CLARIN Annual Conference Proceedings

2024

Edited by

Vincent Vandeghinste and Thalassia Kontino

15 – 17 October 2024
Barcelona, Spain

Please cite as:
CLARIN Annual Conference Proceedings, 2024. ISSN 2773-2177 (online).
Eds. Vincent Vandeghinste and Thalassia Kontino.
Barcelona, Spain, 2024.

Preserving Privacy in Small Communities: Tailored Anonymization Techniques for Icelandic Conversational Data

Elena Callegari
University of Iceland, Reykjavík
ecallegari@hi.is

Agnes Sólmundsdóttir
University of Iceland, Reykjavík
ags46@hi.is

Anton Karl Ingason
University of Iceland, Reykjavík
antoni@hi.is

Abstract

We examine the challenges and methodologies of anonymizing a dataset of Icelandic conversations, emphasizing the need for language-specific strategies due to Iceland's small, interconnected population and the morphological richness of the language. We discuss the importance of preserving grammatical elements such as case and gender to maintain data utility for linguistic research. The study proposes an anonymization technique that balances data utility with privacy, resorting to pseudonyms that match the original phrase's linguistic properties to protect individual identities while preserving the structural integrity of the Icelandic language.

1 Introduction

In the era of data-driven decision-making, the collection and analysis of conversational data have become essential for advancing linguistic research and possibly improving language technologies. However, the use of such data raises significant privacy concerns, particularly in the context of small, tightly-knit communities where individuals are easily identifiable even from seemingly innocuous information. This paper explores the anonymization of a dataset of audio-recorded Icelandic conversations, underlining the necessity of language-specific anonymization protocols to protect participant privacy while maintaining the utility of the data.

The conversations in question have been audio-recorded with the intention of creating the very first Icelandic Dementia corpus. We have been collecting speech samples from Icelandic individuals suffering from various stages of Alzheimer's Disease (AD) as well as from healthy, age-matched individuals Callegari et al., 2023, 2024. We plan to release the transcriptions of the conversations in the form of a publicly accessible dataset, so that any researcher working on AD, clinical applications for NLP, or both, may also make use of the data we are collecting.

Anonymization involves processing personal data to remove or obscure identifying details, ensuring that individuals cannot be identified directly or indirectly by the retained data. Anonymization helps mitigate the risks of unauthorized data re-identification which could lead to privacy invasions, discrimination, or other forms of harm. Effective anonymization allows researchers to share and analyze datasets without compromising individual privacy, thereby adhering to ethical standards and legal requirements, such as Europe's General Data Protection Regulation (GDPR). Iceland, with its population of roughly 370,000, exemplifies a scenario where simple anonymization methods, such as merely removing proper names from data files, may not suffice. The Icelandic language is used by a relatively small speaker community. In such environments, these elementary anonymization approaches might leave sufficient linguistic and contextual clues that could potentially lead to the identification of individuals. This risk is particularly high in cases where unique personal references, local dialects, or specific sociolects are prevalent.

Moreover, the morphological richness of Icelandic—characterized by a complex system of inflections and derivations—means that anonymizing content without distorting linguistic structures requires careful consideration. Preserving the grammatical integrity of the language is essential for ensuring that the anonymized data remains valuable for linguistic research and the development of natural language processing tools.

2 Data to Anonymize

In the broader scope of data privacy and protection, certain types of information commonly require anonymization across various contexts and geographies. These include personal identifiers, such as full names, addresses, social security numbers; these are always anonymized to prevent the straightforward recognition of an individual's identity. Information that is also generally anonymized are health records, financial data, and personal communications, i.e. any sensitive information that could impact an individual's privacy and security if exposed.

In smaller countries or communities, the effectiveness of these basic anonymization techniques may diminish due to the increased likelihood of identifying individuals through indirect means. For example, in small communities, certain cultural practices, local events, or affiliations (e.g., membership in specific local organizations) can serve as identifiers. In Icelandic data, references to participation in local festivals, or membership in less-common local clubs could inadvertently reveal someone's identity. Moreover, while anonymizing a city's name might not be necessary for countries with larger populations, in Iceland, the name of a town or district might need to be anonymized due to the small number of inhabitants and the resulting ease of individual identification. Iceland also has a notable interest in genealogy, facilitated by extensive public and private records that trace family genealogies. Any data hinting at familial connections or lineage, such as particular names or patronymics, which might be benign in larger populations, could lead to individual identification in Iceland. In smaller or more specialized professional communities, detailing someone's educational background or employment history (specific roles, small or niche industries) can be particularly revealing. In Iceland, mentioning a person's role in a specific sector, like fisheries or geothermal energy, might narrow down the identity of individuals far more than in larger economies. Moreover, in small populations, even anonymized health data might be re-identifiable if it includes rare health conditions or treatments that are unique to a few individuals within the community. Finally, in datasets with fine-grained demographic segmentation, details that might be individually harmless, such as gender, employment, specific names of towns and festivals, can be used in combination with other data to re-identify individuals. Consider for instance examples (1) and (2):

(1)
Ég var alin upp á Hólmavík hjá foreldrum mínum, Árna og Jónu,
I was raised.FEM in Hólmavík at parents.DAT mine.DAT, Árni.DAT and Jóna.DAT

'I was raised in Hólmavík at my parents, Árni and Jóna.'

