# Guiding Text-to-Text Privatization by Syntax

**Stefan Arnold** and **Dilara Yesilbas** and **Sven Weinzierl**
Friedrich-Alexander-Universität Erlangen-Nürnberg
Lange Gasse 20, 90403 Nürnberg, Germany
(stefan.st.arnold, dilara.yesilbas, sven.weinzierl)@fau.de

## Abstract

*Metric Differential Privacy* is a generalization of differential privacy tailored to address the unique challenges of text-to-text privatization. By adding noise to the representation of words in the geometric space of embeddings, words are replaced with words located in the proximity of the noisy representation. Since embeddings are trained based on word co-occurrences, this mechanism ensures that substitutions stem from a common semantic context. Without considering the grammatical category of words, however, this mechanism cannot guarantee that substitutions play similar syntactic roles. We analyze the capability of text-to-text privatization to preserve the grammatical category of words after substitution and find that surrogate texts consist almost exclusively of nouns. Lacking the capability to produce surrogate texts that correlate with the structure of the sensitive texts, we encompass our analysis by transforming the privatization step into a candidate selection problem in which substitutions are directed to words with matching grammatical properties. We demonstrate a substantial improvement in the performance of downstream tasks by up to $4.66\%$ while retaining comparative privacy guarantees.

## 1 Introduction

From compliance with stringent data protection regulations to building trust, privacy emerged as a formidable challenge to applications that build on user-generated data, and consensus exists regarding the need to safeguard user privacy.

In the context of text analysis, privacy is typically protected by sanitizing personally identifiable information from the text via ad-hoc filtering or anonymization. The literature is replete with naïve approaches that either redact words from the text or insert distractive words into the text. Using generalization and suppression on quasi-identifiers, an intuitive way of expressing privacy is presented by $k$-anonymity (Sweeney, 2002) and its notable adaptations for text data (Jiang et al., 2009; Sánchez and Batet, 2016).

However, these approaches are fundamentally flawed. Incapable of anticipating an adversary's side knowledge, most anonymization schemes are vulnerable to re-identification and thus provably non-private. As text conveys seemingly innocuous information, researchers demonstrated that this information can be leveraged to identify authorship (Song and Shmatikov, 2019) or disclose identifiable information (Carlini et al., 2020; Pan et al., 2020; Song and Raghunathan, 2020; Thomas et al., 2020). Carlini et al. (2020), for instance, recovered verbatim text from the training corpus using black-box querying to a language model.

Building upon noise calibration, *Differential Privacy* (DP) (Dwork et al., 2006b) attracted considerable attention for their robust notion of privacy. For text analysis, DP is applied to the vector-valued representation of text data (Coavoux et al., 2018; Weggenmann and Kerschbaum, 2018; Vu et al., 2019).

We focus on *Metric Differential Privacy* (Chatzikokolakis et al., 2013), in which data is processed independently, similar to the setting of randomized response (Kasiviswanathan et al., 2011). To avoid the curse of dimensionality of randomized response, noise is scaled by a general distance metric. For text-to-text privatization, Feyisetan et al. (2020) adopted a distance metric so that words that are close (*i.e.* more similar) to a word are assigned with a higher substitution probability than those that are more distant (*i.e.* less similar). This requires that the text is mapped onto a continuous embedding space (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017). Proceeding from the embedding, each word in the text is privatized by a three-step protocol: (1) retrieving the vector representation of the word, (2) perturbing the vector representation of the word with

noise sampled from a multivariate distribution, and (3) projecting the noisy representation of the word back to the discrete vocabulary space. As the noisy representations are unlikely to exactly represent words in the embedding space, a nearest neighbor approximation is returned.

Since text-to-text privatization operates directly on embeddings and words in the embedding space are mapped based on co-occurrences, words tend to be substituted by words that stem from a common semantic context. However, there is no guarantee that words are substituted by words that serve similar roles within the grammatical structure of a text. Motivated by the example of sentiment analysis, in which sentiment is typically expressed by adjectives and forms of adjectives (Benamara et al., 2007), we hypothesize that substitutions strictly based on co-occurrences may degrade downstream performance. This hypothesis is in line with linguists finding repeated evidence for the relevance of grammatical properties for language understanding (Myhill et al., 2012).

