Integrierte Lexikographische Dienste zur Unterstützung der digitalen Geisteswissenschaften (aiLEs)

Karlheinz Mörth, Hannes Pirker

Austrian Center for Digital Humanities (ACDH) der Österreichischen Akademie der Wissenschaften

Digitale Wörterbücher stellen ein wichtiges Hilfsmittel für die geistes- und sozialwissenschaftliche Forschung dar, denn sie bieten nicht nur Unterstützung bei der schriftlichen Formulierung von Forschungsresultaten, sondern werden insbesondere als grundlegende Informationsquellen für die automatische Annotation und Analyse unterschiedlicher Textdaten benötigt. Indem sie als unabdingbare Basis verschiedener Verfahren der automatischen Textanalyse fungieren, bilden sie die Grundlage für die Erschließung textueller Informationsquellen, einem der zentralen Beiträge der *Digital Humanities* im gesamtwissenschaftlichen und -gesellschaftlichen Kontext.

Während in der Lexikographie, von der hier die Rede sein soll, digitale Methoden seit langer Zeit Anwendung finden, stehen frei verfügbare und qualitativ hochwertige Lexika nach wie vor nur in sehr begrenztem Maß zur Verfügung. Es werden immer mehr digitale Sprachdaten verfügbar, vertrauenswürdige lexikographische Daten, die von digital arbeitenden Linguisten und Lexikographen für ihre Forschungen verwendet werden können, gibt es nach wie vor kaum. Der Zugang zu existierenden Lexika wird sowohl durch technische als auch durch kommerzielle Hürden behindert. Der größte Teil derartiger Materialien wurde unter kommerziellen Gesichtspunkten erzeugt, und ist, wenn im Netz verfügbar, nur über Schnittstellen zugänglich, die es nicht erlauben, direkt mit diesen Daten zu arbeiten. Selbst für gut dokumentierte und digital gut erschlossene Sprachen wie Deutsch, Französisch, Spanisch usw. steht es um die Verfügbarkeit von lexikalischen Ressourcen nicht gut.

In diesem Beitrag wird die Erstellung einer neuen lexikalischen Datensammlung diskutiert, des Weiteren werden die organisatorischen und technischen Strategien aufgezeigt, durch die die so entstandenen Daten dauerhaft verfügbar gemacht werden.

AiLEs – *Austrian Integrated Lexicographic System*

aiLEs steht für *Austrian Integrated Lexicographic System* und ist ein Versuch, die eingangs beschriebene Situation zumindest ansatzweise zu verbessern. Im Rahmen von *aiLEs* sollen lexikalische Daten zur Verfügung gestellt werden (*xBaffle*). Die *technische* Verfügbarkeit der Lexika wird durch die Verwendung standardisierter XML-Formate für die Datenrepräsentation garantiert, der programmatische Zugriff auf die Daten durch die Bereitstellung von REST basierten Schnittstellen ermöglicht werden. Auf organisatorischer Ebene sollen die Daten gemäß der *open access* Philosophie barrierefrei zugänglich sein. Durch die Einbettung des Projekts in den Kontext der europäischen Infrastrukturkonsortien CLARIN und DARIAH soll sowohl die *technische* Langzeitstabilität – konkret durch die Bereitstellung der Ergebnisse im *Language Resources Portal* des CLARIN-Zentrums Wien[1] – als auch die nicht minder wichtige *institutionelle* Stabilität des Unterfangens garantiert werden.

Im Rahmen von *aiLes* ist mit *xBaffle* eine Reihe lexikalischer Ressourcen für unterschiedliche Sprachen konzeptioniert, wobei *Baffle* für *Basic Austrian Fullform Lexicon* steht, und das *x* im Namen jeweils durch den entsprechenden zweibuchstabigen Sprachidentifikator zu er-

1 http://www.oeaw.ac.at/icltt/ccv

setzen ist. Das erste digitale Wörterbuch, das aus diesem Projekt hervorgehen soll, ist *deBaffle*, ein deutsches Vollformenlexikon mit morphologischen Basisdaten, das eine möglichst gute Abdeckung der deutschen Gegenwartssprache zum Ziel hat.

Es wurden für das Deutsche in der Vergangenheit bereits umfangreiche morphologische Datenbestände (Morphy, Canoo) aufgebaut. Die meisten dieser Ressourcen sind jedoch nicht zugänglich, unvollständig oder nur gegen teures Geld verfügbar. Die Dokumentation der für die Erzeugung von *deBaffle* angewendeten Methoden und Werkzeuge soll es ermöglichen, in der Zukunft auch über das Deutsche hinaus ähnliche Sprachressourcen aufzubauen.

