Building and Annotating a Corpus of German-Language Newsgroups

Jasmin Schröck
Institut für Deutsche Sprache
Mannheim
schroeck@direktion.ids-mannheim.de

Harald Lüngen
Institut für Deutsche Sprache
Mannheim
luengen@ids-mannheim.de

Abstract

Usenet is a large online resource containing user-generated messages (news articles) organised in discussion groups (newsgroups) which deal with a wide variety of different topics. We describe the download, conversion, and annotation of a comprehensive German news corpus for integration in DeReKo, the German Reference Corpus hosted at the Institut für Deutsche Sprache in Mannheim.

1 Introduction

Usenet news are an instance of the genre of computer-mediated communication (CMC) which is of interest in many current research questions (cf. Beißwenger & Storrer 2008). Recent initiatives for the creation of CMC corpora have co-operated firstly in the DFG research network empirikom (Beißwenger 2012), and since 2013 within the TEI Special Interest Group on CMC, amongst other things to create a TEI-based standard for the encoding and annotation of CMC corpora for use in empirical linguistics research. Several CMC corpora based on versions of the encoding scheme provided by the TEI CMC SIG have been compiled so far, e.g. (German) chat and wikipedia discussions (Beißwenger et al. 2012; Margaretha & Lüngen 2014) and (French) corpora of various CMC subgenres in the project CoMeRe (Chanier et al. 2014). Currently the Dortmund Chatkorpus (Beißwenger 2013) is being prepared along the lines of the TEI CMC SIG for integration in CLARIN research infrastructures. Consequently, the aim of the work described in this paper was to close another gap by creating an edited Usenet corpus containing all newsgroups from the de. hierarchy and annotating relevant CMC phenomena according to the principles proposed by the TEI CMC SIG. The news corpus has been marked up for metadata and text structure according to I5, which is the TEI customization (Lüngen & Sperberg-McQueen 2012) used for the encoding of texts in DeReKo and which incorporates features of the TEI CMC SIG.

2 Usenet

Usenet originated in 1979 and is based on the NNTP internet protocol (Horton & Adams 1987). The features of news messages include rich formatted metadata (the NNTP header) with fields for the sender, the posting date, the subject, the reply history of a message and other types of information. Header fields are obligatory or optional.

In the message body, many textual features also found in emails or letters prevail, such as salutations (openers and closers), postscripts, or signatures. Another characteristic feature is the highly recursive usage of quotations from previous articles, often introduced by an automatically generated line containing the e-mail address and name of the author and the posting date of the quoted article. Finally, the language used in news messages contains many familiar netspeak phenomena such as the use of emoticons, interjections, and inflectives (cf. Feldweg et al. 1995; Gausling 2005).

A newsgroup works similar to a web discussion forum, one difference being that all newsgroups are organised in a universal, topic-based hierarchy. Newsgroups are stored world-wide on so-called news servers, and everyone is free to set up such a server. All news servers are regularly synchronised with each other so that they offer the same amount of news messages in each newsgroup sooner or later. Similarly, everyone is free to connect to a news server using news client software to subscribe to newsgroups to read

Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media
pages 17–22, University of Duisburg-Essen, September 28, 2015.

messages and to post one's own news messages to the server.

Usenet communication has had its heydays in the 1990s, consequently it is a pre-Web 2.0 form of CMC. But Usenet lives on, as a dedicated community has constantly been using it.

3 Related Work

A previous German Usenet corpus initiative was undertaken in the ELWIS project (*Corpus-based development of lexical knowledge bases*), where a corpus of contemporary German was compiled, beginning in 1992. All messages of the year 1993, containing altogether 433,000 articles in 647 newsgroups, served as a base for the investigation of the language use in newsgroups (cf. Feldweg et al. 1995). More recently, the WestburyLab at the University of Alberta collected English-language Usenet data in the project *A reduced redundancy Usenet Corpus* from 2005 until 2011 (Shaoul & Westbury 2013). Their corpus covers 47,860 newsgroups containing more than seven billion words and seems to be the largest news archive ever prepared as a linguistic corpus. Another English-language corpus was created by Matt Mahoney in the project *Usenet as a text corpus* in 2000, containing 53,247 articles from 9,359 newsgroups (Mahoney 2000). Neither of the previous corpora seems to have been marked up using XML/TEI, nor have they been annotated for CMC phenomena.

4 Creation of the Corpus

Following a common strategy in the construction of TEI corpora from text archives (cf. e.g. Fankhauser et al. 2013; Margaretha & Lüngen 2013), we divided the corpus creation into several steps: In a first step, all German-language Usenet data currently available on the newsserver news.individual.de was downloaded and converted into a well-formed XML version of the NNTP format (dubbed nntpXML) in a straightforward way. In a second stage, the nntpXML data was filtered and converted into the TEI-based I5 target format. In the third stage, we applied heuristics for the annotation of CMC-phenomena typical of Usenet articles to the newly created corpus, creating I5 with annotations (see also Figure 1).

