Anonymisation of the Dortmund Chat Corpus 2.1

Harald Lüngen[1], Michael Beißwenger[2], Laura Herzberg[3], Cathrin Pichler[2]
[1]Institut für Deutsche Sprache, R5 6-13, D-68161 Mannheim
[2]Universität Duisburg-Essen, Institut für Germanistik, Berliner Platz 6-8, D-45127 Essen
[3]Universität Mannheim, Germanistische Linguistik, Schloss, D-68131 Mannheim
E-mail: luengen@ids-mannheim.de, michael.beisswenger@uni-due.de, lherzber@mail.uni-mannheim.de, cathrin.pichler@tu-dortmund.de

Abstract

As a consequence of a recent curation project, the Dortmund Chat Corpus is available in CLARIN-D research infrastructures for download and querying. In a legal expertise it had been recommended that standard measures of anonymisation be applied to the corpus before its republication. This paper reports about the anonymisation campaign that was conducted for the corpus. Anonymisation has been realised as categorisation, and the taxonomy of anonymisation categories applied is introduced and the method of applying it to the TEI files is demonstrated. The results of the anonymisation campaign as well as issues of quality assessment are discussed. Finally, pseudonymisation as an alternative to categorisation as a method of the anonymisation of CMC data is discussed, as well as possibilities of an automatisation of the process.

Keywords: Corpora, Computer-mediated communication, Anonymisation

1. Introduction

In the CLARIN-D curation project "Integration of the Dortmund Chat Corpus into CLARIN-D" (Lüngen et al., 2016), a legal expertise was sought to clarify issues concerning the possibility to republish the material, which had been collected between 2004-2008 partly without written consent of the participants, in the CLARIN-D infrastructures (Lüngen et al., 2016; Beißwenger et al., 2017). The legal expertise was composed by the company iRights.law (Berlin), which specialises on legal issues concerning digital media. Below is a summary of the recommendations that were given to the hosting institutions. They follow from considerations of personality and data protection rights. Other legal statuses like copyright and intellectual property rights were also considered in the expertise, but are not discussed here.

1. Remove the chats which originally came from psycho-social counselling platforms completely (10 out of 480 logfiles)
2. Grant access to chats collected from closed platforms only for authorised scientific use
3. Apply "standard measures" of anonymisation to all chat files
   a. Randomise/replace host names, nicknames, place names, and platform names
   b. Remove or permute the time stamps

Medlock (2006) distinguishes between categorisation and pseudonymisation. The latter is a procedure of permutation or replacement of the sensitive references with instances of the same ontological category (e.g. replacing occurrences of the male name *Holger* with the male name *Werner*. To get an idea of possible types and categories of sensitive references in chat and CMC, we also looked at the anonymisation in previous CMC corpus projects. Among them was no project dealing with chat data, however an email corpus (Medlock, 2006), a Facebook corpus (DiDi, 2015), two SMS corpora (Panckhurst, 2013; Ueberwasser, 2015), as well as one spoken conversation corpus (FOLK, cf. Winterscheid, 2015). In all the CMC corpora, anonymisation was realised as categorisation, only in the spoken corpus was it realised as pseudonymisation.

2. Anonymisation by categorisation

Categorisation preserves some of the information so that a corpus can still reasonably be used for linguistic analyses. It implies the replacement of a sensitive reference with a placeholder string that indicates its ontological category, such as *Person_name* or *Place_name*. Since most of the references that had to be anonymised in the chat corpus are names, we firstly included the five named entity categories PER, ORG, LOC, GPE, OTH from the TüBa-D/Z treebank (Telljohann et al., 2004), which had already been used in NER experiments with DeReKo (Bingel & Haider, 2015) in our category inventory. Because these five NER cateogories are relatively coarse-grained, and because the annotations in the original chat corpus resource contained already more specific information, we extended the set by the categories NICK (for nickname, a subcategory of PER) and ROOM (for chat room). Moreover, we added the category GEO_DE for a noun or adjective derived from a LOC or a GPE (a union of the categories _GeoNE_ and _GeoADJA_ in DiDi, 2015). Besides these, three categories for more formal references were added: URL (for a web address), email (for an email address), and NUMBER (for any kind of referencing number, see Table 1 for examples). Following Winterscheid (2015) and DiDi (2015), we also introduced the two rarer categories IMPLICIT (for an implicit reference), and CITATION (for a quote by which an individual might be identified). The 13 anonymisation categories used are shown in Table 1 with their definitions, example(s), and the source of or inspiration for the category.

