Offene Forschung

Medienlinguistische Methodik

Arne Rubehn
arne.rubehn@uni-passau.de

20.01.2026

Zusammenfassung

In dieser Sitzung wollen wir uns mit einigen Grundlagen der offenen Forschung befassen und dabei vor Allem auch auf konkrete Lösungsvorschläge eingehen, die es ermöglichen, die eigene Forschung selbst offen zu gestalten.

1 Grundlagen

Mit **offener Forschung** (*Open Science*) bezeichnet man eine Haltung in der Forschung, die nach maximaler Offenheit alle Prozesse, die zu neuen Erkenntnissen führen, strebt. Die offene Forschung wird von vielen Menschen auf der Welt inzwischen mehr oder weniger radikal propagiert, wobei man häufig auch argumentiert, dass Forschung an sich unabhängig von Nationalitäten und anderen Aspekten durchgeführt werden sollte, weil sie mit ihren grundlegenden Prinzipien über die Idee von Nationalstaaten und ähnliche politische Strukturen hinausreichen sollte. Natürlich lässt sich diese Haltung angesichts von miteinander in Konkurrenz oder sogar im Krieg stehender Nationen, wie wir sie zu Beginn des 21. Jahrhunderts vorfinden, kaum realistisch aufrecht erhalten. Vor allem im Zusammenhang mit der Konkurrenz zwischen den großen militärischen Mächten und dem vermeintlichen Siegeszug der Forschungs zur künstlichen Intelligenz ist Forschung in bestimmten Bereichen weit davon entfernt, *offen* zu sein.

In kleinen, in ihrer Ausrichtung weitgehend harmlosen Forschungsbereichen, wie auch der multilingualen Computerlinguistik oder der digitalen Geisteswissenschaften, lässt sich dieses Ideal der offenen Forschung, die maximal vielen Menschen frei zugänglich sein sollte, jedoch in vielen Fällen auch in seiner radikalen Form aufrecht erhalten. Dass Forschung dabei jedoch strukturelle Ungleichheiten über Ländergrenzen hinweg auch praktisch zu überbrücken vermag, muss jedoch leider stark bezweifelt werden, auch wenn es natürlich wünschenswert wäre, wenn dies wirklich realisiert werden könnte.

Im Zusammenhang mit der offenen Forschung werden verschiedene Aspekte diskutiert. Dazu gehört bspw. die freie Verfügbarkeit von Materialien (Artikeln, Lehrbüchern), die im Rahmen der *Open Access*-Bewegung verfolgt wurde. Diese wurde inzwischen von den Verlagen jedoch so vereinnahmt, dass zwar viele Artikel frei verfügbar sind, während sich die Ungleichheiten durch die komplizierten Publikationsstrukturen jedoch weiter erhöht haben. Denn für Open Access müssen jetzt die Institutionen oftmals horrende Summen zahlen, während die Verlage ihre Services gleichzeitig drastisch heruntergefahren haben.

Da Open Access heute grundlegend auf sehr einfache Art gewährleistet werden kann, bspw. indem man seine Artikel auf offenen Server archiviert, ist es inzwischen wichtiger, die offenen Aspekte, die vor der Publikation von Artikeln zum Tragen kommen, oder jenseits der Artikel wichtig sind, unter die Lupe zu nehmen. Dazu gehören vor allem auch die offenen Daten (*Open Data*) und der offene Code (*Open Code*), da ja viele Untersuchungen

auf Daten und Code beruhen, wobei Daten und Code nach wie vor sehr häufig nicht offen geteilt werden.

Wir können als Grundziel der offenen Forschung festhalten, dass Forschung wiederholbar, transparent und integrierbar sein sollte. Im Folgenden gehen wir gesondert auf die drei Punkte ein.

1.1 Wiederholbarkeit

Wiederholbarkeit bezieht sich darauf, dass eine Forschung, die man selbst durchführt, auch von anderen durchgeführt können werden sollte. Das Wiederholen ist dabei in zweierlei Hinsicht wichtig. Zum einen gibt es anderen Menschen die Möglichkeit, von einem bestimmten Ansatz zu lernen und diesen zu erweitern. Zum anderen kann man durch das Wiederholen sicherstellen, dass die Messungen, die vorgenommen wurden, auch tatsächlich unabhängig von der Person die forscht, Bestand haben.

