

## Project Report

Michael Beißwenger\*, Wolfgang Imo, Marcel Fladrich and Evelyn Ziegler

# ***https://www.mocoda2.de*: a database and web-based editing environment for collecting and refining a corpus of mobile messaging interactions**

<https://doi.org/10.1515/eujal-2019-0004>

**Abstract:** This paper reports on findings from the *MoCoDa2* project which is creating a corpus of private CMC interactions from smartphone apps based on donations by their users. Different from other projects in the field, the project involves users not only as donators but also as editors of their data: In a web-based editing environment which provides users with access to their raw data, they are supported in pseudonymising their data and enhancing them with rich metadata on the interactional context, meta-data on the interlocutors and their relations, and on embedded media files. The resulting corpus will be a useful resource not only for quantitative but also for qualitative CMC research. For representation and annotation of the data the project builds on best practices from previous projects in the field and cooperates with a language technology partner.

**Keywords:** computer-mediated communication, CMC, corpora, data collection

---

\*Corresponding author: **Michael Beißwenger**, University of Duisburg-Essen, Essen, Germany,  
E-Mail: michael.beisswenger@uni-due.de

**Wolfgang Imo**, University of Hamburg, Hamburg, Germany,  
E-Mail: wolfgang.imo@uni-hamburg.de

**Marcel Fladrich**, University of Hamburg, Hamburg, Germany,  
E-Mail: marcel.fladrich@uni-hamburg.de

**Evelyn Ziegler**, University of Duisburg-Essen, Essen, Germany,  
E-Mail: evelyn.ziegler@uni-due.de

# 1 Introduction

In recent years, there has been an increasing amount of work dedicated to the creation of corpora of computer-mediated communication (CMC) that shall be made available as resources for the scientific community through established corpus infrastructures (e.g. CLARIN<sup>1</sup>) and through adapting to standards in the field of Digital Humanities (Beißwenger et al. 2017a, Beißwenger et al. 2017b).

A desideratum in the CMC corpora landscape are resources that represent CMC discourse from the private sphere and that allow for research of discourses found in applications such as WhatsApp and similar mobile chat and messaging services which are frequently used by adolescents (cf. e.g. KIM 2016, JIM 2016). Data of that type as well as the metadata needed for an adequate interpretation (topic and context of the interaction; age, sex, languages used in familial contexts, and socio-demographic background of interlocutors; social relations between interlocutors) can only be collected with the help of the users themselves.

We report on findings from the project *MoCoDa2 (Mobile Communication Database)*<sup>2</sup>, which is funded by the Ministry for Innovation, Science, Research and Technology of the German federal state North Rhine-Westphalia and in which a team of researchers from two universities has created a database and web front-end for the repeated collection of written CMC from mobile messaging services.

## 2 Related work

Collections of CMC data from the private sphere have been addressed in several previous projects: In the DiDi project (Frey et al. 2014), Facebook users were asked to give their permission to collect CMC data from their profile pages via a web application. However, the collection of data from private mobile phones can only be achieved when users actively donate their data by submitting them to a project platform. Practices for donation-based collections of SMS and WhatsApp data have been developed in the *sms4science* project (Dürscheid and Stark 2011)<sup>3</sup>, in the projects “What’s up, Switzerland?” (*WuS*, Ueberwasser and Stark 2017)<sup>4</sup> and “What’s up Deutschland?” (*WuD*), and in the predecessor project *MoCoDa1* (cf. Sect. 4 and Imo 2015, Imo 2017).

---

1 <http://clarin.eu/>

2 <https://www.mocoda2.de>

3 <http://www.sms4science.org/>

4 <http://www.whatsup-switzerland.ch/>

### 3 Data collection and editing design

MoCoDa2 adopts a donation-based collection strategy. Different from the aforementioned projects, the CMC users are not only involved as donators but also as editors of their donated data: The data collection component allows users to donate selected parts of their private WhatsApp interactions via email and then to log into a web-based editing environment in which they can edit their donations, pseudonymise the data, and enhance them with relevant metadata to transform them into valuable contributions to a corpus that can be a useful resource both for quantitative and qualitative research on CMC. The Language Technology Lab at the University of Duisburg-Essen, the language technology partner in the project, provides an infrastructure which splits the data up into tokens (tokenisation) and adds part-of-speech information to the data. The tokenisation task is performed as a preparatory task for the editing process, the part-of-speech task is performed after donators have finished editing their donated data.

