Proceedings

of the 11th Conference on computer-mediated communication and social media corpora

September, 5-6, 2024
University Côte d'Azur
France

Céline Poudat
Mathilde Guernut (eds.)

BASES CORPUS LANGAGE
UMR 7320 CNRS / UNS / UCA

UNIVERSITÉ CÔTE D'AZUR

CORLI

UNIVERSITÉ CÔTE D'AZUR | ÉCOLE UNIVERSITAIRE DE RECHERCHE ARTS ET HUMANITÉS

Université Côte d'Azur FRANCE 2030 Initiative d'Excellence

CMC 2024

11th Conference on Computer-Mediated Communication and Social Media Corpora

Proceedings of the Conference

September 5-6, 2024

https://cmc-corpora-nice.sciencesconf.org/

Proceedings of the 11th International Conference on CMC and Social Media Corpora for the Humanities

05-06 September 2024, Université Côte d’Azur, Nice, France

Editors: Céline Poudat, Mathilde Guernut

Published by: University Côte d’Azur and CORLI

Nice, 2024

Conference website: https://cmc-corpora-nice.sciencesconf.org/

Preface

Following the great success of the tenth conference held in Mannheim, Germany, in 2023, we are very pleased to present the proceedings of the eleventh edition of the International Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC2024). The main focus of the conference is to explore the collection, annotation, processing, and analysis of corpora from computer-mediated communication and social media.

Our general aim is to serve as the meeting place for a wide range of language-oriented investigations into CMC and social media, drawing from linguistics, philology, communication sciences, media studies and foreign language teaching and learning, with research questions stemming from corpus and computational linguistics, language technology, text technology, and machine learning.

The 11th Conference on CMC and Social Media Corpora was held at the Maison des Sciences de l'Homme et de la Société Sud-Est (MSHS) on September 5th and 6th at the University Côte d'Azur in Nice, France. This volume contains 19 accepted papers and the abstracts of the 13 posters presented at the event. Each submission was reviewed by the members of the scientific committee. The contributions were presented across three sessions with two parallel streams, along with a poster session. They cover a broad range of topics, from corpus construction to analysis, including the methods employed in that context.

The program also included two invited talks: an international keynote by Susan Herring (Indiana University, USA), who did us the great honor of attending in person, on the pros and cons of upscaling the model of Computer-Mediated Discourse Analysis she developed; and a national keynote by Marco Cappellini (University of Lyon 1) on corpora in telecollaboration and virtual exchange. This volume contains the abstracts of the invited talks. Additionally, the conference featured a community-building and metadata session, and participants were invited to attend two training workshops on Inception and Iramuteq.

We wish to thank all our colleagues who contributed to the conference and to this volume with their papers and posters. Our thanks also go to the members of the international scientific committee and to our local organizing committee, without whom the conference would not have been possible. We would also like to express our gratitude to the MSHS for kindly hosting us. We are very grateful to the University Côte d'Azur, which provided administrative and financial support: we received several grants, from the University, the Academy of Excellence 5, the Creates Graduate School and our BCL Lab. Thanks also to the national CORLI consortium, which co-organized the conference and provided both administrative and financial support.

We hope that the Nice 2024 conference will foster vibrant exchanges and contribute to strengthening the community of researchers building and using CMC and social media corpora for research in the humanities and beyond.