1.2 Transparenz

Um Wiederholbarkeit zu gewährleisten, muss insbesondere auch Transparenz vorausgesetzt werden. Transparenz zielt hier darauf ab, dass die Forschung so beschrieben wurde, dass man klar erkennen kann, welche Schritte durchgeführt und welche Daten zugrunde gelegt wurden. Transparenz wird leider auch unabsichtlich oft vernachlässigt, indem wichtige Aspekte einer Untersuchung verschleiert werden. Bei Computeranalysen sind dies vielleicht bestimmte Parameter, mit denen eine Software aufgerufen wird, oder auch die Form, die Daten haben müssen. Transparenz ist in vielen Fällen eine Grundlage von Wiederholbarkeit, jedoch nicht ausschließlich, da wir uns auch Fälle vorstellen können, wo ein Forschungsansatz wiederholbar ist, aber nicht besonders transparent gestaltet wurde. Dies trifft vor allem dann zu, wenn Software eine Forschung begleitet und auch von anderen verwendet werden kann, die Software mit ihren Grundannahmen selbst aber nur unzulänglich beschrieben wird.

1.3 Integrierbarkeit

Unter Integrierbarkeit wird hier verstanden, dass Forschung, die zu einem Thema durchgeführt wurde, sich nicht nur in den Ergebnissen sondern auch in den Methoden mit anderen Forschungsansätzen integrieren lassen können sollte. Hier ist es vor allem wichtig, dass man es mit der eigenen Forschung anderen ermöglicht, diese in neue Forschungsansätze einzubauen. Grundlage ist dabei die Transparenz eines Forschungsansatzes, aber auch die Nutzung von Standards, oder die Nutzung von modularem Programmiercode,

oder von Tools, die grundsätzlich auch anderen zugänglich sind. Es verbietet sich die Nutzung proprietärer Formate, die von anderen Menschen nicht verwendet werden können, da sie beispielsweise teure Lizenzen voraussetzen.

2 Offene Daten

2.1 FAIRness von Daten

Wilkinson et al. (2016) haben den Begriff der FAIRness von Daten etabliert, der sich auf deren Auffindbarkeit (*findability*), Zugänglichkeit (*accessibility*), Integrierbarkeit (*interoperability*) und Wiederverwertbarkeit (*reusability*) beziehen. Bei Wilkinson werden diese vier Aspekte unterschiedlich beschrieben, sie sind jedoch stark miteinander verzahnt und auch nicht immer eindeutig voneinander zu trennen.

2.2 Daten auffindbar machen (*Findability*)

Um Daten auffindbar zu machen, muss man sie so gestalten, dass sie so gut beschrieben werden, dass man sie beispielsweise über Suchmaschinen finden kann. Dabei ist es auch wichtig, dass man die Daten in Archiven anbietet, die leicht durchsucht werden können. Auffindbarkeit kann man als das schwächste Glied in der FAIRnis-Kette ansehen, weil es nicht nur von denen abhängt, die Forschung betreiben, sondern auch von der Qualität von zur Verfügung stehenden Archiven und dergleichen.

2.3 Daten zugänglich machen (*Accessibility)*

Dass man auf Daten zugreifen kann ist eine Grundvoraussetzung für offene Wissenschaft. Dabei gibt es natürlich zuweilen auch rechtliche und ethische Einschränkungen bei personenbezogenen oder sicherheitspolitisch relevanten Daten. In der Linguistik und den Geisteswissenschaften allgemein haben wir es oft mit freien Daten zu tun, die weder sicherheitspolitisch relevant sind, noch Angaben zu Personen enthalten, die noch leben. Daher lässt sich dieser Aspekt der Daten oft ausschließen, was dann aber auch heißt, dass es kaum eine Entschuldigung geben kann, wenn Daten nicht frei zugänglich gemacht werden.