We summarize our contributions as follows:

- We investigate text-to-text privatization via metric differential privacy in terms of its capability to preserve the grammatical properties of words after substitution. We find that privatization produces texts that consist to a large extent of incoherent nouns.

- We incorporate grammatical categories into the privatization step in the form of a constraint to the candidate selection. We demonstrate that broadening the candidate pool to $k > 1$ (instead of $k = 1$) and selecting a substitution with matching grammatical properties amplifies the performance in downstream tasks while maintaining an equivalent level of privacy.

## 2 Preliminaries

### 2.1 Differential Privacy

*Differential Privacy* (DP) (Dwork et al., 2006b) emerged as a robust notion for privacy applied in privacy-preserving data mining and machine learning. Due to its composability and robustness to post-processing regardless of an adversary's side knowledge, it formalizes privacy without the critical pitfalls of previous anonymization schemes. To ensure a consistent understanding of the algorithmic foundation of differential privacy, we present a brief taxonomy and a formal definition of the variants used for text analysis.

Formally, a randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies $\varepsilon$-indistinguishability if any two adjacent inputs $d, d' \in \mathcal{D}$ and for any subset of outputs $S \subseteq \mathcal{R}$ it holds that:

$$\frac{\mathbb{P}[\mathcal{M}(d) \in S]}{\mathbb{P}[\mathcal{M}(d') \in S]} \leq e^{\varepsilon}. \tag{1}$$

At a high level, a randomized mechanism is differentially-private if the output distributions from two adjacent datasets are (near) indistinguishable, where any two datasets are considered adjacent that differ in at most one record. An adversary seeing the output can therefore not discriminate if a particular observation was used. This notion of indistinguishability is controlled by the parameter $\varepsilon$ acting as a privacy budget. It defines the strength of the privacy guarantee (with $\varepsilon \to 0$ representing strict privacy and $\varepsilon \to \infty$ representing the lack of privacy). To enhance the accounting of the privacy budget, several relaxations exist (Dwork et al., 2006a; Mironov, 2017; Dong et al., 2019).

Depending on the setting, DP can be categorized into *global* DP (Dwork et al., 2006b) and *local* DP (Kasiviswanathan et al., 2011).

Global DP addresses the setting in which privacy is defined with respect to aggregate statistics. It assumes a trusted curator who can collect and access raw user data. The randomized mechanism is applied to the collected dataset to produce differentially private output for downstream use. With noise drawn from a predetermined distribution, the design of the randomized mechanism builds upon an additive noise mechanism. Commonly used distributions for adding noise include Laplace and Gaussian distribution (Dwork et al., 2014). The noise is further calibrated according to the function's sensitivity and the privacy budget. This technique is useful for controlling the disclosure of private information of records processed with real-valued and vector-valued functions.

Local DP addresses the setting in which privacy is defined with respect to individual records. In contrast to global DP, local DP does not rely on a trusted curator. Instead of a trusted curator that applies the randomized mechanism, the randomized mechanism is applied to all records independently to provide plausible deniability (Bindschaedler et al., 2017). The randomized mechanism to achieve local DP is typically *Randomized*

*Response* (RR) (Warner, 1965), which protects private information by answering a plausible response to the sensitive query.

Since we aim for text-to-text privatization, formulating DP in the local setting through RR appears to be a natural solution. However, the strong privacy guarantees constituted by RR impose requirements that render it impractical for text. That is, RR requires that a sentence $s$ must have a non-negligible probability of being transformed into any other sentence $s'$ regardless of how unrelated $s$ and $s'$ are. This indistinguishability constraint makes it virtually impossible to enforce that the semantics of a sentence $s$ are approximately captured by a privatized sentence $s'$. Since the vocabulary size can grow exponentially large in length $|s|$, the number of sentences semantically related to $s$ becomes vanishingly small probability under RR (Feyisetan et al., 2020).

