Towards shared standards for pseudonymization of research data

Elena Volodina[1],*, Simon Dobnik[2], Therese Lindström Tiedemann[3], Ricardo Muñoz Sánchez[1], Maria Irena Szawerna[1], Lisa Södergård[3] and Xuan-Son Vu[4]

[1] *Språkbanken Text, SFS, University of Gothenburg, Sweden*
[2] *FLOV, University of Gothenburg, Sweden*
[3] *Department of Finnish, Finno-Ugrian and Scandinavian Studies, University of Helsinki, Finland*
[4] *Lund University and DeepTensor AB, Sweden*

Abstract

Pseudonymization has attracted a lot of attention recently due to legislation (e.g. the GDPR), the European Guidelines on Pseudonymization, the increased need for high-quality ethical data for the training of large language models as well as the desire to be able to share data with other researchers. This article introduces key concepts in pseudonymization, summarizes the half-way findings in the intradisciplinary research environment Mormor Karl, and proposes ways to unify and standardize the field of pseudonymization.

Keywords

Open data, GDPR, ethical AI, Mormor Karl, pseudonymization, anonymization, linguistics, Swedish, large language models, privacy

1. Introduction

There is a legal obligation to protect the privacy of data subjects as well as other people mentioned in research data [1, 2]. There exist various techniques to do so, such as encryption, authorization, data minimization, anonymization, pseudonymization [3, 4, 5]. Although none of these approaches can guarantee *absolute protection* of personal privacy [6, 7], they lower the risk of reidentification [8, 7], leading to continuous development of such approaches.

The field of pseudonymization has attracted a lot of attention lately due to legislation like the GDPR [1] and the European Guidelines on Pseudonymization[1], as well as the increased need of high-quality expert data for ethical training of language models [9, 10, 11]. According to the GDPR [1, Art.4:5], pseudonymization is a technique where *Personally Identifiable Information* (PII) have been replaced with substitutes and it is only possible to re-identify a person through additional information, such as name-id keys, which are kept separate from the data. This is a critical difference from the *anonymous* data, where no such keys exist and no reidentification is feasible. This potentially means that once the project destroys the keys, the data should no longer be under the jurisdiction of the GDPR and can be made open to the public. However,

Huminfra Conference 2025, Stockholm, 12–13 November 2025.

*Corresponding author.
mormor.karl@svenska.gu.se (E. Volodina)

[1] https://www.edpb.europa.eu/system/files/2025-01/edpb_guidelines_202501_pseudonymisation_en.pdf

0. orig.	Hi my name is Deniz Kaya, I live in Sweden in Jämtland and I speak Turkish
1. detect.	Hi my name is Deniz Kaya, I live in in Sweden Jämtland and I speak Turkish
2. label.	Hi my name is @firstname_male.1 @surmale.2, I live in @country.3 in @region.4 and I speak @lang.5
3. pseudo	Hi my name is Alex Bax, I live in Spain in Andalusia and I speak Spanish

Table 1
Example of a sentence containing Personally Identifiable Information (PII) and the processing steps.

the national legal landscape may obstruct that step, in our particular case, the Swedish Ethical Review Authority requires the original data (including the keys) to be preserved for the first ten years after data release; and the Swedish Archives Act, Arkivlagen (SFS 1990:782), protects the documentation (including the above-mentioned keys) from being destroyed in an unauthorized way.

Despite much attention the field is still not unified, and it is challenging to compare the results achieved by different research groups. As a community we need to understand the extent of pseudonym effects on research conclusions, define the tolerance levels and to find a compromise that may be acceptable for both sides - the research in a discipline, and the privacy protection. In the research environment group *Mormor Karl* (Eng. 'Grandma Karl')[2] we work on this particular approach to privacy protection, *pseudonymization*, and with a focus on text-based (Swedish) linguistic research data. In the project, we delve into both the methodological and practical issues related to pseudonymization as a way to secure open access to research data [12]. The *practical issues* cover approaches to detect, label and replace personal information - both manually and automatically [e.g. 13]. The *methodological issues* cover, among others, the effects of pseudonymization on research conclusions [e.g. 14, 15, 16]; effectiveness of privacy protection through pseudonymization; as well as the semantic and cultural value of original tokens versus their pseudonyms [e.g. 17, 18, 19]. In this paper, we propose a few directions towards the unification of the field, after shortly outlining the research context for pseudonymization.

2. Pseudonymization: the basics

We define *pseudonymization* as "the process of replacing an individual's personal data with a pseudonym, which is not related to the original data" [12]. An important notion for pseudonymization is *Personally Identifiable Information* (PII). PII is any data that can be used to distinguish, trace or identify an individual, directly, indirectly or in combination with other information sources. PII is conventionally split into categories, such as names, institutions, geographical names and similar. Table 1:2 exemplifies some of PII categories, e.g. @surname, @country, etc. Another related notion is *sensitive information*, such as sexual orientation, political views, religion, ethnical background and medical condition.

