Adnominale Possession in einem Bibel-Parallelkorpus

Fleischmann, Florian
Florian.Fleischmann@itg.uni-muenchen.de
LMU München, Germany

Empirische Sprachwissenschaft erarbeitet Erkenntnisse auf Basis sprachlichen Original-Materials. Für Untersuchungen über Unterschiede zwischen verschiedenen diachronen Stufen oder dialektalen Varietäten einer Sprache muss sprachliches Material aus diversen Quellen miteinander verglichen werden. Textsorte, Inhalt und Art der Quellen addieren hierbei weitere Variablen zur Untersuchung. Eine Möglichkeit, diesem Umstand zu begegnen, bieten Parallelkorpora. Sie haben den Vorteil, dass hier nur ein einziger Text in verschiedenen Sprachstufen bzw. -varietäten verglichen wird. Varianz in der Textsorte scheidet so als Erklärung für etwaige beobachtete linguistische Unterschiede aus. Um ein umfassendes Parallelkorpus zu erstellen, muss der untersuchte Text in möglichst vielen diachronen und/oder dialektalen Fassungen vorliegen. Einen Text, für den dies zutrifft, stellt die Bibel dar. Bibelübersetzungen ins Deutsche existieren bereits seit dem Althochdeutschen (z.B. Evangelienharmonie des Tatian, ca. 830 (Sievers 1872)) und sind lückenlos bis zum Neuhochdeutschen überliefert. Auch dialektal besitzen sie synchron eine weite Verbreitung. Es existieren bspw. (Teil-)Übersetzungen ins Hessische (Mieth 2011), Kärntnerische (Bünker 2007), Niederdeutsche (Jessen 1984), Pennsylvania Dutch (o. A. 2002), Schwäbische (Paul 1997), Walliserdeutsche (Theler 2011) oder Zürichdeutsche (Weber 2011). Ein weiterer Vorteil der Bibel liegt darin, dass hier eine möglichst textgetreue Wiedergabe im Sinne des Übersetzers liegt. Anpassungen an die individuelle Varietät erfolgen möglichst behutsam und konservativ, insbesondere was Syntax und Lexik angehen. Das Erstellen eines derartigen Parallelkorpus ist Kern meiner Promotionsschrift.

Besonderes Augenmerk liegt dabei auf einer nachhaltigen Aufbereitung der zugrundeliegenden Daten im Sinne der FAIR-Prinzipien. In einem ersten Schritt müssen die Bibeln digitalisiert werden. In vielen Fällen ist dies bereits durch Bibliotheken geschehen, meistens jedoch nur als Bild-PDFs. Zur weiteren Verarbeitung werden diese mittels automatischer Texterkennung (OCR) in maschinenlesbare Form gebracht. Bei älteren Texten stellt dies aufgrund der verwendeten gebrochenen Schriftarten (wie Fraktur) ein Problem dar. Vortrainierte Texterkennungsmodelle versagen hier oft (Baierer 2020, Springmann 2017: 3), so dass selbstständig neuronale Netze zur Erkennung trainiert werden müssen. Mit OCR4all (https://www.ocr4all.org) steht ein leistungsfähiges Software-Paket hierfür zur Verfügung. OCR4all basiert auf dem OCRopus-Derivat Calamari, das im Vergleich zu anderen OCR-Lösungen wie OCRopy. Tesseract oder OCRopus die besten Erkennungsraten zeigt (Wick 2020). OCR4all vereint weiterhin die Schritte von Pre-Processing, Character Recognition und Post-Processing in einem Tool und sorgt so für einen effizienten Arbeitsablauf. Bei Pilotversuchen ließen sich für ausgewählte historische Bibeln Erkennungsraten von 98 % und mehr erzielen. Mit OCR-D ist eine vergleichbare Lösung in Entwicklung, zumindest momentan liegen dessen Erkennungsraten jedoch noch niedriger (Baierer 2020).

Die Ausgabe liegt zunächst als reiner Text vor. Diese werden in eine SQL-Datenbank importiert und nach ihren Bibelversen annotiert. Auf diese Weise lässt sich in einem späteren Schritt eine einfache Weboberfläche entwerfen, die eine alignierte Darstellung von Bibelversen zulässt. Dadurch können auch Forschungszweige außerhalb der Sprachwissenschaften (im Rahmen von UrhG § 60d) auf das Parallelkorpus zugreifen. Im Kern soll das Parallelkorpus als Grundlage für linguistische Fragestellungen dienen. Hierfür sind weitere Verarbeitungsschritte notwendig, um die Daten entsprechend aufzubereiten. Es ist angezeigt, die Annotation der Texte um POS-Tags zu erweitern. Für Nicht-Standard Varietäten muss hierfür wieder auf das Training eigenständiger Tagger durch neuronale Netze zurückgegriffen werden. Obwohl diese sich im NLP bewährt haben, sind die Voraussetzungen für einen erfolgreichen Einsatz im linguistischen Kontext nicht vollständig klar. Im Rahmen dieser Arbeit sollen Einflussvariablen identifiziert und deren Auswirkung auf die Arbeit mit neuronalen Netzen vermessen werden. Ziel der Untersuchung ist es deshalb weiterhin, Verfahren zu verbessern und Parameter einzugrenzen, wie Neuronale Netze optimal zur Erkennung sprachlicher Strukturen genutzt werden können.