## 2.2 Metric Differential Privacy

*Metric Differential Privacy* (Chatzikokolakis et al., 2013) is a generalization of differential privacy that originated in the context of location-based privacy, where locations close to a user are assigned with a high probability, while distant locations are given negligible probability. By using word embeddings as a corollary to geo-location coordinates, metric differential privacy was adopted from location analysis to textual analysis by Feyisetan et al. (2020).

We follow the formulation of Xu et al. (2021) for metric differential privacy in the context of textual analysis. Equipped with a discrete vocabulary set $\mathcal{W}$, an embedding function $\phi : \mathcal{W} \rightarrow \mathbb{R}$, where $\mathbb{R}$ represents a high-dimensional embedding space, and a distance function $d : \mathbb{R} \times \mathbb{R} \rightarrow [0, \infty)$ satisfying the axioms of a metric (*i.e.*, identity of indiscernibles, symmetry, and triangle inequality), metric differential privacy is defined in terms of the distinguishability level between pairs of words. A randomized mechanism $\mathcal{M} : \mathcal{W} \rightarrow \mathcal{W}$ satisfies metric differential privacy with respect to the distance metric $d(\cdot)$ if for any $w, w', \hat{w} \in \mathcal{W}$ the output distributions of $\mathcal{M}(w)$ and $\mathcal{M}(w')$ are bounded by Equation 2 for any privacy budget $\varepsilon > 0$:

$$\frac{\mathbb{P}[\mathcal{M}(w) = \hat{w}]}{\mathbb{P}[\mathcal{M}(w') = \hat{w}]} \leq e^{\varepsilon d\{\phi(w), \phi(w')\}}. \quad (2)$$

This probabilistic guarantee ensures that the log-likelihood ratio of observing any word $\hat{w}$ given two words $w$ and $w'$ is bounded by $\varepsilon d\{\phi(w), \phi(w')\}$ and provides plausible deniability (Bindschaedler

et al., 2017) with respect to all $w \in \mathcal{W}$. We refer to Feyisetan et al. (2020) for a complete proof of privacy. For $\mathcal{M}$ to provide plausible deniability, additive noise is in practice sampled from a multivariate distribution such as the *multivariate Laplace distribution* (Feyisetan et al., 2020) or *truncated Gumbel distribution* (Carvalho et al., 2021b).

We recall that differential privacy requires adjacent datasets that differ in at most one record. Since the distance $d(\cdot)$ captures the notion of closeness between datasets, metric differential privacy instantiates differential privacy when Hamming distance is used, *i.e.*, if $\forall x, x' : d\{\phi(w), \phi(w')\} = 1$. Depending on the distance function $d(\cdot)$, metric differential privacy is therefore generally less restrictive than differential privacy. Intuitively, words that are distant in metric space are easier to distinguish compared words that are in close proximity. Scaling the indistinguishability by a distance $d(\cdot)$ avoids the curse of dimensionality that arises from a large vocabulary $\mathcal{W}$ and allows the mechanism $\mathcal{M}$ to produce similar substitutions $\hat{w}$ for similar $w$ and $w'$. However, this scaling complicates the interpretation of the privacy budget $\varepsilon$, as it changes depending on the metric employed.

## 2.3 Related Work

Grounded in metric differential privacy, text-to-text privatization implies that the indistinguishability of substitutions of any two words in the vocabulary is scaled by their distance.

Fernandes et al. (2018) achieve this indistinguishability by generating a bag-of-words representation and applying the *Earth Mover's distance* to obtain privatized bags.

In contrast to a bag-of-words representation, Feyisetan et al. (2020) formalized text-to-text privatization to operate on continuous word embeddings. Word embeddings capture the level of semantic similarity between words and have been popularized by efficient embedding mechanisms (Mikolov et al., 2013; Pennington et al., 2014). This mechanism was termed MADLIB.

