Digitale Workflows in Langzeitprojekten am Beispiel einer Infrastruktur zur Dokumentation indigener nordeurasischer Sprachen (INEL)

,
hanna.hedeland@uni-hamburg.de
Universität Hamburg, Deutschland

,
timm.lehmberg@uni-hamburg.de
Universität Hamburg, Deutschland

,
beata.wagner-nagy@uni-hamburg.de
Universität Hamburg, Deutschland

Zusammenfassung

Gegenstand des Beitrages sind die Arbeiten zu digitalen Workflows und infrastruktureller Einbindung im Rahmen des Langzeitprojektes *INEL* (Grammatical Descriptions, Corpora, and Language Technology for Indigenous Northern Eurasian Languages). Das Projekt wurde von Prof. Dr. Beata Wagner-Nagy (Institut für Finnougristik/Uralistik, Hauptantragstellerin) sowie von Dr. Michael Rießler (Skandinavisches Seminar Albert-Ludwigs-Universität Freiburg) und der Geschäftsführung des Hamburger Zentrums für Sprachkorpora (Hanna Hedeland und Timm Lehmberg) beantragt. Ziel des Projektes ist es, über den Zeitraum von 18 Jahren die dringend erforderliche Erschließung der sprachlichen Ressourcen des genealogisch diversen nordeurasischen Sprachraums (s. Abbildung 1) zu leisten. Durch den Einsatz von State-of-the-Art-Methoden und -Werkzeugen der linguistischen Datenaufbereitung, die bisher nur für gut erforschten Sprachen und Varietäten zum Einsatz kamen, wird eine Lücke in diesen für die empirische Sprachwissenschaft bisher schlecht zugänglichen Arealen der Welt nachhaltig geschlossen.

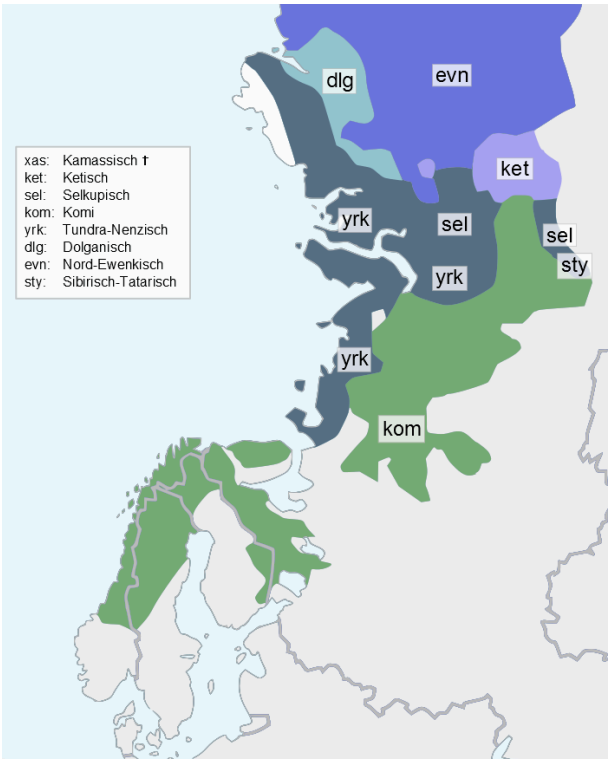

Abb. 1: Der geographische Skopus des Projekts.

Dieses ehrgeizige Ziel stellt hohe Anforderungen an die Organisation der Projektworkflows und erfordert zudem die Schaffung einer eigenen nachhaltigen und international vernetzten digitalen Forschungsinfrastruktur. Das Projekt leistet somit jenseits seiner linguistischen Ausrichtung einen wichtigen Beitrag für die Digital Humanities.

Anforderungslage

Aufgrund des drohenden Verfalls der zum großen Teil auf obsoleten analogen Originalträgern (Wachswalzen, Schellackplatten, Mikrofiche etc.) vorhandenen Audio-Aufnahmen, Niederschriften und Beschreibungen, schließt sich in absehbarer Zeit das Fenster für einen Erhalt dieser Daten. Gleichzeitig gehen die Sprecherzahlen vieler Sprachen und Varietäten stetig zurück. Indem existierende Materialien zu digitalen Korpora aufbereitet und der bisherige Gesamtbestand um neue Ressourcen ergänzt wird, kann dieses Erbe als wertvolle empirische Basis für vielfältige Forschungsvorhaben erhalten werden. (Eine detaillierte Beschreibung des Standes der linguistischen Erfassung befindet sich im Förderantrag, S. 3 ff.)

