Forschungsdaten spielen in den letzten Jahren in der Linguistik eine immer größer werdende Rolle. Der Fachinformationsdienst (FID) Linguistik verfolgt deshalb im laufenden Projekt (DFG-Förderung 2017–2019) schwerpunktmäßig das Ziel, Sichtbarkeit, Auffindbarkeit und Verfügbarkeit linguistisch relevanter, digitaler Sprachressourcen zu erhöhen. Als Ausgangspunkt dient das Linguistik-Portal, das bereits über einschlägige Suchfunktionen verfügt. Durch die Anwendung innovativer Technologien im Bereich Linked Open Data wird der Suchraum sowohl qualitativ als auch quantitativ erweitert. Dies geschieht auf dem Wege der Vernetzung des Linguistik-Portals mit Terminologie-Repositorien in der Cloud. Darüber hinaus werden Verfahren zur automatisierten Metadaten-Anreicherung von digitalen Ressourcen entwickelt. Weitere Services wie die Verknüpfung von Sekundärliteratur mit Forschungsdaten ergänzen das Angebot des FID Linguistik.

In the last few years, research data has played an increasing role in linguistics. The Specialist Information Service (SIS) Linguistics is therefore pursuing the goal of increasing the visibility, traceability and availability of linguistically relevant digital language resources in the current project (DFG Funding 2017–2019). The starting point is the Linguistik portal, which already features suitable search functions. The search space is extended both qualitatively and quantitatively by applying innovative Linked Open Data technologies. This is done by linking the Linguistik portal to terminology repositories in the cloud. In addition, methods are being developed for the automated metadata enrichment of digital resources. Further aspects such as the linking of secondary literature with research data complete the SIS Linguistics services.

HEIKE RENNER-WESTERMANN

Fachinformationsdienst Linguistik zwischen Innovation und Tradition

Forschungsdaten in der Linguistik

Nicht nur Forscherinnen und Forscher aus dem Bereich der angewandten Linguistik oder Computerlinguistik, sondern alle empirisch arbeitenden Geisteswissenschaftlerinnen und Geisteswissenschaftler rekurrieren zunehmend auf große textbasierte Datenmengen, deren Analyse durch automatisierte Verfahren erleichtert wird. Das daraus erwachsene, breite Forschungsfeld der Digital Humanities bedient sich dabei linguistischer Methoden. Der Fachinformationsdienst (FID) Linguistik hat sich zum Ziel gesetzt, solche korpuslinguistischen Vorhaben durch verschiedene Maßnahmen zu fördern. Der Fokus liegt dabei weniger auf digitalen Texteditionen, sondern auf Sprachressourcen, wie Korpora oder elektronischen Wörterbüchern, die maschinenlesbar zur Verfügung stehen und im besten Fall bereits linguistisch annotiert sind (Tokenisierung, Lemmatisierung, Part-of-Speech-Tagging). Eine zentrale Aufgabe des FID Linguistik besteht darin, die Sichtbarkeit, Auffindbarkeit und Verfügbarkeit dieser Sprachressourcen zu erhöhen.

Der FID Linguistik wird an der Universitätsbibliothek Frankfurt am Main in Kooperation mit der Forschungsgruppe Angewandte Computerlinguistik am Institut für Informatik der Goethe-Universität Frankfurt am Main aufgebaut. Als zentrale Plattform dient das Linguistik-Portal[1], welches bis April 2017 mit DFG-Förderung als Virtuelle Fachbibliothek entstanden ist und im Rahmen des laufenden Projekts weiterentwickelt wird. In der ersten Ausbaustufe[2] des Portals waren Sprachressourcen auf konventionelle Art und Weise im Linkverzeichnis hinsichtlich Sprache und Zugriffsmöglichkeiten intellektuell erschlossen und nachgewiesen worden. Dieses traditionelle Verfahren eignet sich jedoch im Grunde nur für solche Ressourcen, die noch nicht in andere überregionale Nachweisportale – allen voran LingHub[3], DataHub[4], CLARIN-VLO[5] oder Meta-Share[6] – Eingang gefunden haben. Bereits in der zweiten Ausbaustufe des Linguistik-Portals und schwerpunktmäßig im laufenden FID-Projekt wurden und werden deshalb zusätzliche Anstrengungen unternommen, um die Verbesserung der Sichtbarkeit und Auffindbarkeit von frei verfügbaren Sprachressourcen zu erreichen.

