Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5557–5564
Marseille, 20-25 June 2022

Anonymising the SAGT Speech Corpus and Treebank

Özlem Çetinoğlu, Antje Schweitzer
Institute for Natural Language Processing (IMS), University of Stuttgart
{ozlem.cetinoglu,antje.schweitzer}@ims.uni-stuttgart.de

Abstract

Anonymisation, that is identifying and neutralising sensitive references, is a crucial part of dataset creation. In this paper, we describe the anonymisation process of a Turkish-German code-switching corpus, namely SAGT, which consists of speech data and a treebank that is built on its transcripts. We employed a selective pseudonymisation approach where we manually identified sensitive references to anonymise and replaced them with surrogate values on the treebank side. In addition to maintaining data privacy, our primary concerns in surrogate selection were keeping the integrity of code-switching properties, morphosyntactic annotation layers, and semantics. After the treebank anonymisation, we anonymised the speech data by mapping between the treebank sentences and audio transcripts with the help of Praat scripts. The treebank is publicly available for research purposes and the audio files can be obtained via an individual licence agreement.
Keywords: code-switching, Turkish, German, anonymisation

1. Introduction

Dataset sharing is important for replicability and so that other researchers can make use of existing data. When making datasets available, anonymisation is crucial both ethically and legally. Medlock (2006) defines anonymisation as "the task of identifying and neutralising sensitive references within a given document or set of documents." Sensitive references could be direct such as a person's name. They could also be indirect, in that they contribute to revealing identities when combined with other information. Example 1 demonstrates an indirect sensitive reference.

As a consequence of anonymisation then, direct and indirect references in a dataset have to be altered. However it is also important to keep the data as intact as possible so that anonymised data is still useful for research purposes. Our motivation here is to find a balance in anonymisation.

Our work describes the anonymisation stages of a code-switching corpus, namely SAGT, which is a collection of interviews with Turkish-German bilinguals (Çetinoğlu and Çöltekin, 2022). The interviews were collected as audio recordings and transcribed. The transcriptions were then turned into a treebank with various layers of annotation (Çetinoğlu and Çöltekin, 2019). The interviews were conducted in an informal setting to help increase the frequency of code-switching. Common conversation themes include studies, work, travel, future plans, and free time activities such as sports, books, and TV. Hence the data contains private information about participants. Anonymisation took place both on the resulting treebank side and on the audio files. We asked two main research questions during this process: *(i)* Which anonymisation strategy should we follow? *(ii)* How can we anonymise in one medium (i.e., text) and transfer it to the other one (i.e., audio).

In data collection, processing, and maintenance we closely follow EU General Data Protection Regulation (GDPR)[1]. The trade-off between privacy protection policies and keeping data as intact as possible has shaped our anonymisation approach. We paid attention to syntactic and semantic plausibility, that is, anonymisation preserves the syntactic and semantic properties of the original expression so that the annotation layers are still meaningful. We also changed the semantics of the data as minimally as possible so that common world knowledge from external resources (e.g., via word embeddings) could still be utilised in automatically processing the corpus. For instance, if we anonymised the sentence in (1) below, we would replace *Liverpool* by a comparable city that also has an airport, so that the sentence is still faithful to common world knowledge, and that word embeddings for the replacement should be reasonably similar to those of the original.

(1) **Original:** She flew to Liverpool.
Anonymised: She flew to Manchester.

This strategy resulted in a *selective pseudonymisation* approach, that is, we manually decided what references should be anonymised instead of anonymising a whole class, i.e., *blanket* anonymisation. And once we identified sensitive references, we replaced them with similar surrogate values (*pseudonymisation*) as opposed to category labels, i.e., *categorisation* (Section 4.1). We then used a mapping between the transcriptions and audio to mark the sensitive intervals on the audio side. Finally we replaced the marked intervals with a beep sound in the audio files (Section 4.2). The SAGT treebank is publicly available for research purposes. The audio files and corresponding annotated transcriptions are available to researchers via a licence agreement.