2.4 Daten integrierbar machen (*Interoperability*)

Integrierbarkeit wurde schon erwähnt und bezieht sich vor allem darauf, dass man Daten, die in einem Ansatz verwendet wurden, gern in anderen Ansätzen weiterverwenden will,

oder dass man Daten verschiedener Quellen kombinieren möchte. Hier ist es wichtig, dass Daten kombinierbar und integrierbar gemacht werden, was vor allem mit Hilfe verbindlicher Standards erreicht werden kann. Wer seine Daten interoperabel machen möchte, tut daher gut daran, sich mit klaren Standards zu beschäftigen, die für den Datentypen in der Wissenschaft etabliert worden sind. Nur mit solchen Standards ist das Aggregieren verschiedener Datensätze dann auch in transparenter Form möglich.

2.5 Daten wiederverwertbar machen (*Reusability*)

Die Wiederverwertbarkeit von Daten setzt vor allem voraus, dass sie offene Lizenzen bekommen, die das Wiederverwerten erlauben. So gibt es Datensätze, die zwar frei zur Verfügung stehen, aber eben mit einer Lizenz geteilt werden, die es schlicht verbietet, Teile aus diesen Daten herauszunehmen, oder sie in egal welcher anderen Art modifizierend zu übernehmen. Solche Daten sind aus legaler Perspektive unbrauchbar für weitere Forschungsarbeiten, weshalb man sie besser gleich und bewusst ignorieren sollte.

Im Kontext von Forschungsdaten fällt die Wahl zumeist auf *Creative Common*-Lizenzen (CC-BY), die grundlegend die Weiterverwendung von Daten erlauben. Urheber:innen können hierbei modular festlegen, welche Formen der Weiterverwendung gestattet sind – z.B. ob die kommerzielle Nutzung gestattet ist oder Derivate aus den Daten erstellt werden dürfen. Unter CC-BY-Lizenzen müssen die Urheber:innen bei jeglicher Weiterverwendung zitiert werden.

3 Offener Code

Auch in Bezug auf Code sollten wir uns der Offenheit verschreiben. Das heißt auch hier, dass wir den FAIR-Prinzipien versuchen sollten zu genügen. In bestimmten Details unterscheidet sich Programmiercode hier aber vom Forschungsdaten, die wir uns im Folgenden kurz anschauen.

3.1 Code auffindbar machen (*Findability*)

Bei der Auffindbarkeit trifft auch zu, dass man darauf achten sollte, dass Code entdeckt werden kann. Die Zielportale sind hier jedoch andere. Während man bei Daten gern öffentliche Archive verwendet, sind es bei Code oft die Package-Indizes, die für verschiedene Programmiersprachen kuratiert und zur Verfügung gestellt werden. So sollte man, wenn man Python programmiert, zum Beispiel den Code nach Möglichkeit als Bibliothek im Python Package Index (PyPI, `https://pypi.org`) zur Verfügung stellen.

3.2 Code nutzbar machen (*Accessibility*)

Zugänglichkeit bei Code heißt vor allem, dass man ihn frei zur Verfügung stellt, so dass er heruntergeladen und auch direkt installiert werden kann. Das das Installieren oftmals die größte Hürde bei Code ist, ist es hier besonders wichtig, auf verschiedenen Plattformen zu testen und sicherzustellen, dass der Code auch ohne Probleme dort installiert werden kann und auch funktioniert.

3.3 Code integrierbar machen (*Interoperability*)

Für die Integrierbarkeit von Code ist es wichtig, ihn in Form von Packages anzubieten (*Software Libraries*). Denn wenn man eine Library installieren kann, dann kann man sie danach auch im eigenen Code frei verwenden und in die eigenen Arbeiten integrieren. Hier gibt es einige Regeln zu beachten, die das Package-Management von Softwarebibliotheken betreffen, die wir uns in diesem Zusammenhang aber nicht genauer anschauen müssen.

3.4 Code nachhaltig machen (*Reusability*)

Auch bei der Wiederverwendbarkeit von Code spielen Lizenzen eine wichtige Rolle. Zusätzlich nutzt fast jeder Code, den man schreibt, den Code von anderen, weshalb es hier auch wichtig ist, zu schauen, welchen Lizenzen der Code, den man verwendet, underliegt. Allgemein sollte kein Code ohne Nutzungslizenz geteilt werden – hier greifen nämlich standardmäßig sehr restriktive Regelungen des Urheberrechts, die die Weiterverwendung durch Dritte rechtlich nahezu unmöglich macht. Im Falle offener Forschung sollte die Lizenz maximal frei sein, auch hier eignen sich meistens CC-BY-Lizenzen.