(2)
og pabbi minn var læknirinn í bænum og mamma vann í móttökunni hjá honum.
and dad.NOM mine.NOM was doctor-the.NOM in town-the.DAT and mom.NOM worked in recept

'and my dad was the town doctor and my mom worked at his reception.'

In datasets of languages spoken by larger communities, the personal details provided in examples (1) and (2) might seem innocuous, yet in a society as small as the Icelandic one, such a combination of geographic, familial, and occupational information may be enough to lead to individual re-identification. This is especially the case for our Dementia dataset, given that we interviewed a very specific demographic group (individuals aged 60 to 80 at the time of the interview, some of whom had a neurocognitive condition). Moreover, part of our interview protocol consists in asking participants to recall their childhood home; this often lead to descriptions of family members, and to mentions to schools, specific places and organizations.

3 Icelandic Morphology & Anonymization Strategies

One of the defining characteristics of the Icelandic language is its inflectional morphology. Icelandic has four cases: nominative, accusative, dative, and genitive. Nouns in Icelandic agree in gender (masculine,

feminine, neuter) and number (singular, plural). Icelandic verbs are conjugated according to mood, tense, voice, person, number, and gender. The richness of this system is a significant aspect of the language's morphology. The morphological complexity of Icelandic has direct implications for data anonymization, particularly when considering the need to maintain linguistic integrity for research purposes. Case usage in Icelandic can reveal subtle demographic or sociolinguistic patterns. For instance, Callegari et al., 2024 have shown variations in the use of the dative case across different age groups of Icelandic speakers. Ideally, anonymization strategies should therefore preserve case information to allow for the study of such linguistic phenomena without compromising the privacy of the individuals involved.

Recently, numerous studies focusing on text anonymization across various languages have emerged (e.g. Francopoulo and Schaub, 2020; Mozes and Kleinberg, 2021; Strathern et al., 2020; Adams et al., 2019). Currently, there is no default anonymization method, as the choice of method varies across different fields of research and is dependent on the intended purposes of the data. Most studies suggest the following four key requirements for anonymizing text before publication: (i) ensuring the anonymity of participants and individuals mentioned in the text, (ii) allowing in-house semantic data analysis and language analysis through NLP, (iii) proof that an anonymization has taken place, and (iv) the method should be applicable for different European languages (Francopoulo and Schaub, 2020). Several options to anonymize conversational transcripts exist; what can sometimes be challenging is ensuring that all of the four requirements listed above are fulfilled at the same time.

One particularly straightforward method for anonymizing conversational data is to fully redact sensitive information by completely removing personally-identifying information, and replacing it with an "X" token. This can be suitable for the public release of government documents without secondary analysis. However, this method does not fulfill the aforementioned requirement number (ii), as important linguistic information such as semantic cohesiveness, syntax and other lexical properties needed for in-depth linguistic analysis would be lost (Mozes and Kleinberg, 2021).

Another method mentioned in Strathern et al., 2020 is aggregation, where identifiable information units are coarsened or aggregated by creating classes or categories, e.g. replacing someone's age with age classes, or replacing a specific person's name with a relevant but vague role, such as "student" or "employee". Aggregation therefore keeps some semantic cohesiveness and can be used for secondary analysis, i.e. sentiment analysis and topic modelling, but to a limited extent. The drawbacks are that, as in the first anonymization method, too much linguistic information is lost for NLP purposes.

A third and frequently mentioned method is the use of pseudonyms, which refers to renaming identifiable units, such as people, institutions etc. This can be done in two ways: either by using anonymous place-holders (e.g., 'Person-Name', 'City-Name' etc.) or by using unique identifiers, such as choosing a random name with comparable properties. For example, one could replace the name *Björk Gumundsdóttir* with a name like *Ösp Davísdóttir*: both are female names, and both feature a similar number of syllables and hence have a comparable length. The latter method, recommended by Aldridge et al. (2010), is considered more suitable for linguistic data analysis as the chosen pseudonyms match all linguistic properties of the original utterance.

3.1 Current Practices in Icelandic Data Anonymization

In our collaboration with other researchers working on anonymization for Icelandic corpora, several practical strategies have been highlighted. For example, in the court document corpus published by Clarin IS (Barkarson et al., 2022), a Named Entity Recognizer (NER), MIM-GOLD-NER, was used to identify and replace personal names with strings encoding the first letter, gender, and case marking. While this approach preserved grammatical information relevant to Icelandic, the accuracy of the model in recognizing foreign names and professions remained a challenge. Additionally, place names, streets, and organizations were anonymized using similar approaches, although the NER used in this project was not specifically designed to handle foreign names effectively.

These experiences suggest that for certain corpora, particularly those that involve sensitive legal or clinical information, a combination of Named Entity Recognition and Part-of-Speech (PoS) tagging could be an effective strategy for ensuring linguistic integrity.

