Pseudonymization Categories across Domain Boundaries

Maria Irena Szawerna[1], Simon Dobnik[2], Therese Lindström Tiedemann[3]
Ricardo Muñoz Sánchez[1], Xuan-Son Vu[4], Elena Volodina[1]
[1]Språkbanken Text, SFS, University of Gothenburg, Sweden
[2]CLASP, FLoV, University of Gothenburg, Sweden
[3]Department of Finnish, Finnougrian and Scandinavian Studies, University of Helsinki, Finland
[4]Department of Computing Science, Umeå University, Sweden
maria.szawerna@gu.se, mormor.karl@svenska.gu.se

Abstract

Linguistic data, a component critical not only for research in a variety of fields but also for the development of various Natural Language Processing (NLP) applications, can contain personal information. As a result, its accessibility is limited, both from a legal and an ethical standpoint. One of the solutions is the pseudonymization of the data. Key stages of this process include the identification of sensitive elements and the generation of suitable surrogates in a way that the data is still useful for the intended task. Within this paper, we conduct an analysis of tagsets that have previously been utilized in anonymization and pseudonymization. We also investigate what kinds of Personally Identifiable Information (PII) appear in various domains. These reveal that none of the analyzed tagsets account for all of the PII types present cross-domain at the level of detailedness seemingly required for pseudonymization. We advocate for a universal system of tags for categorizing PIIs leading up to their replacement. Such categorization could facilitate the generation of grammatically, semantically, and sociolinguistically appropriate surrogates for the kinds of information that are considered sensitive in a given domain, resulting in a system that would enable dynamic pseudonymization while keeping the texts readable and useful for future research in various fields.

Keywords: pseudonymization, anonymization, privacy, deidentification, universal tagset

1. Introduction

It is hard to avoid coming across information that could be considered private and/or sensitive when working with linguistic data. For example, we have the text in Example 1[1] and need to ensure that it cannot reveal the writer's identity once it is made accessible for researchers outside the project. According to the GDPR (EU Commission, 2016, Art.4), we need to locate and handle all Personally Identifiable Information (PIIs) that can reveal a data subject through direct and indirect clues. After the first inspection, we hypothesize that some PIIs can lead to re-identification more easily (**underline**) than other (***italics***), and we mark them accordingly.

> The really bad day for me when I ***lost*** my ***sister***, **Natanya**, when my ***sister dieded***. She was **four** years ***only***, she was ***little*** when she ***dieded***. One day she became very sick, ***doctor*** **Harif** come and tell us **Nata** have **canser**. My ***mother*** became very sad, lied in bed, no food. My ***sister die five months*** later. My ***brother*** **Karim** and my ***father*** **Malik** cried. My ***mother*** got so bad she ***ended in a hospital*** with **brakedown**. It happened ***nine*** years ago.

Example 1: Original text

We need to somehow process this text to make it open for other researchers and at the same time make it useful for various research tasks. We therefore experiment with anonymization (Example 2), labeling (Example 3), and pseudonymization (Example 4).

Upon inspection, it becomes evident that *anonymization* makes texts that are personal in nature rather inappropriate for any use. *Labeling* PIIs may help to a certain degree – especially if we label only the underscored PIIs; however, the utility of the text for potential research questions is best preserved, in relative terms, in the *pseudonymized* version.

We further note that:

1. Through all of these manipulations we risk changing the value of the data. For example, by replacing misspelled '`canser`'(=cancer) with any of the techniques (i.e. `XX` vs `medcondition` vs `stroke`), we lose information about the learner's knowledge of spelling, grammar or level of vocabulary.

2. Most of the italicized PIIs, taken on their own, present only minimal risk for re-identification of a person – but can help re-identify a person in combination with other clues.[2] It might be reasonable to consider leaving the italicized PIIs as in the original, depending on the research

[1]A modified and translated learner essay.

[2]More examples of various types and combinations of personal data can be found in section 4 and section 12.

LREC-COLING 2024, pages 13303–13314

field and research questions, while the underscored ones should always be pseudonymized. We see this as a strong impetus to work towards dynamically configurable approaches to pseudonymization and privacy protection.

3. Pseudonymization presents challenges with respect to semantic aspects of the context. For example, the selection of a pseudonym for `fam-member` needs to be in accordance with factors such as gender (`sister` vs `cousin`), age (especially in relation to the author) and timeline in the essay (i.e. `it happened XX years ago` vs `Now I am XX years old`). In this respect, the more detailed the tagset is, the better a pseudonym can be generated, which in turn denounces taxonomies that are too general.

The really bad day for me when I XX my ***XX***, **XX**, when my ***XX XX***. She was **XX** years ***XX***, she was ***XX*** when she ***XX***. One day she became very sick, ***XX*** **XX** come and tell us **XX** have **XX**. My ***XX*** became very sad, lied in bed, no food. My ***XX XX XX*** later. My ***XX*** **XX** and my ***XX*** **XX** cried. My ***XX*** got so bad she ***XX*** with **XX**. It happened ***XX*** years ago.

Example 2: Anonymized text

The really bad day for me when I ***event*** my ***fam-member***, **name**, when my ***fam-member event***. She was **age** years ***other***, she was ***other*** when she ***event***. One day she became very sick, ***profession*** **name** come and tell us **name** have **med-condition**. My ***fam-member*** became very sad, lied in bed, no food. My ***fam-member event time*** later. My ***fam-member*** **name** and my ***fam-member*** **name** cried. My ***fam-member*** got so bad she ***event*** with **med-condition**. It happened ***time*** years ago.

Example 3: Text with labelled PIIs

The really bad day for me when I ***lost*** my ***cousin***, **Frankie**, when my ***cousin dieded***. She was **six** years ***only***, she was ***little*** when she ***dieded***. One day she became very sick, ***doctor*** **Hank** come and tell us **Frankie** have **stroke**. My ***grandmother*** became very sad, lied in bed, no food. My ***cousin die three months*** later. My ***cousin*** **John** and my ***uncle*** **Harris** cried. My ***grandmother*** got so bad she ***ended in a hospital*** with **arthritis**. It happened ***ten*** years ago.

Example 4: Selectively pseudonymized text

Our aim in this paper is to establish (1) which tagsets have been used to categorize PII (including Personal Health Information, PHI) and in what ways they differ, as well as (2) whether there are kinds of sensitive information that have not previously been covered, in order to facilitate future research on pseudonymization and its effects on research data.