Lexikalischer Ausgangspunkt: Wiktionary

Als Ausgangspunkt für *deBaffle* wurde das deutschsprachigen Wiktionary[2] gewählt. Wiktionary ist ein Schwesterprojekt zur freien Enzyklopädie Wikipedia und arbeitet an Wörterbüchern, die wie die Wikipedia selbst, in kollaborativer Art und Weise manuell von lexikographischen Enthusiasten editiert werden. Die deutschsprachige Version ist im Vergleich zu anderen Wiktionaries verhältnismäßig umfangreich. Die Wörterbucheinträge verfügen oft über bemerkenswert umfangreiche linguistische Informationen. Neben Angaben zur Orthographie finden sich auch Daten zur Flexion, Morphologie, Phonologie und Semantik. Trotz zahlreicher Lücken im Material stellte sich das Material für unsere Zwecke als erstaunlich brauchbar heraus.

Eine Hürde für die Verwendung von Wiktionary-Daten in automatisierten Sprachverarbeitungsprozessen stellt das allen Wikiprodukten zugrunde liegende Repräsentationsformat dar. Hierbei handelt es sich um eine sogenannte wiki Sprache, die je nach Domäne und natürlichsprachlichem Kontext unterschiedliche Vokabulare verwendet und deren Syntax den Einsatz von Standardtools erschwert, in der Regel unmöglich macht. Um zeitgemäße Verfahren des Datenmanagements, wie z. B. automatisches Validieren zu erlauben, ist es nötig, derartige Daten in ein XML Format zu konvertieren. Aus einer Reihe von hierfür in Frage kommenden Zielformaten (LMF, RDF, …) wurde für das konkrete Projekt TEI P5[3] gewählt, da die gesamte verwendete Infrastruktur auf dieses Format zugeschnitten ist und langjährige Expertise mit der Applikation dieses Formats vorliegt (vgl. Budin et al. 2013).

Korpusbasierte Lexikonkonstruktion

Bei der Erstellung von *deBaffle* kommt eine korpusbasierte inkrementelle Methode zur Anwendung. Die Grundidee dabei ist, Informationen aus bereits verfügbaren lexikalischen Ressourcen – in diesem Fall dem Wiktionary – mit Daten aus möglichst umfangreichen Textkorpora zu verknüpfen.

Der Abgleich der Lexikoninhalte mit der tatsächlichen Sprachrealität in Texten bietet die Möglichkeit, den bestehenden Abdeckungsgrad durch das Lexikon laufend zu evaluieren, und so den Prozess der Lexikonerweiterung durch die Identifikation systematischer Lücken zu optimieren. Insbesondere kann aber das Korpus helfen, sowohl die Korrektheit bestehender lexikalischer Informationen zu überprüfen, als auch Informationen über noch unbekannte Wortformen abzuleiten.

deBaffle ist als Vollformlexikon konzipiert, d. h., alle Flexionsformen eines Lemmas sind explizit im Lexikon aufgeführt. Dieses Lexikondesign wird in der automatischen Sprachverarbeitung bevorzugt verwendet, weil es erlaubt, die morphosyntaktische Kategorie einer Wortform durch einfaches Nachschlagen im Lexikon zu ermitteln, und somit auf eine morphologische Analysekomponente zu verzichten. Eine zentrale Aufgabe bei der automatischen

2 http://de.wiktionary.org/

3 http://www.tei-c.org/release/doc/tei-p5-doc/en/html/

Erstellung eines neuen Lexikoneintrags für *deBaffle* ist somit die Festlegung der Wortklasse und des Flexionsparadigmas eines noch unbekannten Wortes, um in der Folge alle Flexionsformen zu einem Lemma zu generieren.

Die korpusbasierte Methodik in *deBaffle* soll anhand der Substantivflexion beispielhaft demonstriert werden. Zur Ermittlung des Flexionsparadigmas bei Substantiven muss immer zumindest das grammatische Geschlecht und eine Pluralform bekannt sein.

Unser Ansatz nutzt nun den Umstand, dass bestimmte morphosyntaktische Muster den Kasus eines Substantivs determinieren können. Beispielsweise kann aus dem Auftreten eines Musters wie „(eines|des) *Subst*“ im Korpus abgeleitet werden, dass das Substantiv männliches oder sächliches Geschlecht aufweist und im Genitiv steht. Ein Muster wie „(viele|einige) *Subst*“ identifiziert Substantive im Plural usw. Durch die inkrementelle Anwendung und schrittweise Verfeinerung solcher Regeln können Hypothesen über die Wortart und das Flexionsparadigma eines Wortes gebildet werden. So können Informationen für neue Lexikoneinträge automatisch identifiziert, aber auch bestehenden Einträge verifiziert werden.