#	Category short form	Category long form and definition	Examples	Source
1	PER	PERSONNAME: A first name or second name or a sequence out of first name and second name	“Erwin”, “Meike”, “Anna Hein”	TüBa-DZ (Telljohann et al., 2004)
2	NICK	NICKNAME: User name chosen by a chat participant, or a variant therof	“superman”, “lela2” ,“Tiger”, “Lan5”, “KainPech”	DO chat corpus
3	ORG	ORGANISATIONNAME: Company (e.g. the employer of a participant), sports club, institute, university etc.	“RUB”, “John Deere”, “ASV Schifferstadt”	TüBa-DZ
4	LOC	LOCATIONNAME: A place or area which is not a GPE, e.g. mountains, valleys, rivers, roads, motorways, etc.	“Augustaanlage”, “Neckar”, “Königstuhl” “A6”	
5	GPE	GEOPOLITICALENTITTYNAME: A geo-political entity, i.e. a place or area of which the borders are officially defined, i.e. cities, municipalities, countries, states, suburbs etc., including their spelling variants and abbreviations	“Mannheim”, “NRW”, “Italien” “doaaadmund” “DO”	
6	GEO_DE	GEODERIVATIONNAME: Noun or adjective that is morphologically derived from a (mostly GPE or LOC) name and which expresses an association or a quality (adjectives) or a group or inhabitants (noun)	“Mannheimer”, “Mannheimerinnen”, “Gelbfüßler”	
7	OTH	OTHERNAME: Residual category for all sensitive names and references that cannot be categorised otherwise	“unicum”	
8	ROOM	CHATROOMNAME: Name of a chatroom	“Welcome”, “blue”	DO chat corpus
9	URL	WWWURL: Web address	“http://www.ids-mannheim.de/”	
10	EMAIL	EMAIL: Email address	“fix@ids-mannheim.de”	
11	NUMBER	NUMBER: Any number or code that can be associated with a person: e.g. house number, serial numer, postal code, telephone number, passport number, account number, IP address, password	“0621/1581418”, “10.0.1.45”, “68161”	
12	IMPLICIT	IMPLICIT: Implicit reference: Revealing descriptions and pieces of information from which the identity of a chat participant or a third party can be inferred (e.g. someone’s job)	„IT-Operator“	FOLK (Wintersche id, 2015)
13	CITATION	CITATION: A quote, e.g. from a song, which can be used to identify a chat participant or a third party		FOLK

Table 1: Anonymisation categories in the Dortmund Chat Corpus 2.1.

3. Anonymisation campaign and results

The major bulk of the anonymisation by categorisation process was carried out by four student assistants of Mannheim University, Duisburg-Essen University and the Institute for the German Language (IDS), Mannheim. The sensible references that had not been pre-annotated were identified and annotated with the category inventory in Table 1 using the "author mode" of the XML editor Oxygen. The campaign lasted from August till December 2016 (five months) and took approximately 625 hours of manual annotation work. Subsequently an XSLT post-processing step was implemented to insert the replacement strings and to provide TEI annotation in terms of the elements <name> and <ref>.

Listings 1-4 contain XML code snippets that show what the result of the anonymisation looks like in CLARIN-D TEI (cf. Lüngen et al., 2016).

```
<particDesc>
    <!-- 1301005 -->
    <listPerson>
      <!-- ... -->
      <person role="celebrity" xml:id="A03">
        <persName type="nickname">Günther Beckstein</persName>
        <sex evidence="estimated">male</sex>
      </person>
      <!-- ... -->
      <person role="moderator" xml:id="A04">
        <persName type="nickname">[_MALE-MODERATOR-A04_]</persName>
        <sex evidence="estimated">male</sex>
      </person>
      <!-- ... -->
      <person role="participant" xml:id="A07">
        <persName type="nickname">[_FEMALE-PARTICIPANT-A07_]</persName>
        <sex evidence="estimated">female</sex>
      </person>
      <person role="participant" xml:id="A08">
        <persName type="nickname">[_PARTICIPANT-A08_]</persName>
        <sex evidence="estimated">unknown</sex>
      </person>
      <!-- ... -->
    </listPerson>
  </particDesc>
```