The issue with this mechanism is that the magnitude of the noise is proportional to the dimensionality of the vector representation. This translates into adding the same amount of noise to any word in the embedding space, regardless of whether this word is located in a dense or sparse region. For words in densely populated areas, adding noise that is large in magnitude renders it difficult for the mech-
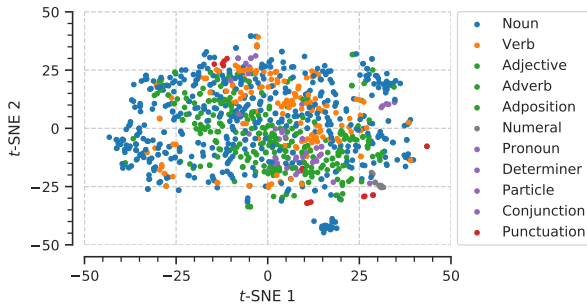
Figure 1: Embedding space of the $1,000$ most frequent words in 100-dimensional `GloVe`, automatically encoded with their universal part-of-speech tags.

anism to select reasonable substitutions, as nearby relevant words cannot be distinguished from other nearby but irrelevant words. For words in sparsely populated areas, adding noise of small magnitude renders the mechanism susceptible to reconstruction, as the word closest to a noisy representation is likely to be the original word.

To tackle some of the severe shortcomings of `MADLIB`, a variety of distance metrics have been employed to scale the indistinguishability, including *Hamming distance* (Carvalho et al., 2021a), *Manhattan distance* (Fernandes et al., 2019), *Euclidean distance* (Fernandes et al., 2019; Feyisetan et al., 2020; Carvalho et al., 2021b; Feyisetan and Kasiviswanathan, 2021), *Mahalanobis distance* (Xu et al., 2020) and *Hyperbolic distance* (Feyisetan et al., 2019).

While related extensions have focused almost exclusively on geometric properties to enhance text-to-text privatization, we focus on linguistic properties. We extend `MADLIB` by a candidate selection that directs substitutions based on matching grammatical properties and demonstrate that multivariate perturbations supported by grammatical properties substantially improve the utility of the surrogate texts in downstream tasks.

## 3  Methodology

Since text-to-text privatization operates directly on geometric space of embeddings, it is necessary to understand the structure of the embedding space. To get an understanding of the embedding space, we selected a subset of $1,000$ most frequent words from the 100-dimensional `GloVe` embedding and manifolded them onto a two-dimensional representation. Enriched by grammatical properties derived from the universal part-of-speech tagset (Petrov

et al., 2011), we chart a $t$-distributed stochastic neighbor embedding (Van der Maaten and Hinton, 2008) in Figure 1.

We note that we derived each word's grammatical category without context, which may explain the general tendency towards *nouns* (presumably misclassified *verbs*). Regardless of potentially misclassified grammatical categories, we can draw the following conclusions: while *nouns*, *verbs*, and *adjectives* are distributed throughout the embedding space, we find distinct subspaces for *numerals* and *punctuation*. This is because word embeddings are trained towards an objective that ensures that words occurring in a common context have similar embeddings, disregarding their syntactic roles within the structure of a text. Considering that text-to-text privatization typically selects the nearest approximate neighbor after the randomized mechanism is queried as substitution, we expect this mechanism to fall short in producing syntactically coherent texts.

We adopt the multivariate Laplace mechanisms of `MADLIB` (Feyisetan et al., 2020). Aimed at preserving the grammatical category of a word after its substitution, we incorporate a constraint into the candidate selection that directs the randomized mechanism towards words with a matching grammatical category. This constraint is incorporated as follows: we create a dictionary that serves as a lookup table for the grammatical category of each word in the vocabulary and generalize the randomized mechanism to return a flexible $k \gg 1$ (instead of $k = 1$) approximate nearest neighbors. If available, a word is replaced by the nearest word (measured from the noisy representation) that matches its grammatical category. Otherwise, the protocol reduces to canonical `MADLIB`. The computational overhead of the candidate selection is $O(\log k)$.

This modification introduces the size of the candidate pool $k$ as an additional hyperparameter. Intuitively, $k$ should be chosen based on the geometric properties of the embedding, *i.e.*, $k$ should be large enough to contain at least one other word with a matching grammatical category.