Korpusdaten: Austrian Media Corpus (AMC)

Als Basis für die korpusbasierte Methodik von *deBaffle* dient das Austrian Media Corpus (AMC), eine von der Austria Presse Agentur (APA) kompilierte Sammlung von Texten aus österreichischen Zeitungen und Magazinen der letzten 20 Jahre (Ransmayr et al. 2013). Mit einem Umfang von derzeit ca. 6 Mrd. Wörtern und seinem kontrollierten Genre-Repertoir bietet das Korpus einen repräsentativen Querschnitt durch den aktuellen schriftlichen Sprachgebrauch in Österreichs Medienlandschaft. Das Korpus wurde mit zwei unterschiedlichen *Part-of-Speech (PoS) Taggern* (RFTagger und TreeTagger) mit Wortarteninformation und auch Flexionsinformation angereichert. Diese automatisch generierten Informationen sind zwar naturgemäß fehlerbehaftet, sollen aber dennoch im oben beschriebenen iterativen Verarbeitungszirkel miteinbezogen werden. Durch Abgleich der Annotationsentscheidungen der beiden PoS-Tagger untereinander, mit Informationen aus dem Wiktionary und den Ergebnissen der morphosyntaktischen Testmuster sollen die verschiedenen Informationsquellen evaluiert, und in der Folge verbessert werden. Das annotierte Korpus bietet also eine Grundlage zur Konstruktion des Lexikons, ist aber gleichzeitig über mittelfristig verbesserte Annotationen wiederum Nutznießer dieses Prozesses.

Statistische Informationen für das Lexikon

Durch den Abgleich zwischen lexikalischen Einträgen und Textkorpus liegen statistische Informationen zur Auftretenshäufigkeit einer Wortform oder eines Lemmas vor. Diese Daten werden in *deBaffle* zur Verfügung gestellt, und dienen sowohl den Lexikographen bei der Erstellung und Pflege des Lexikons, als auch den künftigen Anwendern der Lexika.

Durch den Abgleich von Lexikon und Korpus lassen sich etwa die quantitativ vordringlichsten Lücken in der lexikalischen Abdeckung identifizieren, die freilich immer auch einen korpusspezifisch *bias* aufweisen. So zählen beispielsweise im AMC die Akronyme österreichischer politischer Parteien zu den häufigsten Wörtern, die von den PoS-Taggern nicht identifiziert werden konnten. Auffällige Lücken finden sich auch bei – insbesondere österreichischen – Toponymen. Hier zeigen sich systematische Verzerrungen, verursacht durch die unterschiedliche geografische Herkunft der lexikalischen Ressourcen einerseits und des AMC andererseits. Das Beispiel der mangelhaften Abdeckung der Toponyme legt die Integration zusätzlicher externer Wissensquellen – etwa Ontologien mit Geodaten – zur Ergänzung der Lexika als sinnvollen zusätzlichen Schritt nahe.

Manuelle Verifikation

Alle automatischen Schritte produzieren auch fehlerhafte Daten, die schlussendlich nur durch manuelle Verifikation eliminiert werden können. Für diesen notwendigen Bearbeitungsschritt steht mit dem am ICLTT entwickelten Viennese Lexicographic Editor (VLE)[4] zumindest ein effizientes Werkzeug zur Verfügung, das für die Produktion von TEI-konformen Lexikoneinträgen optimiert wurde (Budin et al. 2013).

Status und Ausblick

Das gegenwärtig im Aufbau begriffene digitale Wörterbuch gehört zu einer Reihe von Tools die im Rahmen des Österreichischen Zentrums für digitale Geisteswissenschaften (ACDH) als Ergänzung des bestehende Toolinventars geplant ist. Es soll in Zukunft als Teil des österreichischen Engagements in den europäischen Infrastrukturkonsortien CLARIN und DARIAH weitergepflegt werden.

Im Moment enthält Baffle nur deutschsprachige Daten. Es wird am ACDH aber bereits an ähnlichen Sprachressourcen für andere Sprachen gearbeitet.

Referenzen

Budin, G., Moerth, K., Ďurčo, M. (2013). European Lexicography Infrastructure Components. In: *Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estland.*

Ransmayr, J., Moerth, K., Ďurčo, M. (2013). Linguistic variation in the Austrian Media Corpus. Dealing with the challenges of large amounts of data. In: *Proceedings of International Conference on Corpus Linguistics (CILC), Alicante, Spanien.*

4 https://clarin.oeaw.ac.at/ccv/vle