Pseudonymization generally includes techniques conventionally divided into several steps, as shown in Table 1. The initial steps comprise detection (Tab.1:1) and labeling (Tab.1:2) of personal (and sensitive) information in unstructured texts, such as essays [20, 21, 22], medical records

[2] https://mormor-karl.github.io/

[23, 24, 25], or court cases [26, 27, 28]. These are followed by the replacement of personal and sensitive information with (neutral) pseudonyms, epithets or codes (Tab.1:3). This process or its parts can be attempted manually [e.g. 20] or automatically [e.g. 29]. It is important to note that, while some studies of automatic pseudonymization treat detection, labelling and replacement separately [13], other approaches merge them together [cf. 30].

2.1. Pseudonymization and linguistics

The early principle in relation to pseudonymization in linguistics has been that we should choose pseudonyms which match *all* linguistic properties of the original, including the number of syllables and other length measures [31]. Recent developments in automatic pseudonymization have often instead stressed both the impossibility of retaining all linguistic properties (e.g. spelling errors and inflectional characteristics) and the need for the pseudonym (cf. the definition above) to be disconnected from the original.

Real names of people (*orthonyms*) are closely related to a person's identity [32], with associations to age, gender, ethnicity, and social background [cf. 33, 34, 35]. While pseudonyms protect the identity of the person, they may have different connotations, thus affecting how the participant is perceived [36] and possibly also how the whole text is interpreted. Wang et al. [37] stress that careful consideration is needed when renaming participants, in order to represent them in an appropriate way, e.g. using Anglo-sounding names for participants with diverse backgrounds implies that the participants' identities and background are not respected. Similarly, placenames (*toponyms*) can be connected to specific historical, cultural and topographical associations [36, 38]. Changing placenames may also affect how the text is perceived. In addition, the language of the placename might be important since in bilingual societies there might be toponyms for one place in both languages and which one you should use could depend on the language you are speaking, hence if you break that norm that might say something about your linguistic proficiency. For instance, in Finland many places have both Finnish and Swedish names, e.g. the capital of Finland is *Helsinki* in Finnish, but *Helsingfors* in Swedish. If you use the Finnish placename when speaking/writing in Swedish this will stand out and a reader/listener might interpret it as (1) not knowing Swedish very well, and/or (2) having certain attitudes regarding dual language placename policy.

But pseudonymization is not only about names, it also covers changing other PIIs in the text. This can involve changing lexical items related to relatives, occupations, health issues, etc. It can also involve changing pronouns and numerals in relation to age, buslines, house numbers etc. Basically, any linguistic item might be affected. This in turn means that the semantic (and pragmatic) context in the text can be affected immensely – *My cousin Tom is 24 years old* is not the same as *My grandma Karl is 27 years old* as we have illustrated in the name of our project.

From the linguistic point of view, we must, therefore, consider how we can retain the original meaning of the text considering lexical semantics but also coreference, contextual semantics and pragmatics. To do this our project uses methods related to linguistic theories from e.g. semantics, pragmatics and grammar. Our results are particularly important for linguistic research, but changes such as PII replacements have profound effects also on other disciplines working on textual data. An analysis in education and social sciences where the connection to certain social groups or regions has been distorted in the data is bound to affect the results [cf. 38].

2.2. Pseudonymization and Natural Language Processing

In automatic pseudonymization (see steps in Table 1) rule-based methods have often been employed [e.g. 39, 23] for the automatic detection and labelling of personal and private information, and this still remains a valid alternative for low-resource settings [40, 41]. However, approaches based on machine learning have been shown to provide the highest performance [42] in some settings. Szawerna et al. [43] argue that the concept of personal or sensitive information and the types of entities that can appear differ between domains, so it is important to be aware that the best approach for a given task or dataset might not necessarily be the same for others. Heterogeneity of classes can also play a major role: a *miscellaneous*-type category (miscellaneous) which encompasses all personal information not covered by other categories is notoriously difficult to detect automatically [44, 13], not to mention issues which then follow as part of replacement.

The most challenging and underexplored step in the automatic pseudonymization process is *pseudonym generation* [45]. This step goes beyond replacing entities with placeholders like `@name`. These pseudonyms should match the context grammatically and semantically to avoid sentences that are non-sensical in context. Besides, in linguistic data it is important to keep as much of the linguistic information as possible, and deciding on what is enough and what is too risky to the person can be extremely difficult, not to mention the possible effects on the usefulness of the data for linguistic research (cf. Section 2.1, *My cousin Tom* vs *My grandma Karl*).