We investigate our modification to `MADLIB` in terms of its capability to preserve grammatical properties and its implications. For reasons of reproducibility, we base all experiments on the 100-dimensional `GloVe` embedding.

To keep the computational effort feasible, we formed a vocabulary that consists of $24,525$ words

reflecting a natural distribution of grammatical categories: 26 *pronouns*, 5,000 *nouns*, 5,000 *verbs*, 5,000 *adjectives*, 4,341 *adverbs*, 92 *adpositions*, 5,000 *numerals*, 6 *conjunctions*, 2 *particles*, 39 *determiner*, and 19 *punctuations*.

Once we determined our sub-vocabulary, we calculated the necessary size of the candidate pool $k$. We counted the number of steps required from each word in our subset until a neighbor with a matching category was found. Averaging this count revealed that each word is linked to another word with a matching category within a neighborhood of 20. We thus parameterized the candidate pool to a fixed $k = 20$ across all experiments.

## 4 Experiments

We conduct a series of experiments at a strategically chosen set of privacy budgets $\varepsilon = \{5, 10, 25\}$ to demonstrate the relevance of directing substitution to words that share similar syntactic roles rather than restricting substitution only to words that appear in a similar semantic context.

These privacy budgets represent three privacy regimes: $\varepsilon = 5$ for high privacy, $\varepsilon = 10$ for moderate privacy, and $\varepsilon = 25$ for low privacy.

### 4.1 Linguistic Analysis

We intend to assess the effectiveness of our constraint to the candidate selection in retaining grammatical properties of words after substitution. We query each word contained in the vocabulary 100 times and record the grammatical category for its surrogate word in the form of a frequency count.

Given a moderate privacy budget of $\varepsilon = 10$, Figure 2 visualizes the calculated frequency counts similar to a confusion matrix. The diagonal represents the preservation capability of grammatical categories, *i.e.*, universal part-of-speech tags. A comparison across $\varepsilon \in \{5, 10, 25\}$ is deferred to Figure A.1 in the Appendix A.

We start with the examination of the baseline mechanism in Figure 3(a). Consistent with the independent and concurrent results of Mattern et al. (2022), our results indicate that the privatization mechanism is likely to cause grammatical errors. Mattern et al. (2022) estimate that the grammatical category changes in $7.8\%$, whereas we calculated about $45.1\%$ for an identical privacy budget. This difference arises from the fact that Mattern et al. (2022) only consider the four most frequent categories of *nouns*, *verbs*, *adjectives*, and *adverbs*,
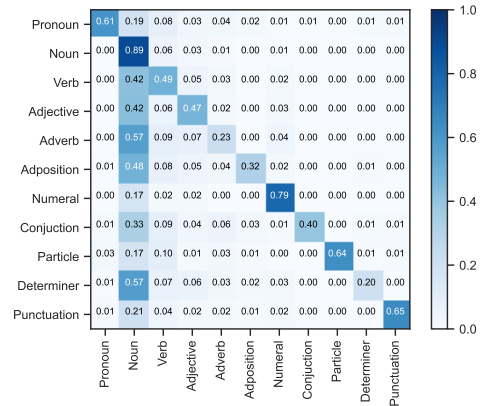


(a) MADLIB with $k = 1$
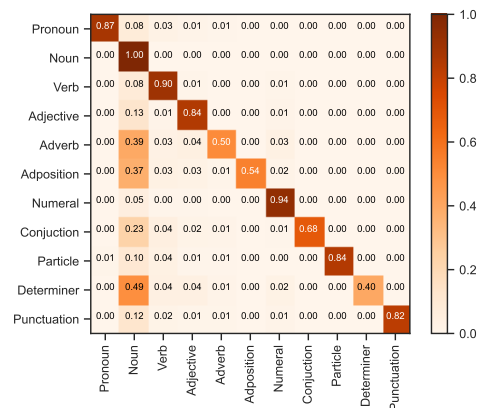


(b) MADLIB with $k = 20$

Figure 2: Approximated frequency counts by querying a subset of words and recording their universal part-of-speech tags before and after substitution. The diagonal represents the ideal preservation of grammatical properties.

while we consider eleven categories according to the universal part-of-speech tagset. In addition to the number of grammatical categories, we indicate the fluctuations between categories, while Mattern et al. (2022) only measures whether a category was changed. Owing to the tracking of the fluctuations, we find a disparate impact on the preservation of the grammatical categories. We find that the preservation of grammatical categories of words declines with growing guarantees for privacy, until the text after privatization consist almost entirely of nouns.