2. Related Work

Rock's (2001) seminal work that addresses questions on what and how to anonymise linguistic data dates

[1] https://gdpr-info.eu/

back to two decades ago. It was followed by studies that discuss anonymisation practices on datasets, with a dominance in the medical domain (cf. (Uzuner et al., 2007; Meystre et al., 2010; El Emam, 2013)). Other domains in which anonymisation is applied include emails (Medlock, 2006; Eder et al., 2019), SMSes (Patel et al., 2013), chat messages (Lüngen et al., 2017), learner corpora (Megyesi et al., 2018), and job postings (Jensen et al., 2021). Among them only two (Lüngen et al., 2017; Jensen et al., 2021) applied categorisation, that is, they replaced sensitive references with categories. The other studies employed pseudonymisation with either completely manual replacements (Megyesi et al., 2018) or automatic replacements followed by manual passes on the data (Medlock, 2006; Patel et al., 2013; Eder et al., 2019).

Similar to our work, (García-Sardiña et al., 2018) applied anonymisation on transcriptions. Their dataset consists of dialogues in Spanish. They identified sensitive references manually and applied substitution automatically. To ensure data privacy they corrected the automatic substitution manually. While their transcriptions are available publicly, the corresponding audio corpus cannot be released due to sensitivity of the data.

Fully and irrevocably anonymising audio data is more complex than text data as discussed in early work (Corti et al., 2000; Rock, 2001). Unlike text, not only the content could be sensitive but also the voice itself is personal data. Therefore sometimes the anonymisation methods centre around speaker de-identification via voice transformations (Jin et al., 2009; Pobar and Ipšić, 2014; Justin et al., 2015; Fang et al., 2019). However since some linguistic research questions may require phonetic analyses that would suffer from such transformations, we chose not to apply de-identification, but to protect the privacy of the speakers by restricting access and requiring a stricter license for the audio recordings than for the treebank annotations of the data.

Interestingly, in the voice transformation cases mentioned above, unlike in our study, removing sensitive information from the utterances is not part of anonymisation. Cohn et al. (2019) solved the problem of sensitive information in utterances with a pipeline approach, assuming named entities as data to be anonymised. Their system first obtained automatic transcriptions of the audio data via Automatic Speech Recognition (ASR), then applied Named Entity Recognition on transcriptions to identify sensitive references and finally mapped identified references to audio intervals using the alignments coming from the ASR system. The marked intervals were then redacted from the original audio file.

(Põldvere et al., 2021) followed an approach quite similar to ours in preparing their London-Lund Corpus 2, which consists of audio files and their transcriptions. During manual transcription, annotators marked references to anonymise and added their surrogate values. They inserted timestamps to the transcriptions to align them to audio and used these alignments in marking sensitive references on the audio side. Marked references are then replaced with a humming sound to retain prosodic characteristics. We differ from their approach in choosing selective pseudonymisation instead of blanket pseudonymisation.

3. The SAGT Corpus

The SAGT Corpus is part of the "Computational Structural Analysis of German-Turkish Code-Switching" project that aims to analyse the mixed language of Turkish-German bilinguals from a computational linguistics perspective. Here we give a brief description of the corpus, more details about data collection and transcription can be found in (Çetinoğlu and Çöltekin, 2022).

3.1. Audio Collection and Transcription

The data collection was done as conversation recordings by Turkish-German bilinguals who study Computational Linguistics. There are 48 conversations from 20 bilingual participants. The transcription and annotations were done using Praat (Boersma and Weenink, 2019). For each audio file there are six annotation tiers: `spk1_verbal`, `spk1_norm`, `spk2_verbal`, `spk2_norm`, `lang`, `codesw`. The first four tiers contain verbal and normalised transcriptions of speakers 1 and 2. The verbal tiers follow speech conventions while the normalised tiers follow edited text conventions. The tier `lang` corresponds to the language of intervals and can have `TR` for Turkish, `DE` for German, and `LANG3` for utterances in other languages.

The first five tiers are interval tiers, while the last one is a point tier that denotes sentence and code-switching boundaries. In this last tier, the labels on the boundaries are `SB` (sentence boundary) when both sides of the boundary are in the same language; `SCS` (sentence code-switching) when the language changes from one sentence to the next (intersentential); and `WCS` (word code-switching) when the switch is between words within a sentence (intrasentential).