Different approaches have been tried for replacement. These include manual pseudonymization [20], rule-based approaches [46, 23, 29, 21] based on ontologies and entity mapping [47], hierarchical word representations [48], statistical models [49, 30], and machine learning, including Large Language Models (LLMs) [30, 50]. Each of these has its pros and cons. *Manual approaches* are more reliable as far as keeping the semantic integrity of the text is concerned, but they are very time-consuming and risk being inconsistent. Among the automatic approaches, *rule-based approaches* and *statistical models* cannot take into account the semantic context and the common knowledge aspects of the surrounding text, whereas approaches based on *machine learning* and *LLMs* in particular tend to be better where surface semantics are concerned. However, there are some major issues which arise with the use of generative language models. They might rewrite the input text to a more "fluent" version and miss the deeper semantics. This is problematic as changes must be minimal when it comes to pseudonimyzing linguistic data collected to study language use. Additionally, larger models are often run on external servers, particularly in academic settings. This risks running into legal issues in case the texts cannot be shared due to the nature of the personal information contained within.

3. Unifying strategies: a proposal

3.1. Universal pseudo-tagset

If researchers in different research domains and languages could agree to use the same standard for pseudonymization, similar to the Universal Dependencies initiative [51], it could ensure comparability of datasets, results and it could promote the development of multilingual solutions.

We are currently exploring two approaches: the first one deals with the proposal of a *detailed* pseudo-tagset, that could cover the needs of all research domains [e.g. 43]. The second is the opposite of the first and deals with *reducing all tags to one "personal" tag* [e.g. 52, 13]. Hypothetically, the ‘one-tag approach’ is feasible for automatic detection of personal information, and having fewer categories to choose between benefits machine-learning approaches. In turn, the ‘detailed-tag approach’ can be crucial for the selection of appropriate pseudonyms e.g. in rule-based approaches.

Some of the major conceptual and practical challenges when it comes to proposing a universal multi-tag tagset are taxonomy choice and interoperability. Szawerna et al. [43] have shown that existing tagsets vary in terms of types of personal and sensitive information that they cover depending on the genre and domain. A proposed universal tagset would have to account for all of the possibilities and feature a way to expand it in case a new kind of category becomes ubiquitous. Simultaneously, a number of annotated corpora already exist (even if they are not all publically available). Being able to easily map between at least some of the categories would help facilitate the re-annotation to a common standard.

3.2. Testing the effects on research data

Research communities and disciplines need to analyze the potential effects of pseudonymization on their research data as well as on conclusions within their disciplines. This is an important but rather neglected aspect of pseudonymization. Within our project we have done some experiments on effects of pseudonyms on language proficiency assessment. The first one showed no effects on automatic assessment when a first name within a learner essay is changed to a name from another sociocultural background [53]. Testing the same experimental setup with human assessors also showed no clear indication of a correlation between the assessment and the sociocultural associations of the first names that were used [54, 17]. The experiment is currently being extended to include more categories.

Another experiment looked into the linguistic analysis of PII strings to uncover what information may be lost if the original strings were replaced with pseudonyms [19, 55, 18]. Our analysis shows that a non-negligible number of PII are misspelled - information that is lost in the automatic pseudonymization process. However, misspellings carry important information, e.g. they can be linked to the native language of a learner e.g. *Danska* which literally means ’Danish’ but where the intention is clearly Denmark and it is influenced by the Finnish name *Tanska* ’Denmark’. Pseudonymization will lose the connection to the mother tongue and also the knowledge implied in the use of <d> instead of <t> in the spelling.

We suggest to standardize the practice of testing effects of pseudonymization on research conclusions through a few typical tasks within the target domain and report the results.

3.3. Testing effectiveness of pseudonymization

Advances in machine learning (ML) privacy and security reveal critical vulnerabilities in deep neural networks deployed in sensitive domains. There are three interconnected threat landscapes: (1) adversarial attacks that manipulate model behavior through input perturbations, (2) reidentification attacks that extract sensitive training data through model inversion and membership inference, and (3) motivated intruders - individuals who use publicly available or background information to attempt re-identifying individuals in pseudonymized text.

Collectively, they demonstrate that modern ML systems face severe privacy and security risks, with re-identification succeeding against 50–90% of vulnerable models [56], and anonymized datasets showing 10–40% re-identification rates under motivated intruder scenarios [57]. The convergence of these threats necessitates integrated defense strategies combining robust architectures, privacy-enhancing computations, and rigorous validation protocols.

Pseudonymization techniques show limited resilience against motivated intruders, as empirical studies highlight vulnerabilities in datasets thought to be de-identified when public data sources are exploited. This emphasizes the urgent need for advanced pseudonymization methods to mitigate such risks and robust methods to diagnose such vulnerabilities. The increasing focus on privacy-enhancing technologies under frameworks like the GDPR [1] and the EU AI Act [58] reflects the importance of regulatory alignment in strengthening data protection. In connection to that we encourange the community to standardize the use of reidentification tests as a way to evaluate effectiveness of pseudonymization, in addition to the standard preformance tests.