Vielmehr als nur ein digitales Archiv entsteht im Rahmen von *INEL* jedoch eine umfassende virtuelle Forschungsumgebung, die durch die Integration in supranationale Forschungsinfrastrukturen der wissenschaftlichen Öffentlichkeit dauerhaft zugänglich gemacht wird. Ein primäres Ziel des Projektes besteht zunächst darin, existierende Beschreibungen

einzelner nordeurasischer Sprachen und Varietäten, die aufgrund der bisher begrenzten Auswahl von verfügbaren Sprechern und Genres eher partikuläre Idiolekte dokumentieren, zusammenzutragen und mit ergänzenden Korpora als umfangreiche digitale Ressource zugänglich zu machen. Durch die so geschaffene, der Vielfalt der Sprache angemessenen, Datenbasis werden für zukünftige Generationen von Forschenden erstmalig varietätenübergreifende Analysen möglich, etwa die Erforschung kontaktinduzierter Sprachveränderungen, Anwendungen aus dem Bereich der Dialektometrie oder sprachsoziologische Untersuchungen. Die unterschiedlichen Erhebungszeiten der Sprachdaten erlauben zudem erstmalig datengestützte Untersuchungen von diachronem Sprachwandel sowie Grammatikalisierungsprozessen. Ebenso bedeutend sind die Art des Zugangs zu den Sprachdaten und die damit verbundenen Analysemöglichkeiten. Die Sprachdaten können in der entstehenden Forschungsumgebung kollaborativ und dezentral um beliebige weitere Beschreibungsebenen angereichert werden, die dann für verschiedene Auswertungsszenarien zur Verfügung stehen. Auf diese Weise wird die virtuelle Forschungsumgebung modular aufgebaut und in vielen Fällen so generisch sein, dass auch die Resultate technologischer und methodologischer Entwicklungen der akademischen Öffentlichkeit als Best Practices und als konkrete Grundlage für vergleichbare Vorhaben zur Verfügung stehen werden.

Modularisierung und Workflows

Die oben beschriebene Ausgangslage determiniert zwei Dimensionen der Entwicklung der entstehenden Ressourcen, denen durch entsprechende Modularisierung von Workflows begegnet werden muss.

1. Entwicklung hinsichtlich der arealen Abdeckung durch Erschließung der Einzelsprachen.
2. Entwicklung hinsichtlich der Komplexität der Daten infolge von hinzuzufügenden Glossierungen und Mehrebenenannotationen, die sowohl innerhalb des Erfassungsprozesses jeder Einzelsprache als auch über den gesamte Laufzeit erfolgen.

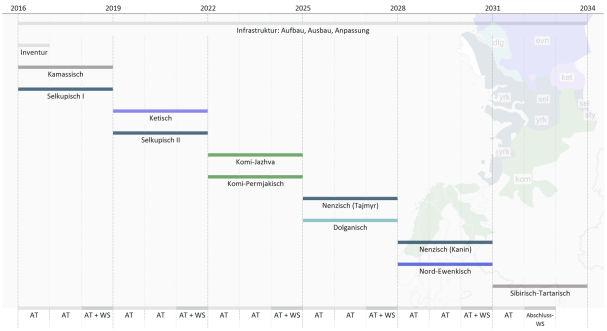

Abb. 2: Teilprojekte.

Der folgtende Abschnitt basiert auf den den im Förderantrag (S. 17 ff) beschriebenen konzeptionellen Vorarbeiten zur Modularisierung der Projektworkflows.