Another type of code-switching is alternating languages at morpheme boundaries within a word, namely intra-word code-switching. We observe it in the corpus mainly as non-Turkish common and proper nouns with Turkish suffixes due to the agglutinative nature of Turkish. We marked such words in the `codesw` tier as an intra-word switch and use the symbol `$` following (Çetinoğlu, 2016). Example (2) shows the representation of a mixed word where the German noun *Fleugzeug* 'plane' is followed by the Turkish instrumental suffix *-la* 'with'.[2] Figure 1 demonstrates the Praat representation of this word as part of a full sentence. The `$` and `WCS` boundaries, and tiers could also be observed from the same figure.

[2]This and all following enumerated examples are from the SAGT corpus. German words are represented in bold.

(2) **Flugzeug** § -la
plane -INS
'with a plane'

The treebank contains only sentences with intrasentential and intra-word CS. Due to limited time and funding, we transcribed only sentences that are included in the treebank.

3.2. Treebank Annotation

The treebank sentences to annotate consist of the normalised tier of the transcriptions. We followed the Universal Dependencies (UD) framework (Nivre et al., 2016; Nivre et al., 2020; de Marneffe et al., 2021) as the annotation scheme. The annotation layers are language IDs, lemmas, POS tags, morphological analyses, and dependency relations.

The treebank sentences extracted from the Praat annotations were segmented following UD rules. The language IDs assigned to intervals in Praat were automatically mapped to token-level language IDs with a more fine-grained tag set. The tokens in `TR`, `DE`, and `LANG3` intervals inherited their language ID tags, then two rules override these values. The tokens with intra-word CS were assigned the `MIXED` tag. Punctuation and special symbols get an `OTHER` tag. We kept a mapping between the treebank sentences and the original transcripts from the Praat files to maintain alignment. Other annotation layers are then applied manually on the treebank sentences. The annotated version of the sentence in Figure 1 is given in Figure 2. The glossed version can be found in Example (5). In total, the treebank contains 2184 sentences and 36940 tokens.

4. Anonymisation

GDPR Article 4 defines pseudonymisation as "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without use of additional information, provided that such additional information is kept separately [...]"[3] We started privacy protection at the very first step of data collection. The participants signed a printed consent form that included their names and contact information. Then they were given a unique participant ID. The consent forms were stored separately and only the project coordinator could access them. The participants were also given the contact information of the project coordinator so that they can reserve the right to have their recordings and transcriptions deleted any time.

During interviews we also collected metadata. Each recording has an accompanying metadata file where participant IDs are used as speaker identifiers. We asked for mother tongues, dialects, and foreign languages of participants, as they are crucial information in our research, but we used vague values for other types of information. For instance we asked for their age range instead of the actual age and we presented them a list of occupation categories to choose from, instead of directly asking for their occupation.

4.1. Transcript Anonymisation

The most common categories that have undergone anonymisation in the corpus are proper names, specifically person, company, university, and location names, and course and job titles. There are few cases of personal information, such as hair colour, as well.

Proper names make up the majority of the anonymisation cases. All person names were anonymised, with the exception of celebrity names. For instance, in Example (3) from the SAGT corpus, we did not anonymise *Christiano Ronaldo*.

(3) Zamanında o **der war besser als**
time.POSS3SG.LOC he he was better than
Christiano Ronaldo.
Christiano Ronaldo.
'Back in the day he was better than Christiano Ronaldo.'

In contrast to person names, titles and location names do not always carry sensitive information. However, when combined with other information in context, they can help identify the speaker. The standalone sentence in Example (4) for instance does not reveal any sensitive information. But combined with other sensitive references in the original version of the corpus the course name *Konstruktionstechnik* 'Engineering Design' could narrow down the identity of the speaker. To avoid this potential risk, we replaced it with *Messtechnik* 'Measurement Technology' in the treebank.[4]

(4) Vallaha şimdi yaptığımız ders **ist**
Honestly now do.PART.2PL course is
Konstruktionstechnik.
construction technology.
'Well the course we take now is Engineering Design.'