To answer these questions we conduct two analyses. We compare some of the existing tagsets and inspect the types of PIIs present in different textual domains. Based on that, we advocate for the creation of a list of core PII categories, domain-specific ones, and other types as the first input to creating a dynamically configurable approach to pseudonymization. We propose that a universal list of pseudo-categories be used as a standard by the community for comparable results in the field of pseudonymization, e.g. for use in multilingual shared tasks or similar.

2. Prior Research

As Eder et al. (2022) point out, the detection of "privacy-sensitive information" has progressed at different paces within various domains, with Protected Health Information (PHI), which we consider to be a subtype of PII, having historically received more focus. However, it is worth noting that as far as PHI is concerned, some information that could be considered private, such as health-related information, is not classified as private and is not subject to removal or replacement. This is especially interesting given that this type of information is generally considered a special (sensitive) category of personal data within the GDPR (EU Commission, 2016, Art. 9).

What further highlights the differences in sensitivity between various pieces of information is the division of personal identifiers suggested by Pilán et al. (2022), where information can either be a direct identifier (a "value unique to a given individual", such as a name or a social security number) or a quasi-identifier (information that "may [lead to reidentification] combined with other quasi-identifiers and associated with background knowledge). However, as shown in section 1, where all of the private information included in the example should be considered a quasi-identifier according to this definition, various pieces of information appear to have varying levels of sensitivity.

Lison et al. (2021) define pseudonymization as the "process of replacing direct identifiers with pseudonyms or coded values." They also point out that while the detection of many kinds of PII has already been tackled, the pseudonym generation stage has not received nearly as much attention. Volodina et al. (2023) draw attention to the challenges in the generation of appropriate pseudonyms, as the replacement needs to have the right grammatical form, as required by the context, but also it needs to fit into various contextual and linguistic constraints, such as semantic restrictions or lexical variation and frequency if the reading of

Anonymization		
ID	**Paper and domain**	**Tagset**
1	Adams et al. (2019) Chat	**PII:** Person, Address, Zip Code, Location, Email, UID, IP Address, (Date) **Corporate Identifying Information (CII):** Organization, Product, Facility, URL **Other:** Nationality, Geographical, Event, Work of Art, Language, Unit, Misc, Med/Chem, Sports Team, Known Group, Known Figure, Fictional Figure, Date
2	Pilán et al. (2022) Legal	PERSON, CODE, LOC, ORG, DEM, DATETIME, QUANTITY, MISC, BELIEF, POLITICS, SEX, ETHNIC, HEALTH, NOT_CONFIDENTIAL
3	Accorsi et al. (2012) SMS	PRE *(first name)*, NOM *(last name)*, SUR *(nickname)*, ADR *(address)*, LIE *(place)*, TEL *(telephone number)*, COD *(code)*, URL, MAR *(brand name)*, MEL *(e-mail)*, Other
4	Bråthen et al. (2021) Medical	First_Name, Last_Name, Age, Health_Care_Unit, Phone_Number, Social_Security_Number, Date_Full, Date_Part, Location
Pseudonymization		
ID	**Paper and domain**	**Tagset**
5	Megyesi et al. (2018) Megyesi et al. (2021) L2 essays	**Personal name:** firstname_male, firstname_female, firstname_unknown, initials, middlename, surname **Institution:** school, work, other_institution **Geographic:** area, city, geo, country, place, region, street_nr, zip_code, foreign **Transportation:** transport_name, transport_nr **Age:** age_digits, age_string **Dates:** date_digits, day, month_digit, month_word, year **Other tags:** phone_nr, email, url, personid_nr, account_nr, license_nr, other_nr_seq, extra, prof, edu, fam, sensitive, gen, def, pl
6	Eder et al. (2019, 2020) Eder et al. (2022) E-mail	**ACTOR:** ORG, USER, **PERSON:** FAMILY, **GIVEN:** FEMALE, MALE **LOC:** CITY, ZIP, STREET, STREETNO DATE **FID *(formal identifier)*:** PASS, UFID *(unique formal identifier)* **ADD *(address)*:** EMAIL, PHONE, URL
7	Alfalahi et al. (2012) Medical	Age, Full_Date, Date_Part, First_Name, Last_Name, Location, Health_Care_Unit, Phone_Number
8	Dalianis (2019) Medical	Female First Name, Male First Name, Gender-neutral First Name, Last Name, Age, Phone Number, Location, Full Date, Date Part, Health Care Unit

Table 1: Tagsets, approaches, and domains in prior research. Some of the tags' meanings are included in italics, while words in bold indicate hierarchical category names.

the sentence and its usefulness in future research to be as close to the original as possible. It is our belief that appropriate PII labeling can facilitate better pseudonym generation.

A large part of the relevant research in the field can be divided into two categories, namely, papers discussing the removal of PII or the replacement of PII with pseudonyms. It is worth pointing out that there exist other approaches, e.g. differential privacy; however, since they do not remove or replace PII, they are not suitable for the comparisons conducted in this paper (Danezis et al., 2015; Lison et al., 2021). The tags presented in the papers mentioned below are collected in Table 1.

Anonymization Adams et al. (2019) focused on anonymizing unstructured chat data using a hierarchical classification with the overarching categories of PII, Corporate Identifying Information (CII), and other. While they anonymized only the first two categories, they acknowledged that the information included in the third category could also be used for re-identification in some situations. Pilán et al. (2022) introduced a corpus and evaluation framework for text anonymization based on court cases from the European Court of Human Rights. The tags used by the authors are more general than many of the other tagsets discussed in this paper, as the definitions are rather broad. Accorsi et al. (2012) attempted to anonymize a collection of French text messages, focusing on replacing names and addresses (both physical and digital). Bråthen et al. (2021) designed a Norwegian medical corpus for de-identification tasks, with the tags corresponding to the most common types of personal data in this domain.

Pseudonymization When building their Swedish corpus of learner language, Megyesi et al. (2018) decided to pseudonymize the data with the help of

a large set of tags. The structure of the tagset is hierarchical. More precise definitions of the tags can be found in the pseudonymization guidelines for pseudonymizing the corpus (Megyesi et al., 2021). Another corpus, CodE Alltag 2.0, composed of German e-mails, was also pseudonymized, although using a smaller tagset (Eder et al., 2019, 2020, 2022). Alfalahi et al. (2012) focused on pseudonymizing Swedish medical data, once more with a less varied tagset. Very similar categories were used by Dalianis (2019) when working with the same domain.

