Projekt erstellte Korpus wird der Forschergemeinde voraussichtlich durch CLARIN-D zur Verfügung gestellt.

Vom historischen Dokument zum Toolkit

1.) Korpus und Transkription

Unser Korpus besteht aus Rechtstexten und Arzneibüchern aus dem 13. bis 17. Jahrhundert, aufgeteilt in (I.) eine Zusammenstellung mittelniederdeutscher Texte von 1227 bis 1650 (1,2 Mio. Token) und (II.) eine Dokumentensammlung frühneuhochdeutscher Texte, die im ehemals mittelniederdeutschen Sprachraum verfasst wurden (400.000 Token). Nach Möglichkeit versuchen wir, nur Primärquellen zur Transkription zu benutzen. In den meisten Editionen wurden bereits – von uns unerwünschte – Normalisierungen bezüglich des Layouts vorgenommen. Unsere Transkriptionen sollen diplomatisch sein, d.h. textstrukturierende Elemente des historischen Dokuments, etwa Rubrizierungen, Leerzeilen etc., sollen 1:1 in das Transkript übernommen werden. Diese Elemente liefern wichtige Informationen im Hinblick auf strukturelle Änderungen und entscheidende Hinweise zum historischen Gebrauch dieser Texte. In Abbildung 1 zeigen wir zwei Auszüge aus unseren Primärquellen. In der linken, älteren Quelle wurden zur Strukturierung des Textes Rubrizierungen und Majuskeln genutzt. Rubrizierte Textstellen übernehmen dabei u. a. die Funktion einer Überschrift, während Majuskeln den Anfang eines Paragraphen kennzeichnen. Die rechte, jüngere Quelle hingegen nimmt Zentrierungen, Einrückungen und Absätze zur Hilfe.



Links: Stadtrecht von Lübeck (1294); rechts: Landrecht von Dithmarschen (1667).

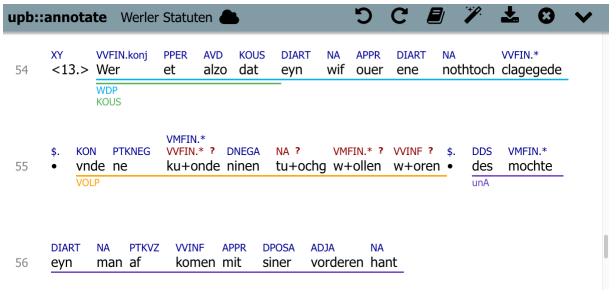
2.) *XML*-Format

Jedes Transkript wird in unser *XML*-Format transformiert, das aus den drei Hauptkomponenten (i) *metadata* , (ii) *layoutinfo* und (iii) *token* besteht. Im *metadata* werden der Name des Dokuments, Entstehungsort, Entstehungsjahr und Texttyp (Rechtstext oder Arzneibuch) gespeichert. Diese Informationen sind für geplante temporale und lokale Visualisierungen des Sprachausbaus nötig. In *layoutinfo* werden all die Informationen bezüglich des Layouts gespeichert, die für eine 1:1-Abbildung im Tool benötigt werden. Jedes Wort des Textes wird als ein *token*-Element gespeichert, welches wiederum aufgeteilt wird in (i) diplomatische Ebene und (ii) moderne Ebene. Eine Illustration und Erläuterung folgen am Ende von 3.).

3.) upb::annotate

Das graphische Nutzer- *Interface* stellt die *token* der *XML*-Datei entsprechend der *layoutinfo* dar, d.h. dem Layout der Primärquelle entsprechend. Dies hat den

Vorteil, dass die Annotatorinnen und Annotatoren sehr einfach erkennen können, was z.B. eine Überschrift ist oder ob am Zeilenende eine nicht-markierte Worttrennung stattfand. Weiterhin kommt dem Annotationsfluss zugute, dass die Daten in Leserichtung annotiert werden und der Kontext unmittelbar sichtbar ist. Das Toolkit ermöglicht, beide Annotationen (*POS*/Konstruktionen) im gleichen Nutzer-Interface auszuführen, der Wechsel zur anderen Ebene erfolgt durch Klick auf den entsprechenden Trigger. Mit Blick auf die Nutzerpraxis ist es einfacher, zunächst Phrasen mit Konstruktion-Tags auszuzeichnen und dann den der Phrase zugehörigen Token *POS*-Tags entsprechend ihres Kontextes zuzuweisen. Wir zeigen einen Screenshot in Abbildung 2.



POS-Tag-Annotation über den Token; Konstruktion-Tag-Annotation unter den Token.

Eine wichtige Funktion des Tools ist das Editieren von Token. Aufgrund der historischen Schreibvariation ist es teilweise nötig, zwei Token zusammenzuziehen oder ein Token zu trennen, um konsistente *POS*-Tag-Annotationen vornehmen zu können. Für beide Operationen wird in der zugrundeliegenden *XML*-Datei jeweils ein eindeutiger Marker gesetzt, der die manuellen Editierungen nachvollziehen lässt. So sagt uns die diplomatische Ebene für „in|dhere“ (*t290* in Abbildung 3), dass es im Originaldokument als ein Wort geschrieben wurde. Jedoch besteht es aus zwei Wörtern, die jeweils ein eigenes *POS*-Tag auf moderner Ebene erhalten. Für „screi#mannen“ (*t302* in Abbildung 3) sagt uns die diplomatische Ebene, dass es als zwei Wörter im Originaldokument geschrieben wurde. Jedoch ist es ein Wort mit einem *POS*-Tag auf moderner Ebene.

```
<token id="t290" trans="in|dhere">
 <dipl id="t290_d1" trans="in|dhere" utf="indhere"/>
 <mod id="t290_m1" trans="in|" utf="in">
  <pos tag="APPR"/>
 </mod>
 <mod id="t290_m2" trans="dhere" utf="dhere">
  <pos tag="DDART"/>
 </mod>
</token>

<token id="t302" trans="screi#mannen">
 <dipl id="t302_d1" trans="screi#" utf="screi"/>
 <dipl id="t302_d2" trans="mannen" utf="mannen"/>
 <mod id="t302_m1" trans="screi#mannen" utf="screimannen">
  <pos tag="NA"/>
 </mod>
</token>
```

t290: Manuelle Trennung durch den Marker ‚|’. t302: Manuelle Zusammenfügung durch den Marker ‚#’.

Posterpräsentation

Auf dem Poster präsentieren wir einen Überblick über unser Projekt und beleuchten den obigen Prozess genauer. Zudem werden wir auf unsere Tagsets eingehen sowie auf die Generierung automatischer Vorschläge für POS- und Konstruktion-Tags durch *Machine Learning* und *Pattern Matching*.

Bibliographie

Bollmann, Marcel / Petran, Florian / Dipper, Stefanie / Krasselt, Julia (2014): *“CorA: A web-based Annotation Tool for Historical and other non-standard Language Data”*, in: *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)* 86—90.

Traugott, Elizabeth / Trousdale, Graeme (2013): *Constructionalization and Constructional Change*. Oxford: Oxford University Press.

Yimam, Seid Muhie / Biemann, Chris / de Castilho, Richard / Gurevych, Iryna (2014): *“Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno“*, in: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* 91—96.