Selective pseudonymisation is subjective. While in some cases such as participant names it is clear that anonymisation is necessary, some other cases are subject to discussion. In Example (5) we decided to keep *İstanbul* as it is, but to anonymise *Adana*. Istanbul is the largest hub for flying to Turkey[5] and combined with the fact that the rest of the recording does not contain or reveal any personal information relevant to the flight,

[3] `https://gdpr-info.eu/art-4-gdpr/`

[4] The *original* examples are also anonymised here, not to reveal any protected information that is actually replaced in the treebank.

[5] More than 90,000 domestic and 180,000 international passengers per day travelled via Istanbul's two airports at the time of the recordings. `en.wikipedia.org/wiki/Istanbul_Airport`, `en.wikipedia.org/wiki/Istanbul_Sabiha_Gokcen_International_Airport`

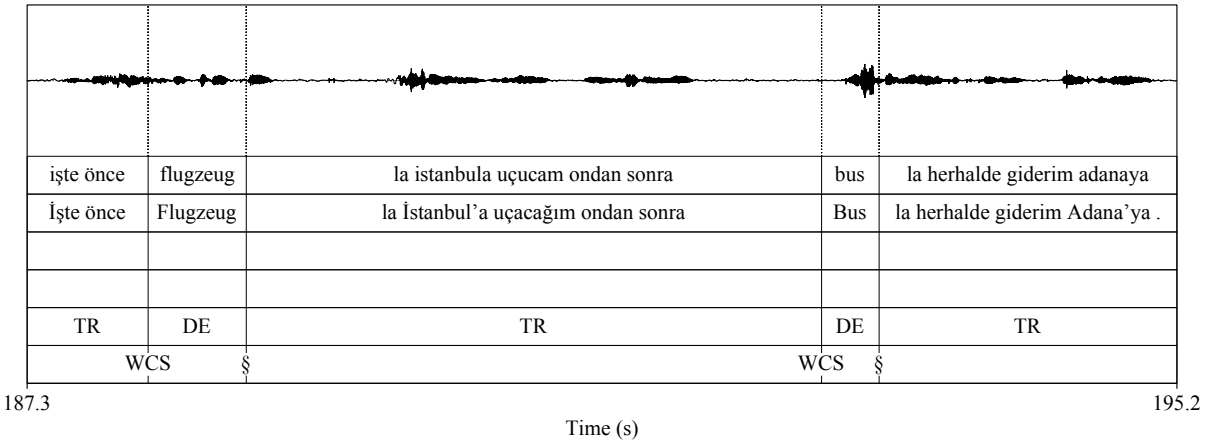

Figure 1: The original Praat annotation of the sentence *İşte önce Flugzeugla İstanbul'a uçacağım ondan sonra Busla herhalde giderim Adana'ya.* 'Well, I will first fly to Istanbul with a plane, then I go to Adana probably with a bus'. The six annotation tiers are described in Section 3.1

.

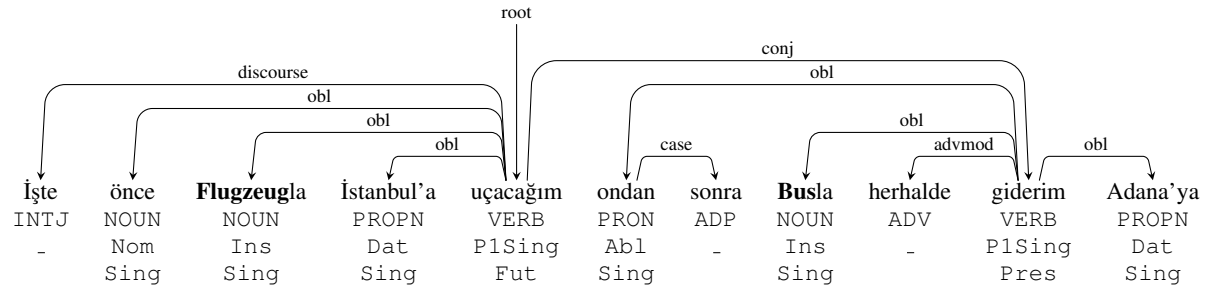

Figure 2: The treebank representation of the sentence *İşte önce Flugzeugla İstanbul'a uçacağım ondan sonra Busla herhalde giderim Adana'ya.* 'Well, I will first fly to Istanbul with a plane, then I go to Adana probably with a bus'. Some features are removed for the sake of brevity. All features could be found in UD's CoNLL-U representation in Figure 4 in the Appendix.

it is generic enough to retain as is. The final destination however could perhaps help identify the speaker (together with the rest of the recording), therefore it is anonymised.