3.4. Customizable pseudonymization

Another issue to consider is whether pseudonymization should be *customizable* to different types of users and use cases. Imagine a *sociolinguist*, who will look for linguistic details of interest to describe a certain variety or issues in relation to language and power, or *a data scientist*, who will not inspect the actual data manually but as properties of a dataset as a whole, or *a forensic linguist* who will try to identify the person behind the writing. If we start applying different types of pseudonymization to different use cases and users, we *need to standardize different pseudonymization criteria and methods* so that they can be objectively reported in research (e.g. we used data X with pseudonymization Y). Thorough analysis of re-identification risks is extremely urgent in this case since having access to different pseudonymized versions of the same data increases the risk of re-identification.

A good way forward is to explore a dynamic approach for pseudonymization that would identify per text and context whether the text can be matched to a situation within the context. Hence, to achieve this one can simply compare the text with the context: is the psedonymized text consistent with the context of a research domain it is used for? The task could perhaps be treated as *Natural Language Inference Task (NLI)* [59].

Another implication of this approach is that pseudonymization would then be a tool to create several versions of a given dataset which would be adjusted to a given task. However, the approach still does not answer the question of possible re-identification in cases where several pseudonymized versions of the same text could be merged together to reconstruct the information from the original text.

3.5. Evaluation benchmarks

Another field-unifying strategy that we look into relates to the possibility of organizing shared tasks on the topic of automatic detection and pseudonymization of personal information. The community needs standardized evaluation benchmarks for pseudonymization tasks, and we expect the data from shared tasks to fulfil this function. The immediate concerns are:

(a) What data to use – much of the original data containing authentic personal information is under protection and cannot be released or used for model training. The existing publically available PII-annotated corpora [26, 60, 61, 25, 62, 63] are likely to have already been a part of the training data for LLMs, as well as cover only limited number of languages. *Synthetic data* [e.g. 64] may be good enough for development of automated detection methods, as has been indicated by Vakili et al. [65], but, since it is not authentic human-produced data it should not be used for anything beyond training models.

(b) The second concern is how to perform *automatic evaluation of the pseudonymization step*, where there are no agreed-upon standard metrics. A part of the ongoing work in the project is aimed at testing the validity of various automatic evaluation approaches, attempting to approximate the human judgements of grammatical and semantic acceptability [cf. 66].

Our approach to circumventing the legal limitations and ethical concerns surrounding the use of authentic personal information for a shared task consist of creating a corpus of fictive texts. Unlike purely synthetic data, our texts are written by human respondents. However, they are not written about natural persons, but about fictive characters invented for the sake of writing. This minimizes the risk to any natural person and approximates the way that authentic texts are written. Our data collection so far consists of fictive personal stories and fictive legal case descriptions, but more domains (e.g. medical, social media) are planned.

3.6. Standardized venues

There is a clear need for venues for meetings both within disciplines and in interdisciplinary groups to discuss pseudonymization challenges and share findings. We have initiated two venues for meeting researchers working with similar questions within *computational linguistics, computer science and privacy*, in particular: (a) the CALD-pseudo workshop[3] with its first edition at EACL 2024 and which we hope to make a standard recurrent venue on this topic. (b) the AI Trust workshop[4] with the first edition in 2024 at the WASP conference in Gothenburg. This workshop has a slightly broader focus than CALD-pseudo, but both workshops indicated signficant interest and an expressed need to meet and discuss these issues both within and across disciplines. In 2025 we also reached out to *linguists* working on the Swedish language and held a workshop at Svenskans beskrivning 40 (2025)[5] which similarly proved a joint sense of a need for more research on pseudonymization techniques and their effects on linguistic research.

Our experience with interacting with different audiences through these workshops shows that there is an increasing need for unified procedures and guidelines related to dealing with pseudonymization of research data that does not only affect linguistics and computational linguistics but also has implications for other domains, some of which may not strictly use typical natural language processing data.

[3] https://mormor-karl.github.io/events/CALD-pseudo/#cald-pseudo-workshop-at-eacl-2024

[4] https://mormor-karl.github.io/events/AITrust-Workshop/

[5] https://www.su.se/institutionen-for-svenska-och-flersprakighet/forskning/konferenser-och-seminarier/konferens-2025-svenskans-beskrivning-40-1.730200

4. Concluding remarks and future outlook

There are still many open issues in relation to pseudonymization of research data and what it means for our disciplines as well as for research participants (e.g. writers of essays, people who are interviewed) or people that are mentioned in our research data. The results of our research have clear benefits to research infrastructures on a practical level, and important implications for research on the methodological level.

Research on pseudonym generation is in its initial stages. A promising solution for the semantics-based issues might be to *generate fake or non-existent entities*. That is, names, cities, etc. that would look like real ones but not contain any semantic or pragmatic value for the reader, although preserving the grammatical features. However, both real and fake names with similar structure can have very different associations and reference.

