MIDEUKO - Mittelhochdeutsches Korpus für die NFDI

Eckhart Arnold (Bayerische Akademie der Wissenschaften),
Stefan Müller (Bayerische Akademie der Wissenschaften)

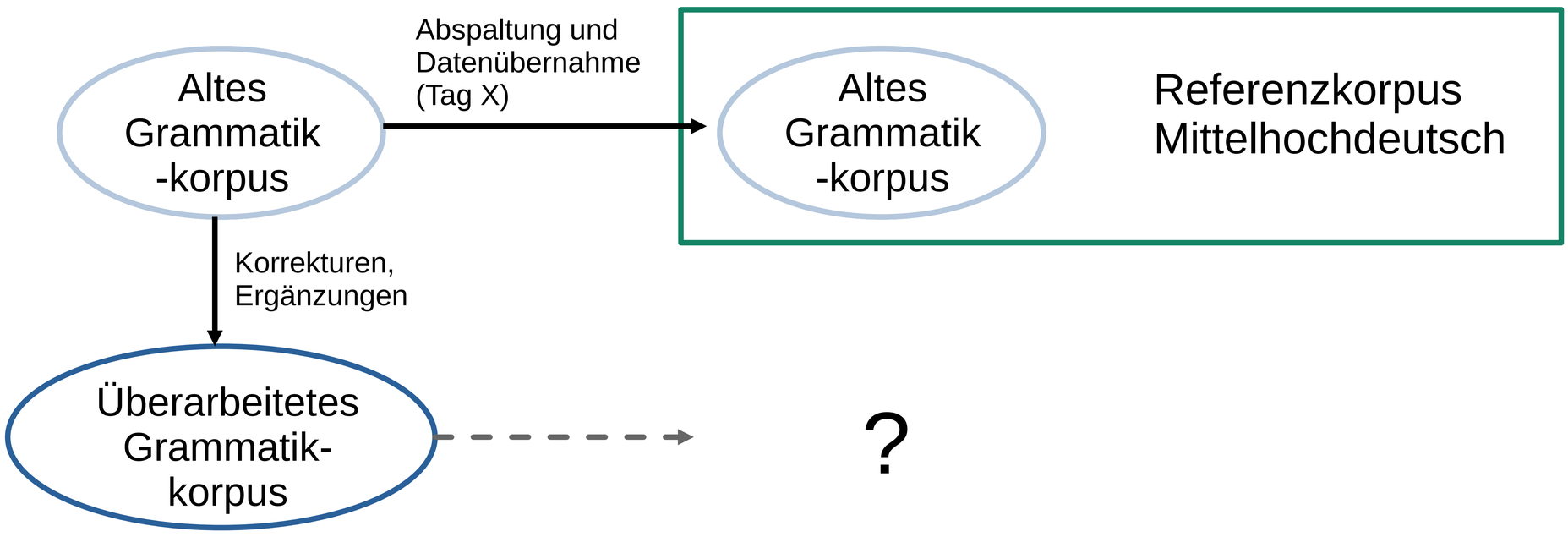
Die Ausgangslage: Ein verwaistes Korpus und seine Abspaltungen

Das **Mittelhochdeutsch-Korpus, auch "Grammaitkkorpus"** ist ein Sprachkorpus mittelhochdeutscher Schriften. Ursprünglich für die Erarbeitung einer Grammatik des Mittelhochdeutschen zusammengestellt, gibt es eine erweiterte aber technisch und inhaltlich vereinfachte Abspaltung in Form des **Referenzkorpus Mittelhochdeutsch**, kurz **"Referenzkorpus"** das an verschiedenen Stellen im Netz zu finden ist.

Das **Grammatikkorpus** war *bis vor kurzem verwaist* (es befand sich nur noch auf der privaten Festplatte seines inzwischen emeritierten Schöpfers Thomas Klein), ist aber i*nzwischen in einem gitlab-Repositorium der Bayerischen Akademie der Wissenschaften unter CC-BY-SA-Lizenz der Öffentlichkeit zugänglich* und durch die daran angeschlossenen Archivierungs- und Backupsysteme des Leibniz-Rechenzentrums vor dem Datenverlust gesichert.

Das **Refernzkorpus** existiert in (mindestens) zwei Abspaltungen, von denen die eine von der Ruhr-Uni-Bochum, die andere vom beim Deutschen Textarchiv angeboten wird.

Das Grammatikkorpus wurde auch nach der Abspaltung des Referenzkorpus über viele Jahre hinweg weiterbearbeitet und fehlerkorrigiert. Der Teil des Referenzkorpus, der dem Grammatikkorpus entspricht, ist daher veraltet und entspricht so gesehen möglicherweise nicht dem Stand der Forschung, da spätere *Änderungen am Grammatikkorpus nicht mehr in das Referenzkorpus eingeflossen* sind. Für Nutzer und Nutzerinnen des Referenzkorpus bleibt dieser Zusammenhang mehr oder weniger unsichtbar.



Die Datenlage

Grammatikkorpus:

Datenformat:	wohldokumentierte Eigennotation
Auszeichnungstiefe:	sehr detailliert
Datenqualität:	handverlesen
Speicherung:	passiv (gitlab)

Referenzkorpus:

Datenformat:	XML
Auszeichnungstiefe:	weniger detailliert
Datenqualität:	handverlesen, aber großzügigere Kriterien, teilweise veraltet
Speicherung:	Datenbank mit Suchfunktion

Die meisten Unterschiede sind unproblematisch und ergeben sich aus der unterschiedlichen Zielsetzung der Datenbestände. Problematisch ist jedoch, dass es keine Übernahme aktualisierter Daten vom „Flußoberlauf" gibt.

Konsolidierung der Daten und Integration in Text+

Primär-Ziele des Projektes:

1. **Aufbereitung der aktuellen Korpusdaten des Grammatikkorpus für den Export**, so dass sie im Referenzkorpus aktualisiert werden können. Im Wesentlichen erfordert dies eine XML-Konvertierung der Daten. (Die Eigennotation des Grammatikkorpus bleibt dabei aber ein „Bürger erster Klasse", da sie auf den besonderen Zweck dieses Korpus (mittelhochdeutsche Grammatik) zugeschnitten ist und sich dafür wie auch für einige andere Anwendungsfelder besser eigent als XML.)

2. **Integration der Daten in das Referenzkorpus**

3. **Verzeichnung der Daten, insbesondere des weitgehend unbekannten Grammatikkorpus in Suchsystemen und Katalogen**, insbesondere der Sammlungen von Text+

Sekundärziele:

- Etablierung eine dynamischen Datenverbindung zwischen Geber („Flussoberluaf") und Nehmer („Flussunterlauf"), um künftige Aktualisierungen ggf. per Knopfdruck durchführen zu können.
- Definition von guten Praktiken für die danymische Datenübernahme bzw. Zusammenführung aus unterschiedlichen Quellen.

Eine grundlegende Herausforderung:
Fehlende Verbindung von Datengebern und Datennehmern in dynamischen Datenentwicklungskontexten

Ein Lücke, die die FAIR-Prinzipien und Open-Access-Lizenzen noch offen lassen: *Wie kann das (ungewollte) Auseinanderdriften von Datenbeständen nach (freundlichen) Abspaltungen verhindert werden?*

Lösungen aus dem Abhängigkeitsmanagement in der Software-Entwicklung könnten evtl. als Vorbild dienen.

Das NFDI-Konsortium Text+ wird gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 460033370