(5) İşte önce **Flugzeug**la İstanbul'a
There first plane.INS Istanbul.DAT
uçacağım ondan sonra **Bus**la herhalde
fly.FUT.1SG that.ABL after bus.INS probably
giderim Adana'ya.
go.PRES.1SG Adana.DAT
'Well, I will first fly to Istanbul with a plane, then I go to Adana probably with a bus.'

Identifying sensitive references goes hand in hand with finding surrogate values. Our focus in this corpus is studying code-switching, hence we wanted to keep the language ID alternations intact. It means all references were replaced with surrogates of the same language. In addition, we paid attention to three criteria in choosing replacements:

Phonological Parallelism In Turkish phonology surface realisation of suffixes change according to vowel harmony and consonant alternations (Göksel and Kerslake, 2005, p. 21,44). For instance, the Genitive forms of the male names *Ahmet* and *Tolga* are *Ahmet'in* and *Tolga'nın*, respectively. Note the surface differences in suffixes *-in* and *-nın*. If we anonymised *Ahmet*, we chose a surrogate that would follow the same alternation pattern against phonological rules. For instance *Mehmet* would be a proper surrogate for *Ahmet*. This parallelism ensures that when we anonymise a word we can anonymise its lemma only, and all its inflected forms preserve the correct orthography.[6] Since many German words also have Turkish suffixes due to intra-word code-switching, when they were anonymised, their surrogates also followed phonologi-

[6]This was required because the lemma and the surface form are stored as two separate layers in the annotation representation (cf. Figure 4) in the Appendix.

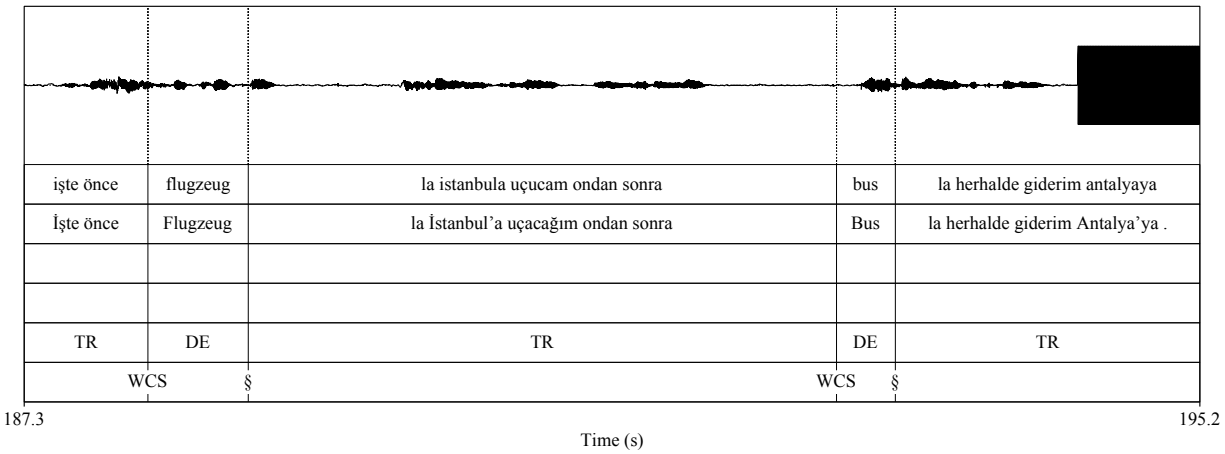

Figure 3: The anonymised Praat annotation of the sentence *İşte önce Flugzeugla İstanbul'a uçacağım ondan sonra Busla herhalde giderim Antalya'ya.* 'Well, I will first fly to Istanbul with a plane, then I go to Antalya probably with a bus'.

cal parallelism.