Vernetzung des Linguistik-Portals mit Linked Open Data (LOD)

Die Linguistic Linked Open Data Cloud (LLOD-Cloud)[7] vereint unterschiedliche linguistische Ressourcen, wie Korpora, Wörterbücher, Terminologie-Repositorien oder fachspezifische Datenbanken, und zielt darauf ab, durch die Interoperabilität vieler Ressourcen eine automatisierte Informationsgewinnung zu ermöglichen. Die Vernetzung des Linguistik-Portals mit der Cloud basiert auf der Bereitstellung des Schlagwort-Thesaurus der Bibliography of Linguistic Literature (BLL) für LOD. Der BLL-Thesaurus existiert seit 1971 und

wird seitdem im Zuge der Arbeiten für die gleichnamige Bibliografie kontinuierlich an die laufende Entwicklung des Faches angepasst. Der Thesaurus umfasst über 7.800 Begriffe und ist mit mehr als 400.000 Titeldaten verknüpft. Darüber hinaus liefert er die Grundlage für die sachliche Navigation und die inhaltliche Erschließung in den Modulen des Linguistik-Portals.

Für die Bereitstellung des BLL-Thesaurus für LOD werden automatische mit intellektuellen Verfahren kombiniert: Der Thesaurus wird automatisch in ein LOD-konformes Format konvertiert, intellektuell nach Semantic-Web-Standards remodelliert sowie mit geeigneten Terminologie-Repositorien in der Cloud verlinkt. Die LOD-Edition[8] des BLL-Thesaurus enthält alle Modelle und Mappings und wird unter einer freien Lizenz veröffentlicht. Dazu gehört als wesentliche Ergänzung auch die Verknüpfung der Titeldatensätze mit den Schlagwort-Normdatensätzen, in denen sich die inhaltliche Erschließung der BLL widerspiegelt.[9] Durch die LOD-Edition wird die Nachnutzung der Daten durch andere – insbesondere nichtbibliothekarische – Akteure ermöglicht.

Im Fokus der zweiten Ausbaustufe standen die Sachschlagwörter, die der Beschreibung einzelner objektsprachlicher Phänomene dienen.[10] Aufbereitet wurden zunächst die Thesaurus-Bereiche Syntax, Morphologie und Lexikologie mit Begriffen wie *Adjektiv, Adverb, Kausalsatz, Finalsatz, Kasus, Tempus* usw. Es folgte die Verknüpfung dieser Begriffe mit den entsprechenden Konzepten aus dem Referenzmodell der Ontologies of Linguistic Annotations (OLiA)[11], einer frei verfügbaren linguistischen Ontologie, die als Mediator zwischen verschiedenen Klassifikationssystemen beziehungsweise Annotationsschemata fungiert.

Für die Recherche nach Sprachressourcen in der Cloud wurde eine innovative Suchfunktion entwickelt: Die Metadaten der Sprachressourcen, die ein eigens dafür implementierter Webcrawler findet, werden mit BLL-Schlagwörtern angereichert und nahtlos in die bestehende Katalogsuche integriert. Die Eingabe von BLL-Schlagwörtern in die Suchmaske des Linguistik-Portals führt unter Ausnutzung der LOD-Vernetzung auf portalexterne Ressourcen. Sucht man zum Beispiel nach *Auxiliarverb,* dann findet man neben der im Linguistik-Portal verzeichneten Sekundärliteratur auch frei verfügbare Sprachkorpora aus der LLOD-Cloud, deren Annotationsschema eine Entsprechung zu *Auxiliarverb* enthält. Umgekehrt ist es möglich, ausgehend von linguistischen Ressourcen in der Cloud auf thematisch qualifizierte Publikationen im Portal zu verlinken.

Die so erreichte Vernetzung des Portals mit der LOD-Cloud gilt es nun, im Rahmen des FID Linguistik voranzutreiben. Im laufenden Projekt wird der Bestand an Sprachschlagwörtern[12] für LOD aufbereitet und mit Online-Repositorien verlinkt. Als passende Anker in der LLOD-Cloud wurden Glottolog[13] und Lexvo[14] ausgewählt. Es handelt sich dabei um zwei frei verfüg-bare Metadaten-Repositorien, die nach LOD-Prinzipien aufbereitet sind, sich jedoch konzeptionell und in Bezug auf ihren Umfang grundsätzlich unterscheiden. Glottolog liefert eine detaillierte Klassifikation, die Verwandtschaftsbeziehungen zwischen den Sprachvarietäten abbildet und auf Quellenmaterial basiert. Lexvo hingegen definiert persistente Identifier für die Codes aus der ISO-639-Gruppe, ohne sich auf bibliografische Nachweise zu beziehen, und verzichtet dabei auf eine systematische Sprachklassifikation nach genealogischen Prinzipien. Beide Repositorien liefern zusätzliche Informationen wie geospatiale Daten oder fremdsprachliche Benennungen und sind mit weiteren Ontologien und Terminologien vernetzt.