Research on effects of pseudonymization on research conclusions has hardly begun and there is still much to be explored in relation to automatization, bias and privacy preservation. Further analysis is necessary, especially in relation to actual research questions on actual research data for us to ascertain that our research will still be reliable after pseudonymization, that the rights of our participants will be protected both in relation to privacy preservation and in terms of their right to their data, their culture and the correctness of research findings.

One open question is how to handle ‘privacy guarantees’ when research data comes in *several modalities*, i.e. not only as text datasets, but also speech/audio, video, pictures. In our project we are focusing on written (and transcribed) data, but there is a definite need to look at other modalities also and this too will need to be done in relation to different research disciplines.

Limitations

The work is focused on text modality of research data only, which means we do not look into other types of privacy protection, where video, audio, graphics (including, for example, handwritten versions of texts in our collection) present further challenges.

Ethics Statement

To conduct this research, we follow all legal and ethical practices and are in constant contact with university lawyers.

Acknowledgements

This work has been funded through the research environment grant from the Swedish Research Council: *Grandma Karl is 27 years old: Automatic pseudonymization of research data* (nr.2022-02311).

References

[1] E. EU Commission, General data protection regulation., Official Journal of the European Union, 59, 1-88., 2016. URL: https://gdpr-info.eu/.

[2] M. R. C. MRC, GDPR Guidance note 5: Identifiability, anonymisation and pseudonymisation, 2019. URL: https://mrc.ukri.org/documents/pdf/gdpr-guidance-note-5-identifiability-anonymisation-and-pseudonymisation/, (Accessed 2025-09-26).

[3] ENISA, Privacy Enhancing Technologies: Evolution and State of the Art. A Community Approach to PETs Maturity Assessment, 2017.

[4] ENISA, A tool on Privacy Enhancing Technologies (PETs) knowledge management and maturity assessment, 2018. URL: https://www.enisa.europa.eu/publications/pets-maturity-tool, (Accessed 2025-09-26).

[5] G. Danezis, J. Domingo-Ferrer, M. Hansen, J.-H. Hoepman, D. Le Métayer, R. Tirtea, S. Schiffner, Privacy and Data Protection by Design – from policy to engineering, 2014. URL: https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design.

[6] L. Rocher, J. M. Hendrickx, Y.-A. De Montjoye, Estimating the success of re-identifications in incomplete datasets using generative models, Nature communications 10 (2019) 1–9.

[7] L. G. G. Charpentier, P. Lison, Re-identification of de-identified documents with autoregressive infilling, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 1192–1209. URL: https://aclanthology.org/2025.acl-long.60/.

[8] B. Manzanares-Salor, D. Sanchez, P. Lison, Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack, Data Mining and Knowledge Discovery 38 (2024) 4040–4075.

[9] Z. Ji, Y. Shen, K. R. Koedginer, J. Lin, Enhancing the de-identification of Personally Identifiable Information in educational data, Journal of Educational Data Mining 17(2) (2025) 55–85. URL: https://doi.org/10.5281/zenodo.17114271.

[10] I. Pilán, B. Manzanares-Salor, D. Sánchez, P. Lison, Truthful text sanitization guided by inference attacks, arXiv preprint arXiv:2412.12928 (2025). URL: https://doi.org/10.48550/arXiv.2412.12928.

[11] J. Zhang, Z. Tian, M. Zhu, Y. Song, T. Sheng, S. Yang, Q. Du, X. Liu, M. Huang, D. Li, DYNTEXT: semantic-aware dynamic text sanitization for privacy-preserving LLM inference, in: Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 20243–20255.

[12] E. Volodina, S. Dobnik, T. Lindström Tiedemann, V. Xuan-Son, Grandma Karl is 27 years old - research agenda for pseudonymization of research data, in: Proceedings of 2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService), the 2023 Workshop on Big Data and Machine Learning with Privacy Enhancing Tech, 2023.

[13] M. I. Szawerna, S. Dobnik, R. Muñoz Sánchez, E. Volodina, The devil's in the details: the detailedness of classes influences personal information detection and labeling, in: R. Johansson, S. Stymne (Eds.), Proceedings of the Joint 25th Nordic Conference on

Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), University of Tartu Library, Tallinn, Estonia, 2025, pp. 697–708. URL: https://aclanthology.org/2025.nodalida-1.70/.

[14] L. Södergard, Deltagare 1, K1989, Lova eller Latife – hur forskare benämner personer som förekommer i forskningsmaterialet, in: Svenskan i Finland, submitted.

[15] L. Södergard, Pseudonymisering av orter, skolor och organisationer. Hur gör språkforskare i praktiken?, in: Svenskans beskrivning 40, Stockholms universitet, in progress.

[16] R. Muñoz Sánchez, S. Dobnik, M. I. Szawerna, T. Lindström Tiedemann, E. Volodina, Did the names I used within my essay affect my score? Diagnosing name biases in automated essay scoring, in: E. Volodina, D. Alfter, S. Dobnik, T. Lindström Tiedemann, R. Muñoz Sánchez, M. I. Szawerna, X.-S. Vu (Eds.), Proceedings of the workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024), 2024, pp. 81–91.