3. Tagset Analyses

The objective of the two analyses presented in this section has been to evaluate the existing tagsets for language data pseudonymization and explore whether a universal tagset would be desirable. To do that, we have first performed a systematic comparison of several available tagsets (Analysis 1, subsection 3.1) to identify common categories, domain-specific categories, and other types of sensitive categories, zooming into the coverage and semantic overlaps of the tagsets. In Analysis 2 (subsection 3.2), we applied one selected tagset to data samples representing several domains with personal and/or sensitive information, to test the representativeness of the tagset across domains and to identify currently missing categories.

3.1. Analysis 1: Comparison of Existing Taxonomies

The first analysis is concerned with the tagsets utilized in prior research on anonymization and pseudonymization.

Selection of taxonomies Not all projects report the tagsets used for pseudonymization or anonymization. Therefore, we have limited this analysis to eight taxonomies reported in literature[3], where the tags' coverage was explained, as shown in Table 1. Since up until recently the terms anonymization and pseudonymization were used synonymously (pseudonymization as a separate field being almost non-represented), we conflate the two types of tagsets in this study, even though we are well aware of the fact that they have essentially different objectives with data manipulation.

Comparison of taxonomies To compare, we listed all taxonomies in a spreadsheet, one column per tagset, trying to align the different categories between the tagsets. Our choice of medium in this case was motivated by the ease of collaboration and this being merely a preliminary study or comparison, not aimed at constructing our own tagset based on the analyzed ones. The number of tags per taxonomy ranged from 8 to 39 categories. In certain cases the tags could be mapped 1:1 between the tagsets; however, in many cases, the mapping was only partial, the semantics and coverage of tags being different (e.g. `PERSON` in one taxonomy versus `NAME_FEMALE, NAME_MALE, SURNAME` in another. This is partly due to some of the taxonomies having a more hierarchical structure, whereas our comparisons only pertained to the most detailed levels thereof, in or-

[3]The tagsets come from both reports, papers, and dataset descriptions.

Tag	Domains
firstname_male	**1, 2, 3, 4, 5, 6, 7, 8**
firsname_female	**1, 2, 3, 4, 5, 6, 7, 8**
firstname_unknown	**1, 2, 3, 4, 5,** 6, **7, 8**
initials	**1, 2,** 3, 4, **5,** 6, **7, 8**
middlename	**1, 2,** 3, 4, **5,** 6, **7, 8**
surname	**1, 2, 3, 4, 5, 6, 7, 8**
foreign	1, 2, 3, 4, **5,** 6, 7, 8
area	**1, 2, 3, 4, 5,** 6, **7, 8**
city	**1, 2, 3, 4, 5, 6, 7, 8**
geo	**1, 2, 3, 4, 5,** 6, **7, 8**
country	**1, 2, 3, 4, 5,** 6, **7, 8**
place	**1, 2, 3, 4, 5, 6, 7, 8**
region	**1, 2, 3, 4, 5,** 6, **7, 8**
street_nr	**1, 2, 3, 4, 5, 6, 7, 8**
zip_code	**1, 2, 3, 4, 5, 6, 7, 8**
school	**1, 2,** 3, 4, **5,** 6, 7, 8
work	**1, 2,** 3, 4, **5,** 6, 7, 8
other_institution	**1, 2, 3,** 4, **5, 6,** 7, 8
transport_name	1, 2, 3, 4, **5,** 6, 7, 8
transport_nr	1, 2, 3, 4, **5,** 6, 7, 8
age_digits	1, **2,** 3, **4, 5,** 6, **7, 8**
age_string	1, **2,** 3, **4, 5,** 6, **7, 8**
date_digits	**1, 2,** 3, **4, 5, 6, 7, 8**
day	**1, 2,** 3, **4, 5, 6, 7, 8**
month_digit	**1, 2,** 3, **4, 5, 6, 7, 8**
month_word	**1, 2,** 3, **4, 5, 6, 7, 8**
year	**1, 2,** 3, **4, 5, 6, 7, 8**
phone_nr	**1, 2, 3, 4, 5, 6, 7, 8**
email	1, 2, **3,** 4, **5, 6,** 7, 8
url	**1,** 2, **3,** 4, **5, 6,** 7, 8
personid_nr	**1, 2,** 3, **4, 5, 6,** 7, 8
account_nr	1, **2,** 3, 4, **5, 6,** 7, 8
license_nr	1, **2,** 3, 4, **5, 6,** 7, 8
other_nr_seq	1, **2,** 3, 4, **5,** 6, 7, 8
extra	**1, 2, 3,** 4, **5,** 6, 7, 8
prof	1, 2, 3, 4, **5,** 6, 7, 8
edu	1, 2, 3, 4, **5,** 6, 7, 8
fam	1, 2, 3, 4, **5,** 6, 7, 8
sensitive	1, **2,** 3, 4, **5,** 6, 7, 8

Table 2: SweLL tags and the papers in Analysis 1 they correspond to (in bold). The numbers representing the domains follow the numeration in Table 1.

Tag	Domains
username	1, **2,** 3, 4, 5, **6,** 7, 8
password	1, 2, **3,** 4, 5, **6,** 7, 8
IP address	**1,** 2, 3, 4, 5, 6, 7, 8
product	**1,** 2, **3,** 4, 5, 6, 7, 8
facility	**1,** 2, 3, 4, 5, 6, 7, 8
nationality	**1, 2,** 3, 4, 5, 6, 7, 8
work of art	**1,** 2, 3, 4, 5, 6, 7, 8
language	**1, 2,** 3, 4, 5, 6, 7, 8
unit	**1,** 2, 3, 4, 5, 6, 7, 8
med/chem entity	**1,** 2, 3, 4, 5, 6, 7, 8
sports team	**1,** 2, 3, 4, 5, 6, 7, 8
known group	**1,** 2, 3, 4, 5, 6, 7, 8
known figure	**1,** 2, 3, 4, 5, 6, 7, 8
fictional figure	**1,** 2, 3, 4, 5, 6, 7, 8
healthcare unit	1, 2, 3, **4,** 5, 6, **7, 8**
demographic attribute	1, **2,** 3, 4, 5, 6, 7, 8
duration	1, **2,** 3, 4, 5, 6, 7, 8
quantity, value	1, **2,** 3, 4, 5, 6, 7, 8
nickname	1, **2, 3,** 4, 5, 6, 7, 8
belief	1, **2,** 3, 4, 5, 6, 7, 8
political views	1, **2,** 3, 4, 5, 6, 7, 8
sexuality, gender identity	1, **2,** 3, 4, 5, 6, 7, 8
ethnicity	1, **2,** 3, 4, 5, 6, 7, 8
health	1, **2,** 3, 4, 5, 6, 7, 8
patronymic/other name	1, 2, 3, 4, 5, 6, 7, 8