Nice, August 20th 2024

On behalf of the organizing committee

Céline Poudat, Cécile Angella, Marie Chandelier, Maëlle Debard, Morgane George, Mathilde Guernut, Minerva Rojas, Simona Ruggia

Committees

Organizing Committee

Céline Poudat	BCL, CORLI
Cécile Angella	CORLI
Marie Chandelier	BCL
Maëlle Debard	BCL
Morgane George	BCL
Mathilde Guernut	CORLI
Minerva Rojas	BCL
Simona Ruggia	BCL

International Steering Committee of the Conference series

Steven Coats	University of Oulu
Julien Longhi	Cergy Paris Université
Lieke Verheijen	Radboud University
Reinhild Vandekerckhove	Universiteit Antwerpen

Scientific Committee

Adrien Barbaresi	Berlin-Brandenburgische Akademie der Wissenschaften
Michael Beißwenger	UDE
Mario Cal Varela	Universidade de Santiago de Compostela
Marie Chandelier	Université Côte d'Azur
Steven Coats	University of Oulu
Louis Cotgrove	Leibniz-Institut für Deutsche Sprache
Orphée De Clercq	Ghent University
Susana Doval Suárez	Universidade de Santiago de Compostela
Annamária Fábián	University of Bayreuth
Klaus Geyer	University of Southern Denmark
Francisco Javier Fernández Polo	Universidade de Santiago de Compostela
Jennifer-Carmen Frey	European Academy of Bozen
Aivars Glaznieks	Eurac Research
Jan Gorisch	Leibniz-Institut für Deutsche Sprache
Iris Hendrickx	Radboud University
Axel Herold	Berlin-Brandenburgische Akademie der Wissenschaften
Laura Herzberg	Universität Mannheim
Mai Hodac	Université de Toulouse
Pawel Kamocki	Leibniz-Institut für Deutsche Sprache
Alexander König	CLARIN ERIC
Florian Kunneman	Vrije Universiteit Amsterdam
Marc Kupietz	Leibniz-Institut für Deutsche Sprache
Gudrun Ledegen	Université Rennes 2
Els Lefever	Ghent University
Julien Longhi	Cergy Paris Université
Paula López Rúa	Universidade de Santiago de Compostela
Harald Lüngen	Leibniz-Institut für Deutsche Sprache

Jean-Philippe Magué	ENS Lyon
Elsa María González Álvarez	Universidade de Santiago de Compostela
Maja Miličević Petrović	University of Bologna
Nelleke Oostdijk	Radboud University
Ignacio Palacios Martínez	Universidade de Santiago de Compostela
Céline Poudat	Université Côte d’Azur
Thomas Proisl	Friedrich-Alexander Universität Erlangen-Nürnberg
Ines Rehbein	Universität Mannheim
Sebastian Reimann	Ruhr-Universität Bochum
Minerva Rojas	Université Côte d’Azur
Simona Ruggia	Université Côte d’Azur
Tatjana Scheffler	Ruhr-Universität Bochum
Stefania Spina	Università per Stranieri di Perugia
Egon W. Stemle	Eurac Research
Angelika Storrer	Universität Mannheim
Caroline Tagg	The Open University
Ludovic Tanguy	Université de Toulouse
Erik Tjong Kim Sang	Netherlands eScience Center
Simone Ueberwasser	University of Zürich
Reinhild Vandekerckhove	Universiteit Antwerpen
Lieke Verheijen	Radboud University
Ciara Wigham	Université Clermont Auvergne

Table of Contents

Keynote Speakers

Corpora in Telecollaboration and Virtual Exchange
Marco Cappellini 1

Large-Scale Analyses of Small-Scale Research Questions: Pros and Cons of Upscaling Computer-Mediated Discourse Analysis
Susan C. Herring 2

Talks

Annotating the IDA Corpus: Misogynistic and Stereotypical Content across Two Transnational Incels' Communities
Selenia Anastasi 3

Beware of the Hoover : Examining How Users of the r/BPDlovedones Subreddit Use Language to Transform Personal Experiences into Lay Diagnostic Criteria of Borderline Personality Disorder
James Adrian Balfour 9

Linguistic Variations within French Wikipedia
Nelly Bonhomme 12

A Framework for Analysis of Speech and Chat Content in YouTube and Twitch Streams
Steven Coats 16

The Analysis of "Inclusion" and "Accessibility" in Computer-Mediated-Communication for an Inclusive Transformation in Digital Societies
Annamaria Fabian, Igor Trost, Kevin Altmann and Mara Schwind 20