Die LOD-Modellierung der BLL-Sprachbezeichner ist dabei mit spezifischen Problemen verbunden. Insbesondere die Beziehungen zwischen den einzelnen Sprachen und die Gruppierung von Varietäten werden in der wissenschaftlichen Literatur häufig kontrovers diskutiert. Die Herausforderung für den FID Linguistik besteht darin, eine wissenschaftlich adäquate Klassifikation der Sprachvarietäten zu erstellen, die traditionelle Beschreibungen und etabliertes Wissen abbildet, im Kontext des Linguistik-Portals von praktischem Nutzen ist und gleichzeitig den Semantic-Web-Standards entspricht. Im Zuge dieser Arbeiten wird auch intellektuell geprüft, ob Sprachbezeichner in allen drei Systemen deckungsgleich sind oder in welchem Verhältnis sie jeweils zueinanderstehen. Dies schlägt sich auf formaler Ebene in unterschiedlichen Link-Stärken nieder.

Die dadurch erzielte Interoperabilität zwischen den Repositorien wird zu einer wesentlichen Erweiterung und Optimierung der bestehenden LOD-Suche führen – dies reicht von der Einbindung zusätzlicher Sprachressourcen bis zur kombinierten Facettierung der Suchergebnisse hinsichtlich Annotation und Sprache. Durch die LOD-Vernetzung werden die BLL-Sprachbegriffe darüber hinaus um Informationen angereichert, die weitere Anwendungsszenarien ermöglichen. So könnten zum Beispiel die von Glottolog verzeichneten geospatialen Informationen als Basis für eine geografisch geleitete Suche nach Ressourcen zu konkreten Sprachen dienen.

Automatisierte Sacherschließung digitaler Sprachressourcen

Die im vorangegangenen Abschnitt beschriebene Suchfunktionalität ist auf Ressourcen beschränkt, die schon LOD-Formalismen zur Repräsentation der Daten verwenden. Obwohl deren Zahl kontinuierlich steigt, handelt es sich hierbei noch immer lediglich um einen Bruchteil der elektronisch verfügbaren Sprachressourcen, die in den letzten 50 Jahren angelegt wurden.

Deshalb soll die für das Linguistik-Portal entwickelte, LOD-basierte Suchfunktionalität auf Sprach-

ressourcen ausgedehnt werden, deren Dateninhalte zwar noch nicht als LOD verfügbar sind, deren Metadaten jedoch bereits LOD-konform bereitgestellt werden. Zur Erreichung dieses Ziels werden Methoden entwickelt, mit denen geeignete inhaltliche Metadaten hinsichtlich der verwendeten linguistischen Terminologie automatisch zugeordnet werden können. Dabei wird von folgender Prämisse ausgegangen: Wenn für ein gegebenes Korpus bekannt ist, welches Modell zur linguistischen Annotation benutzt wurde, dann folgt daraus, dass Begriffe der diesem Modell zugrundeliegenden Terminologie als potentielle Sucheinstiege für dieses Korpus betrachtet werden können.

Als Ausgangsbasis dienen Korpora, die in etablierten Metaportalen[15] verzeichnet sind. Wenn hier für ein Korpus keine Angabe über das verwendete Annotationsmodell in den Metadaten zu finden ist, wird eine Datenprobe gezogen und analysiert. Im besten Fall wird ein bereits in das bestehende LOD-Netz integriertes Annotationsmodell erkannt, sodass dem Korpus die zugehörigen linguistischen Begriffe unmittelbar als Sucheinstiege zugeordnet werden können. Für den Fall, dass die automatisierte Erkennung nur mit einer relativen Wahrscheinlichkeit erfolgen kann, wird dies im Resultat gekennzeichnet und für eine intellektuelle Nachkontrolle vorgemerkt. In den Fällen, in denen keine unmittelbare Erkennung erfolgen kann, wird experimentell eine Technologie erprobt, bei der vorliegende Annotationen mithilfe maschinellen Lernens automatisch auf bereits in das bestehende Netz integrierte Annotationsmodelle abgebildet werden.