[17] T. Lindström Tiedemann, L. Södergard, R. Muñoz Sánchez, S. Dobnik, M. I. Szawerna, Names, pseudonyms and biases in language assessment, TBA (in progress).

[18] T. Lindström Tiedemann, L. Södergard, E. Volodina, S. Dobnik, M. Szawerna, R. Muñoz Sánchez, X.-S. Vu, Om mormor Karl sägs vara 27 år gammal, vad säger det om skribenten? En presentation om att identifiera och ersätta identifierande element i språkvetenskapliga forskningsdata [=If Grandma Karl is said to be 27 years old, what does that say about the writer? A presentation about identifying and replacing identying elements in linguistic research data], in: Svenskans beskrivning 40, Stockholms universitet, in progress.

[19] L. Södergard, T. Lindström Tiedemann, Att ansvarsfullt skydda och dölja identitet i språkforskning [=to protect and obscure identity responsibly in linguistic research], in: Kielitieteen paivat Spraakvetenskapsdagarna The Finnish Conference of Linguistics, Helsinki 12–14 May 2025, 2025.

[20] B. Megyesi, L. Granstedt, S. Johansson, J. Prentice, D. Rosén, C.-J. Schenström, G. Sundberg, M. Wirén, E. Volodina, Learner Corpus Anonymization in the Age of GDPR: Insights from the Creation of a Learner Corpus of Swedish, in: Proceedings of the 7th NLP4CALL, Swedish Language Technology Conference, SLTC 2018, 2018, pp. 47–56.

[21] E. Volodina, Y. A. Mohammed, S. Derbring, A. Matsson, B. Megyesi, Towards privacy by design in learner corpora research: A case of on-the-fly pseudonymization of Swedish learner essays, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020.

[22] M. I. Szawerna, S. Dobnik, R. Muñoz Sánchez, T. Lindström Tiedemann, E. Volodina, Detecting personal identifiable information in Swedish learner essays, in: E. Volodina, D. Alfter, S. Dobnik, T. Lindström Tiedemann, R. Muñoz Sánchez, M. I. Szawerna, X.-S. Vu (Eds.), Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 54–63. URL: https://aclanthology.org/2024.caldpseudo-1.7/.

[23] H. Dalianis, Pseudonymisation of Swedish electronic patient records using a rule-based approach, in: L. Ahrenberg, B. Megyesi (Eds.), Proceedings of the Workshop on NLP and Pseudonymisation, Linköping Electronic Press, Turku, Finland, 2019, pp. 16–23. URL: https://aclanthology.org/W19-6503/.

[24] P. Ngo, M. Tejedor, T. Olsen Svenning, T. Chomutare, A. Budrionis, H. Dalianis, Deidentifying a Norwegian clinical corpus - an effort to create a privacy-preserving Norwegian large clinical language model, in: E. Volodina, D. Alfter, S. Dobnik, T. Lindström Tiedemann, R. Muñoz Sánchez, M. I. Szawerna, X.-S. Vu (Eds.), Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 37–43. URL: https://aclanthology.org/2024.caldpseudo-1.5/.

[25] M. Marimon, A. Gonzalez-Agirre, A. Intxaurrondo, J. A. L. Martin, M. Villegas, Automatic De-Identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results (2019).

[26] I. Pilán, P. Lison, L. Øvrelid, A. Papadopoulou, D. Sánchez, M. Batet, The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization, Computational Linguistics 48 (2022) 1053–1101.

[27] M. Sierro, B. Altuna, I. Gonzalez-Dios, Automatic detection and labelling of personal data in case reports from the ECHR in Spanish: Evaluation of two different annotation approaches, in: E. Volodina, D. Alfter, S. Dobnik, T. Lindström Tiedemann, R. Muñoz Sánchez, M. I. Szawerna, X.-S. Vu (Eds.), Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 18–24. URL: https://aclanthology.org/2024.caldpseudo-1.3/.

[28] T. Allard, L. Béziaud, S. Gambs, Publication of court records: circumventing the privacy-transparency trade-off, in: AICOL 2020 - 11th International Workshop on Artificial Intelligence and the Complexity of Legal Systems, in conjunction with JURIX 2020, Virtual, Czech Republic, 2020. URL: https://inria.hal.science/hal-03225201, a version of this work was presented at the Law and Machine Learning workshop at ICML 2020 (no proceeding).

[29] E. Eder, U. Krieg-Holz, U. Hahn, De-identification of emails: Pseudonymizing privacy-sensitive data in a German email corpus, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), INCOMA Ltd., Varna, Bulgaria, 2019, pp. 259–269. URL: https://aclanthology.org/R19-1030/.