Making Suggestions in Students' Forum Discussions
Francisco Javier Fernández Polo 26

Analysis of Socially Unacceptable Discourse with Zero-shot Learning
Rayane Ghilene, Dimitra Niaouri, Michele Linardi and Julien Longhi 30

"I Think Your Translation Is Great, but I Have Some Suggestions That May Help You (or Not)". the Use of Concessives as Politeness Devices in Asynchronous Online Discussion Forums
Elsa María González Álvarez and Susana Doval-Suárez 35

Corpus-based Didactics in Higher Educational Settings: Empirically Investigating Online German Youth Language Phenomena
Laura Herzberg and Louis Cotgrove 41

Testing the Weak-tie Hypothesis with Social Media
Mikko Laitinen and Masoud Fatemi 46

Studying Digital Communication of Multilingual Communities: How to Strive towards Sustainability in CMC Studies?
Martti Mäkinen 52

Spatial and Temporal Deixis in Digital Asynchronous Discussions: Where Is "Here", When Is "Now"?
Michel Marcoccia 56

Gay Slang on Facebook: Subversion or Stigmatization?
Yonatan Marik and Hadar Netz 60

AI Device for Deradicalization Process
Andrea Russo 65

Spoken vs. Written Computer Mediated Communication
Hannah J. Seemann, Sara Shahmohammadi, Manfred Stede and Tatjana Scheffler 70

Collecting Metadata for Social Media Corpora in the Face of Ever-changing Social Media Landscapes
Egon Stemle, Alexander König and Lionel Nicolas 75

Talking to Oneself in CMC: A Study of Self Replies in Wikipedia Talk Pages
Ludovic Tanguy, Céline Poudat and Lydia-Mai Ho-Dac 79

Who Cares about Correct Spelling? Spelling Discourse in Social Media Conversations
Reinhild Vandekerckhove 84

Language Style Accommodation in Computer-Mediated Communication: Exploring the Effects of Age and First Language
Lieke Verheijen 89

Posters

Contested Landscapes: Scripts as Graphic and Semiotic Tools to Social Meaning
May Ahmar 95

"Are There Any 'Must Attend' Lectures?": Initial Results from a Reddit Corpus of cross-UK University Student Discussions
Marc Alexander 96

Writing Oral Languages Online: Ettounsi and Tamazight Challenging Standard Ideologies
Soubeika N. Bahri 97

Politicizing Public Health: The Discourses around Public Health Organizations
Tatiana Schmitz de Almeida Lopes, Fernanda Peixoto Coelho and Renata Sant'Anna Lamberti Spagnuolo 100

Harnessing Twitter Corpus for Neural Machine Translation in Low-Resource Languages: A Case Study of Spanish-Galician
María do Campo Bayón 101

From the Web to the Street, and Back: A Semiotic Approach to the Circulation of Militant Writings
Claire Doquet and Chenyang Zhao 103

Bringing CMC Corpora to the People: Improving the Usability of the French CoMeRe Collection
Achille Falaise 104

Lexical Variation of the Albanian Language Used in Computer-mediated Communication and the Challenge for Processing
Besim Kabashi 106

Texting in Time: Approaching Processualities in Everyday Mobile Messaging Interaction
Jasmin Lallo 108

Expressing Laughter in Twitter Conversations: The "xD" Emoticon and the "Face with Tears of Joy" Emoji
Sandra Marion 109

The 3DSeTwitch Corpus – a Three-dimensional Corpus Annotated for Sexist Phenomena
Ariane Robert and Paola Pietrandrea 110

Communication Dynamics in 'No Vax' Groups During Deradicalization Phases
Andrea Russo 113

Exploring Discursive Multidimensionality and Multimodality on Twitter: Analyzing Xenophobic Representations Targeting China During the COVID-19 Pandemic.
Cicero SOARES Da Silva 116