As a matter of fact, the amount of data donated by a single user is expected to be smaller than the data collected in the WuS and WuD projects mentioned in Sect. 2, where the complete logfiles stored on the donators' mobile phones were submitted to the project API and integrated into the corpora. The goal of the MoCoDa2 project is not to create a corpus which is intended as a competitor to the WuS and WuD corpora; instead, the goal is to create a corpus of interaction sequences which have been manually selected and edited by their donators to provide corpus users with all metadata needed to use the corpus for qualitative and quantitative research. During the editing process users keep full control of their donations and may crop the donated log file to a certain selection. Additionally and in contrast to the aforementioned projects, MoCoDa2 does not perform a one-time collection. Rather, the front-end is used repeatedly so that the size of the corpus will gradually grow over time and – as a long-term objective – will allow for micro-diachronic research on language variation and change in mobile messaging discourse.

Fig. 1 and 2 given below illustrate how the online data-editing process is organised. The language of the interface is German. The import component is able to deal with German, English, French, Dutch, Russian, Turkish, Portuguese, Greek, Italian and Arabic data; so far the language-technology component is able to do tokenization and part-of-speech tagging for German data only.

*Specification of metadata:* The system automatically extracts the names of all participants of an interaction from the donated logfile and encourages the donator to add metadata for each interlocutor. The metadata include information on age, sex, place of residence (city, state, country), place of birth (city, state, country), educational level, profession, language(s) used on a daily basis. The specification of these metadata is voluntary, but donators are encouraged on the project

website to put effort in adding relevant metadata in order to make the donated sequences a valuable basis of analysis for other users of the resource.

In a second step the donator is asked to define the social relations between the individuals that are detected as chat participants pair-by-pair. Relations can be specified according to several predefined dimensions of which more than one may apply and can be assigned to chat partners.

Fig. 1 illustrates how the values can be assigned to a pair of individuals (Stefan and Susanne) via selection from a dropdown menu. The predefined values can be enhanced with textual descriptions (e.g. as given in Fig. 1 “know each other from kindergarten” as an additional explanation of the predefined value ‘is good friends with’). Fig. 2 shows an overview generated by the system on how many relations between the detected chat participants have been specified by a donator within a donated interaction. Fig. 3 shows how the metadata specified for one participant in a group chat (Alexa) is presented on the MoCoDa2 user interface together with a donated sequence. Rows 1–12 contain individual metadata, row 13 includes information on the relations of the individual with the other participants (Olivia, Anna).

The screenshot shows the MoCoDa2 user interface. At the top, there is a navigation bar with the following items: Mobile Communication Database 2, Start, Hinweise, Dateneingabe, Recherchieren, Kontakt, and Blog. Below the navigation bar, the main content area is titled "Stefan -> Susanne". On the left side, there is a vertical menu with the following categories: "Familiäre Beziehungen", "Berufliche Beziehungen", "Beziehungen in Ausbildung und Studium", and "Beziehungen im Bereich Freizeitgestaltung und außerberufliches Engagement". A dropdown menu is open, showing the following options: "ist eng befreundet mit", "ist gut befreundet mit", "ist gut bekannt mit", "andere", "ist Expartner von", "in einer festen Partnerschaft mit", "ist Cousin/Cousine von", "ist Neffe/Nichte von", and "ist eng befreundet mit". To the right of the dropdown menu, there are three text input fields, each with a "Weitere Informationen" label. The first field contains the text "kennen sich seit dem Kindergarten". At the bottom left of the form, there is a "Speichern" button.

**Fig. 1:** Screenshot: Using the assistant for the specification of social relations between interlocutors. The image shows dropdown menus which allow the donator to specify the social relation between the two interlocutors Stefan and Susanne in several dimensions: type of familial relation (*Persönliche Beziehung*, e.g. “is closely befriended with”, “is nephew/niece of”, “is cousin of”, “is in a stable relationship with”, “is ex-partner of” etc.), type of professional relation (*Berufliche Beziehung*), type of relation in an educational context (*Beziehungen in Ausbildung und Studium*) and type of relation w.r.t. leisure activities and voluntary commitment (*Beziehungen im Bereich Freizeitgestaltung und außerberufliches Engagement*).



**Fig. 2:** Screenshot: Overview of relations specified for all couples of interlocutors in a donated sequence. The image shows how many types of relations have been specified for five participants of a group chat; e.g. two for Stefan-Susanne and one for Stefan-Peter.