Syntactic Structure Some sensitive references consist of multiple tokens with a syntactic structure, e.g., titles, organisations, or locations. When they were anonymised, we refrained from changing their annotation layer properties. That is, we pay attention to have the same number of tokens, with the same POS tags and morphological features, so that the dependency relations between them still hold. For instance, the course name *Current Advances in Mathematics* is a noun phrase that consists of a sequence of an adjective, a noun, a preposition and another noun. A good surrogate for this reference could be *Selected Topics in Physics*.

Semantic Consistency While we kept the syntactic structure of the anonymised references we also took into account the broader context in multiple levels, beginning with semantic consistency within a sentence, as in (5). Similar to the reference, the surrogate is a large city in Turkey, and it is plausible to travel there by bus from Istanbul. The consistency criterion extends to the entire conversation and can span to the other recordings of the same speaker if they participated in more than one. For instance if other sentences of the conversation explicitly mentioned that a city is by the seaside and has a small population, we found a surrogate city with similar properties.

All anonymisation was done manually. After preparing guidelines and training annotators, we gave them ten of the corpus files for dual annotation. One of the annotators identified 93 references, while the other had 78 candidates. 64 of the references were identified by both annotators. The first annotator identified 29 references that the second annotator did not identify. The second annotator differed from the first one in 14 references. When we looked at decisions more closely, we noticed that the first annotator was more conservative about anonymising indirect references, e.g., course names. The disagreements were resolved via meetings. For the remaining files, one annotator identified the references and the other annotator controlled them. Similarly, disagreements and sometimes the surrogate values were discussed as part of weekly annotation meetings. We also used heuristics to verify anonymisation decisions, e.g. filtering proper names that are not anonymised to determine if they were sensitive references or not. This check was manually done by the first author to ensure quality control by multiple viewpoints.

4.2. Speech Anonymisation

Once we completed the treebank anonymisation, we moved to anonymising audio files. During transcription the audio files and transcripts were aligned to the sentence boundaries and code-switching points. That is, there was no token-level alignment. Since anonymisation happens in the token level, the first step was to automatically create alignments via a Praat script. The script uses Praat's built-in alignment tool that supports both Turkish and German. We switched between language settings during alignment based on the labels in the `lang` tier (cf. Figure 1). The alignment created a point tier for token boundaries and an empty tier `anon` to be used in the next step.

Once the token alignments are ready, we manually went over the corpus in Praat and replaced the references to be anonymised with the surrogates used in the treebank files, both in `_verbal` and `_norm` tiers. During this process we also checked if the alignment boundaries around the references were correct and if not, we corrected them manually. We also used the `anon` tier to mark the references to be anonymised with a `BEEP` label.

In the next step, we converted the intervals with the BEEP label to a beep sound in the audio files. At this point the intermediate tiers that were introduced for token-level alignment were no longer necessary, therefore we deleted them. The resulting anonymised transcripts and corresponding audio files are exemplified in Figure 3. The tiers are identical to the original representation in Figure 1. Due to anonymisation, the references *adanaya* in tier spk1_verbal and *Adana'ya* in tier spk1_norm are replaced with *antalyaya* and *Antalya'ya* respectively. The sound corresponding to the token *Adana'ya* is replaced with a beep sound, specifically with a sine wave of 500 Hertz. The approximately 400 oscillations of the wave within the 0.82 seconds of the anonymised word appear as a black rectangle in the figure.

As mentioned in Section 3.1, we have not transcribed monolingual sentences in the recordings. Since they might also contain sensitive information, we muted the sound of such sentences automatically by identifying them thanks to boundary markers annotated in the codesw tier. Finally, the anonymised files went through a manual verification step. We asked another bilingual Computational Linguistics student to listen to all the audio files and to compare them to accompanying Praat annotations to ensure that anonymisation and transcript-sound alignments were correctly done.

5. Conclusion

In this paper we presented the anonymisation stages of the SAGT speech corpus and treebank. We have employed a selective pseudonymisation strategy for anonymisation. In identifying sensitive references our goal was to find a balance between preserving the privacy of participants and preserving the data integrity of the original corpus. Thus we selected references to be anonymised based on context. In choosing surrogates for those references, we took into account their languages to retain code-switching properties and also paid attention to phonological parallelism, syntactic structure, and semantic consistency.