POSTERS

Lexical Variation of the Albanian Language used in computer-mediated communication and the challenge for processing

Besim Kabashi
Friedrich-Alexander-Universität Erlangen-Nürnberg
Bismarckstraße 6, 91054 Erlangen, Germany
besim.kabashi@fau.de

Abstract

In addition to the standard variant of a language, a lot is also spoken and written in non-standard variants. The processing of data that is available in a non-standard variant is associated with many difficulties because the resources and tools were initially created and developed for standard variants and are either missing or insufficient and not good enough for non-standard variants. However, this data is very diverse and prove to be richer than the standard language data, i.e. is also important and must be taken into account and processed. Here we present our work dealing with the processing of lexical variants in the Albanian language used in computer mediated communication. In particular, we are concerned with the normalization of lexical variants and their tagging. We have collected texts from social media that are written in non-standard language, i.e. variants. We discuss them, the phenomena and the steps of processing.

Keywords: The Albanian language, Lexical Variation, Computer-mediated Communication, Normalisation, PoS-Tagging

1. Introduction

There is a lot of online communication these days. Communication in text form makes up a large part of it. The machine processing of these texts still poses a challenge in many cases.[1] This is the case with many languages, even with languages that have a large number of speakers who have had good resources and tools for decades. The number of resources and tools for the Albanian language is far from satisfactory. First and foremost, there is a lack of language resources. Even basic resources are missing, e.g. a free full-form lexicon. Most of them could be developed by linguists of the traditional schools. In this sense, the Albanian language is still considered an under-resourced language.

A large number of social media users use non-standard (lexical) variants of words instead of using the standard Albanian language. This often makes the text difficult to process, even to read and understand for the humans themselves.

We focus here on the use of non-standard lexical variants. Since there are many difficulties involved and they cannot be solved quickly, we are planning a somewhat longer-term project and are presenting an outline of the preliminary work here.

2. The lexical variation of Albanian

The Albanian language is mainly spoken in two varieties (dialect groups), in Gheg, spoken in the north, and Tosk, spoken in the south of the river Shkumbin. The dialects differ mainly phonetically, but are mutually intelligible, i.e. the respective speakers understand each other without difficulty. But that is only one aspect of lexical variation. Since Albanian is spoken in several countries and is taught under different school systems, it has developed several variations. The cultural and economic exchange of the respective countries where Albanian is spoken with neighboring countries or other countries has also contributed to lexical variation, e.g. in loan words (from Greek, Italian, South Slavic languages, as well as English, French and German) as well as in technical and scientific terminology. For example, a tool in Albania may have a different name than one in Kosovo – even if this has been standardized (in many cases) in the dictionary. In addition, it is well known that the social background and education of the language users contribute to lexical diversity. There are other factors that can play a role, but these cannot be dealt with here.

3. Computer-mediated communication

Computer-mediated communication (CMC), online communication, or communication at a distance, refers to *instant messaging*, *e-mailing*, *chatting*, *online forums*, *social networks* and similar services, or *social media* for short. It is easier and more common for users from different cities, regions, dialects and countries to come together and meet online than to have to meet in person (i.e. not online).

3.1. The use of the Albanian language in CMC

On the basis of empirical data, i.e. corpus data, and the word lists, frequency lists, n-grams, collocations, etc. extracted from them, we have identified various phenomena and analyzed them and developed machine language processing methods to deal with them. Some frequent ones are, above all, *dialect* or *idiosyncratic variants* of words, *abbreviations*, *contractions*, *emoticons*, and *creative spellings*. In the Albanian language used in the social media such word forms, e.g. the abbreviation *flm* (short form for *ju/të/... falemnderit*, engl. *thank you*), e.g. the contraction *ti* (engl. *you*) instead of *t'i* (i.e. subjunctive'accusative or dative object clitic), and e.g. the creative spelling *(e) ver8* instead of *(e) vertetë* (engl. (*it is*) *true*), i.e. *8* stands for *tetë* (engl. *eight*), can no longer be omitted in the communication texts.