Ein möglicher beispielhafter Untersuchungsgegenstand ist die Possession, eine grundlegende, sprachübergreifende Kategorie, um Besitzverhältnisse auszudrücken. Possession kann durch unterschiedliche Konstruktionen realisiert werden. Diese können in Konkurrenz stehen oder parallel existieren. Eine Unterkategorie dieser Möglichkeiten bilden die adnominalen Konstruktionen. Im Deutschen umfassen diese (vgl. Kasper 2017: 300):

- possessiver Genitiv: Marias Kind, das Kind Marias .
- possessiver Dativ: Maria ihr Kind .
- von -Konstruktion: das Kind von Maria .
- Possessivpronomen: ihr Kind .

Es liegen zahlreiche Arbeiten zur Possession allgemein (Seiler 1983; Heine 1997; Stolz et al. 2008; McGregor 2009; Börjars et al. 2013) oder zu Teilaspekten vor: ihre Verwendung in einzelnen Dialekten des Deutschen (z.B. für Hessen: Kasper 2017), die Konkurrenz zwischen Genitiv und von -Konstruktionen in Abhängigkeit der Textsorte (Lang 2018), den frühkindlichen Erwerb possessiver Phrasen (Eisenbeiß et al. 2009) oder Possession im Sprachvergleich – beispielsweise Deutsch und Koreanisch (Shin 2004). Nicht vorhanden ist hingegen eine longitudinale Studie, die die gesamte deutsche Sprachgeschichte abdeckt. Und obwohl für Einzeldialekte Untersuchungen zur Possession existieren, fehlt eine umfassende empirische Studie, die eine größere Anzahl an dialektalen Varietäten des Deutschen abdeckt. Die adnominale Possession eignet sich deshalb besonders, um anhand des geschaffenen Parallelkorpus hinsichtlich ihrer diachronen Entwicklung und der Realisierung in dialektalen Varietäten untersucht zu werden.

Bibliographie

Baierer, Konstantin et al. (2020): “OCR-D kompakt: Ergebnisse und Stand der Forschung in der Förderinitiative“, in: *Bibliothek Forschung und Praxis* 44/2: 218-230.

Börjars, Kersti / Denison, David / Scott, Alan, (eds.) (2013): *Morphosyntactic Categories and the Expression of Possession*. Amsterdam, Philadelphia: John Benjamins.

Bünker , Michael / Lagger , Sepp (Übers.) (2007): *Es wead ana kemmen.* Das Markusevangelium auf Kärntnerisch. Übersetzt von Michael Bünker und Sepp Lagger. Klagenfurt: Heyn.

Eisenbeiß, Sonja / Matsuo, Ayumi / Sonnenstuhl, Ingrid (2009): „Learning to encode possession“. In: McGregor, William (ed.): *The expression of possession*. Berlin [u.a.]: de Gruyter, S. 143–213.

Heine, Bernd (1997): *Possession. Cognitive sources, forces, and grammaticalization*. Cambridge: Cambridge University Press.

Jessen, Johannes (Übers.) (1984): *Dat Ole un dat Nie Testament in unse Moderspraak.* Übersetzt von Johannes Jessen. Göttingen: Vandenhoeck & Ruprecht.

Kasper, Simon (2017): „Adnominale Possession“. In: *SyHD-Atlas*.

Lang, Kristine (2018): *Possession. Empirisch-funktionale Untersuchungen zu Genitivattribut und Präpositionalphrase mit von*. München: Iudicium.

McGregor, William (2009): *The expression of possession*. Berlin [u.a.]: de Gruyter.

Mieth, Klemens (Übers.) (2011): *Das Neue Testament uff Hessisch.* Übersetzt von Klemens Mieth. Norderstedt: Books on Demand GmbH.

o. A. (2002): *Es Nei Teshtament. Mitt Di Psaltah un Shpricha*. South Holland, IL: The Bible League.

Paul, Rudolf (Übers) (1997): *D Bibel für Schwoba. s Matthäus-Evangeliom*. Ens Schwäbische übersetzt vom Pfarrer Rudolf Paul. Tübingen: Silberburg.

Shin, Yong-Min (2004): *Possession und Partizipantenrelation. Eine funktional-typologische Studie zur Possession und ihren semantischen Rollen am Beispiel des Deutschen und Koreanischen*. Bochum: Brockmeyer.

Seiler, Hansjakob (1983): *Possession as an Operational Dimension of Language*. Tübingen: Narr.

Sievers, Eduard (ed.) (1872): *Tatian. Lateinisch und altdeutsch mit ausführlichem Glossar*. Paderborn: Schöningh.

Springmann, Uwe / Lüdeling, Anke (2017): “OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus”, in: *Digital Humanities Quarterly* 11/2.

Stolz, Thomas / Kettler, Sonja / Stroh, Cornelia / Urdze, Aina (2008): *Split possession. An areal-linguistic study of the alienability correlation and related phenomena in the languages in Europe.* Amsterdam, Philadelphia: Benjamins.

Theler, Hubert (Übers.) (2011): *Ds Niww Teschtamänt uf Wallisertitsch.* Übersetzt von Hubert Theler. Visp: Rotten.

Weber, Emil (Übers.) (2011): *S Nöi Teschtamänt Züritüütsch*. Us em Griechische. Übersetzt von Emil Weber. Zürich: Jordan.

Wick, Christoph / Reul, Christian / Puppe, Frank (2020): “Calamari - A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition”, in: *Digital Humanities Quarterly* 14/2.