During annotation, we first manually identified and replaced sensitive references on the treebank side and then semi-automatically mapped them to corresponding audio intervals on the speech side. We carried out quality assurance by double annotation and group discussions for treebank anonymisation and manual verification for speech anonymisation. Clearly, there is room for improvement in the evaluation of the anonymisation quality. As we discussed with examples in previous sections, selective pseudonymisation is subjective, both in identifying sensitive references and in deciding surrogate values. Therefore by definition, selective pseudonymisation is hard to reproduce. Nevertheless, more fine-grained guidelines could be a step towards more systematic pseudonymisation. We believe the strong point in our case is to have multiple passes on the dataset by different people, which reduces the chances of mistakes and increases consistency. However this approach is not applicable to larger datasets with relatively small annotation budgets. Thus, quality assurance and annotation efficiency are two important topics to be explored in the future.

The treebank is publicly available for research purposes in the Universal Dependencies repositories[7]. The audio files, corresponding annotated transcripts, and the accompanying metadata files are available for research purposes via an individual licence agreement. The licence agreement can be obtained by contacting the email address sagt-audio@ims.uni-stuttgart.de.

6. Acknowledgements

This work is funded by DFG via project CE 326/1-1 "Computational Structural Analysis of German-Turkish Code-Switching" (SAGT). We thank Reha Sakızlı for adapting Praat scripts to the SAGT project and Gökçe Taban for verifying speech anonymisation.

7. Bibliographical References

Cohn, I., Laish, I., Beryozkin, G., Li, G., Shafran, I., Szpektor, I., Hartman, T., Hassidim, A., and Matias, Y. (2019). Audio de-identification - a new entity recognition task. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 197–204, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Corti, L., Day, A., and Backhouse, G. (2000). Confidentiality and informed consent: Issues for consideration in the preservation of and provision of access to qualitative data archives. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 1(3), Dec.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 07.

El Emam, K. (2013). *Guide to the de-identification of personal health information*. CRC Press.

Fang, F., Wang, X., Yamagishi, J., Echizen, I., Todisco, M., Evans, N., and Bonastre, J.-F. (2019). Speaker anonymization using x-vector and neural waveform models. *arXiv preprint arXiv:1905.13561*.

Göksel, A. and Kerslake, C. (2005). *Turkish: A comprehensive grammar*. Routledge.

Jin, Q., Toth, A. R., Schultz, T., and Black, A. W. (2009). Voice convergin: Speaker de-identification by voice transformation. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3909–3912.

[7] https://github.com/UniversalDependencies/UD_Turkish_German-SAGT/tree/master

Justin, T., Štruc, V., Dobrišek, S., Vesnicer, B., Ipšić, I., and Mihelič, F. (2015). Speaker de-identification using diphone recognition and speech synthesis. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 04, pages 1–7.

Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., and Samore, M. H. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10(1):1–16.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France, May. European Language Resources Association.

Pobar, M. and Ipšić, I. (2014). Online speaker de-identification using voice transformation. In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1264–1267.

Rock, F. (2001). Policy and practice in the anonymisation of linguistic data. *International Journal of Corpus Linguistics*, 6(1):1–26.

Uzuner, Ö., Luo, Y., and Szolovits, P. (2007). Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

8. Language Resource References

Boersma, P. and Weenink, D. (2019). Praat: doing phonetics by computer [computer program]. http://www.praat.org/. Version 6.1, retrieved: 2019-08-13.

Çetinoğlu, Ö. (2016). A Turkish-German code-switching corpus. In *The 10th International Conference on Language Resources and Evaluation (LREC-16)*, Portorož, Slovenia.

Çetinoğlu, Ö. and Çöltekin, Ç. (2019). Challenges of annotating a code-switching treebank. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 82–90, Paris, France, August. Association for Computational Linguistics.

Çetinoğlu, Ö. and Çöltekin, Ç. (2022). Two languages, one treebank: building a Turkish–German code-switching treebank and its challenges. *Language Resources and Evaluation*, pages 1–35.