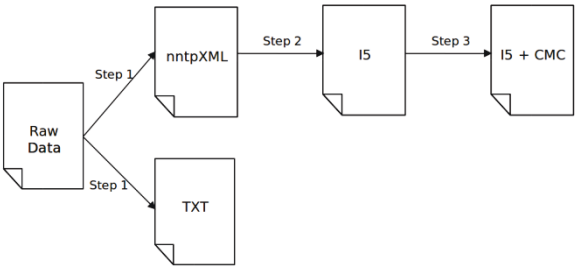

Figure 1: Workflow

4.1 Download and conversion to nntpXML

In the first stage, all currently available Usenet articles of all 379 German-language newsgroups from the de-hierarchy were downloaded from the newsserver news.individual.de (run by FU Berlin, with a retention time of 621 days) on 1 June 2015 using the Python client nntplib. The downloaded data was preprocessed by converting it to Unicode and to well-formed nntpXML, using the Python library lxml. For each newsgroup, a separate file was generated. The original Usenet structure was mostly preserved, only the header lines were ordered in obligatory, optional and others (see Horton & Adams 1987).

4.2 Conversion from nntpXML to I5

The generated nntpXML files were then converted to the TEI format I5 which is used for DeReKo. An I5 corpus file is structured according to the three levels *corpus* (<idsCorpus>, the root element), *document* (<idsDoc>), and *text* (<idsText>). All news articles of one calendar year were stored in a separate <idsDoc> while each article was included in a separate <idsText> document. Note that this corpus structure differs from previous CMC corpora where one thread or logfile containing a set of postings usually corresponds to one corpus text (Beißwenger et al. 2012; Margaretha & Lüngen 2014; Chanier et al. 2014). With news articles (and similarly emails), the messages come neither grouped in a self-contained document (like e.g. a Wikipedia page), nor is news a synchronous type of communication like chat, hence we do not consider threads or logfiles as suitable corpus units for news. Each of the three levels received its own header containing the metadata that were appropriate and could be extracted from the messages. The I5 structure was created using python with lxml, and XSLT stylesheets.

A major task was to identify TEI elements for the encoding of the metadata of the original header of each article. One issue in this area was how to represent the reply history of a message as contained in the NNTP "References" header

line. From this header field, news readers like Mozilla Thunderbird derive the threaded view of the messages. Since neither the TEI Guidelines, nor the TEI CMC SIG provides metadata elements for the reply history, we resorted to a recent proposal by the TEI Correspondence SIG (2015). This SIG develops, amongst other things, TEI elements for the encoding of correspondence-specific metadata applying to all kinds of correspondence such as letters, telegrams, diaries, e-mails and blogs. Consequently, we added the elements <correspDesc> and <correspContext> to I5. The latter serves the encoding of information about previous and following messages in a correspondence.[1]

Furthermore, we adopted the suggestion by Beißwenger et al. (2012) to create a list of participants in the newsgroup (<listPerson>) and a timeline (<timeline>), but stored them in separate files as a step in the anonymisation of the data. The list of persons contains the e-mail address for each participant and their name if available. Also, following Beißwenger et al. (2012), the text of an article was wrapped in a <posting> element whose attributes @who and @synch refer to the corresponding ID in the list of persons and the timeline, respectively.

4.3 Annotation of CMC phenomena and quality assessment of the annotations

For an annotation of CMC phenomena as introduced in Section 2, we developed several heuristics and implemented them in an XSLT 2.0 stylesheet, creating a regular expression for each phenomenon and tagging the matching strings with a suitable TEI element. The following CMC phenomena were annotated: quotations, as well as the lines introducing them, links to the World Wide Web, links to other newsgroups, salutations (openers and closers), postscripts, user signatures, and emoticons. Apart from these CMC categories, paragraphs were annotated. Table 1 shows the phenomena and the respective elements used for their annotation.