Table 3: Generic non-SweLL tags and the papers in Analysis 1 they correspond to (in bold). The numbers representing the domains follow the numeration in Table 1.

der to see what kinds of distinctions the taxonomies are capable of making. This was probably one of the most critical aspects we looked into – the semantic overlap between the tags, i.e. whether they cover the same PIIs or whether they are semantically the same. The overlaps between the tagsets are shown in Table 2 and Table 3.

Distribution of categories (a case study) We have additionally compared the PII counts reported in the background literature for various domains in order to see whether there are any differences in the distribution of analogous tags. We looked into `PERSON` (Figure 1), `LAST_NAME, FIRST_NAME` (Figure 2) and `FEMALE, MALE, GENDER_NEUTRAL, LAST` (Figure 3); these labels are generalizations of the tags present in the papers, accounting for differences in label names. Figure 2 includes papers that featured more fine-grained distinctions (e.g. male and female first names), but where those have been collapsed into one for the sake of the comparison. The results are discussed in section 5.

Insights The first study has outlined the difference in approaches to pseudonymization between the domains and projects. Our takeaway is that we should potentially distinguish between at least

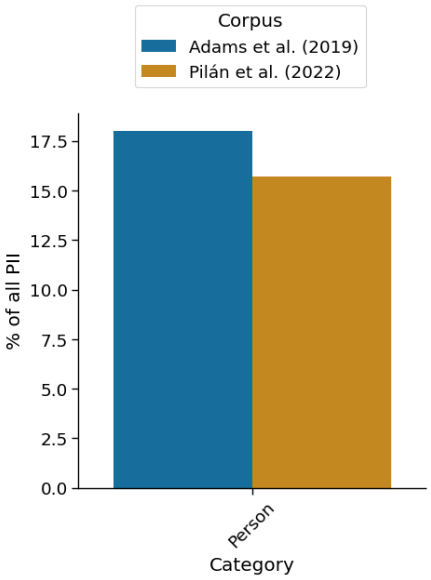

Figure 1: A comparison of the distribution of tags between papers that only included "person" as a category of PIIs.

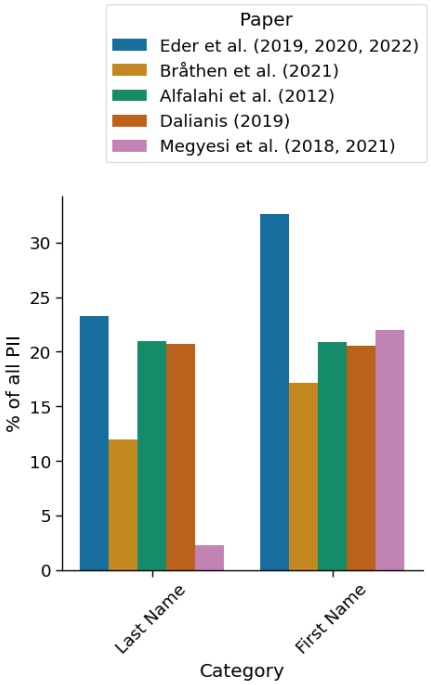

Figure 2: A comparison of the distribution of tags between papers that included a distinction into first and last names. Keep in mind that the "first name" category may combine subcategories relative to gender.

three types of PIIs: (1) general core ones, that should be replaced or masked in all text types, (2) domain-specific ones, that should always be replaced in texts from certain domains and for certain uses, and (3) extra ones, that can be replaced on a case-to-case basis, depending on the application

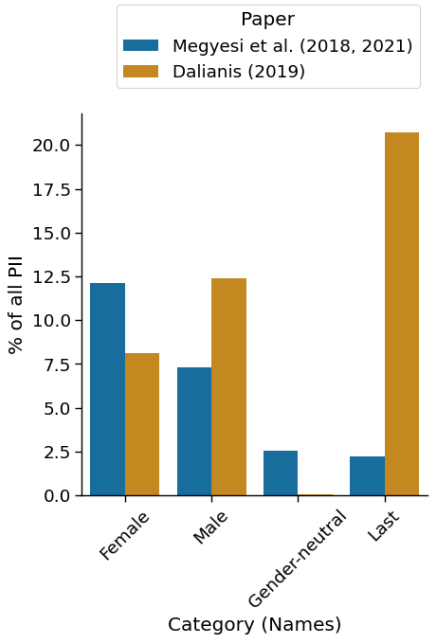

Figure 3: A comparison of the distribution of tags between papers that included female, male, and gender-neutral first names as well as last names.

scenario and other circumstances.

However, we need a more extensive study to see whether all potential PIIs could be classified under the three headings given that some of the PIIs will be context-dependent, whereas some will require new labels, unaccounted from previous annotation experiences.

Sierro et al. (2024) showed that more fine-grained categorization of privacy-sensitive information can lead to better performance at the detection stage. We have, therefore, selected the most extensive and semantically least ambiguous taxonomy from this study for the analysis of data re-annotation (subsection 3.2), namely the SweLL pseudo-taxonomy (Megyesi et al., 2018, 2021).

3.2. Analysis 2: Data Re-annotation with One Tagset

The second analysis consists of a comparison of the types of PII present in texts from various domains, as outlined in Table 4. For that, we have applied the same tagset to all data samples based on the results from Analysis 1 (subsection 3.1), namely the SweLL tagset with 39 labels listed in Table 5.