Das Projekt gliedert sich dem entsprechend in insgesamt zwölf Teilprojekte (s. Abbildung 2), von denen elf jeweils die Erschließung einer der zu erfassenden Sprachvarietäten zum Gegenstand haben. Das zwölfte technisch-infrastrukturelle Teilprojekt läuft im Gegensatz zu den elf jeweils auf drei Jahre angelegten Erschließungsprojekten durchgängig über die gesamte Projektlaufzeit und schafft somit die notwendige Kontinuität für die Entwicklung, Anpassung und Vermittlung der Funktionalitäten der technischen Infrastruktur. Das Arbeitsprogramm wird für die jeweiligen Teilprojekte aber auch teilprojektübergreifend in Form von methodischen Arbeitspaketen umgesetzt:

Arbeitspaket 1: Korpusaufbau

In diesem Arbeitspaket werden existierende Ressourcen erschlossen, kuratiert und aufbereitet sowie ggf. um neu zu akquirierende und zu erhebende Daten ergänzt. Die dabei erforderlichen Verarbeitungsschritte sind, wie in Abbildung 3 dargestellt, modularisiert. Die Module Digitalisierung, linguistische Modellierung, Annotation / Glossierung und Finalisierung/Integration entsprechen jeweils Aufbereitungsschritten, die abhängig vom Ausgangszustand der einzelnen Ressource für die Integration in den Gesamtbestand erforderlich sind.

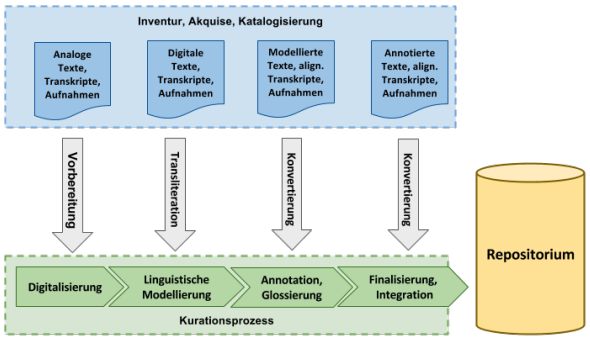

Abb. 3: Modularisierung und Anpassung der Verarbeitungsschritte.

In der Praxis handelt es sich jedoch keineswegs um einen linearen Prozess, den einzelne Texte einmalig durchlaufen und an dessen Ende die Speicherung und Zugänglichkeit einer in sich geschlossenen Ressource in einem digitalen Repositorium steht. Vielmehr müssen die zu planenden Workflows der Aufbereitung und Speicherung den

tatsächlichen Gegebenheiten der Datenaufbereitung in Projekten dieser Art Rechnung tragen (s. Abbildung 4):

- Es ist in vielen Fällen wünschenswert, Versionen von Ressourcen bereits in einem frühen Stadium der Aufbereitung (beispielsweise noch vor Abschluss der Annotation und Glossierung) der wissenschaftlichen Öffentlichkeit zugänglich zu machen.
- Bei dem Arbeitsschritt der Annotation und Glossierung handelt es sich wiederum um iterative Prozesse, in deren Rahmen mehrere Ebenen der Auszeichnung, möglichwerweise sogar zeitlich überlappend, zu den Primärdaten hinzugefügt werden, was hohe Anforderungen an Koordination und Qualitätskontrolle stellt.
- Insbesondere bei Langzeitvorhaben entstehen oft neue Versionen von Korpora aus bereits bestehenden Ressourcen, indem diese mit zusätzlichen Annotationsebenen ausgezeichnet werden.

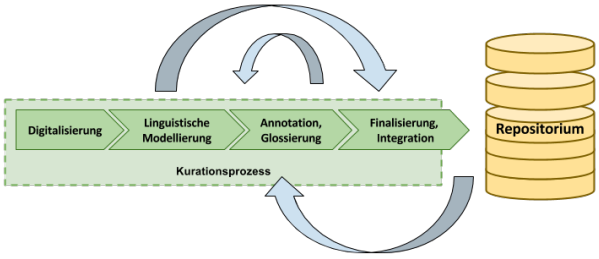

Abb. 4: Nicht-lineare Abfolge der Verarbeitungsschritte.

Diesen Gegebenheiten kann in der Praxis durch ein ausdifferenziertes Versionierungskonzept begegnet werden. Im Rahmen des Vortrages werden einige konkrete Workflows aus dem Projekt vorgestellt und ihre Implementierung unter Verwendung eines Git -basierten Repositoriums ausführlich erläutert.