*Pseudonymisation:* When entering metadata for the interaction participants, donators are asked to assign a pseudonym to each of the participants following the following rule:

“Please make sure to choose a realistic name as a pseudonym that resembles the gender and origin of the original name. A person’s place of birth or residence only needs to be anonymized if it is a very small place.”

The authors’ names which are automatically rendered as part of the metadata for each post in the chat logfile are automatically replaced by a pseudonym specified by the donator. Mentions of authors’ names in the content of posts have to be manually replaced with the respective pseudonym by the donator. In addition, the donators are asked to assign pseudonyms to any other personal names mentioned in the posts. To do so, they can click on a token of their choice and replace it with a pseudonym. Once a certain character string (e.g. “Matthis”) has been replaced by a pseudonym (e.g. “Manuel”), for every other occurrence of the respective character string in all other posts of the sequence the system suggests to replace them by the same pseudonym accordingly. The donator can agree with the suggestion by clicking on the green “check” symbol that appears along with the automatically generated suggestion (Fig. 4: system suggests to replace the character string “Matthis” by “Manuel” because the donator has replaced a pre-

vious occurrence of “Matthis” by “Manuel”). During the process the system stores a temporary list of all pseudonyms previously used by the donator; this list is deleted from the database when the donator declares the editing process as finished. Besides the pseudonymisation of person and – if needed – city names, the donators are asked to anonymise URLs by replacing them with the category ‘URL’ and add the type of target resource the original URL has been pointing at (e. g. “social website”, “shopping website”).

Further editing steps include the formulation of textual description for media objects (images, videos) included in the original sequence (which, due to copyright restrictions, are not stored in the MoCoDa2), a transcription of audio posts (which are not stored in the database due to IPR restrictions), a textual description of the interaction context and the specification of a brief title for the donated sequence.

Fig. 5 gives an example of how a donated and post-edited sequence is presented on the MoCoDa2 user interface together with the title (“Skatergirls”) and the textual description of the interaction context entered by the donator (left column). The sequence in the screenshot contains a textual description of an image that was embedded in the original data (“Zu sehen ist ein kleiner Pool, ...”). The pseudonyms of the participants given in the bottom of the left column under “Teilnehmer” link to representations of the participant metadata as given in Fig. 3.

**Fig. 3:** Presentation of metadata for one interlocutor on the MoCoDa2 user interface (German).

Name:	<b>Alexa</b>
<i>Name:</i>	
Alter:	26–30
<i>Age:</i>	
Geschlecht:	weiblich
<i>Sex:</i>	
Wohnort Land:	Deutschland
<i>Place of residence (country):</i>	
Wohnort Bundesland:	Nordrhein-Westfalen
<i>Place of residence (federal state):</i>	
Wohnort Stadt:	Essen
<i>Place of residence (city):</i>	
Geburtsort Land:	Deutschland
<i>Place of birth (country):</i>	
Geburtsort Bundesland:	Nordrhein-Westfalen
<i>Place of birth (federal state):</i>	

Muttersprache(n): <i>First language(s):</i>	Polnisch
Alltagsprache(n): <i>Other languages used in everyday life:</i>	Polnisch, Deutsch
Höchster Bildungsabschluss: <i>Highest educational level:</i>	Abitur
Berufsgruppe: <i>Professional group:</i>	Student/in
Weitere Informationen zum Beruf: <i>Further information on profession:</i>	Arbeitet außerdem seit 3 Monaten Vollzeit als wissenschaftliche Mitarbeiterin an ihrer Universität.
Beziehungen: <i>Relations with other participants of the chat:</i>	<p><i>Olivia:</i></p> <ul style="list-style-type: none"> <li>– Alexa ist befreundet mit Olivia</li> <li>– Alexa ist Kollege oder Kollegin von Olivia</li> </ul> <p><i>Anna:</i></p> <ul style="list-style-type: none"> <li>– Alexa ist eng befreundet mit Anna</li> <li>– Alexa ist Kollege oder Kollegin von Anna</li> </ul>



Fig. 4: Screenshot: Using the pseudonymisation assistant.