Table 1: Annotated phenomena, used elements and their source

Phenomenon	Source	Example
Link to www	TEI	<ref type="www" target="URL"> URL </ref>
Link to newsgroup	TEI	<ref type="newsgroup" target="de.rec.fahrrad"> de.rec.fahrrad </ref>
Opener	TEI	<seg type="opener"> Hallo, </seg>
Closer	TEI	<seg type="closer"> Ciao, NAME </seg>
Postscript	TEI	<seg type="postscript"> P.S. dürften sich eigentlich links der durchgezogenen Linie in dieser Fahrradstraße noch Fahrräder aufhalten? </seg>
Signature	TEI	<trailer> -- Life's a road, not a destination. </trailer>
Emoticon	Beißwenger et al. (2012)	<interactionTerm> <emoticon> :-(</emoticon> </interactionTerm>
Quotation, with or without introductory line	TEI	<cit type="replyCit"> <bibl type="introQuote"> Am 23.09.2013 12:33, schrieb NAME: </bibl> <quote> Die Zukunft ist da, seilzuglose Rennräder sind möglich geworden. </quote> </cit>

We assessed the quality of the CMC annotations by conducting a small evaluation of each annotated CMC feature on 200 articles from five newsgroups, which according to their topics seemed reasonably diverse: de.etc.sprache.deutsch (the German language),

[1] Apparently, these elements have been added to the official TEI Guidelines in the meantime, cf. TEI Consortium (2015).

de.rec.mampf (food/eating) de.comp.os.ms-windows.misc (windows operating system), at.gesellschaft.politik (society and politics, Austria) and de.soc.senioren (senior citizens) (cf. Schröck 2015). For this test set, correct reference annotations were created manually by an expert familiar with the TEI elements used for the annotation of the CMC categories and their TEI (or SIG) definitions. The reference set eventually contained 3,291, the test set 3,438 annotation instances (TEI elements marking up CMC phenomena). Comparing the test set with the reference, the micro average precision over all eleven annotation categories was found to be 79%, and the micro average recall was 82%. The categories that were identified best by the regular expressions were signature and emoticon. The categories most difficult to identify were postscript and opener. However, these two categories, and also links to newsgroups, didn't occur very frequently in the test and reference sets, which were relatively small.

Furthermore, the results for openers, closers and introduction lines of quotes, which often contain names, could be improved by using Named Entity Recognition.

The results for all elements and the overall results are shown in Table 2.

Table 2: Number of annotation instances for each element in reference set (# ref) and test set (# test) and micro-averaged precision (P), recall (R) and F-measure (F)

I5 Tag	# ref	# test	P	R	F
<cit>	618	661	.69	.74	.71
<bibl>	359	263	.94	.69	.80
<quote>	606	661	.80	.87	.83
<p>	1186	1358	.79	.91	.85
<ref type= ”www”>	109	106	.88	.85	.86
<seg type= “signature“>	89	86	1	.97	.98
<emoticon>	99	89	.93	.84	.88
<ref type= ”newsgroup”>	13	28	.46	1	.63
<seg type= ”postscript”>	2	16	.13	1	.23
<seg type= ”closer”>	182	140	.81	.63	.71
<seg type= ”opener”>	28	30	.4	.43	.41
	Σ = 3291	Σ = 3438	Ø = .79	Ø = .82	Ø = .80

5 The Corpus

Download was carried out on 1 June 2015 and took 12 hours, using four threads on a linux machine with an AMD Opteron 8439 SE processor with 48 cores at 2.8GHz, and 256G RAM. The news server potentially contained 1,004,157 articles in 379 newsgroups in the *de* hierarchy; however, 62,878 articles were discarded because their X-No-Archive field was set to yes, and another 70,376 because their encoding could not be determined and hence not be converted to UTF-8. The conversion-to-I5 phase took 2:20 hours, and the annotation phase took 54 hours. Four newsgroups contained no messages, and with one newsgroup, the annotation did not terminate. From the remaining 374 groups, 11 messages were discarded because they contained an error in the Date field. The resulting, annotated full corpus in I5 format contains 374 newsgroups comprising 870,892 news articles with 128.78 million word tokens. It takes up 7.2G of disk space. The corpus contains messages posted between 24/9/2013 and 1/6/2015. The biggest newsgroup (929MB) is de.soc.politik.misc containing 117,950 messages (16.8 million word tokens). However, the size of the corpus will be further reduced in the deduplication and cleaning step.

6 Conclusion

The news corpus described in this paper is currently being further anonymised, cleaned, and de-duplicated, mostly according to the principles described in Shaoul & Westbury (2013). The resulting version is scheduled to be included in the upcoming DeReKo release DeReKo-2015-II. However, the question of whether the corpus can be shared with the linguistic community remains to be solved. CMC texts, like all other texts, are subject to copyright, and in principle each author

of an article contained in the corpus would have to give his or her consent first. On the other hand we think that a news article is not the same as a text on the W3C, as someone who posts to a newsgroup is aware of the fact (in fact wants) that his/her message will be distributed to many servers all over the world in the first place. We are currently seeking legal advice in this matter in cooperation with the CLARIN-D curational project Chatkorpus2CLARIN.[2] Until further notice, the news corpus will be accessible from the premises of the IDS Mannheim only.