Listing 1: Anonymisation of metadata (participant list). Mentions of celebrities and politicians are from the public sphere and are not anonymised.

```
<post auto="false" rend="color:lime" type="event" who="#A14" xml:id="m487">
    <name corresp="#A14" type="nickname">
      <w lemma="[_PARTICIPANT-A14_]" type="NE" xml:id="m487.t1"> [_PARTICIPANT-
      A14_]</w>
    </name>
    <w lemma="werden" type="VAFIN" xml:id="m487.t2">wird</w>
    <w lemma="schlecht" type="ADJD" xml:id="m487.t3">schlecht</w>
  </post>
```

Listing 2: Anonymisation of a nickname without role entry in participant list.

```
<w lemma="auch" type="ADV" xml:id="m40.t1">auch</w>
<w lemma="bei" type="APPR" xml:id="m40.t2">bei</w>
<w lemma="die" type="ART" xml:id="m40.t3">den</w>
<name type="GEO_DE">
    <w lemma="[_GEODERIVATIONNAME_]" type="NN"
xml:id="m40.t4">[_GEODERIVATIONNAME-4_]</w>
</name>
```

Listing 3: Anonymisation of a derivation of a name (like *Düsseldorfern*).

```
<post auto="false" rend="color:#808080" synch="#t427" type="standard" who="#A26" xml:id="m576">
    <w lemma="wollen" type="VMFIN" xml:id="m576.t1">willst</w>
    <w lemma="du" type="PPER" xml:id="m576.t2">du</w>
    <w lemma="eine" type="ART" xml:id="m576.t3">ne</w>
    <w lemma="Therapie" type="NN" xml:id="m576.t4">therapie</w>
    <ref corresp="#A31" type="addressingTerm">
      <w lemma="@" type="ADRIND" xml:id="m576.t5">@</w>
      <w type="NE" xml:id="m576.t6">[_MALE-PARTICIPANT-A31_]</w>
    </ref>
    <w lemma="ich" type="PPER" xml:id="m576.t7">ich</w>
    <w lemma="studieren" type="VVFIN" xml:id="m576.t8">studier</w>
    <ref type="IMPLICIT">
      <w type="NE" xml:id="m576.t9">[_IMPLICIT-1_]</w>
    </ref>
  </post>
```

Listing 4: Anonymisation of an implicit reference.

Below are the stats of the annotation of categories in the hole corpus - remember that the chat corpus contains roughly 1 million tokens

Category	# Occurrences
NICK:	30,022
ROOM:	2,409
OTH:	1,819
URL:	1,742
GPE:	1,309
PER:	838
ORG:	741
GEO_DE:	178
NUMBER:	169
IMPLICIT:	130
LOC	107
EMAIL:	50
CITATION:	5
∑ =	39,519

Table 2: Occurrences of categories for sensitive references in Dortmund Chat Corpus 2.1.

4. Quality assessment

To ge some impression of the agreement between our coders, we asked all four of them to annotate the chat logfile with the ID 1102001 immediately after the training session. The file contains 675 chat posts, and the union of the sensitive references identified by the four coders contained 126 references. This figure was subsequently used as N (number of items to be coded) in the Kappa calculation described in the following. We calculated Fleiss' Kappa using the IRR package for the programming language R (function *kappam.fleiss*)[1]. The agreement between the four coders was κ=0.582. According to the interpretation scale by Landis & Koch (1997), this corresponds to "moderate" agreement.

A closer inspection of the disagreements revealed that

[1] Cf. https://cran.r-project.org/package=irr [22.06.2017].

Coder 1 was the source of an unusual great deal of the mismatches. For instance, in cases where the name of a person was given fully as first name + last name, Coder 1 had, contrary to the training instructions, always annotated them as separate instances. There were at least 15 such first name + last name combinations. (Coder 1 was subsequently made aware of this error i.e. before anonymising her share of the corpus.)

