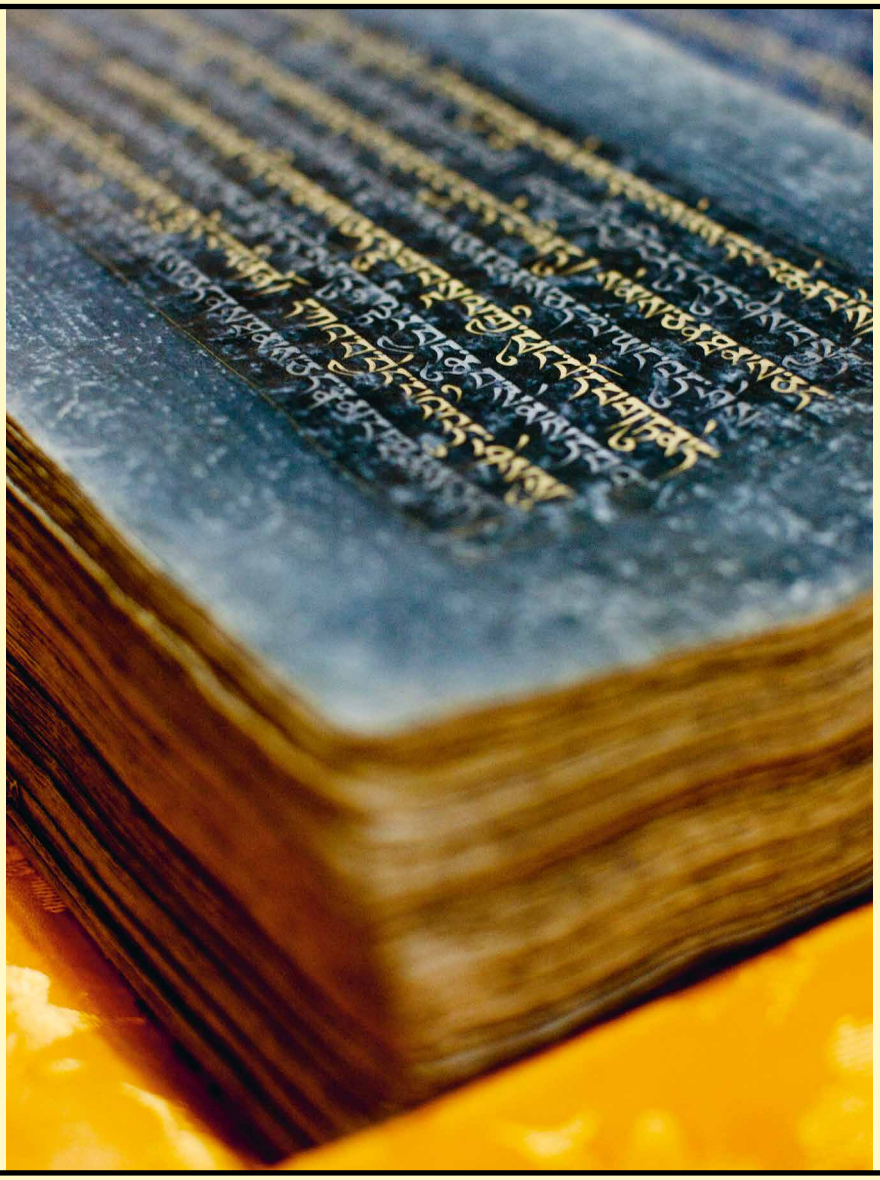
Tausend Jahre Sprachentwicklung

Ein digitales Wörterbuch zeichnet die **Entwicklung des tibetischen Verbs** vom 8. Jahrhundert bis heute nach und dient auch dem Erhalt dieser bedrohten Sprache.



Buch über den tibetischen Buddhismus aus dem 12. Jahrhundert, heute aufbewahrt in Dharamshala, Indien.

Von **Samyo Rode-Hasinger** und **Nikolai Solmsdorf**

Unter der Leitung von Ulrich Pagel an der SOAS University of London und in Kooperation mit dem Projekt „Wörterbuch der tibetischen Schriftsprache“ der Bayerischen Akademie der Wissenschaften hat im Juli 2017 das Forschungsvorhaben „Lexicography in Motion: A History of the Tibetan Verb“ mit einer dreijährigen Laufzeit begonnen. Das Projekt wird vom britischen Arts and Humanities Research Council finanziert und stellt der BAdW Personalmittel im Umfang einer Mitarbeiterstelle zur Verfügung.

Ziel des Projektes ist ein digitales Wörterbuch des tibetischen Verbs, das auf der Grundlage eines umfangreichen Textkorpus erstellt wird. Das Textkorpus umfasst dabei ebenso Literatur aus der alttibetischen Periode des 8. bis 10. Jahrhunderts wie aus dem klassischen (11.–19. Jhdt.) und dem modernen Tibetisch (20./21. Jhdt.). Somit wird die Sprachentwicklung in einem Zeitraum von mehr als tausend Jahren nachgezeichnet. Für das Wörterbuch werden verschiedene Textgattungen, Transkripte von Online-Magazinen, Zeitungen, Chats und Videos ausgewertet.

