

Michael Beißwenger (Duisburg-Essen)/Marcel Fladrich  
(Hamburg)/Wolfgang Imo (Hamburg)/Evelyn Ziegler  
(Duisburg-Essen)

## **Die *Mobile Communication Database 2* (*MoCoDa 2*)**

Die MoCoDa 2 (<https://db.mocoda2.de>) ist eine webbasierte Infrastruktur für die Erhebung, Aufbereitung, Bereitstellung und Abfrage von Sprachdaten aus privater Messenger-Kommunikation (WhatsApp und ähnliche Anwendungen). Zentrale Komponenten bilden (1) eine Datenbank, die für die Verwaltung von WhatsApp-Sequenzen eingerichtet ist, die von Nutzer/innen gespendet und für linguistische Recherche- und Analysezwecke aufbereitet wurden, (2) ein Web-Frontend, das die Datenspende/innen dabei unterstützt, gespendete Sequenzen um analyse-relevante Metadaten anzureichern und zu pseudonymisieren, und (3) ein Web-Frontend, über das die Daten für Zwecke in Forschung und Lehre abgefragt werden können. Der Aufbau der MoCoDa-2-Infrastruktur wurde im Rahmen des Programms „Infrastrukturelle Förderung für die Geistes- und Gesellschaftswissenschaften“ vom Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen gefördert. Ziel des Projekts ist es, ein aufbereitetes Korpus zur Sprache und Interaktion in der deutschsprachigen Messenger-Kommunikation bereitzustellen, das speziell auch für qualitative Untersuchungen eine wertvolle Grundlage bildet.



**Abb. 1:** MoCaDa 2-Logo

Eine erste Version der *Mobile Communication Database (MoCoDa 1)* wurde von Marcel Fladrich und Wolfgang Imo 2011 initiiert. Zu Beginn der Datenerhebungen mit der MoCoDa 1 überwogen noch Interaktionen über den Kurznachrichtendienst SMS, ab 2012 wurden zunehmend auch WhatsApp-Daten erfasst. Da die MoCoDa zunächst als autonome ‚Insellösung‘ mit geringen Mitteln aufgebaut wurde und da sich zudem die Messenger-Kommunikation (v. a. in WhatsApp und vergleichbaren Anwendungen) zunehmend von einer primär schriftbasierten Kommunikation zur multimodalen Kommunikation wandelte, war eine grundlegende Neukonzeption der Datenbank erforderlich. Dies wurde möglich durch die Erweiterung

<https://doi.org/10.1515/9783110679885-018>

des Projektteams um Michael Beißwenger und Evelyn Ziegler und die Einwerbung der Infrastrukturförderung durch das Land Nordrhein-Westfalen. Im Rahmen der Neukonzeption wurde das Web-Frontend, über das die Spender/innen in den Prozess der Datenaufbereitung einbezogen werden, deutlich erweitert. Als Neuerungen gegenüber der Vorgängerversion wurde weiterhin ein unterstützter Datenimport<sup>1</sup> aus dem WhatsApp-Messenger umgesetzt; zudem kann die Datenbank nun auch Gruppenchats abbilden. Die Möglichkeiten für die Erfassung von Metadaten zu einzelnen Chats und zu den Interaktionsbeteiligten wurden erheblich erweitert; u.a. lassen sich in der MoCoDa 2 die Beziehungsrelationen sämtlicher Beteiligter spezifizieren. Darüber hinaus erkennt die MoCoDa 2 integrierte Medienobjekte (Bilder, Videos, Sprachnachrichten, Standorte, Sticker etc.) und repräsentiert diese aus datenschutz- und urheberrechtlichen Gründen anhand typisierter Platzhalter, die anschließend über das Web-Frontend um textuelle Beschreibungen (im Falle der Sprachnachrichten Wortlautabschriften oder Transkripte) ergänzt werden können. Die manuelle Pseudonymisierung der Chats durch die Spender/innen wird durch Assistenzfunktionen unterstützt. Über eine Kooperation mit dem *Language Technology Lab* der Universität Duisburg-Essen (Torsten Zesch und Mitarbeiter) wurden zudem Verfahren für die sprachtechnologische Aufbereitung der in der MoCoDa 2 erfassten Daten (Part-of-Speech-Annotation) in die Infrastruktur integriert. Darüber hinaus nutzt die MoCoDa 2 Expertise und Ressourcen aus dem CLARIN-D-Kurationsprojekt „ChatCorpus2CLARIN“ (Lüngen et al. 2016), in dessen Rahmen 2016/17 das Dortmunder Chat-Korpus in die CLARIN-D-Korpusinfrastrukturen integriert und in diesem Zusammenhang um zusätzliche Annotationen erweitert wurde. Eine ausführliche Beschreibung der Konzeption und des aktuellen Stands der MoCoDa 2 bieten Beißwenger et al. (2019).