4 Standardisierung von Daten

Auch das Thema der Standardisierung haben wir bereits kurz angeschitten. Hierunter verstehen wir im Bezug auf Daten eine unabhängige Vorgabe, nach der man sich richtet, wenn man Daten bearbeitet. Einfache Beispiele für Standards sind Dateiformate wie CSV oder JSON, die klar vorschreiben, wie Daten zu formatieren sind. Dadurch wird gewährleistet, dass verschiedene Programme die Daten problemlos einlesen können. Neben formalen Vorgaben können Standards auch definieren, welche Datenpunkte oder Metadaten inkludiert werden müssen.

4.1 Referenzkataloge

Wenn wir in unserer Forschung mit Forschungsobjekten arbeiten, die als Konstrukte klar definiert und auch von der Forschungsgemeinschaft als klare Einheiten aufgefasst werden, dann lohnt es sich oft nicht, die Information, die wir über diese Objekte haben, unabhängig zusammenzutragen. Stattdessen ist es leichter, auf einheitliche Kataloge zu verweisen, die diese Informationen bereits anbieten. Ein solches Beispiel ist Glottolog (Hammarström, Forkel, Haspelmath & Bank, 2025) als Katalog von Sprachen, der Informationen zu Sprachen zusammenträgt und frei zur Verfügung stellt. Weitere Beispiele, die hier in Passau am Lehrstuhl für Multilinguale Computerlinguistik angesiedelt sind, sind das Concepticon (`https://concepticon.clld.org`, List et al. 2025) und die *Cross-Linguistic Transcription Systems* (`https://clts.clld.org`, Anderson et al. 2018). Im ersten Fall liefert das Concepticon Definitionen und Beispiele von semantischen Glossen in sprachwissenschaftlichen Sammlungen (vor allem Wortlisten). Im zweiten Fall liefert der CLTS-Katalog die Möglichkeit, Sprachlaute, die einem aus dem IPA abgeleiteten Standard entsprechen, zu generieren.

4.2 Cross-Linguistic Data Formats

Im Rahmen der „Cross-Linguistic Data Formats“-Initiative (CLDF, Forkel et al. 2018, `https://cldf.clld.org`) versuchen wir, explizit Datensätze anzulegen, die Sprachdaten standardisiert zur Verfügung stellen. Dabei haben wir begon- nen, größere Datensätze anzulegen, die vor allem Wortlisten (Blum et al., 2025) und Sammlungen grammatischer Merkmale umfassen (Skirgård et al., 2023). Grundlage von CLDF ist der Standard CSVW (`https://csvw.org`), welcher es ermöglicht, CSV-Tabellen mit Metadaten zu versehen, was dazu führt, dass Datentypen genau festgelegt und auch entsprechend festgelegt werden können. Wenn man Daten dann anlegt, die den Datentypen nicht folgen, so kann eine Validierung diese Fehler direkt aufzeigen.

5 Archivierung von Daten

Wenn wir Datensätze angelegt haben und diese veröffentlichen wollen, müssen wir dies so tun, dass deren Langlebigkeit gesichert ist. Hier ist es wichtig, darauf zu achten, dass viele Formen, die einem auf den ersten Blick sinnvoll zu sein scheinen, nicht langlebig sind, weil das Bestehen der Daten nicht in öffentlicher Hand liegt. Dazu gehört es zum Beispiel, Da- ten auf GoogleDrive zu speichern und öffentliche Links zur Verfügung zu stellen. Auch DropBox oder andere Cloud-Speicherdienste sind nicht lang- lebig, geschweige denn eigene Webseiten. Darüber hinaus sind natürlich wieder die bereits diskutierten FAIR-Prinzipien zu beachten.

5.1 Versionierung

Wenn man Daten bearbeitet, dann sollte man sie *versionieren*. Dabei geht es darum, dass man sicherstellen kann, welchen Zustand die Daten zu bestimmten Zeitpunkten aufweisen, wie sie sich entwickeln, aber auch, um sicherzustellen, dass man weiß, welche Analysen mit welchem Zustand der Daten durchgeführt wurden. Oft werden Datensätze zwar im Web ohne Version veröffentlicht, es ist allerdings sehr schlechte Forschungspraxis, dies zu tun, denn auf diese Art verliert man die Möglichkeit, FAIR zu arbeiten und zusätzlich replizierbare Ergebnisse zu liefern.