## 4 Resources and technology

*MoCoDa2* builds on the expertise and resources from three preceding projects:

- *MoCoDa1* (Imo, 2015; 2017) – a corpus project with a similar profile which had been initiated to collect a corpus for the qualitative analysis of CMC<sup>5</sup>. Since 2012, this project has collected a (relatively small) data set of 2,198 interactions with 19,161 user posts with ~193,000 tokens. For *MoCoDa2*, the database and web front-end have been re-implemented from scratch, and especially the editing environment was supplemented with a lot of additional functions and features.
- *ChatCorpus2CLARIN* (Lüngen et al., 2016) – a curation project in the context of the German CLARIN-D initiative in which the Dortmund Chat Corpus (Beißwenger, 2013), a well-established CMC corpus for German, has been remodelled following up-to-date standards for corpus resources in the digital humanities and integrated into the CLARIN language resources infrastructure<sup>6</sup>. The project has developed guidelines for anonymisation of CMC resources (Lüngen et al., 2017) and a schema for the representation of CMC corpora building on the TEI-P5 encoding guidelines of the Text Encoding Initiative (TEI)<sup>7</sup>. The schema and the anonymisation guidelines are adopted for representing and editing CMC data in *MoCoDa2*.
- *EmpiriST 2016* – a shared task on tokenisation and part-of-speech tagging of German CMC/social media data (Beißwenger et al., 2016) which developed guidelines for tokenisation and a part-of-speech tag set for German CMC ('STTS 2.0'<sup>8</sup>) and which resulted in a number of tokenisers and taggers which had been adapted or retrained to fit the linguistic peculiarities of CMC discourse. The tools and tag set used for the annotation of the *MoCoDa2* data are built on the EmpiriST resources and results.

---

<sup>5</sup> <http://mocoda.spracheinteraktion.de/>

<sup>6</sup> The corpus can be retrieved via the repositories of the Institute for the German Language (IDS) at <http://hdl.handle.net/10932/00-0379-FDFE-CC30-0301-E> and of the Berlin-Brandenburg Academy of Sciences (BBAW) at <http://hdl.handle.net/11858/00-203Z-0000-002D-EC85-5>. It can be queried online via the COSMAS II interface of the German Reference Corpus (DEREKO) and – after a free registration – via the text corpora component of <http://www.dwds.de>, the online information system on German language provided by the BBAW.

<sup>7</sup> The schema can be retrieved via <http://wiki.tei-c.org/index.php?title=SIG:CMC/clarindschema>. A detailed description is given in Beißwenger (2018).

<sup>8</sup> The tag set and annotation guidelines can be retrieved via <https://sites.google.com/site/empirist2015/home/annotation-guidelines>.



The technological backbone of the project is a *mongoDB* database which performs fast enough to execute processing operations relevant during the online editing process in real-time. In view of the large amount, data have to be processed in short time (raw text, tokenisation, annotations, metadata). The technology has completely been built on a *JavaScript* base with *Node* using the *Angular* framework. The system uses different microservices in order to handle certain operations like parsing, tokenising, annotating and indexing as different processes. The *docker* technology allows us to use load balancing and thus provide an optimal performance while importing larger amounts of data or handling several donations and editing processes simultaneously. Our beta tests have shown that the system can import even long logfiles quite fast and provide users with all information needed for editing their donations in (almost) real time.

## 5 Results so far

A beta version of the data collection and editing component has been subject of testing and optimisation in several university classes at three German universities during the summer term 2018 and winter term 2018/19. Currently, an open beta version is online which can be used and rested by everybody. Up to now we have collected 331 sequences from WhatsApp logs with 26,879 user posts and 225,357 tokens. Based on the user experience feedback collected from students it is planned to develop extensions which provide enhanced query options and to extend the set of ethnographical meta data which are collected during the editing process.

Sequences submitted to the database are checked for unethical content by the project members on a regular basis so that potentially offending content or sequences which were obviously not completely pseudonymised can be removed if necessary.

As a mid-term goal it is planned to integrate the data into the German Reference Corpus *DEREKO* at the Institute for the German Language (IDS) in Mannheim (Lüngen, 2017). Once the corpus has become part of *DEREKO* users will be able to use and query it via the query interfaces available for the IDS corpora collection. In the meantime, querying the corpus is possible via the query interface provided as part of the open beta version at <https://db.mocoda2.de/#/search>.

The screenshot displays the MoCoDa2 interface, which is divided into two main sections: a metadata sidebar on the left and a chat window on the right.