Data The selection of the data samples was partly motivated by our common sense combined with literature review, and partly by the availability of the data we were interested in analyzing. The diversity of domains was extremely important since we wanted to reveal, among others, different aspects of privacy. The obvious candidates were, of course, medical data, data from social media, blogs, and emails, but also learner essays, reviews, memoirs, and court cases. The corpora that we have utilized in this analysis were also representative of various languages, namely Swedish, English, and Polish. The final set of domains is presented in Table 4.

ID	Domain	Source and Language
1	Medical	Private source[3] *Swedish*
2	E-mails	Enron Corpus Enron Corp and Cohen *English*
3	Legal	ECHR Corpus Pilán et al. (2022) *English*
4	Memoirs	Juliusz Czermiński's Memoir Szawerna (2023) *Polish*
5	Blogs	Swe-NERC Ahrenberg et al. (2020) *Swedish*
6	Tweets	Twitter Mix n/a (2022) *Swedish*
7	Web News	Swe-NERC Ahrenberg et al. (2020) *Swedish*
8	L2 Essays	SweLL-gold Volodina et al. (2022) *Swedish*
9	Reviews	Amazon Fine Food Reviews McAuley and Leskovec (2013) *English*

Table 4: Different domains and sources featured in the comparison.

Sample sizes From most of the datasets in Table 4, we have randomly selected sentences or paragraphs amounting to at least 5000 tokens, with the exception of Twitter and medical data and the assumption that all of the categories from the SweLL tagset would be present in the SweLL data. When it comes to tweets, we have looked at 2000 sentences as obtained from the concordance search tool Korp which was our way of accessing the data (Borin et al., 2012). This data consists of sentences from tweets that come in a randomized order; we have selected 1000 results of searches for the lemma *du* "you" and another 1000 for *jag* "I". As for the medical data, we have only had access to a physical sample of records[4].

[4]Since non-anonymized medical data was extremely difficult to find in open access, we have used a sample kindly provided to us by personal contact at Västra Götalandsregionen, where the medical records contained

Annotation The SweLL taxonomy (Megyesi et al., 2018, 2021), with its 39 categories,[5] was used for the annotation of the data samples from the selected nine domains. Each data sample was manually processed using the development version of the Svala annotation tool (Wirén et al., 2019) which enables the assignment of SweLL pseudonymization labels to the analyzed strings.

We have then analyzed the samples for all of the domains with the SweLL tagset in mind, making note of (a) which kinds of accounted-for PII data could be found in the samples across domains, and (b) which other types of sensitive information the SweLL tagset is lacking.

Insights The annotation made it possible for us to determine which categories are represented in which domain, as summarized in Table 5, where certain SweLL categories are present in all domains (e.g. CITY), some are present in 8 out of 9 domains (e.g. OTHER_INSTITUTION, EXTRA, FAM), etc.

We can also summarize categories, lacking in the SweLL taxonomy, by domains. As far as the medical data is concerned, we have not observed any new PII types, with the potential exception of medical conditions and medical procedures – however, these are traditionally not anonymized or pseudonymized in medical records. As one of the subsequent examples shows, though, this kind of information can appear in other domains, where it could lead to reidentification. The legal proceedings included durations and events that could be revealing. In the memoir, we have noted an abundance of professional and social titles or descriptions of unusual physical appearance. Blogs included names of people's pets, routine events, day names, and descriptions of looks or character. Tweets featured @-names and hashtags, the latter of which could refer to e.g. specific events. In the web news samples, we have identified demonyms, decades, as well as works of art as potentially sensitive information. Finally, in the online reviews, people shared their or their relatives' medical conditions.

The analysis made it possible for us to observe what kinds of personal information are present in these domains, and which potential new categories need to be added to the pseudonymization taxonomy if we were to aim at a universal taxonomy.

only descriptions without any metadata (such as patient names, age, date of the visit, etc.).

[5]This number excludes tags used to mark features such as definiteness, plurality, case, and other functional tags.

Tag	Domains
firstname_male	1, **2, 3, 4, 5, 6, 7, 8,** 9
firsname_female	1, **2,** 3, **4, 5, 6, 7, 8,** 9
firstname_unknown	1, 2, 3, 4, 5, **6, 7, 8,** 9
initials	1, **2, 3, 4, 5,** 6, 7, **8,** 9
middlename	1, 2, 3, **4,** 5, 6, **7, 8,** 9
surname	1, **2, 3, 4, 5, 6, 7, 8,** 9
foreign	1, 2, **3,** 4, 5, 6, **7, 8,** 9
area	1, 2, 3, **4,** 5, 6, 7, **8,** 9
city	**1, 2, 3, 4, 5, 6, 7, 8, 9**
geo	1, 2, 3, 4, 5, 6, **7, 8,** 9
country	1, **2, 3,** 4, 5, **6, 7, 8, 9**
place	**1, 2,** 3, 4, 5, **6,** 7, **8,** 9
region	1, **2,** 3, **4, 5, 6, 7, 8, 9**
street_nr	**1, 2,** 3, 4, 5, **6,** 7, **8,** 9
zip_code	**1, 2,** 3, 4, 5, 6, 7, **8,** 9
school	**1,** 2, 3, **4,** 5, **6,** 7, **8,** 9
work	1, **2, 3,** 4, 5, **6, 7, 8,** 9
other_institution	**1, 2, 3,** 4, **5, 6, 7, 8, 9**
transport_name	1, 2, 3, 4, 5, **6, 7, 8,** 9
transport_nr	1, 2, 3, 4, 5, **6,** 7, **8,** 9
age_digits	1, 2, 3, **4,** 5, **6, 7, 8, 9**
age_string	1, 2, 3, 4, 5, 6, 7, **8,** 9
date_digits	**1, 2,** 3, 4, 5, 6, 7, **8,** 9
day	1, **2, 3, 4,** 5, **6, 7, 8,** 9
month_digit	1, **2,** 3, 4, 5, 6, 7, **8,** 9
month_word	1, **2, 3, 4,** 5, **6, 7, 8,** 9
year	1, **2, 3, 4,** 5, **6, 7, 8,** 9
phone_nr	**1, 2,** 3, 4, 5, 6, 7, **8,** 9
email	1, **2,** 3, 4, 5, **6,** 7, **8,** 9
url	1, **2,** 3, 4, **5, 6,** 7, **8,** 9
personid_nr	1, 2, 3, 4, 5, 6, 7, **8,** 9
account_nr	1, 2, 3, 4, 5, 6, 7, **8,** 9
license_nr	1, 2, 3, 4, 5, 6, 7, **8,** 9
other_nr_seq	1, 2, **3,** 4, 5, 6, 7, **8,** 9
extra	1, **2, 3, 4, 5, 6, 7, 8, 9**
prof	1, 2, **3, 4,** 5, **6, 7, 8,** 9
edu	1, 2, **3,** 4, 5, 6, 7, **8,** 9
fam	**1,** 2, **3, 4, 5, 6, 7, 8, 9**
sensitive	**1,** 2, **3,** 4, 5, 6, 7, **8,** 9

Table 5: SweLL tags and the domains in Analysis 2 they correspond to (in bold). The numbers representing the domains follow the numeration in Table 4.