4 Our Chosen Anonymization Practice

The corpus we will release will contain manually transcribed speech samples elicited from individuals of Icelandic nationality who are aged 60 to 80 and who are either healthy or suffering from various degrees of Alzheimer's Disease. Seeing as we are working with sensitive information, e.g. clinical diagnostic information, anonymity is of the utmost importance. However, as the main focus of our study is on the specific effects AD can have on language production, it is important not to lose any linguistic information relevant to the study.

The initial step in the anonymization of our data will involve establishing a comprehensive list of basic entity types that could contain identifiable information. This includes full names, addresses, social security numbers, and other direct identifiers. Additionally, as outlined in section 2, we will identify a subset of entity types that may be traceable within smaller communities like Iceland. This subset will potentially include names of organizations, schools, cities, towns, regions, and employment roles, among others. Following the creation of this detailed inventory of elements requiring redaction, we will proceed to anonymize our dataset.

We intend to follow the "pseudonym" anonymization method discussed in Section 3, by which personal identifiers are replaced with pseudonyms (or made-up numbers in the case of numerical information, such as particular dates or mentions of one's age), therefore retaining all grammatical information while protecting anonymity. Annotators will carefully mark the linguistic properties, morphological and syntactic information of each phrase to be anonymized and choose a random pseudonym that features the same linguistic properties. To illustrate our anonymization process, consider the fragment sentence in (3), in which a participant's enrollment to a local Icelandic school is discussed:

(3) *Já ég var í sa- sama skólanum ee í Langholtsskóla*
Yes I was in sa- same school uh in Langholtsskóli-DAT.
'Yes I went to the same school, Langholtsskóli (*note: this is a school in Reykjavík*).'

The example in (3) contains possibly identifiable information, i.e. the name of a specific school. This sentence could for example be published as in (4), where the original school name has been replaced with a pseudonym -a fake school name-, maintaining the grammatical properties of the original sentence.

(4) *Já ég var í sa- sama skólanum ee í Borgarskóla*
Yes I was in sa- same school uh in Borgarskóli-DAT.
'Yes I went to the same school, The City School (=*a fake school that does not exist*).'

Similarly, examples (1) and (2) from Section 2 could be anonymized as (5) and (6) respectively:

(5)
Ég var alin upp á Ólafsvík hjá foreldrum mínum, Bjarna og Önnu.
I was raised.FEM in Ólafsvík at parents.DAT mine.DAT, Bjarni.DAT and Anna.DAT
'I was raised in Ólafsvík at my parents, Bjarni and Anna.'

(6)
og pabbi minn var bæjarstjórinn í bænum og mamma vann í móttökunni hjá honum.
and dad.NOM mine.NOM was mayor-the.NOM in town-the.DAT and mom.NOM worked in recepti
'and my dad was the mayor and my mom worked in his reception.'

In addition to replacing the town name "Hólmavík" and the personal names of the parents, "Árni" and "Jóna", with lexically similar names, the occupation "doctor" has been substituted for "mayor". This preserves the grammatical properties of the original utterance and maintains semantic cohesiveness while avoiding possible de-identification of the participant.

References

Adams, A., Aili, E., Aioanei, D., Jonsson, R., Mickelsson, L., Mikmekova, D., Roberts, F., Valencia, J. F., & Wechsler, R. (2019, September). AnonyMate: A toolkit for anonymizing unstructured chat data. In L. Ahrenberg & B. Megyesi (Eds.), *Proceedings of the workshop on nlp and pseudonymisation* (pp. 1–7). Linköping Electronic Press. https://aclanthology.org/W19-6501

Aldridge, J., Medina, J., & Ralphs, R. (2010). The problem of proliferation: Guidelines for improving the security of qualitative data in a digital age. *Research Ethics*, *6*(1), 3–9.

Barkarson, S., Steingrímsson, S., & Hafsteinsdóttir, H. (2022). Evolving large text corpora: Four versions of the icelandic gigaword corpus. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2371–2381.

Callegari, E., Nowenstein, I. E., Kristjánsdóttir, I. J., & Ingason, A. K. (2024). Automatic extraction of language-specific biomarkers of healthy aging in icelandic. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 1915–1924.

Callegari, E., Sólmundsdóttir, A., & Ingason, A. K. (2023). The acode project: Creating a dementia corpus for icelandic. *CLARIN Annual Conference Proceedings*, 100.

Francopoulo, G., & Schaub, L.-P. (2020). Anonymization for the GDPR in the Context of Citizen and Customer Relationship Management and NLP. *workshop on Legal and Ethical Issues (Legal2020)*, 9–14. https://hal.science/hal-02939437

Mozes, M., & Kleinberg, B. (2021, March). *No intruder, no validity: Evaluation criteria for privacy-preserving text anonymization*.

Strathern, W., Issig, M., Mozygemba, K., & Pfeffer, J. (2020). *Qualianon - the qualiservice tool for anonymizing text data* (tech. rep. No. TUM-I2087).