Statement der *Gesellschaft für Angewandte Linguistik* (GAL e. V.) zum Aufbau wissenschaftsgeleiteter Forschungsinfrastrukturen für die Geistes- und Kulturwissenschaften

Gesellschaft für Angewandte Linguistik e.V.

Michel Beißwenger[1], *Markus Bieswanger*[2], *Matthias Knopp*[3] *und Bernd Meyer*[4] im Namen der Gesellschaft für Angewandte Linguistik (GAL e.V.), September 2018

Relevanz digitaler Forschungsinfrastrukturen im Bereich der angewandtlinguistischen Forschung

Gerade die angewandt-linguistische Forschung bezieht ihre Erkenntnisse und ihre Theorie- und Modellbildung zu fachlichen Gegenständen in hohem Maße aus der empirischen Arbeit mit authentischen Sprachdaten. Die Bandbreite untersuchter Daten reicht dabei von Text- und Audiodaten (z. B. Gesprächsaufzeichnungen) bis hin zu komplexen multimodalen Datentypen (videografierte Interaktionen, Webgenres/Social Media etc.). Qualitative und quantitative Ansätze sind gleichermaßen von Interesse. Existierende Forschungsinfrastrukturen (beispielsweise Text- und Gesprächskorpora und Korpora internetbasierter Kommunikation sowie darauf bezogene Abfrage- und Analysewerkzeuge), wie sie u. a. im Rahmen von CLARIN und DARIAH bereitgestellt werden, bilden bereits eine gute Grundlage für eine Reihe von datengestützten Untersuchungen.

Zu vielen der im Bereich der Angewandten Linguistik behandelten Themen und interessierenden Datentypen gibt es allerdings noch keine frei verfügbaren Ressourcen, so dass in vielen Fällen Forschende darauf angewiesen sind, selbst geeignete Datensets zu erheben. Dies bedeutet einen hohen konzeptionellen und auch technischen Aufwand, der als Voraussetzung für die eigentliche Forschungsarbeit zu leisten ist. Die Konzeption und der Aufbau eines Korpus kann zwar selbst als wichtige Forschungsarbeit und -leistung verstanden werden; für viele NachwuchswissenschaftlerInnen im Bereich der Angewandten Linguistik, die mit begrenztem Zeit- und Finanzbudget ein Forschungsprojekt voranbringen möchten, werden in aller Regel aber die auf die Analyse ihrer Daten bezogenen linguistischen Forschungsfragen im Vordergrund stehen. Auch verfügen viele NachwuchswissenschaftlerInnen, je nach thematischer Ausrichtung, nicht vollumfänglich über die Expertise und Ressourcen, derer es bedarf, um ein Datenset/Korpus so zu repräsentieren und für Forschungszwecke aufzubereiten, dass es später auch für Dritte von Wert und Nutzen sein könnte (z. B. Erschließung durch Metadaten, Repräsentation in Übereinstimmung mit Standards im Bereich der Digital Humanities, linguistische Annotation).

Datensets/Korpora, die im Rahmen von Forschungsarbeiten entstehen, zu deren Gegenstand noch keine im Rahmen von Forschungsinfrastrukturen verfügbaren Datensets/Korpora existieren, können aber einen wichtigen Ausgangspunkt für die Erweiterung der Ressourcenlandschaft bilden. Um das zu leisten, ist es aus Sicht der GAL wichtig, Unterstützungsangebote von Verbundprojekten wie CLARIN und DARIAH, die auf die Einbindung von Nutzerinnen und Nutzern aus den adressierten Scientific Communities zielen, zu stärken und in den jeweiligen Communities bekannt zu machen und zu bewerben.