Zur Versionierung von Daten gibt es verschiedene Tools. Inzwischen ist GIT (`https://git-scm.com/`) eines der am weitesten verbreiteten Tools geworden, vor allem auch, weil es prominent durch Projekte wie GitHub (`https://github.com`) oder GitLab (`https://gitlab.com`) unterstützt wird.

5.2 Open Science Framework

Das Open Science Framework (`https://osf.io`) ist eine gemeinnützige Organisation, die verschiedene Services zur Verfügung stellt, um Offene Forschung zu betreiben. Dazu gehört auch die Möglichkeit, Daten online zu speichern und sie beispielsweise anderen Forschenden per anonymem Link zur Verfügung zu stellen. Dies ist dann sinnvoll, wenn man Artikel einreicht und anonyme Gutachten benötigt. Das Projekt kann dadurch nicht von Suchmaschinen gefunden werden, der Link erlaubt aber direkten Zugriff. Das Open Science Framework empfiehlt sich also besonders für das Teilen von Daten, wenn Artikel begutachtet werden sollen.

5.3 Zenodo

Zenodo (`https://zenodo.org`) ist einer der größten Services für das Teilen von Forschungsdaten, den wir in Europa haben. Finanziert wird Zenodo von der Europäischen Union und ist angesiedelt an das CERN. Es wird davon ausgegangen, dass der Service von Zenodo mindestens für die Laufzeit des CERNs bestehen wird. Viele Forschungsprojekte, die öffentlich gefördert werden, müssen sich an strikte Vorgaben in Bezug auf offene Daten halten. In den meisten Fällen erfüllt man diese Forderungen direkt, wenn man seine Daten über Zenodo teilt. Anonymes teilen ist auch über Zenodo möglich, allerdings ist das weniger einfach und wird seltener verwendet. Daher empfiehlt sich Zenodo eher für die Langzeitarchivierung, die man dann benötigt, wenn man Daten öffentlich teilen möchte. Zenodo bietet inzwischen auch spezielle Communities an, also Gruppen, die Daten veröffentlichen und für eine gewisse Qualität der Daten einstehen, da Services wie Zenodo auch häufig für Spam missbraucht werden. In Communities kann man Forschungsergeb-

nisse teilen oder auch thematisch bestimmte Daten organisieren. Zenodo ermöglicht es darüber hinaus, Daten in GitHub zu kuratieren und dann automatisch bestimmte Versionen in Zenodo zu archivieren.

Literatur

Anderson, C., Tresoldi, T., Chacon, T., Fehn, A.-M., Walworth, M., Forkel, R. & List, J.-M. (2018). A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznań Linguistic Meeting*, *4* (1), 21–53.

Blum, F., Barrientos, C., Englisch, J., Forkel, R., Greenhill, S. J., Rzymski, C. & List, J.-M. (2025). Lexibank 2: pre-computed features for large-scale lexical data [version 2; peer review: 3 approved]. *Open Research Europe*, *5* (126), 1-19. doi: 10.12688/openreseurope.20216.2

Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., ... Gray, R. D. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, *5* (180205), 1-10. doi: 10.1038/sdata.2018.205

Hammarström, H., Forkel, R., Haspelmath, M. & Bank, S. (2025). *Glottolog 5.2.1*. Jena: Max Planck Institute for the Science of Human History. Zugriff auf `https://glottolog.org`

List, J.-M., Tjuka, A., Blum, F., Kučerová, A., Barrientos Ugarte, C., Rzymski, C., ... Forkel, R. (2025). *CLLD Concepticon [Dataset, Version 3.4.0]*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Zugriff auf `https://concepticon.clld.org/`

Skirgård, H., Haynie, H. J., Blasi, D. E., Hammarström, H., Collins, J., Latarche, J. J., ... Gray, R. D. (2023). Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, *9*. doi: 10.1126/sciadv.adg6175

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... others (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 1-9. doi: 10.1038/sdata.2016.18