Arbeitspaket 2: Infrastruktur und Best Practices

Die zu errichtende Infrastruktur basiert in vielerlei Hinsicht auf den Vorarbeiten des HZSK sowie generell auf vorangegangenen Erkenntnissen aus dem Aufbau digitaler Forschungsumgebungen. Große Teile der gewünschten Funktionalitäten wurden bisher in Form von Standalone-Werkzeugen (bspw. EXMARaLDA) oder Webapplikationen (bspw. als Teil der CLARIN-D-Infrastruktur) am HZSK entwickelt und eingesetzt. Einen integralen Bestandteil der Infrastruktur bildet ein Repositorium, mit dessen Hilfe die nachhaltige Datenvorhaltung gewährleistet werden kann. Weitere Komponenten, die unmittelbar daran anknüpfen, bilden zusätzliche relevante Aspekte der gewünschten Funktionalität der Infrastruktur ab, wie etwa die kollaborative Bearbeitung und Aufbereitung von Daten in der Arbeitsumgebung, die Auslieferung von Metadaten an Kataloge und Archive, welche die Auffindbarkeit der Ressourcen für andere Forscher ermöglicht, sowie Schnittstellen für die Exploration und Analyse der vorgehaltenen Ressourcen. Auch die fortlaufende Anbindung an bzw. Vernetzung mit weiteren bestehenden Forschungsinfrastrukturen wird als essentielles Merkmal des Projektes betrachtet.

Arbeitspaket 3: Evaluation und Dissemination

Neben der organisatorischen und der technisch-infrastrukturellen Organisation und Vernetzung, die mit den Arbeitspaketen 1 und 2 abgedeckt ist, erfordert ein Langzeitprojekt wie *INEL* zudem umfassende Arbeiten im Bereich der Dissemination und den Austauschs im supranationalen interdisziplinären Kontexten mit anderen Forschenden. Um neue Arbeitsinstrumente zu entwickeln und zu diskutieren, Arbeitspläne zu koordinieren, aktuelle Forschungsfragen und die mit der Projektarbeit zusammenhängenden praktischen Fragestellungen zu diskutieren sowie zu Zwecken der Fortbildung werden im Rahmen von *INEL* jährliche Workshops abgehalten, an denen ausgewählte Konsultanten sowie Kooperationspartner aus dem Ausland und die Projektmitarbeiter selbst teilnehmen.

Ausblick

Der Beitrag basiert auf den Planungen zu der *INEL*-Langzeitprojekt, dessen 18-jährige Laufzeit im Januar 2016 beginnen wird. Die Vorarbeiten zu dem Projekt lieferten wichtige Erkenntnisse hinsichtlich der Modularisierung von Projektabläufen sowie der Priorisierung von Verarbeitungsschritten der Datenaufbereitung. So wird beispielsweise der technische Fokus nicht auf der Neuentwicklung von weiteren Werkzeugen und Standards der Datenaufbereitung, sondern der Entwicklung von modularisierten Infrastrukturen und Workflows liegen, die auf die Interoperabilität, Interaktion und Integration existierende Komponenten abzielen. Nur so kann ein flexibler Betrieb der *INEL*-Infrastruktur in einer sich permanent wandelnden Ressourcen- und Infrastrukturlandschaft gewährleistet werden.

Von der Entwicklung und Erprobung kontrollierter und modularisierter Workflows der Datenaufbereitung, die beispielsweise eine transparente Dokumentation und Publikation entstehender Versionen von Forschungsdatensammlungen erlauben, sind zudem wichtige Beiträge auf dem Feld des Forschungsdatenmanagements von Projekten in den Digital Humanities zu erhoffen.

Bibliographie

Git-Hub (o.J.): *Git*. Local branching on the cheap http://git-scm.com/ [letzter Zugriff 16. Februar 2016].

INEL (2015): *Grammatiken, Korpora, Sprachtechnologie für indigene nordeurasische Sprachen*. Förderantrag, eingereicht bei der Union der deutschen Akademien der Wissenschaften.

Hedeland, Hanna / Lehmberg, Timm / Schmidt, Thomas / Wörner, Kai (o.J.): *EXMARaLDA*. Werkzeuge für mündliche Korpora http://www.exmaralda.org/ [letzter Zugriff 16. Februar 2016].

HZSK (o.J.): *Hamburger Zentrum für Sprachkorpora* https://corpora.uni-hamburg.de/drupal/ [letzter Zugriff 16. Februar 2016].

Universität Hamburg (o.J.): *Institut für Finnougristik / Uralistik* https://www.slm.uni-hamburg.de/ifuu [letzter Zugriff 16. Februar 2016].