**Metadata Sidebar (Left):**

- Titel:** Skatergirls
- Kontext:** Olivia, Alexa und Anna kennen sich von ihrer ehemaligen Arbeit. Olivia arbeitet immer noch dort und muss manchmal bis 2 Uhr nachts arbeiten, so wie auch an dem Tag dieses Gesprächs. Alle drei wohnen in einem Stadtteil und knüpfen an ihre Fähigkeit zu skateboarden an, was sie seit etwa 10 Jahren nicht mehr getan haben. Sie sind also alle Anfänger, aber Olivia fährt am häufigsten und auch am besten. Mit "Sandsturm" meint Olivia ihre Bluetooth-Musikbox, die sie aufgrund eines Insiders so getauft hat. Olivia kommt außerdem aus Mülheim und war dort zwischendurch, Freunde und ihre Eltern zu besuchen und weil ihre Eltern einen kleinen Pool haben, ist sie bei dem sehr heißen Wetter dort geblieben. Nach dem schwülen und heißen Wetter freut sie sich auf das Unwetter - in der Hoffnung, dass sich die Luft verbessert.
- Messenger:** WhatsApp
- Anzahl Nachrichten:** 40
- Zeitraum:** 27.05.2018 - 29.05.2018
- Teilnehmer (3):**
  - Olivia
  - Anna
  - Alexa

**Chat Window (Right):**

- Gruppenchat - #qaw95**
- Participants:** Alexa
- Search:** Schnellsuche
- Message #1 (11:10):** Olivia: Ich fahr mal los
- Message #2 (11:17):** Alexa: Bin schon da
- Message #3 (11:18):** Olivia: Sandsturm mitnehmen?
- Message #4 (11:18):** Olivia: Ist Ma drin
- Message #5 (11:18):** Olivia: Bg
- Message #6 (11:24):** Anna: Ich geh so in 10 min raus hatte keine Motivation aufzustehen
- Date Separator:** Mo. 28. Mai 2018
- Message #7 (10:30):** Olivia:  Zu sehen ist ein kleiner Pool, der von der Sonne angestrahlt wird. Im Wasser schwimmt ein blauer Wasserreifen. Bin jetzt erstmal bei meinen Eltern Sonnen, meldet euch heute Abend mal was ihr macht 🍷

Fig. 5: Presentation of a donated sequence on the MoCoDa2 user interface.

## 6 References

- Beißwenger, Michael. 2013). Das Dortmunder Chat-Korpus. *Zeitschrift für germanistische Linguistik* 41(1), 161–164.
- Beißwenger, Michael. 2018. Internetbasierte Kommunikation und Korpuslinguistik: Repräsentation basaler Interaktionsformate in TEL. In Henning Lobin, Roman Schneider & Andreas Witt (eds.), *Digitale Infrastrukturen für die germanistische Forschung*, 307–349. Berlin & New York: de Gruyter.

- Beißwenger, Michael, Sabine Bartsch, Stefan Evert & Kay-Michael Würzner. 2016. EmpiriST 2015: A Shared Task on the Automatic Linguistic Annotation of Computer-Mediated Communication and Web Corpora. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task* (ACL Anthology W16–2606), 44–56. Stroudsburg: Association for Computational Linguistics. <http://aclweb.org/anthology/W/W16/W16–2606.pdf> (accessed 29 January 2019).
- Beißwenger, Michael, Thierry Chanier, Tomaž Erjavec, Darja Fišer, Axel Herold, Nikola Lubešić, Harald Lungen, Céline Poudat, Egon Stemle, Angelika Storrer & Ciara Wigham. 2017a. Closing a Gap in the Language Resources Landscape: Groundwork and Best Practices from Projects on Computer-mediated Communication in four European Countries. In Lars Borin (ed.), *Selected papers from the CLARIN Annual Conference 2016. Aix-en-Provence, 26–28 October 2016* (Linköping University Electronic Conference Proceedings 136), 1–18. Linköping. <http://www.ep.liu.se/ecp/contents.asp?issue=136> (accessed 29 January 2019).
- Beißwenger, Michael, Ciara Wigham, Carole Etienne, Darja Fišer, Holger Grunt Suárez, Laura Herzberg, Erhard Hinrichs, Tobias Horsmann, Natali Karlova-Bourbonus, Lothar Lemnitzer, Julien Longhi, Harald Lungen, Lydia-Mai Ho-Dac, Christophe Parisse, Céline Poudat, Thomas Schmidt, Egon Stemle, Angelika Storrer & Torsten Zesch. 2017b. Connecting Resources: Which Issues Have to be Solved to Integrate CMC Corpora from Heterogeneous Sources and for Different Languages? In Egon W. Stemle & Ciara R. Wigham (eds.), *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora 2017). Bolzano, Italy, Oct 03–04, 2017*, 52–55. <https://cmc-corpora2017.eurac.edu/proceedings/> (accessed 29 January 2019).
- Chanier, Thierry, Céline Poudat, Benoit Sagot, Georges Antoniadis, Ciara Wigham, Linda Hriba, Julien Longhi & Djamel Seddah. 2014. The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *Journal of Language Technology and Computational Linguistics*, 29(2), 1–30. <https://jclcl.org/content/2-allissues/6-Heft2-2014/1Chanier-et-al.pdf> (accessed 29 January 2019).
- Dürscheid, Christa & Elisabeth Stark. 2011. sms4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland. In Crispin Thurlow & Kristine Mroczek (eds.), *Digital Discourse. Language in the New Media*, 299–320. Oxford, UK: Oxford University Press.
- Frey, Jennifer-Carmen, Egon W. Stemle & Aivars Glaznieks. 2014. Collecting Language Data of Non-Public Social Media Profiles. In Gertrud Faaß & Josef Ruppenhofer (eds.), *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, 11–15. Hildesheim: Universitätsverlag Hildesheim.
- Imo, Wolfgang. 2015. Vom Happen zum Häppchen... Die Präferenz für inkrementelle Äußerungsproduktion in internetbasierten Messengerdiensten. *Networx* 69, 1–35. <http://www.mediensprache.net/de/networx/networx-69.aspx> (accessed 29 January 2019).
- Imo, Wolfgang (2017): Interaktionale Linguistik und die qualitative Erforschung computervermittelter Kommunikation. In Michael Beißwenger (ed.), *Empirische Erforschung internetbasierter Kommunikation* (Empirische Linguistik / Empirical Linguistics 9), 81–108. Berlin & New York: de Gruyter.
- [JIM 2016] Medienpädagogischer Forschungsverbund Südwest (ed.). 2016. *Jugend, Information, (Multi-)Media. Basisuntersuchung zum Medienumgang 12-19-Jähriger*. <http://www.mpfs.de/de/studien/jim-studie/2016/> (accessed 29 January 2019).