4. Results

In our attempts to align the various tagsets to each other, we have discovered that they take rather diverse approaches to categorizing certain types of information, especially geographical and temporal. A few of the tagsets covered a wide variety of PII with just one general tag (e.g. Pilán et al. (2022): "PERSON Names of people, including nicknames/aliases, usernames, and initials"), and some ignored certain types of possibly sensitive information altogether (e.g. Accorsi et al. (2012) not including dates). While having rather general categories appears to be something that tagsets used

in anonymization have in common, some used in pseudonymization were rather similar with regard to that. While already diverse, the SweLL tagset was still missing certain types of PII included elsewhere, such as e.g. other names (usernames, nicknames, forms of address, etc.), demographic features (nationality, language, medical conditions, etc.), other entities (works of art, sports teams, medications, etc.), other alphanumerical values (IP addresses, passwords, units, quantities, etc.).

The analysis of the samples from various domains reveals that an overwhelming majority of them (with the exception of online reviews) tend to include PII such as `city` or `personal names`. All of the domains contain some kind of `geographical information`, at varying degrees of specificity, as well as mentions of `institutions` such as workplaces or schools. The presence of categories such as `age` is not guaranteed in every domain, and few domains contain mentions of `public transportation` that could be sensitive. `Dates` appear everywhere except for blogs and online reviews, but there they should be present in the accompanying data (posts tend to include accompanying information such as the date and time of posting and the author in addition to the body of the post), if not in the text itself. Data such as `e-mail` addresses, `phone numbers`, and other numerical PII do not appear in web news or online reviews; they are also absent from the memoir, but it is important to note that the samples we have worked with for this domain were old, and this absence may be dictated by the age of the text and not be domain-specific. Finally, most of the domains feature mentions of `profession, education`, or `family relations`. This distribution is illustrated in Table 5.

What is worth pointing out is that the analysis of the samples has revealed the presence of possibly sensitive information that has, at best, been included by the umbrella categories in tagsets such as the one presented by Pilán et al. (2022) or by categories such as `extra` or `Misc`. For instance, specific physical characteristics or formal titles, as seen in:

> Był w Rawie **Komornikiem** Pan Gniewosz, **malutki wzrostem** [...]
> *The **bailiff** in Rawa was Mr. Gniewosz, **short in stature** [...]* (Szawerna, 2023)

Pet names could also potentially be sensitive:

> Jag kan inte åka hit o dit som jag vill, utan jag måste alltid tänka på **Teddy**.
> *I cannot go wherever I want whenever I want, instead I always have to think about **Teddy**.* (Ahrenberg et al., 2020)

Another example could be mentions of events, including relevant hashtags in social media:

> RT @ mathiasgaunitz : Ska du till Växjö och **#mat2011** idag [...]
> *RT @mathiasgaunitz: If you are coming to Växjö for **#mat2011** today [...]* (n/a, 2022)

Elements such as the ones presented above could be used for reidentification when combined with other PIIs - this is also dependent on external factors, such as the reader's personal knowledge, and more potential categories can be found in the preceding subsection.

While the distribution of more general categories (e.g. `Person`) seems to be similar in the two papers that reported the use of such a division (Figure 1), differences are more visible when comparing categories with more fine-grained distinctions. Most papers feature at least the distinction between first and last names. As shown in Figure 2, the distribution of these two categories can vary significantly, although papers working on data from the same domain (Bråthen et al. (2021); Alfalahi et al. (2012); Dalianis (2019)) feature less diverging distributions. Figure 3 shows the comparison between the only two papers that included `gender-neutral` or `unknown` as a possibility for given names. The proportions shown in the figure are quite different from each other, reflecting how different kinds of data show up in different domains; while the differences are less drastic, Figure 1 and Figure 2 can be used to draw similar consequences. An additional potential factor for varied distribution could be different interpretations of the meaning of a label.

5. Discussion

Pseudonymization appears to offer a **reasonable tradeoff** between the legal requirement to protect the data subject and the need to preserve the utility of research data for open access. As in any case where "tradeoff" is used, the solution is not ideal but involves a certain compromise. While removing PIIs from a collection of texts enables its sharing, which, in turn, allows for and encourages new research, the quality of the data in question is altered. Changing the semantics, social level, ethnic group, or other sociolinguistic associations of the tokens can also affect how the text can be used for future research. This can have consequences when it comes to e.g. building word frequency lists or training distribution-based language models. Developing **reliable pseudonymization methods** seems essential for minimizing the negative effect of the procedure itself, and the goal should be maximizing privacy while minimizing information loss.

It can also be noted that personal information can take in principle any form, and is dependent on the previous knowledge of the reader, such as a reader would probably recognize their neighbor (the writer or the patient) through more intimate knowledge

about them than a generic reader who has never been in contact with that person, deciding what information to treat as PII much more complicated.

Is a universal pseudo-tagset possible?

It is clear from the analyses that none of the discussed taxonomies encompasses all of the kinds of PII that have been noted to occur in various domains and the desired level of detail. The taxonomy for pseudonymization should, ideally, be relatively detailed, so that a detected PII can easily be detected and replaced with a semantically and contextually appropriate surrogate - the degree of detail that most of the analyzed taxonomies are lacking. We have reached a conclusion that any existing tagset would need to be extended for it to be applicable to a wide array of genres of linguistic data; alternatively, the existing tagsets can serve as an inspiration for the creation of a new, universal one. We suggest that the categories be organized **hierarchically** so that varying levels of detailedness can be applied depending on the needs. We believe that the definition of new tags or categories of PII previously not covered by a given tagset should happen in an **empirical** fashion, by evaluating the tagset on data from various domains, and **dynamically** defining new tags based on the actual needs, since the differences of what is considered PII between various tagsets hints at the possibility that not all categories of PII are equally critical for re-identification. As we envision a set of tags that could be adapted to a given domain, it would also be important to develop a **universal format** for reporting what combination of categories has undergone pseudonymization or anonymization in the corpus, e.g. In this study we use data X with Y-type pseudonymization.

