Volume 29 — Number 2 — 2014 — ISSN 2190-6858

JLCL
Journal for Language Technology
and Computational Linguistics

Building and Annotating Corpora of Computer-Mediated Communication: Issues and Challenges at the Interface of Corpus and Computational Linguistics

Herausgegeben von/Edited by
Michael Beißwenger, Nelleke Oostdijk,
Angelika Storrer, Henk van den Heuvel

GSCL Gesellschaft für Sprachtechnologie & Computerlinguistik

Contents

Impressum

Herausgeber	Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL)
Aktuelle Ausgabe	Band 29 – 2014 – Heft 2
Gastherausgeber	Michael Beißwenger, Nelleke Oostdijk, Angelika Storrer, Henk van den Heuvel
Anschrift der Redaktion	Lothar Lemnitzer Berlin-Brandenburgische Akademie der Wissenschaften Jägerstr. 22/23 10117 Berlin lemnitzer@bbaw.de
ISSN	2190-6858
Erscheinungsweise	2 Hefte im Jahr, Publikation nur elektronisch
Online-Präsenz	www.jlcl.org

Editorial

Computer-mediated communication (CMC) is an umbrella term used for interpersonal communication mediated through computer networks and accessed via personal computers and/or mobile devices. Examples of CMC genres are written conversations in chats, online forums or instant messaging applications, tweets, comments on weblogs, conversations on Wikipedia talk pages and on "social network" sites (*Facebook* etc.), interactions in multimodal communication environments such as Skype, online role-playing games (MMORPGs) or *SecondLife*, SMS messages, or conversations via smart phone "apps" such as *WhatsApp* or *Threema*.

In the past two and a half decades, the use of CMC genres has become an important part of everyday communication. To support empirical research on these new forms of communication, standard text corpora need to be supplemented by linguistically annotated corpora covering the language use in CMC. Nevertheless, there have been no standards thus far for the representation of the structural peculiarities of CMC genres. In addition, it has become consistently apparent that NLP tools trained on written standard language (e.g. on newspaper corpora) do not perform in a satisfactory manner on CMC data.

This special issue of the JLCL gathers five contributions of scholars and projects who aim to close the "CMC gap" in the corpora landscape for several European languages: the French CoMeRe project (Chanier et al.), the South Tyrolian DiDi project (Glaznieks & Stemle), the German Wikipedia corpus (Margaretha & Lüngen), the Dutch VU Chat corpus (Spooren & van Charldorp), and the research on language variation in Dutch Twitter data (van Halteren & Oostdijk).

Thierry Chanier, *Céline Poudat*, *Benoit Sagot*, *Georges Antoniadis*, *Ciara Wigham*, *Linda Hriba*, *Julien Longhi* and *Djamé Seddah* present a TEI representation schema centered on the model of 'interaction space' that has been applied to four French corpora. The schema extends the scope of annotation issues to environments in which several CMC modalities (e.g. chat, email, and forums) are used simultaneously.

Aivars Glaznieks and *Egon W. Stemle* describe the DiDi project where they work with German CMC corpora of internet users from the Italian province of Bolzano – South Tyrol. One focus of this project is on the question of how L1 German speakers of South Tyrol from different age groups are using variants of German and other languages when communicating on social networking sites. The authors describe an approach for collecting Facebook postings and report on experiments to improve POS tagging results on their data by means of normalization.

Eliza Margaretha and *Harald Lüngen* present an approach developed at the IDS Mannheim for the transformation of Wikipedia articles and talk pages into TEI-based corpora for integration in the *German Reference Corpus* (Deutsches Referenzkorpus, DeReKo). The article's focus lies on issues of representing the conversations on talk pages in TEI. The authors describe a method for automatically segmenting these conversations into user postings and discuss the findings that arise from evaluating the segmentation results.

Wilbert Spooren and *Tessa van Charldorp* describe the design and data collection strategies of a chat corpus which is part of the SoNaR reference corpus for contemporary Dutch. To avoid the problem of collecting chat data "in the wild" with unclear legal status, the

authors created a setting in which data could be collected from a chatroom for secondary school students with the consent of both the participating pupils and their parents. The authors explain the logistical, ethical and technological challenges they encountered during the collection of the data and discuss general considerations regarding CMC data collection that can be derived from their experiences.

Hans van Halteren and *Nelleke Oostdijk* present results from their experiments in automatically estimating the proportions of word tokens in Dutch tweets that are not covered by standard resources and can therefore be expected to cause problems for standard NLP applications. Based on a fully annotated pilot corpus, the authors present a detailed typology of types of non-word tokens, out-of-vocabulary tokens and in-vocabulary tokens whose form deviates from standard Dutch. The annotated corpus was used to calibrate automatic estimation procedures which were then applied to about 2 billion Dutch tweets. The discussion of their results is an important foundation for getting a better picture of challenges that are faced in adapting NLP tools to the peculiarities of (Dutch) CMC data.

The idea for this special issue developed from an international workshop "*Building Corpora of Computer-Mediated Communication: Issues, Challenges, and Perspectives*", which was held at TU Dortmund University in February 2013 in connection with the German DFG network *"Empirical research of Internet-Based Communication"* (*Empirikom*)[1] and with financial support from the *Global Young Faculty* program of the Mercator Research Center Ruhr. The workshop brought together CMC corpus projects from several European countries. The participants discussed issues in collecting, representing, annotating and processing CMC data with the common goal of improving interoperability between CMC resources for different languages on the one hand, and between CMC corpora and standard text corpora on the other hand. The workshop resulted in the formation of a network of CMC corpus projects and a joint application for the installation of a special interest group (SIG) "Computer-mediated communication" in the *Text Encoding Initiative* (*TEI*)[2], which was approved by the TEI Council in the autumn of 2013. The articles by *Eliza Margaretha* and *Harald Lüngen* and *Thierry Chanier et al.* are explicitly related to this initiative; the ongoing work in this SIG is documented on the SIG pages in the TEI wiki[3].

Our gratitude goes to the colleagues who contributed to this special issue as external reviewers. We also thank Lothar Lemnitzer for supporting us while editing and finalizing the issue.

Dortmund, Mannheim and Nijmegen, December 2014

Michael Beißwenger
Nelleke Oostdijk
Angelika Storrer
Henk van den Heuvel

[1] http://www.empirikom.net
[2] http://tei-c.org
[3] http://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication