Webbasierte Morphemannotation Diachroner Korpora: Ein Weg zu mehr Nachhaltigkeit?

Peukert, Hagen
hagen.peukert@uni-hamburg.de
Universität Hamburg, Deutschland

Die Annreicherung historischer Texte mit derivationsmorphologischen Informationen ist aus der Sichtweise automatisierter Verfahren eine doppelte Herausforderung. Im Gegensatz dazu zeigt die automatische Erkennung von Flexionen bereits gute Ergebnisse (Dipper 2011, Bollmann et al 2014a,b). Die Herausforderungen lassen sich auf zwei wesentliche Unterschiede zurückführen. Erstens ist die Identifikation eines Derivationsmorphems aufgrund der vielzähligen Wortbildungsmechanismen und daraus folgender Überlappungsprobleme bei nicht agglutinierenden Sprachen algorithmisch nicht exakt zu bestimmen (vgl. Givón 1971, Dryer et al 2011, Lehmann 1973) und derzeit nur durch Abgleich mit einem a priori vorhandenen Lexikon überhaupt in Annäherung möglich. Zweitens ändert sich sowohl Form als auch Bedeutung eines Morphems über die Zeit hinweg, sodass sich daraus eine weitere Art der Überschneidung von Form sowie Inhalt (Bedeutung) einzelner Morpheme ergeben kann (vgl. Berg 1998, Faiß 1992, Kastovsky 2009). Vorausgesetzt man lässt eine Komplexitätsreduktion durch die Einführung von Zeitintervallen zu und vernachlässigt so die relativ langen Zeiträume, in denen sich Morpheme in einer Übergangsphase hin zu neuer Form und Inhalt befinden, folgt daraus immer noch, dass entsprechende Lexika für jede festgelegte Zeitperiode vorhanden sein müssen, um größere Textkorpora automatisch bearbeiten zu können. Je feinerkörniger die Zeitintervalle gewählt werden, desto größer wird die Anzahl an benötigten Lexika (proportional zur Anzahl der Zeitperioden). Feine Unterteilungen in den Zeitintervallen sind oft notwendig, um in der Folge die beobachteten Sprachwandelmechanismen genauer und ursächlich erklären zu können.

Die Lösung dieses Problems liegt demnach in der effizienten Erstellung einer entsprechenden Ressource, welche neben dem Lemma mit der Ausweisung der morphologischen Bestandteile auch die Zeit erfasst. Neben den morphologischen Informationen (Wurzel, Position und Anzahl von Präfixen und Suffixen) werden auch die Wortklasse und das Korpus erfasst. Bei Composita gehören zudem Kopf und semantische Kategorien (dvandva, bahuvrihi, appositional) zum Annotationsschema. Effizienzgewinne können dabei einerseits durch eine möglichst geschickte Aufteilung von standardisierbaren Routineaufgaben, welche Automaten abarbeiten können, und komplexeren Entscheidungsaufgaben, die ein Bearbeiter manuell treffen muss, erzielt werden. Andererseits kann dem Bearbeiter bei der Entscheidungsfindung mit der Bereitstellung von wichtigen Informationen und Komfortabilität bei der Bedienung und Präsentation geholfen werden.

Ein *Use Case* eines solchen Wortanalysewerkzeugs konnte mit dem Morphilo-Toolset als *Stand- A lone*-Anwendung ausprogrammiert werden. Diese Software berücksichtigt beim Abgleich großer Textkorpora mit dem Lexikon die Zeitspanne. Sind für das angegebene Zeitintervall Einträge vorhanden, werden diese automatisch zugewiesen. Die übrigen (unbekannten) Typen des Textes werden als neue Lemmata angelegt. Falls ein Lemma in der Vorgängerperiode bereits existiert, wird der aktuelle Eintrag mit den Informationen der Vorgängerperiode belegt und zur Bearbeitung präsentiert. Andernfalls (d.h. der Eintrag ist auch in keiner Vorgängerperiode registriert) wird das Token mit einer generischen Zerlegung automatisiert in seine morphemischen Bestandteile aufgeteilt. Der Nutzer bestätigt eine dieser Zerlegungen oder nimmt über entsprechende Menüs Änderungen vor. Erst jetzt werden diese Informationen persistent abgelegt (vgl. Peukert 2012).

