Armin Hoenen, Alexander Mehler, Jost Gippert

Editorial

Im Februar 2016 fand an der Goethe Universität Frankfurt der Workshop "Corpora and Resources for Low Resource Languages with a Special Focus on Historical Languages" oder kurz CRiLL-HL statt.[1] Er wurde in Kooperation der GSCL-Arbeitskreise *Korpuslinguistik* und *Historisch-Vergleichende Sprachwissenschaft* mit dem *Centrum für Digitale Forschung in den Geistes-, Sozial- und Bildungswissenschaften* (CEDIFOR) der Goethe-Universität veranstaltet. Während für viele, vor allem für größere Sprachen mittlerweile eine gute bis sehr gute technologische Infrastruktur bereitsteht (d.m. in Bezug auf die Verfügbarkeit von Ressourcen einerseits und die Verfügbarkeit von grundlegenden Technologien andererseits), ist dies im Bereich so genannter *Low Resource Languages* (LRL), solcher Sprachen also, welche aus unterschiedlichsten Gründen wenig Zugang zu Ressourcen wie Korpora aller Art, Lexika, Grammatiken usw. aufweisen, noch nicht der Fall. Dies steht im Gegensatz zur großen Bedeutung dieser Sprachen, welche nicht nur in Europa selbst für einen Großteil der linguistischen Diversität verantwortlich sind. Die Situation verbessert sich in mancher Hinsicht stetig durch große Infrastrukturprojekte und Initiativen, sowie Organisationen, welche sich der Erschließung von LRL verschrieben haben. So sind beispielsweise CLARIN[2] mit seinem *Language Resource Inventory* oder die ELRA[3] zu nennen, welche für einen stetig besser werdenden Zugang zu Sprachressourcen sorgen. Unter anderem bedrohte Sprachen werden durch Projekte wie DOBES[4] noch einmal besonders ins Auge gefasst, da ihr unmittelbares Verschwinden droht.
WissenschaftlerInnen, die in einem dieser Kontexte zu LRL Sprachen forschen, sehen sich teilweise aber noch immer mit einer Reihe spezieller, schwer lösbarer Probleme konfrontiert, für deren Diskussion der Workshop ein Forum bieten und sich so in die Bestrebungen um eine bessere Verarbeitbarkeit der genannten Sprachen einreihen wollte. Besonders im Bereich der historischen Sprachen, welche innerhalb der LRL noch einmal eine besondere Stellung einnehmen, fand ein reger wissenschaftlicher Austausch statt. Dies betraf nicht nur die Präsentationen entsprechender Beiträge, sondern auch die Arbeit in themenorientierten Arbeitsgruppen, in welchen die TeilnehmerInnen spezielle Verfahrensweisen (wie z.B. die Lemmatisierung historischer Texte) intensiv diskutierten. In Bezug auf Annotationen korrespondieren einige der diskutierten Themen mit Fragestellungen, wie sie das kürzlich erschienene *Handbook of linguistic Annotation* thematisiert, was einmal mehr die Aktualität des Workshop-Themas unterstreicht.
Das vorliegende Heft des JLCL versammelt im Nachgang zu diesem Workshop nunmehr ausgewählte Beiträge, welche in diesem Kontext entstanden sind:

1. Der erste Beitrag von Florian Petran, Thomas Klein, Stefanie Dipper und Marcel Bollmann stellt mit ReM ein Referenzkorpus des Mittelhochdeutschen vor. Dabei

1 Informationen zum Workshop findet der interessierte Leser auch unter der Adresse http://gscl-ak-korpuslinguistik.hucompute.org
2 https://www.clarin.eu/content/language-resource-inventory
3 http://www.elra.info/en
4 http://dobes.mpi.nl

werden Korpusgenese, Quellen, Struktur und Annotationen genau beschrieben und mit Beispielen ausgeführt. ReM ist Teil eines bundesweiten Projektes zur Schaffung von Referenzkorpora für historische Sprachstufen des Deutschen. Es wurde semi-automatisch mit Annotationen angereichert, so u.a. im Hinblick auf Tokenisierung, Normalisierung, Parts of Speech, morphologische Analyse, Lemmata, wodurch eine Vielzahl weitergehender Analysen ermöglicht wird. Insgesamt umfasst ReM ca. 2,5 Millionen Token (in ca. 400 Texten).

2. Der zweite Beitrag von Roland Mittmann beschreibt eine Methode zur automatischen dialektalen Einordnung althochdeutscher Wortformen. Der Autor stellt das Konzept dieser Methode vor, welches auf aus Grammatiken extrahierten relativen Lautentsprechungen und deren grammatikalischen Funktionen beruht, und demonstriert erste Ergebnisse, welche auf die vielversprechende Anwendbarkeit seiner regel-basierten Methode zum Zwecke der automatischen Einordnung althochdeutscher Texte in Zeit-Dialekträume schließen lassen.
3. Der dritte Beitrag von Armin Hoenen und Lela Samushia stellt ein altgeorgisches Inschriftenkorpus vor, welches im Format der TEI (EpiDoc) codiert wurde, und erörtert die spezifischen Probleme, die dieser Texttyp an die technologische Verarbeitung stellt. Als *proof-of-concept* präsentieren Hoenen und Samushia Front- und Backend eines Tools, das aufzeigt, welche Eigenschaften wichtig sind, um bei der Entschlüsselung und Rekonstruktion der Botschaft von oft nur fragmentarisch überlieferten Inschriften zu helfen. Dabei kommen *language models*, *word embeddings* und frequenzbasierte Statistiken zum Einsatz.

Wir danken allen Gutachtern, der GSCL, den Herausgebern des JLCL sowie dem CEDIFOR für die gewährte Unterstützung und wünschen den LeserInnen ein angenehmes und hoffentlich erkenntnisreiches Leseerlebnis.

Armin Hoenen, Alexander Mehler und Jost Gippert
(Juli 2017, Frankfurt am Main)

Literatur

Ide, N., & Pustejovsky, J. (Eds.). (2017). *Handbook of Linguistic Annotation*. Springer.