Emergence and incorporation of new categories

It is important to note that such a taxonomy may need to be adapted depending on legal requirements or social trends; interestingly, among previously discussed papers on medical data, only one of the more recent ones saw the need for including a gender-neutral name tag, possibly reflecting the **trend in the society** for the inclusion of more diverse gender identities and expressions. Simultaneously, differences in what kinds of personal data may be present can vary **cross-culturally**, where the definitions of what constitutes a name or a surname may be different (e.g. patronymic surnames).

The need to adapt to societal development sheds light on another problem: none of the taxonomies will ever be complete, if based on manual annotations (which reflect past truths in the society). We therefore see a need to establish measures that would let us update the pseudonymization taxonomy dynamically, as new types of personal information are uncovered in the data.

6. Conclusions and Future Work

We strongly believe that there is a need for a universal tagset for privacy-preserving procedures, especially for pseudonymization, but more work is needed to assess exactly what distinctions should be featured in such a tagset. We are convinced that the work done by Megyesi et al. (2018, 2021) lays a solid foundation for the development of such a tagset, but needs to be expanded with the general categories mentioned in **??** and include distinctions into sub-types. Some of the ways to proceed with this could include a project within ISO/TC 37/SC 4 or a community-built ontology,[6] with organizational inspiration drawn from communities such as Universal Dependencies (Nivre et al., 2017).

More work is also needed when it comes to deciding what level of specificity the categories should have for optimal pseudonym generation, and what other kinds of information (e.g. grammatical features) need to be preserved. It would also likely be beneficial to know what kinds of PII prevail in the data, both across and within domains. Simultaneously, it would be interesting to see how language models can contribute to the detection and division of various types of PII.

At the same time, very little is known about the effect pseudonymization may have on a variety of NLP and linguistic tasks. More work is needed to establish the usability of data modified in such a fashion.

Nguyen and Vu (2023) and Holmes et al. (2023) discussed the complexity of standard in data privacy. An individual's data may need to be shared differently depending on where it is shared and for what purpose (importantly, Sierro et al. (2024) and Cabrera-Diego and Gheewala (2024) highlight the differences in what needs to be de-identified based on the jurisdiction). This is especially relevant in an era where personal information is increasingly collected and used (e.g. to train large-language models).

We suggest a solution that could be called "dynamic privacy" in order to address the problem of personalized privacy issues by discovering all personal elements in language data but only removing or replacing the ones that have to be handled given the purpose for which the data is shared. The selection of what should be removed could be guided by appropriate tags. By creating such an adaptable tool we could ensure that individuals are given the means and transparency they need to protect their digital identities, and such a tool could rely on the categorization advocated in this paper.

[6] We thank one of the anonymous reviewers for their suggestions.

7. Limitations

It is important to acknowledge that the analyses conducted in this paper are not exhaustive and do not discuss all of the tagsets that have ever been used for anonymization or pseudonymization. Similarly, only a limited number of samples from different domains have been inspected, and the absence of a given type of PII from them does not mean that that PII type never occurs in a given domain – but perhaps it is only less likely, or the sample size was not big enough. The analyses themselves were preliminary studies, and their intention was to raise this issue of the categorization of PIIs rather than to provide solutions, as we believe it is too early to draw any definite conclusions. Nevertheless, we believe that these investigations have shed some light on the issue and serve to support the point, as well as to indicate further directions for research.

8. Ethical Concerns

As mentioned in section 1, methods such as anonymization and pseudonymization are prompted by both legal and ethical reasons. Minimizing the potential negative impact that gathering and sharing linguistic data may have on its authors or other individuals mentioned in the text is paramount – which is what has prompted this paper. Since we have relied on previously published data, we cannot attest to its representativeness, nor to the way it was originally sourced.

9. Bibliographical References

Pierre Accorsi, Namrata Patel, Cédric Lopez, Rachel Panckhurst, and Mathieu Roche. 2012. Seek&Hide: Anonymising a French SMS corpus using natural language processing techniques. *Linguisticae Investigationes*, 35:163–180.

Allison Adams, Eric Aili, Daniel Aioanei, Rebecca Jonsson, Lina Mickelsson, Dagmar Mikmekova, Fred Roberts, Javier Fernandez Valencia, and Roger Wechsler. 2019. AnonyMate: A Toolkit for Anonymizing Unstructured Chat Data. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 1–7, Turku, Finland. Linköping Electronic Press.

Alyaa Alfalahi, Sara Brissman, and Hercules Dalianis. 2012. 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*.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of språkbanken. In *Proceedings of LREC 2012. Istanbul: ELRA*, volume Accepted, page 474–478.

Synnøve Bråthen, Wilhelm Wie, and Hercules Dalianis. 2021. Creating and Evaluating a Synthetic Norwegian Clinical Corpus for De-Identification. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 222–230, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Luis Adrián Cabrera-Diego and Akshita Gheewala. 2024. PSILENCE: A pseudonymization tool for international law. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 25–36, St. Julian's, Malta. Association for Computational Linguistics.

Hercules Dalianis. 2019. Pseudonymisation of Swedish electronic patient records using a rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 16–23, Turku, Finland. Linköping Electronic Press.

George Danezis, Josep Domingo-Ferrer, Marit Hansen, Jaap-Henk Hoepman, Daniel Le Metayer, Rodica Tirtea, and Stefan Schiffner. 2015. Privacy and data protection by design-from policy to engineering. *arXiv preprint arXiv:1501.03726*.

Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 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. INCOMA Ltd.

Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2020. CodE alltag 2.0 — a pseudonymized German-language email corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4466–4477, Marseille, France. European Language Resources Association.

Elisabeth Eder, Michael Wiegand, Ulrike Krieg-Holz, and Udo Hahn. 2022. "Beste Grüße, Maria Meyer" — Pseudonymization of Privacy-Sensitive Information in Emails. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 741–752, Marseille, France. European Language Resources Association.