Darüber hinaus wird für Ressourcen ohne Angaben zur Objektsprache die Sprache des Korpus mit existierenden Verfahren ermittelt. Die so erzeugten inhaltlichen Metadaten (Annotationsmodell und Objektsprache) werden in die Recherchefunktion des Linguistik-Portals eingebaut und unter einer CC0-Lizenz zur Verfügung gestellt. Zusätzlich erlaubt die Einrichtung einer geeigneten Schnittstelle die Nachnutzung der erzeugten Daten in maschinenlesbarer Form, sodass die gewonnenen Informationen insbesondere auch an die Metaportale zurückfließen können, die als Ausgangspunkt der Recherche dienen.

Auf diese Weise wird die Auffindbarkeit von Sprachressourcen unter Ausnutzung von LOD-Mechanismen verbessert. Die im Hintergrund wirkenden, aufwendigen Verfahren resultieren in einer einfach zu bedienenden, nutzerfreundlichen Suchabfrage im Linguistik-Portal.

Sichtbarkeit von Forschungsdaten und wissenschaftlichen Publikationen

Zur Verbesserung der Sichtbarkeit von linguistischen Forschungsdaten und ihrer wissenschaftlichen Analyse wurde im FID Linguistik ein bibliografisches Teilprojekt aufgesetzt: Einschlägige Publikationen aus der Bibliography of Linguistic Literature werden auf Metadatenebene mit den jeweils behandelten Sprachressourcen verlinkt. Im Zuge dessen werden für die einzelnen Ressourcen Normdatensätze erstellt. Diese Normdatensätze sollen direkt auf die Webpräsenz der Sprachressource verlinken – im Idealfall geschieht dies über einen Persistent Identifier wie URN, DOI oder Handle, um dauerhafte Nutzbarkeit zu gewährleisten. Eine Herausforderung stellen alle Fälle dar, in denen keine stabile Adressierbarkeit gegeben ist. Hier bedarf es kontinuierlicher Linkpflege, um die Funktionalität aufrechtzuerhalten. Die Frage der (lizenz-)rechtlichen Verfügbarkeit steht in diesem Teilprojekt erst an zweiter Stelle: Zugriffsbeschränkte Ressourcen werden, sofern sie online adressierbar sind, genauso wie frei verfügbare bearbeitet. Forschungsdaten, die jedoch isoliert lokal gespeichert sind, bleiben bei der Normdatenerzeugung außen vor.

Das Resultat dieser Arbeiten soll in einem neuen Modul des Linguistik-Portals präsentiert werden, das es beispielsweise erlaubt, anhand der gewünschten Objektsprache nach Ressourcen zu browsen und zu einer gewählten Ressource unmittelbar die dazugehörige Sekundärliteratur anzuzeigen.

FID-Lizenzen und Literaturversorgung

Eine im Vorfeld der Konzeptionierung des FID-Projektes durchgeführte Bedarfsabfrage in der Fachcommunity brachte zutage, dass weitere Maßnahmen im Bereich der Verfügbarkeit von Forschungsdaten erforderlich sind. Dazu gehört die überregionale Lizenzierung kommerzieller Sprachkorpora.

Die Nutzung von Korpora inkludiert per definitionem Text-Mining-Verfahren – bevorzugt basierend auf der lokalen Speicherung des betreffenden vollständigen Korpus auf dem Computer des Endnutzers. Aufgrund dieser spezifischen Nutzungsart sind die gängigen Lizenzierungsmodelle für E-Books und E-Journals nur bedingt für die überregionale Korpus-Lizenzierung geeignet. Neue Lösungen sind gefragt – die Verhandlungen mit Anbietern multilingualer Sprachressourcen dauern an.

Jenseits aller innovativen Unternehmungen besteht jedoch bei Wissenschaftlerinnen und Wissenschaftlern der Wunsch, dass die Tradition der Literaturversorgung fortgeführt wird – allerdings unter der Maßgabe der bedarfsorientierten Umsetzung.[16] Dies bedeutet für den FID Linguistik, die Selektionsrichtlinien des bisherigen Sondersammelgebiets »Allgemeine und Vergleichende Sprachwissenschaft. Allgemeine Linguistik« anzupassen. Die Veränderungen beim fachlichen Zuschnitt betreffen in erster Linie die reduzierte Beschaffung von Literatur zu interdisziplinären Themen und Randbereichen der Linguistik sowie aus wenig etablierten Publikationsregionen. Zur Schärfung des Erwerbungsprofils gehört die Fokussierung auf die Kernbereiche der Linguistik, jedoch ebenso die Öffnung für hochspezielle Veröffentlichungen einzelphilologischer Natur.