Tibetisch – eine bedrohte Sprache

Tibetisch ist heute eine bedrohte Sprache; sie sieht sich einer Verdrängung durch die chinesische Mehrheit ausgesetzt, wird

lediglich in der tibetischen Exilgemeinde kultiviert und vornehmlich in Universitäten und Forschungseinrichtungen außerhalb Tibets untersucht. Das Wörterbuchprojekt soll der Wissenschaftsgemeinde und der Öffentlichkeit den Zugang zur tibetischen Sprache erleichtern und damit ihrem Erhalt dienen.

Das Tibetische gehört der tibeto-birmanischen Sprachfamilie an. Die Besonderheit der Sprache wird durch seine grammatische Struktur betont, die selten in anderen Sprachfamilien zu finden ist. Dabei spielen Verben eine zentrale Rolle; das Wissen um die Bedeutung eines Verbs führt zu dem Verständnis der Argumente, die es erfordert, und der semantischen Rollen, die diese Argumente übernehmen. Unser Lexikon stützt sich auf diese Verbindungen und wird es ermöglichen, Rückschlüsse vom Prädikat auf die vollständige grammatische Struktur zu ziehen. Darüber hinaus werden die morphologischen und semantischen Veränderungen der Verben aus den ersten tibetischen Aufzeichnungen des 8. Jahrhunderts bis heute deutlich.

Korpusannotation und Maschinelles Lernen

In einem ersten Schritt annotiert ein internationales Team aus Wissenschaftlerinnen und Wissenschaftlern mit Arbeitsstellen in England (Alttibetisch), Deutschland (Klassisches Tibetisch) und Indien (Modernes Tibetisch) die tibetischen Dokumente manuell. Das Korpus liegt in digitaler Form vor, die Auszeichnung der Verbstrukturen erfolgt mittels der Browser-gestützten Anwendung BRAT, die als Open Source zur Verfügung steht und es den Teammitgliedern ermöglicht, gleichzeitig an den Texten zu arbeiten. Die Richtlinien für die Annotierung basieren auf den Vorgaben des international anerkannten Auszeichnungsschemas für Dependenzgrammatik „Universal Dependencies“ und wurden in einer für das Projekt eigens angefertigten Dokumentation festgelegt.

Der Annotierungsprozess sei an dem folgenden Beispiel illustriert: Das tibetische Verb *gnang* (geben, gewähren; erlauben) ist ein ditransitives Verb, also ein Verb, das zwei Objekte hat. Entsprechend seiner Valenz fordert es in einem wohlgeformten Satz drei Argumente und wird hier im Sinne von „jemand gewährt jemandem etwas“ verwendet. Der Beispielsatz stammt aus der im 15. Jahrhundert verfassten Biografie des Marpa Chos kyi blo gros, eines wichtigen Geistlichen des 11. Jahrhunderts.

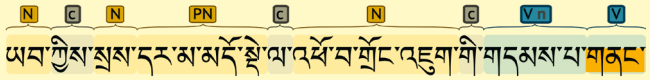
yab kyis
sras dar ma mdo sde la
'pho ba grong 'jug gi
gdams pa
gnang

Vater – [erg]
Sohn Darma mDo sde – [dat]
Bewusstseinsübertragung – [gen]
Unterweisung – [abs]
gewähren

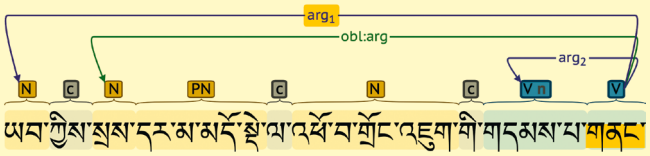
„Der Vater gewährte seinem Sohn Darma mDo sde die Unterweisung der Bewusstseinsübertragung.“

WWW https://tibetan-nlp.github.io/lim-annodoc [Richtlinien für die Annotierung im Projekt]

Subjekt (arg1), direktes Objekt (arg2) und indirektes Objekt (obl:arg) werden im tibetischen Satz vom Prädikat ausgehend anhand von Pfeilen mit den entsprechenden Bezeichnungen markiert.



Durch Wortarten markierter tibetischer Satz aus Rus pa'i rgyan cans (1452–1507) Biografie des Marpa Chos kyi blo gros (1012?–1097).



Tibetischer Satz mit ausgezeichneter Verb-Argument-Struktur.

Basierend auf computerlinguistischen Modellen und statistischer Validierung soll schließlich das handannotierte Textkorpus zum Training für Algorithmen aus dem Bereich des Maschinellen Lernens verwendet werden, womit weitere, in digitaler Form frei verfügbare tibetische Texte ausgezeichnet und erschlossen werden können.

Nach einer umfangreichen Endredaktion werden die Ergebnisse des Projekts in einem Valenzwörterbuch online und als Open Access zugänglich sein. Es wird allen Nutzerinnen und Nutzern die Möglichkeit bieten, die tibetischen Verben samt ihren Bedeutungen, Strukturen und Häufigkeiten nachzuschlagen. Die einzelnen Einträge werden anhand von tibetischen Beispielsätzen mit englischer Übersetzung präsentiert.

Samyo Rode-Hasinger M.A. und **Dr. Nikolai Solmsdorf** sind wissenschaftliche Mitarbeiter am Projekt „Wörterbuch der tibetischen Schriftsprache“ der Bayerischen Akademie der Wissenschaften und am Forschungsvorhaben „Lexicography in Motion: A History of the Tibetan Verb“ der SOAS University of London.

Foto: Matthew Wakem/laif; SOAS-Projekt