Eder, E., Krieg-Holz, U., and Hahn, U. (2019). De-identification of emails: Pseudonymizing privacy-sensitive data in a German email corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 259–269, Varna, Bulgaria, September. INCOMA Ltd.

García-Sardiña, L., Serras, M., and Del Pozo, A. (2018). Es-port: a spontaneous spoken human-human technical support corpus for dialogue research in spanish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Jensen, K. N., Zhang, M., and Plank, B. (2021). De-identification of privacy-related entities in job postings. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 210–221.

Lüngen, H., Beiswenger, M., Herzberg, L., and Pichler, C. (2017). Anonymisation of the dortmund chat corpus 2.1. Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora17), 3-4 October 2017, Eurac Research, Italy, pages 21 – 24, Bolzano.

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, Italy, May. European Language Resources Association (ELRA).

Megyesi, B., Granstedt, L., Johansson, S., Prentice, J., Rosén, D., Schenström, C.-J., Sundberg, G., Wirén, M., and Volodina, E. (2018). Learner corpus anonymization in the age of GDPR: Insights from the creation of a learner corpus of Swedish. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pages 47–56, Stockholm, Sweden, November. LiU Electronic Press.

Patel, N., Accorsi, P. A., Inkpen, D., Lopez, C., and Roche, M. (2013). Approaches of anonymisation of an sms corpus. In *CICLing*.

Põldvere, N., Frid, J., Johansson, V., and Paradis, C. (2021). Challenges of releasing audio material for spoken data: The case of the london–lund corpus 2. *Research in Corpus Linguistics*, 9(1):35–62, Jun.

Appendix

Universal Dependencies uses the tabular CoNLL-U format[8] for data representation. Figure 4 gives the complete annotation features of the sentence in Example (5 in this format. The sensitive reference in this example is *Adana'ya* 'to Adana'. It is marked to be anonymised with the surrogate value *Antalya* in the last column, denoted with the feature `AnonLater`. After sensitive references and surrogates were identified in the entire corpus, the string *Adana* was automatically replaced with *Antalya* both in the surface form and in the lemma.

[8] `https://universaldependencies.org/format.html`

```
1  İşte         işte     INTJ  _ _                          5  discourse _ LangID=TR
2  önce         önce     NOUN  _ Case=Nom|Number=Sing      5  obl       _ LangID=TR
3  Flugzeug$la  Flugzeug NOUN  _ Case=Ins|Number=Sing      5  obl       _ LangID=MIXED
                                                                |DeGender=Neut|DeCase=Dat
4  İstanbul'a   İstanbul PROPN _ Case=Dat|Number=Sing      5  obl       _ LangID=TR
5  uçacağım     uç       VERB  _ Aspect=Perf|Evident=Fh    0  root      _ LangID=TR
                                  |Mood=Ind|Number=Sing
                                  |Person=1|Tense=Fut
6  ondan        o        PRON  _ Case=Abl|Number=Sing      10 obl       _ LangID=TR
                                  |Person=3|PronType=Prs
7  sonra        sonra    ADP   _ _                          6  case      _ LangID=TR
8  Bus$la       Bus      NOUN  _ Case=Ins|Number=Sing      10 obl       _ LangID=MIXED
                                                                |DeGender=Masc|DeCase=Dat
9  herhalde     herhalde ADV   _ _                          10 advmod    _ LangID=TR
10 giderim      git      VERB  _ Aspect=Hab|Evident=Fh      5  conj      _ LangID=TR
                                  |Mood=Gen|Number=Sing
                                  |Person=1|Tense=Pres
11 Adana'ya     Adana    PROPN _ Case=Dat|Number=Sing      10 obl       _ LangID=TR
                                                                |AnonLater=Antalya
                                                                |SpaceAfter=No
12 .            .        PUNCT _ _                          5  punct     _ LangID=OTHER
```

Figure 4: The CoNLL-U representation of the sentence *İşte önce Flugzeugla İstanbul'a uçacağım ondan sonra Busla herhalde giderim Adana'ya.* 'Well, I will first fly to Istanbul with a plane, then I go to Adana probably with a bus'. The glosses of the sentence are given in (5) and the tree representation is given in Figure 2.