- [KIM 2016] Medienpädagogischer Forschungsverbund Südwest (ed.). 2016. *KIM-Studie. Kindheit, Internet, Medien. Basisuntersuchung zum Medienumgang 6-13-Jähriger*. <http://www.mpfs.de/de/studien/kim-studie/2016/> (accessed 29 January 2019).
- Lüngen, Harald. 2017. DeReKo – Das Deutsche Referenzkorpus. Schriftkorpora der deutschen Gegenwartssprache am Institut für Deutsche Sprache in Mannheim. *Zeitschrift für germanistische Linguistik* 45(1), 161–170.
- Lüngen, Harald, Michael Beißwenger, Axel Herold & Angelika Storrer. 2016). Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project Chat-Corpus2CLARIN. In Stefanie Dipper, Friedrich Neubarth & Heike Zinsmeister (eds.), *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, 156–164. [https://www.linguistics.rub.de/konvens16/pub/20\\_konvensproc.pdf](https://www.linguistics.rub.de/konvens16/pub/20_konvensproc.pdf) (accessed 29 January 2019).
- Lüngen, Harald, Michael Beißwenger, Laura Herzberg & Cathrin Pichler. 2017. Anonymisation of the Dortmund Chat Corpus 2.1. In Egon W. Stemle & Ciara R. Wigham (eds.), *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora 2017)*. Bolzano, Italy, Oct 03–04, 2017, 21–24. <https://cmc-corpora2017.eurac.edu/proceedings/> (accessed 29 January 2019).
- [TEI P5] TEI Consortium (ed.). 2007. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. <http://www.tei-c.org/Guidelines/P5/> (accessed 29 January 2019).
- Ueberwasser, Simone & Elisabeth Stark. 2017. What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik online* 84(5), 105–126. <https://bop.unibe.ch/linguistik-online/article/view/3849> (accessed 29 January 2019).

This text is made available via DuEPublico, the institutional repository of the University of Duisburg-Essen. This version may eventually differ from another version distributed by a commercial publisher.

**DOI:** 10.1515/eujal-2019-0004  
**URN:** urn:nbn:de:hbz:465-20240604-160415-7

This publication is with permission of the rights owner freely accessible due to an Alliance licence and a national licence (funded by the DFG, German Research Foundation) respectively.

Beißwenger, M., Imo, W., Fladrich, M. & Ziegler, E. (2019). <https://www.mocoda2.de>: a database and web-based editing environment for collecting and refining a corpus of mobile messaging interactions. *European Journal of Applied Linguistics*, 7(2), 333-344. <https://doi.org/10.1515/eujal-2019-0004>

© 2019 Walter de Gruyter GmbH, Berlin/Boston. All rights reserved.