Neben kostenpflichtigem Literaturerwerb spielen Open-Access-Veröffentlichungen in der Linguistik eine immer größer werdende Rolle.[17] Vor diesem Hintergrund möchte sich der FID auch über die Bereitstellung einer technischen Infrastruktur für Open-Access-Zeitschriften als Dienstleister für Sprachwissenschaftlerinnen und Sprachwissenschaftler positionieren. Eingesetzt wird die quelloffene Software Open Journal Systems (OJS)[18] des Public Knowledge Project, die es ermöglicht, den gesamten redaktionellen Begutachtungs- und Publikationsprozess einer elektronischen Zeitschrift browserbasiert abzubilden.

Fazit

Das neue Fördersystem der Fachinformationsdienste bietet eine größere Flexibilität und schafft Raum für innovative Vorhaben, sodass die jeweiligen Bedürfnisse der Wissenschaftlerinnen und Wissenschaftler stärker in die Konzeptionierung der Infrastrukturmaßnahmen einbezogen werden können. Die enge Kooperation mit der Wissenschaft ist dabei nicht nur die Voraussetzung für die Bedarfsermittlung, sondern – wie im vorliegenden Fall die Zusammenarbeit mit dem Institut für Informatik – auch für die Umsetzung der konkreten Projektvorhaben.

Da der FID Linguistik sich im ersten Drittel der Projektlaufzeit befindet, ist es zu früh, um Bilanz zu ziehen. Aber schon jetzt ist klar, dass es sehr wünschenswert wäre, wenn das Förderprogramm verstetigt würde. Denn gerade in den Bereichen Linked Open Data und Forschungsdaten stellt die rasante technische Entwicklung eine Herausforderung dar. Ohne eine Fortsetzung des Programms wird es schwierig sein, die Portale in dem bis dahin geförderten Umfang weiterzuentwickeln. Um den Veränderungen der Forschungslandschaft gerecht zu werden und Wissenschaftlerinnen und Wissenschaftlern eine infrastrukturelle Unterstützung in ihrer Arbeit anbieten zu können, bedarf es kontinuierlicher technischer Aktualisierungen und neuer maßgeschneiderter Services.

Anmerkungen

1 www.linguistik.de/
2 Die erste Ausbaustufe vollzog sich in den Jahren 2012–2014 und die zweite in den Jahren 2015–2017.
3 http://linghub.org/
4 https://datahub.io/
5 CLARIN Virtual Language Observatory https://vlo.clarin.eu
6 www.meta-share.org/
7 http://linguistic-lod.org/llod-cloud
8 BLL – Linguistic Linked Open Data Edition. Erstveröffentlichung am 22.12.2016 unter: http://data.linguistik.de/bll/index.html
9 Die Jahrgänge 1971 bis 2000 sind lizenzfrei.
10 Größenordnung 1.900 Normdatensätze.
11 Vgl. Chiarcos, Christian und Maria Sukhareva: OLiA – Ontologies of Linguistic Annotation. In: Semantic Web Journal 6 (2015) 4, S. 379–386. Verfügbar unter: http://semantic-web-journal.net/system/files/swj518_0.pdf
12 Größenordnung 2.200 Normdatensätze.
13 Vgl. Nordhoff, Sebastian und Harald Hammarström: Glotto log/Langdoc: Defining dialects, languages, and language families as collections of resources. In: Proceedings of the First International Workshop on Linked Science 2011 (LISC2011), Bonn, Germany, October 24, 2011, S. 53–58. Verfügbar unter: www.mpi.nl/publications/escidoc-1752673
14 Vgl. de Melo, Gerard: Lexvo.org: Language-Related Information for the Linguistic Linked Data Cloud. In: Semantic Web Journal 6 (2015) 4, S. 393–400. Verfügbar unter: http://semantic-web-journal.net/system/files/swj521.pdf
15 Vgl. Anmerkungen 3–6.
16 Der bedarfsorientierte Erwerb wird durch ein zu diesem Zweck online geschaltetes Formular für Kaufvorschläge unterstützt, das auch überregional unmittelbaren Einfluss auf den Literaturerwerb im FID erlaubt.
17 Vgl. die erfolgreiche Gründung des Verlags Language Science Press http://langsci-press.org/
18 Vgl. www.ojs-de.net/ beziehungsweise https://pkp.sfu.ca/ojs/

Die Verfasserin
Heike Renner-Westermann M.A., Fachreferentin für Linguistik, Leitung Bibliography of Linguistic Literature, Fachinformationsdienst Linguistik, Universitätsbibliothek Johann Christian Senckenberg,
Bockenheimer Landstraße 134–138,
60325 Frankfurt am Main,
Telefon +49 69 798-39235,
h.renner-westermann@ub.uni-frankfurt.de
Foto: privat