MoCoDa 2 ist seit Sommer 2018 in Betrieb und ersetzt die vorherige Version. Wie die MoCoDa 1 ist auch die MoCoDa 2 (nach einer Online-Registrierung) für Forschung und Lehre frei zugänglich. Aktuell (Stand: Oktober 2019) enthält das noch junge Korpus 361 Chats mit insgesamt 1.759 Beteiligten, die insgesamt 31.189 Nachrichten bzw. 243.350 Token umfassen. Die Datenbank bietet einen einfachen Zugang mit einer intuitiven Recherche sowie Exportfunktionen und ist auf kontinuierliche Erweiterung ausgelegt: Jederzeit können weitere Daten gesendet und anhand des Web-Frontends für die Integration in die Datenbank aufbereitet werden. Die Daten lassen sich beim Recherchezugriff umfassend nach Metadaten (für die Chatbeteiligten: Alter, Geschlecht u.a.; für die einzelnen Chats: Kommunikationsanlass, Teilnehmerzahl) filtern.

---

1 Siehe dazu ausführlich <https://db.mocoda2.de/c/input>.

Für die Zukunft ist eine jährliche Überführung der Daten in das Deutsche Referenzkorpus (DEREKO) am Leibniz-Institut für Deutsche Sprache (IDS) geplant. Dadurch sollen die Daten auch über die Korpusrechercheschnittstellen des IDS zugänglich gemacht und dort mit Sprachdaten aus anderen in DEREKO vorhandenen Textsorten und Kommunikationsbereichen vergleichend ausgewertet werden können.

Der weitere Ausbau der Ressource steht und fällt mit der Zahl der Datenspenden. Alle Leser/innen und möglichen künftigen Nutzer/innen sind herzlich eingeladen, mit der Spende eigener WhatsApp-Sequenzen dazu beizutragen, dass auf der Basis der MoCoDa 2 die linguistische Erforschung von Sprache und Kommunikation in der privaten Messenger-Kommunikation auf breiter und stets aktueller Datengrundlage ermöglicht werden kann.



**Abb. 2:** QR-Code mit der Weiterleitung zur Webseite der MoCoDa 2

## Literatur

- Beißwenger, Michael/Fladrich, Marcel/Imo, Wolfgang/Ziegler, Evelyn (2019): *https://www.mocoda2.de*: A database and web-based editing environment for collecting and refining a corpus of mobile messaging interactions. In: *European Journal for Applied Linguistics* 7, 2, S. 1–12. Internet: <https://doi.org/10.1515/eujal-2019-0004>.
- Lüngen, Harald/Beißwenger, Michael/Herold, Axel/Storrer, Angelika (2016): Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In: Dipper, Stefanie/Neubarth, Friedrich/Zinsmeister, Heike (Hg.): *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*. Bochum, S. 156–164. Internet: [www.linguistics.rub.de/konvens16/pub/20\\_konvensproc.pdf](http://www.linguistics.rub.de/konvens16/pub/20_konvensproc.pdf) (Stand: 16.8.2019).