[30] O. Yermilov, V. Raheja, A. Chernodub, Privacy- and utility-preserving NLP with anonymized data: A case study of pseudonymization, in: A. Ovalle, K.-W. Chang, N. Mehrabi, Y. Pruksachatkun, A. Galystan, J. Dhamala, A. Verma, T. Cao, A. Kumar, R. Gupta (Eds.), Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 232–241. URL: https://aclanthology.org/2023.trustnlp-1.20/.

[31] E. Callegari, A. Sólmundsdóttir, A. K. Ingason, Preserving Privacy in Small Communities: Tailored Anonymization Techniques for Icelandic Conversational Data, in: CLARIN Annual Conference Proceedings, 2024, p. 121.

[32] T. Ainiala, J.-O. Östman, Introduction: Socio-onomastics and pragmatics, in: Socio-onomastics, John Benjamins Publishing Company, 2017, pp. 1–18.

[33] E. Aldrin, Names as resources for gendering: Trends within the field, Nordic Journal of Socio-Onomastics 5 (2025) 5–32.

[34] E. Aldrin, Vad säger väl ett namn?: Reflektioner kring teorin om markerade namn utifrån

exemplet etniska konnotationer till förnamn, in: Norna-rapporter, volume 100, NORNA-förlaget, 2023, pp. 57–79.

[35] M. Frändén, " vi bestämde oss för att skriva namnet på ett svenskt sätt": Förnamnsval i sverigefinska familjer, Studia Anthroponymica Scandinavica (2015) 75–138.

[36] J. Heaton, "* pseudonyms are used throughout": A footnote, unpacked, Qualitative Inquiry 28 (2022) 123–132.

[37] S. Wang, J. M. Ramdani, S. Sun, P. Bose, X. Gao, Naming research participants in qualitative language learning research: Numbers, pseudonyms, or real names?, Journal of language, identity & education (2024) 1–14.

[38] J. L. Seelig, Place anonymization as rural erasure? A methodological inquiry for rural qualitative scholars, International Journal of Qualitative Studies in Education 34 (2021) 857–870.

[39] P. Accorsi, N. Patel, C. Lopez, R. Panckhurst, M. Roche, Seek&Hide: Anonymising a French SMS corpus using natural language processing techniques, Linguisticae Investigationes 35 (2012) 163–180. doi:10.1075/li.35.2.03acc.

[40] R. Blokland, N. Partanen, M. Rießler, A pseudonymisation method for language documentation corpora: An experiment with spoken Komi, in: T. A. Pirinen, F. M. Tyers, M. Rießler (Eds.), Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages, Association for Computational Linguistics, Wien, Austria, 2020, pp. 1–8. URL: https://aclanthology.org/2020.iwclul-1.1/.

[41] N. Ilinykh, M. I. Szawerna, "I need more context and an English translation": Analysing how LLMs identify personal information in Komi, Polish, and English, in: Š. A. Holdt, N. Ilinykh, B. Scalvini, M. Bruton, I. N. Debess, C. M. Tudor (Eds.), Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025), University of Tartu Library, Estonia, Tallinn, Estonia, 2025, pp. 165–178. URL: https://aclanthology.org/2025.resourceful-1.32/.

[42] V. Yogarajan, B. Pfahringer, M. Mayo, A review of automatic end-to-end de-identification: Is high accuracy the only metric?, Applied Artificial Intelligence 34 (2020) 251–269. URL: https://doi.org/10.1080/08839514.2020.1718343.

[43] M. I. Szawerna, S. Dobnik, T. Lindström Tiedemann, R. M. Sánchez, X.-S. Vu, E. Volodina, Pseudonymization categories across domain boundaries, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 13303–13314.

[44] A. Papadopoulou, Y. Yu, P. Lison, L. Øvrelid, Neural text sanitization with explicit measures of privacy risk, in: Y. He, H. Ji, S. Li, Y. Liu, C.-H. Chang (Eds.), Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online only, 2022, pp. 217–229. URL: https://aclanthology.org/2022.aacl-main.18/.

[45] P. Lison, I. Pilán, D. Sanchez, M. Batet, L. Øvrelid, Anonymisation models for text data: State of the art, challenges and future directions, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp.

4188–4203. URL: https://aclanthology.org/2021.acl-long.323/.

[46] A. Alfalahi, S. Brissman, H. Dalianis, Pseudonymisation of Personal Names and other PHIs in an Annotated Clinical Swedish Corpus, in: Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) held in conjunction with LREC 2012, 2012. URL: https://api.semanticscholar.org/CorpusID:6387546.

[47] A. W. Olstad, A. Papadopoulou, P. Lison, Generation of replacement options in text sanitization, in: T. Alumäe, M. Fishel (Eds.), Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), University of Tartu Library, Tórshavn, Faroe Islands, 2023, pp. 292–300. URL: https://aclanthology.org/2023.nodalida-1.30/.

[48] O. Feyisetan, T. Diethe, T. Drake, Leveraging Hierarchical Representations for Preserving Privacy and Utility in Text , in: 2019 IEEE International Conference on Data Mining (ICDM), IEEE Computer Society, Los Alamitos, CA, USA, 2019, pp. 210–219. doi:10.1109/ICDM.2019.00031.