We intend to update the corpus with new news articles regularly. We are also aiming at downloading from a news server with a longer retention time, though as far as we can see, longer retention times are only offered by commercial news servers.

Reference

Beißwenger, Michael (2012): Forschungsnotiz: Das Wissenschaftliche Netzwerk "Empirische Erforschung internetbasierter Kommunikation" (Empirikom). In: *Zeitschrift für germanistische Linguistik* 40 (3), pp. 459-461.

Beißwenger, Michael (2013): Das Dortmunder Chat-Korpus. In: *Zeitschrift für germanistische Linguistik* 41 (1), pp. 161-164.

Beißwenger, Michael; Storrer, Angelika (2008): Corpora of Computer-Mediated Communication In: Lüdeling, Anke; Kytö, Merja (eds.): *Corpus Linguistics. An international Handbook.* Vol. 1, Berlin: de Gruyter, pp. 292-308.

Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika (2012): A TEI Schema for the Representation of Computer-mediated Communication. In: *Journal of the Text Encoding Initiative* [Online] 3.

Beißwenger, Michael; Lemnitzer, Lothar (2013): Aufbau eines Referenzkorpus zur deutschsprachigen internetbasierten Kommunikation als Zusatzkomponente für die Korpora im Projekt "Digitales Wörterbuch der deutschen Sprache" (DWDS). In: *Journal for Language Technology and Computational Linguistics (JLCL)* 28 (2), Special issue on "Webkorpora in Computerlinguistik und Sprachforschung".

Chanier, Thierry; Poudat, Céline; Sagot, Benoît; Antoniadis, Georges; Wigham, Ciara R.; Hriba, Linda; Longhi,_Julien; Seddah, Djamé (2014): The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. In: *Journal of Language Technology and Computational Linguistics (JLCL)* 29 (2), pp. 1-30.

Feldweg, Helmut; Kibinger, Ralf; Thielen, Christine. (1995): Zum Sprachgebrauch in deutschen Newsgruppen. In: Schmitz, U. (ed.): *Neue Medien.* Osnabrücker Beiträge zur Sprachtheorie 50, Oldenburg: Red. OBST, pp. 143-154.

Gausling, Timo (2005): Der Newsgroup-Beitrag - eine kommunikative Gattung? In: *Studentische Arbeitspapiere zu Sprache und Interaktion (SASI)* 4. Series "Arbeitspapiere des Centrum Sprache und Interaktion der Westfälischen Wilhelms-Universität", available online at: http://noam.uni-muenster.de/sasi/Gausling_SASI.pdf (last visited 2015-07-10).

Horbach, Andrea; Thater, Stefan; Steffen, Diana; Fischer, Peter M.; Witt, Andreas; Pinkal, Manfred (2015): Internet Corpora: A Challenge for Linguistic Processing. In: *Datenbank-Spektrum* 15 (1), pp. 41-47.

Horton, M.; Adams, R. (1987): *RFC-1036 Standard for Interchange of USENET Messages.* Available online at: http://tools.ietf.org/html/rfc1036 (last visited 2015-07-10).

Lüngen, Harald; Sperberg-McQueen, C. M. (2012): A TEI P5 Document Grammar for the IDS Text Model. In: *Journal of the Text Encoding Initiative* [Online] 3.

Mahoney, Matt (2000): *Usenet as a text corpus.* Florida Tech, CS Dept., available online at: https://cs.fit.edu/ mmahoney/dissertation/corpus.html (last visited 2015-07-13).

Margaretha, Eliza; Lüngen, Harald (2014): Building Linguistic Corpora from Wikipedia Articles and Discussions. In: *Journal of Language Technology and Computational Linguistics (JLCL)* 29 (2), pp. 59-82.

Schröck, Jasmin (2015): *Erstellung eines deutschsprachigen Usenet-Newsgroup-Korpus und Annotation von Phänomenen internetbasierter Kommunikation.* BA-Thesis, University Heidelberg.

Shaoul, Cyrus; Westbury Chris (2013): *A reduced redundancy USENET corpus (2005-2011).* Edmonton, AB: University of Alberta, available online at: http://www.psych.ualberta.ca/~westburylab/downloads/usenetcorpus.download.html (last visited 2015-07-13).

TEI Consortium (2015): *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Available online at: http://www.tei-c.org/Guidelines/P5/ (last visited 2015-07-10).

TEI Correspondence SIG (2015). Information and examples available online at: http://www.tei-c.org/Activities/SIG/Correspondence/,

[2] http://chatkorpus.tu-dortmund.de/

http://wiki.tei-c.org/index.php/SIG:Correspondence and https://github.com/TEI-Correspondence-SIG/ (last visited 2015-07-10).