Im so etablierten Workflow hat sich zunächst gezeigt, dass die Bearbeitung von englischen Texten aus dem 17. Jahrhundert (PPCMBE, Kroch et al 2010) schnell und effizient zu bewerkstelligen ist, wenn eine kritische Masse an Einträgen bereits vorhanden ist, da das TTR mit zunehmender Textgröße gegen Null strebt, d.h. immer nur wenige unbekannte Wörter in jedem neuen Text vorzufinden sind (Baayen 1996). Dieser Effekt trat bei der Bearbeitung von frühen mittelenglischen Texten aus dem 12. Jahrhundert (PPCME2, Kroch and Taylor 2000), nicht auf. Es zeigte sich, dass die fehlenden Schreibstandards von historischen Texten die notwendige Lemmatisierung scheitern ließen und somit auch ein Abgleich mit dem Lexikon nicht gelingen konnte (vgl. Peukert 2014).

Berücksichtigt man diese beiden Erfahrungen – schnelle Annotation bei kritscher Masse an Einträgen und langsame Annoation bei fehlenden Standards – bei der Entwicklung von Lösungsstrategien, trifft man unweigerlich auf den Nachhaltigkeitsgedanken beim Ressourcenaufbau, der vorgibt, dass die kostenintensiven Annotationsaufgaben möglichst nicht mehrfach erledigt, aber nachgenutzt werden sollen. Dies impliziert eine gemeinschaftlich-synergetische Bearbeitung der Annotationszuweisung, da man die spätere Nutzung der Resource mit eigener (sehr geringer) Annotationsarbeit “bezahlen” kann. Auf diese Weise können annotierte Daten unterschiedlicher Zeiträume gesammelt werden. In der Fortführung dieser Idee ist die Architektur einer webbasierten Komponente entstanden (Abb. 1), bei der ein *Multi-User*-Design die Annotationsarbeit an unterschiedlichen Korpora

verteilt und Zuweisungen aus verschiedenen Lexika aber der passenden Zeitperiode und Sprache erlaubt. Um dem Problem der fraglichen Qualität der Annotationen entgegenzuwirken, ist es möglich, die Lexika, die man zur Bearbeitung benötigt, auszuwählen. Möchte man sein eigenes Korpus mit derivationsmorphologischen Informationen anreichern lassen, fliesst die jeweils eigens geleistete Annotationsarbeit in den Gesamtdatenbestand ein. Inwiefern die gesammelten Annotationsdaten mit weiteren Verfahren hinsichtlich ihrer Qualität getestet, bewertet und weiter bearbeitet werden können, wird Gegenstand einer weitergehenden Diskussion sein.

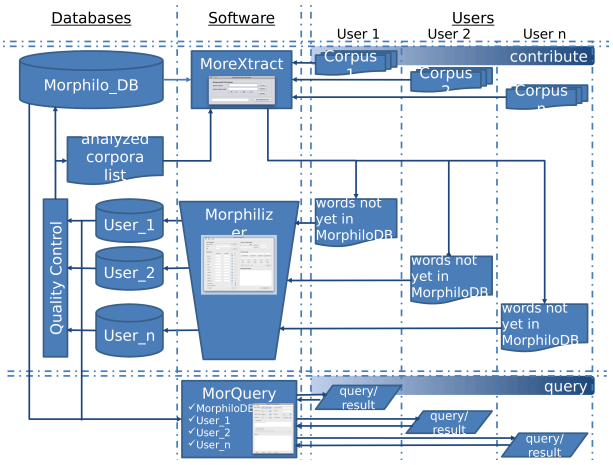

Abb. 1: Architekturentwurf zur Integration des Annotationswerkzeugs in eine webbasierte Anwendung im Mehrnutzerbetrieb