[49] D. Simancek, V. V. Vydiswaran, Handling name errors of a BERT-based de-identification system: Insights from stratified sampling and Markov-based pseudonymization, in: E. Volodina, D. Alfter, S. Dobnik, T. Lindström Tiedemann, R. Muñoz Sánchez, M. I. Szawerna, X.-S. Vu (Eds.), Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 1–7. URL: https://aclanthology.org/2024.caldpseudo-1.1/.

[50] S. Hou, R. Shang, Z. Long, X. Fu, Y. Chen, A general pseudonymization framework for cloud-based llms: Replacing privacy information in controlled text generation, 2025. URL: https://arxiv.org/abs/2502.15233.

[51] M.-C. de Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal dependencies, Computational Linguistics 47 (2021) 255–308. doi:10.1162/coli_a_00402.

[52] M. I. Szawerna, S. Dobnik, R. M. Sánchez, T. Lindström Tiedemann, E. Volodina, Detecting personal identifiable information in swedish learner essays, in: Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024), 2024, pp. 54–63.

[53] R. Muñoz Sánchez, S. Dobnik, M. I. Szawerna, T. Lindström Tiedemann, E. Volodina, Did the names i used within my essay affect my score? diagnosing name biases in automated essay scoring, in: Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024), 2024, pp. 81–91.

[54] R. Muñoz Sánchez, S. Dobnik, M. I. Szawerna, T. Lindström Tiedemann, E. Volodina, Name biases in automated essay assessment, in: International congress of onomastic sciences, ICOS, Helsinki, 19–23 August 2024, 2024.

[55] T. Lindström Tiedemann, L. Södergard, E. Volodina, S. Dobnik, M. Szawerna, R. Munoz Sanchez, X.-S. Vu, En presentation om att ersätta identifierande element i språkvetenskapliga forskningsdata, in: Workshop Pseudonymisering inom språkvetenskap, Svenskans beskrivning 40, Stockholm, 26 May 2025, 2025.

[56] M. Rigaki, S. Garcia, A survey of privacy attacks in machine learning, ACM Computing Surveys 56 (2023) 1–34.

[57] M. Aerni, J. Zhang, F. Tramèr, Evaluations of machine learning privacy defenses are misleading, in: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and

Communications Security, 2024, pp. 1271–1284.

[58] Regulation 2024/1689, The EU Artificial Intelligence Act (2024/1689), Official Journal of the European Union, L series, 2016. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[59] S. Chatzikyriakidis, R. Cooper, S. Dobnik, S. Larsson, An overview of natural language inference data collection: The way forward?, in: C. Gardent, C. Retoré (Eds.), Proceedings of IWCS 2017: 12th International Conference on Computational Semantics, Workshop on Computing Natural Language Inference, Association for Computational Linguistics, Montpellier, France, 2017, pp. 1–6. URL: http://www.aclweb.org/anthology/W/W17/#7200.

[60] A. Stubbs, C. Kotfila, Özlem Uzuner, Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1, Journal of Biomedical Informatics 58 (2015) S11–S19. URL: https://doi.org/10.1016/j.jbi.2015.06.007.

[61] A. Stubbs, M. Filannino, Özlem Uzuner, De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1, Journal of Biomedical Informatics 75 (2017) S4–S18. URL: https://doi.org/10.1016/j.jbi.2017.06.011, supplement: A Natural Language Processing Challenge for Clinical Records: Research Domains Criteria (RDoC) for Psychiatry.

[62] M. Marimon, A. Gonzalez-Agirre, A. Intxaurrondo, H. Rodríguez, J. A. Lopez Martin, M. Villegas, M. Krallinger, MEDDOCAN corpus: gold standard annotations for Medical Document Anonymization on Spanish clinical case reports , 2020. URL: https://doi.org/10.5281/zenodo.4279323.

[63] L. Holmes, Cleaned Repository of Annotated PII, https://www.kaggle.com/datasets/langdonholmes/cleaned-repository-of-annotated-pii, 2024. [Accessed 18-09-2025].

[64] AI4Privacy, ai4privacy/pii-masking-300k · Datasets at Hugging Face — huggingface.co, https://huggingface.co/datasets/ai4privacy/pii-masking-300k, 2024. [Accessed 05-09-2025].

[65] T. Vakili, A. Henriksson, H. Dalianis, End-to-end pseudonymization of fine-tuned clinical bert models: Privacy preservation with maintained data utility, BMC Medical Informatics and Decision Making 24 (2024) 162.

[66] E. Eder, U. Krieg-Holz, U. Hahn, De-Identification of Emails: Pseudonymizing Privacy-Sensitive Data in a German Email Corpus, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), INCOMA Ltd., Varna, Bulgaria, 2019, pp. 259–269. URL: https://aclanthology.org/R19-1030.