We additionally calculated Fleiss' Kappa among the remaining three coders Coder 2, Coder 3, and Coder 4 only. For them, Kappa was found to be κ=0.827 after all, which according to Landis & Koch (1977) can be interpreted as "almost perfect agreement".

Because of these results, we believe that our method is appropriate for achieving an anonymisation of the chat corpus that conforms to legal standards as put forth in the legal expertise. Ideally, one would have had more material annotated by all four coders, and calculated the inter-rater reliability not only at the beginning of the annotation campaign but also in the middle and at the end of it. Moreover, it would have been interesting if we had even checked for *intra*-rater reliability of each or at least some of the coders. Unfortunately, in the present campaign there was no more time for coding, coordination, and evaluation work. But for future projects, this should be kept in mind.

5. Discussion

During the campaign we noticed that for several chats, a more fine-grained category scheme would have been desirable from a discourse linguist's point of view. In some chats, for instance, many locations were mentioned, and in the anonymised version one would have wished to have more information on the kind of location discussed (e.g. restaurant, shop, school) available. On the other hand, a more complex encoding scheme usually affects inter-rater agreement to the negative. A simple solution to this could be to allow coders to add free information strings.

Another way to address all kinds of problems with the category scheme could be to aim for corpus *pseudonymisation* such as in the spoken conversation corpus FOLK. However, to achieve a full pseudonymisation is even more costly than our anonymisation by categorisation method, and besides has its own drawbacks, such as the possibility of introducing inconsistencies in the dialogue.

Finally, it seems obvious that we need an automatisation of the anonymisation process. A campaign like the one described above is simply not feasible for larger corpora, and the need for anonymisation is potentially given with many kinds of CMC, even web corpora. However, the task is non-trivial and comprises more than standard Named Entity Recognition. The anonymised Dortmund Chat Corpus 2.1 can also serve as training data for future developments of corpus anonymisation tools.

6. References

Beißwenger, M., Lüngen, H., Schallaböck, J., Weitzmann, J.H., Herold, A., Kamocki, P., Storrer, A., Wildgans, J. (2017, to appear). Rechtliche Bedingungen für die Bereitstellung eines Chat-Korpus in CLARIN-D: Ergebnisse eines Rechtsgutachtens. In Michael Beißwenger (Ed.), *Empirische Erforschung internetbasierter Kommunikation*. Berlin/Bew York: de Gryuter (Empirische Linguistik).

Bingel, J., Haider, T. (2014). Named Entity Tagging a Very Large Unbalanced Corpus. Training and Evaluating NE Classifiers. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*. Reykjavik: European Language Resources Association (ELRA).

DiDi (2015). *Beschreibung der Anonymisierung im DiDi-Korpus*. Available online http://www.eurac.edu/en/research/autonomies/commul/Documents/DiDi/DiDi_anonymisation_DE.pdf.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. In *Psychological Bulletin* 76 (5), pp. 378–382.

Landis, J.R., Koch, G.G. (1977). The measurement of observer agreement for categorical data. In *Biometrics* 33, pp. 159–174.

Lüngen, H., Beißwenger, M., Herold, A., Storrer, A. (2016). Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In S. Dipper, F. Neubarth & Heike Zinsmeister (Eds.), *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pp. 156-164.

Medlock, B. (2006). An Introduction to NLP-based Textual Anonymisation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC '06)*. Genoa: European Language Resources Association (ELRA). Available online http://www.lrec-conf.org/proceedings/lrec2006/.

Panckhurst, R. (2013). A large SMS Corpus in French: from Design and Collation to Anonymisation, Transcoding and Analysis. In *Procedia - Social and Behavioral Sciences* 95, pp. 96-104.

Telljohann, H., Hinrichs, E., Kübler, S. (2004). The TüBa-D/Z Treebank: Annotating German with a Contet-Free Backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*.

Ueberwasser, S. (2015). *Anonymisation* (SMS4Science.ch). Available online https://sms.linguistik.uzh.ch/bin/view/SMS4Science/Anonymisation.

Winterscheid, J. (2015). GAIS Web: Maskierung. Document in the IDS Project. Available online http://prowiki.ids-mannheim.de/bin/view/GAIS/MasKierung.