EU EU Commission. 2016. *General data protection regulation.* Official Journal of the European Union, 59, 1-88.

Langdon Holmes, Scott Crossley, Harshvardhan Sikka, and Wesley Morris. 2023. Piilo: an open-source system for personally identifiable information labeling and obfuscation. *Information and Learning Sciences*.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation Models for Text Data: State of the art, Challenges and Future Directions. In *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)*, pages 4188–4203, Online. Association for Computational Linguistics.

Beáta Megyesi, Lena Granstedt, Sofia Johansson, Julia Prentice, Dan Rosén, Carl-Johan Schenström, Gunlög Sundberg, Mats Wirén, and Elena Volodina. 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. LiU Electronic Press.

Beáta Megyesi, Lisa Rudebeck, and Elena Volodina. 2021. SweLL pseudonymization guidelines.

Tuan Minh Nguyen and Xuan-Son Vu. 2023. Privacy and trust in iot ecosystems with big data: A survey of perspectives and challenges. In *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 215–222. IEEE.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Maria Sierro, Begoña Altuna, and Itziar Gonzalez-Dios. 2024. Automatic detection and labelling of personal data in case reports from the ECHR in Spanish: Evaluation of two different annotation approaches. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 18–24, St. Julian's, Malta. Association for Computational Linguistics.

Elena Volodina, Yousuf Ali Mohammed, Sandra Derbring, Arild Matsson, and Beata Megyesi. 2020. 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*, pages 357–369, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Elena Volodina, Simon Dobnik, Therese Lindström Tiedemann, and Xuan-Son Vu. 2023. Grandma Karl is 27 years old – research agenda for pseudonymization of research data. In *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService), Athens, Greece, 2023*, Los Alamitos. IEEE Computer Society.

10. Language Resource References

Ahrenberg, Lars and Frid, Johan and Olsson, Leif-Jöran. 2020. *Swe-NERC*. Språkbanken Text, University of Gothenburg. PID https://hdl.handle.net/10794/121.

Enron Corp and Cohen, William W. *Enron Email Dataset*. PID https://hdl.loc.gov/loc.gdc/gdcdatasets.2018487913.

Julian J. McAuley and Jure Leskovec. 2013. *From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews*.

n/a. 2022. *Twitter Mix*. Språkbanken Text. Distributed via SBX/CLARIN. PID https://hdl.handle.net/10794/869.

Pilán, Ildikó and Lison, Pierre and Øvrelid, Lilja and Papadopoulou, Anthi and Sánchez, David and Batet, Montserrat. 2022. *The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization*. MIT Press.

Maria Irena Szawerna. 2023. *IŻ SWÓJ JĘZYK MAJĄ! An exploration of the computational methods for identifying language variation in Polish*.

Volodina, Elena and Granstedt, Lena and Matsson, Arild and Megyesi, Beáta and Pilán , Ildikó and Prentice, Julia and Rosén, Dan and Rudebeck, Lisa and Schenström, Carl-Johan and Sundberg, Gunlög and Wirén, Mats. 2022. *SweLL-gold*.

Språkbanken Text. Distributed via SBX/CLARIN. PID https://hdl.handle.net/10794/846.

Wirén, Mats and Matsson, Arild and Rosén, Dan and Volodina, Elena. 2019. *Svala: Annotation of second-language learner text based on mostly automatic alignment of parallel corpora*. Linköping University Electronic Press.

11. Acknowledgements

This work was possible thanks to the funding of several grants from the Swedish Research Council.

All of the authors are supported by the research environment project *Grandma Karl is 27 years old: Automatic pseudonymization of research data* with the funding number 2022-02311 for the years 2023-2029.

The first, fourth, and sixth authors are also receiving support from the Swedish national research infrastructure *Nationella Språkbanken*, which is funded jointly by contract number 2017-00626 for the years 2018-2024, as well 10 as participating partner institutions.

The second author is also supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the *Centre for Linguistic Theory and Studies in Probability (CLASP)* at the University of Gothenburg.

This work has also been aided by the Swedish national research infrastructure *Huminfra*, funded for the years 2022-2024, contract 2021-00176, and the participating partner institutions.

12. Appendix A: Further examples of sensitive data in various domains

The marking in **bold** of the personal data follows the one provided in the respective source papers, but does not preserve the detailed classification; elements in *italics* are deemed to be less sensitive than the underlined elements.

> The applicant owned a garage in ***Dorset*** and had business connections in ***Spain*** He had two ***Mercedes*** cars each of which had a false compartment in the fuel tank. The false compartments could hold up to 45 kilograms of ***cannabis*** resin. From ***1994*** he was suspected by the police of being involved in drug trafficking. The police also suspected him of being involved in the handling of stolen goods, including stolen vehicles.

Example 5: An example of sensitive data that can be found in a legal documentation, as presented in Pilán et al. (2022)

> Discharge letter ***Huddinge***
> Responsible. specialist / chief physician **Caroline Berg**
> Medical secretary **Marianne Lindgren**
> Print Date ***20120325***
> Care episode **20120311-20120318**
> Main diagnosis according to ICD-10
> History of **52-year-old** woman, well known in the clinic. Treated by **Karin Lundgren** and at the pain clinic. Has a chronic headache without a known origin. Given Methadone, Actiqe and Stesolid. Came to clinic on the ***22/5*** due to unsustainable situation with inadequate pain control. Pat. is frustated over the long waiting time for the discontinuation of opiates which was to be done via ***IVA*** and planned by Dr. **Torbjörn Andreasson**. Pat comes to ***NIVA*** and demands to be admitted to ***IVA*** and threatens to stop taking all drugs. Pat had several conversations with PAL at ***Löwet***, **Sandra Månsson**. Refers to previous notes.

Example 6: An example of (already pseudonymized) sensitive data that can be found in a medical documents, as presented in Dalianis (2019)

> I live in ***Stockholm***. I am **29** years old. I live with my ***boyfriend***. His name is **Cezary**. I have the buss and the ***Stockholm*** train. I lived in ***Danmark*** bifore, in ***Odense***. It was less than ***Stockholm***. But ***Stockholm*** is closser to ***Luxembourg*** than ***Odense***.

Example 7: An example of sensitive data that can be found in a learner essay, as presented in Volodina et al. (2020)