An important feature is that they can be both *synchronous* and *asynchronous*. The synchronous communication is di-

[1] See especially here, among others, the series of CMC conferences, i.e. (Cotgrove et al., 2023) and the previous ones, which deal with different areas and topics, the EmiriST shared task, i.e. (Beißwenger et al., 2016) for linguistic annotation, as well as the *Journal of Computer-Mediated Communication (JCMC)*.

rect, fast and it often causes poorer text quality, e.g. more spelling mistakes. Some participants/users (i.e. 2nd or 3rd generation migrants) do not master the standard of the Albanian language and thus write in a deviant (sometimes, personal) version, which results in *spelling mistakes*. In addition, the speakers of the Albanian language, especially those of the 2nd and 3rd generation of migrated Albanians, very strongly *mix the code* with and use *words* and *idioms of the respective country language*, where they live, which results in *code mixing*. An easy, not difficult example for *dialect words / regionalisms* is *[...] sdi ca ke shkru [...]* instead of *s'di çka ke shkruar*, engl. (*I do not know what you have written*). Also, who writes to whom (all possible relationships can occur) determines the use of language, such as the choice of words, the style, etc. Often, CMC is seen as a very close version to spoken language. This is important for processing the syntax, i.e. for parsing.

The Albanian alphabet is based on the Latin alphabet with the addition of the letters *ë*, *ç* (with diacritics), and ten digraphs *dh*, *gj*, *ll*, *nj*, *rr*, *sh*, *th*, *xh*, and *zh*. The Albanian Language codes are *sq* (ISO 639-1), *alb* (*B*) (ISO 639-2), *sqi* (*T*) (ISO 639-3). The Albanian alphabet is covered by the ISO-8859-1 / Latin-1 character set (West European languages), and other code pages, e.g. by the ISO-8859-3 / Latin-3 character set (Southeast European languages) – and consequently also by Unicode.

The two characters *ë*, *ç* in particular cause problems, as they are often used in the version without diacritical marks. This causes ambiguities and makes language processing (more) difficult. This is because many users use different keyboards and input systems that do not offer the two letters with diacritics directly.

3.2. The collected linguistic data

We have been observing this type of communication ourselves for years and have been collecting texts for years. We have built up three relatively large corpora of Twitter data, i.e. one of them is standard language 9.2 million words, two of them are not-standard language, approx. 3.4 million words and approx. 0.8 million words.[2] We also have large amounts of data in Albanian from Reddit that are currently being processed.

4. The data processing and the NLP tools

We have the following working pipeline for processing the data: (1) *correct encoding* of the source texts/data, (2) *normalization* of the data, (3) *lemmatization*, and (4) the *POS-tagging*.[3] Normalizing the texts is the biggest challenge. We try to normalize the high-frequency cases first. The idiosyncratic variants in particular are very difficult to identify and therefore to normalize. We try to do this manually to create or improve the gold standard, as well as using various training models. The lemmatization (of standard lexical variants) is also not easy, since the Albanian language has a rich morphology and thus has a strong inflection, especially compounds, are difficult, but it is somewhat easier than the normalization. We normalize the non-standard variants first and consequently we lemmatize them. Our goal is to annotate the data according to the suggestions of EmpiriST Corpus 2.0, cf. (Proisl et al., 2020). We use the tagset from (Kabashi and Proisl, 2018), which also offers a mapping to UD tagset and Google tagset. We benefit from the gold standard and models created from it to tag the data. Based on this, we created new models with the collected CMC data. For gold standard creation, which is constantly being improved, two annotators are currently annotating. For variants, annotators need more linguistic knowledge than for standard variants. As this is a work in progress, we do not include preliminary results such as inter-annotator-agreement here. Based on this work and the experience gained, we want to adapt or extend the existing tools for language variants and create (language variant) resources[4] that make processing of CMC data easier and better.