Für die Lösung der noch nicht endgültig fertiggestellten Komponente (in Abb. 1 mit "Quality Control" bezeichnet) werden statistische Verfahren in Anlehnung an das maschinelle Lernen vorgestellt, die sich an zwei unterschiedlichen Strategien ausrichten. Erstens steht die häufigkeitsbedingte Analyse gleicher oder ähnlicher Einträge der verschiedenen Datenbestände der Nutzer (User_1, …, User_n) im Vordergrund. Diese Daten werden genutzt, um Ausreißer und falsche Annotationen mittels automatisierter statischer Signifikanztests zu identifizieren. Dieser Ansatz wird mit einer nutzerorientierten Strategie kontrastiert. Diese zweite Strategie bezieht die Verhaltensdaten der Nutzer ein, d.h. wie oft werden welche Datenbestände anderer Nutzer für die anstehende Annotation ausgewählt. Auch hier basiert der Ausschluss von vermeindlich fehlerhaften Daten mittels eines vorher festgelegten Signifikanzniveaus. Die mit einer dieser Strategie bereinigten Datenbestände könnten danach in den Hauptdatenbestand (Morphilo_DB in Abb. 1) überführt werden.

Bibliographie

Baayen, Harald (1996): „The effects of lexical specialization on the growth curve of the vocabulary", in: *Computational Linguistics* 22: 455–480.

Berg, Thomas (2009): *Structure in language: A dynamic perspective*. New York: Routledge.

Bollmann, Marcel / Petran, Florian / Dipper, Stefanie / Krasselt, Julia (2014a): „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.

Bollmann, Marcel / Petran, Florian / Dipper, Stefanie (2014b): „Applying Rule-Based Normalization to Different Types of Historical Texts — An Evaluation", in: Zygmunt Vetulani and Joseph Mariani (eds.): *Human Language Technology Challenges for Computer Science and Linguistics. 5th Language and Technology Conference*, LTC 2011. Revised Selected Papers. Lecture Notes in Computer Science 8387. Springer 166–177.

Dipper, Stefanie (2011): „Morphological and Part-of-Speech Tagging of Historical Language Data: A Comparison", in: *Journal for Language Technology and Computational Linguistics*. Special Issue 26 (2): 25–37.

Dryer, Matthew S. / Haspelmath, Martin (eds.). (2011): *The world atlas of language structures Online*. München: Max Planck Digital Library.

Faiß, Klaus (1992): *English historical morphology and word-formation: Loss versus enrichment*. Trier: Wissenschaftlicher Verlag.

Givón, Talmy (1971): „Historical syntax and synchronic morphology: An archaeologist's field trip", in: *Chicago Linguistic Society* 7: 394–415.

Kastovsky, Dieter (2009): „Diachronic perspectives", in: Lieber, Rochelle / Štekauer, Pavol (eds.): *The Oxford handbook of compounding*. Oxford: Oxford University Press 321–340.

Kroch, Anthony / Santorini, Beatrice / Diertani, Ariel (2010): *The Penn-Helsinki Parsed Corpus of Modern British English (PPCMBE)*. Department of Linguistics, University of Pennsylvania: CD-ROM, first edition http://www.ling.upenn.edu/hist-corpora/ .

Kroch, Anthony / Taylor, Ann (2000): *The Penn-Helsinki Parsed Corpus of Middle English (PPCME)*. Department of Linguistics, University of Pennsylvania: CD-ROM, first edition http://www.ling.upenn.edu/hist-corpora/ .

Lehmann, Winfred P. (1973): „A structural principle of language and its implications", in: *Language* 49 (1): 47–66.

Peukert, Hagen (2014): „The Morphilo Toolset: Handling the Diversity of English Historical Texts", in: Ammermann, Anne / Brock, Alexander / Pflaeging, Jana / Schildhauer, Peter (eds.): *Facets of Linguistics: Proceedings of the 14th Norddeutsches Linguistisches Kolloquium 2013*. Frankfurt: Peter Lang 161–172.

Peukert, Hagen (2012): „From Semi-Automatic to Automatic Affix Extraction in Middle English Corpora: Building a Sustainable Database for Analyzing Derivational Morphology over Time“, in: Jancsary, Jeremy (ed.): *Empirical Methods in Natural Language Processing, Wien, Scientific series of the ÖGAI* 413–23.