[1] Universität Duisburg-Essen, Institut für Germanistik (*michael.beisswenger@uni-due.de*)
[2] Universität Bayreuth, Englische Sprachwissenschaft *(bieswanger@uni-bayreuth.de)*
[3] Universität zu Köln, Institut für deutsche Sprache und Literatur II *(matthias.knopp@uni-koeln.de)*
[4] Johannes Gutenberg-Universität Mainz, Translations-, Sprach- und Kulturwissenschaft *(meyerb@uni-mainz.de)*

Auswirkungen für Lehre und Weiterbildung

Die Arbeit mit digitalen Sprachdaten spielt immer stärker auch in der angewandt-linguistischen Lehre und Weiterbildung eine Rolle. Digitale Sprachressourcen (Sprachkorpora, lexikalisch-semantische Ressourcen, digitale Nachschlagewerke etc.) bieten exzellente Möglichkeiten für Konzepte des forschenden Lernens einerseits in der Hochschuldidaktik, andererseits, vermittelt über die Lehramtsstudiengänge, für die Sprachreflexion und -analyse in der Schule. Daneben spielt die Arbeit mit authentischen Sprachdaten in der Weiterbildung eine wichtige Rolle, zum Beispiel die Arbeit mit Gesprächsaufzeichnungen und -transkripten in Kommunikationstrainings für einzelne Berufsfelder. Hier wird bislang vor allem qualitativ und mit kleinen, häufig von den Akteuren selbst erstellten Datensets gearbeitet.

Zusätzlich wünschenswert: Die stärkere Heranführung von angehenden (⇨ Lehramtsstudiengänge) und bereits ausgebildeten Lehrerinnen und Lehrern (⇨ Lehrerweiterbildungen) wie auch von Akteuren im Bereich der beruflichen Weiterbildung an die Möglichkeiten von digitalen Sprachressourcen(infrastrukturen) dürfte einerseits den professionellen Akteuren neue Möglichkeiten eröffnen, in der Lehre und Weiterbildung mit authentischen Daten zu arbeiten, andererseits für CLARIN und DARIAH eine Nutzergruppe erschließen, die bislang erst zu Teilen über digitale Infrastrukturen informiert ist. Letztlich betrifft dieser Punkt auch die Frage, wie das Themengebiet Digitale Sprachressourcen und Forschungsinfrastrukturen in die Curricula von Studiengängen integriert werden kann – eine Frage, zu der es in DARIAH und CLARIN ja auch schon Überlegungen und Vorarbeiten gibt.

Auswirkungen für Begutachtung und Antragstellung

Existierende Sprachressourcen und -infrastrukturen bilden eine wichtige Grundlage, um empirisch fundierte Forschungsprojekte auf den Weg zu bringen. Wo mit schon vorhandenen Ressourcen gearbeitet werden kann, müssen Ressourcen nicht erst als Teil der Projektarbeit neu aufgebaut werden. Die Arbeit mit existierenden, frei verfügbaren Ressourcen erhöht zudem die Nachvollziehbarkeit und Reproduzierbarbeit empirischer, wissenschaftlicher Ergebnisse durch Dritte.

Für diverse Forschungsthemen und -felder ist die Ressourcenlage derzeit allerdings unzureichend. Entweder ist die Forschungsfrage oder ist der interessierende Datentyp zu speziell, oder der Datentyp ist – zum Beispiel, weil er zum gegenwärtigen Zeitpunkt nicht ohne Weiteres mit etablierten Standards aus dem Bereich der Digital Humanities erfasst werden kann – noch nicht hinreichend in Ressourceninfrastrukturen präsent. Häufig sind die vorhandenen Daten und/oder die Auswertungsmöglichkeiten hier auch nicht „reichhaltig“ genug, beispielsweise fehlen nicht selten soziolinguistische Parameter. Ein Beispiel wären Korpora zu Social-Media-Genres bzw. Genres internetbasierter Kommunikation. Hier hat sich die Ressourcenlage in den letzten Jahren bereits verbessert und es gibt verschiedene Initiativen (u. a. in CLARIN), die Ressourcenlandschaft zu diesem Thema zu erweitern; da das Thema aber in vielen Disziplinen (neben der Linguistik auch z. B. in den Sozial- und Medienwissenschaften) bereits beforscht wird und sich Kommunikationsformen und -praktiken in diesem Bereich schnell verändern, besteht hier auch künftig ein hoher Bedarf an der Weiterentwicklung existierender und am Aufbau neuer Ressourcen.

Für die Erhebung und Haltung eigener Forschungsdatensets/-korpora in geförderten Projekten hat die DFG als Teil ihrer fachspezifischen Empfehlungen zum Umgang mit Forschungsdaten[5] zwei umfangreiche Handreichungen zu rechtlichen Aspekten[6] und zu datentechnischen Standards[7] bei der Erhebung und Handhabung von Sprachkorpora veröffentlicht, die 2012/2013

[5] http://www.dfg.de/foerderung/antrag_gutachter_gremien/antragstellende/antragstellung/nachnutzung_forschungsdaten/index.html
[6] http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_recht.pdf
[7] http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf

unter Beteiligung u. a. verschiedener Wissenschaftlerinnen und -wissenschaftler aus dem Umfeld der GAL erarbeitet wurden.´

Bezogen auf Datentypen: Dazu ist in den vorangehenden Punkten bereits einiges gesagt. Technisch: Die Infrastruktur sollte auch für Nutzerinnen und Nutzer ohne computerlinguistische bzw. informatische Vorbildung mit möglichst wenig Einarbeitung nutzbar sein. Tutorials sollten verständlich sein für sprachwissenschaftliche Nutzerinnen und Nutzer, einschließlich der Studierenden. Workshops und andere Disseminationsangebote, bei denen an konkreten Beispielen (Forschungsfragen) aus der Scientific Community die Möglichkeiten von Forschungsinfrastrukturen erläutert werden, wären wünschenswert; denkbar wären hier beispielsweise im Umfeld der Tagungen der GAL angesiedelte Workshops und/oder Angebote für Nachwuchswissenschaftlerinnen und Nachwuchswissenschaftler (z. B. Research Schools). Hierbei könnten Werkzeuge und Plattformen für die sprachtechnologische Aufbereitung/automatische linguistische Annotation an Nutzergruppen herangebracht werden, die auf ihrem Karriereweg mit Sprachtechnologie bislang eher nicht in Berührung gekommen sind.

Technisch auch sehr wünschenswert: Nutzerschnittstellen zu Sprachressourcen sollten nicht nur die Abfrage und den Export von Suchanfragen erlauben, sondern auch den weiteren Forschungsprozess – der ja erst nach der Korpusabfrage beginnt – unterstützen. Hier von Interesse: Möglichkeiten zum Speichern und Weiterbearbeiten von Abfrageergebnissen in der Korpusanalyseumgebung und zur Annotation der Ergebnisse; bei Export: Möglichkeit, den Beleg im Korpus wiederzufinden.

Maßnahmen der GAL für eine stärkere Vernetzung der Community der angewandt-linguistisch Forschenden mit dem Bereich der digitalen Forschungsinfrastrukturen

Eine stärkere Vernetzung der angewandt linguistischen Forschung mit dem Bereich Forschungsinfrastrukturen/Sprachressourcen ist aus Sicht der GAL sehr wünschenswert und im Sinne einer Verbesserung der Forschungsmöglichkeiten für empirisch Forschende im Bereich der Sprachwissenschaften. Der Vorstand der GAL hat in seiner Sitzung vom 10.09.2018 beschlossen, die Vernetzung der beiden Bereiche durch Einrichtung eines neuen Forschungsfokus „Digitale Infrastrukturen für die Angewandte Linguistik“ zu unterstützen, der einerseits durch thematisch einschlägige Veranstaltungen und durch die Dokumentation von Beispielen guter Praxis für die Nutzung digitaler Ressourcen in linguistischen Forschungsarbeiten die Potenziale existierender Korpora und Werkzeuge im GAL-Kontext stärker verankern, andererseits die Bedürfnisse linguistischer Nutzerinnen und Nutzer an Akteurinnen und Akteure aus Infrastrukturinitiativen und Ressourcenzentren kommunizieren soll. Dabei soll insbesondere auch der schon erreichte Stand der empirischen Arbeit mit digitalen Ressourcen und Korpora in den verschiedenen thematischen Sektionen der GAL dokumentiert und für die Weiterentwicklung der angewandt-linguistischen Forschung sichtbar gemacht werden. Die Aktivitäten des Forschungsfokus werden von GAL-Mitgliedern koordiniert, die bereits über Expertise, Kontakte und Kooperationen im Bereich der digitalen Forschungsinfrastrukturen verfügen.