We partly use *The IMS Open Corpus Workbench*[5] (*CWB*) tools for processing the data and CQPweb, cf. (Hardie, 2012) as a web platform for using corpora.

5. References

Beißwenger, M., Bartsch, S., Evert, S., and Würzner, K.-M. (2016). EmpiriST 2015: A shared task on the automatic linguistic annotation of computer-mediated communication and web corpora. In Paul Cook, et al., editors, *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 44–56, Berlin. Association for Computational Linguistics (ACL).

Louis Cotgrove, et al., editors. (2023). *Proceedings of the 10th International Conference on CMC and Social Media Corpora for the Humanities 2023 (CMC-2023)*, Mannheim, Germany. Leibniz-Institut für Deutsche Sprache (IDS).

Hardie, A. (2012). CQPweb - Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

Kabashi, B. and Proisl, T. (2018). Albanian part-of-speech tagging: Gold standard and evaluation. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kabashi, B. (2018). A lexicon of Albanian for natural language processing. *Lexicographica*, 34(1):239–248.

Proisl, T., Dykes, N., Heinrich, P., Kabashi, B., Blombach, A., and Evert, S. (2020). EmpiriST corpus 2.0: Adding manual normalization, lemmatization and semantic tagging to a German web and CMC corpus. In Nicoletta Calzolari, et al., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020*, pages 6142–6148, Marseille, France. European Language Resources Association (ELRA).

[2] As far as we know, no other corpora of this type have been created for Albanian so far.

[3] See (Kabashi and Proisl, 2018) for an overview of the annotation tools as well as a tags set for the Albanian standard language. See (Kabashi, 2018) for a lexical resource.

[4] For example an extensive, well-covering list that maps various non-standardized lexical variants to standardized lexical variants.

[5] https://cwb.sourceforge.io.

Author Index

Ahmar, May, 95
Alexander, Marc, 96
Altmann, Kevin, 20
Anastasi, Selenia, 3

Bahri, Soubeika N., 97
Balfour, James Adrian, 9
Bonhomme, Nelly, 12

Cappellini, Marco, 1
Coats, Steven, 16
Coelho, Fernanda Peixoto, 100
Cotgrove, Louis, 41

Da Silva, Cicero SOARES, 116
do Campo Bayón, María, 101
Doquet, Claire, 103
Doval-Suárez, Susana, 35

Fabian, Annamaria, 20
Falaise, Achille, 104
Fatemi, Masoud, 46
Fernández Polo, Francisco Javier, 26

Ghilene, Rayane, 30
González Álvarez, Elsa María, 35

Herring, Susan C., 2
Herzberg, Laura, 41
Ho-Dac, Lydia-Mai, 79

Kabashi, Besim, 106
König, Alexander, 75

Laitinen, Mikko, 46
Lallo, Jasmin, 108
Linardi, Michele, 30
Longhi, Julien, 30
Lopes, Tatiana Schmitz de Almeida, 100

Mäkinen, Martti, 52
Marcoccia, Michel, 56
Marik, Yonatan, 60
Marion, Sandra, 109

Netz, Hadar, 60
Niaouri, Dimitra, 30
Nicolas, Lionel, 75

Pietrandrea, Paola, 110
Poudat, Céline, 79

Robert, Ariane, 110
Russo, Andrea, 65, 113

Scheffler, Tatjana, 70
Schwind, Mara, 20
Seemann, Hannah J., 70
Shahmohammadi, Sara, 70
Spagnuolo, Renata Sant'Anna Lamberti, 100
Stede, Manfred, 70
Stemle, Egon, 75

Tanguy, Ludovic, 79
Trost, Igor, 20

Vandekerckhove, Reinhild, 84
Verheijen, Lieke, 89

Zhao, Chenyang, 103