We compare these results to our constrained mechanism in Figure 2(b). With the introduction of a constrained candidate pool of size $k = 20$, we observe an increased likelihood that surrogate texts retain the grammatical structure of the original texts. This can be seen by the dominance of the vertical line in Figure 3(a) compared to initial signs

```
('it', 'PRON')          ('doing', 'VERB')
('"s', 'VERB')          ('concept', 'NOUN')
('been', 'VERB')        ('courtesy', 'NOUN')
('done', 'VERB')        ('funding', 'NOUN')
('before', 'ADP')       ('1979', 'NUM')
('but', 'CONJ')         ('relatively', 'ADV')
('never', 'ADV')        ('took', 'VERB')
('so', 'ADV')           ('listening', 'VERB')
('vividly', 'ADV')      ('chunky', 'NOUN')
('or', 'CONJ')————————('or', 'CONJ')
('with', 'ADP')         ('vinyl', 'NOUN')
('so', 'ADV')           ('handy', 'ADJ')
('much', 'ADJ')         ('gangsters', 'NOUN')
('passion', 'NOUN')————('fanaticism', 'NOUN')
('.', 'PUNCT')          ('doom', 'NOUN')
```

(a) MADLIB with $k = 1$

```
('it', 'PRON')————————('itself', 'PRON')
('"s', 'VERB')          ('grand', 'ADJ')
('been', 'VERB')————————('having', 'VERB')
('done', 'VERB')        ('make', 'NOUN')
('before', 'ADP')————————('since', 'ADP')
('but', 'CONJ')————————('but', 'CONJ')
('never', 'ADV')————————('otherwise', 'ADV')
('so', 'ADV')————————('back', 'ADV')
('vividly', 'ADV')————('fatally', 'ADV')
('or', 'CONJ')————————('or', 'CONJ')
('with', 'ADP')————————('for', 'ADP')
('so', 'ADV')————————('essentially', 'ADV')
('much', 'ADJ')————————('regular', 'ADJ')
('passion', 'NOUN')————('virtuosity', 'NOUN')
('.', 'PUNCT')————————('(', 'PUNCT')
```

(b) MADLIB with $k = 20$

Figure 3: Example of syntax-preserving capabilities of MADLIB with and without grammatical constraint.



Figure 4: Euclidean distance for word substitutions. We depict default MADLIB ($k = 1$) in blue and MADLIB ($k = 20$) with grammatical constraint in orange.

of a diagonal line in Figure 2(b). Compared to the baseline value $45.1\%$, the preservation capability bounds at $81.4\%$.

We illustrate the alignment of grammatical properties between words from a sensitive text and and their surrogate words with an example sentence in Figure 3. We note that our syntactic guidance prevents words from being misleadingly replaced by numbers (and vice versa), as in the case of *before* being replaced by *1979*.

### 4.2 Geometric Analysis

Intuitive properties for analyzing a mechanism operating on embeddings include magnitude, direction, and orthogonality. Since embeddings capture word co-occurrences, we expect most substitutions to be located in the same region of an embedding space and in the same direction from the embedding origin.

We aim to measure the differences in the Euclidean distance of words with those of their corresponding substitutes generated by baseline $\mathcal{M}(w)$ and our constraint $\mathcal{M}'(w)$. The results capture $\|w - \hat{w}\|$ and $\|w - \hat{w}'\|$, respectively. Since the distances are zero when $w = \hat{w}$ or identical when $\hat{w} = \hat{w}'$, we are only interested in the distances when a substitution has occurred and the mechanisms decided on a distinct candidate for their substitution, *i.e.*, $\mathcal{M}(w) \neq \mathcal{M}'(w) \neq w$.
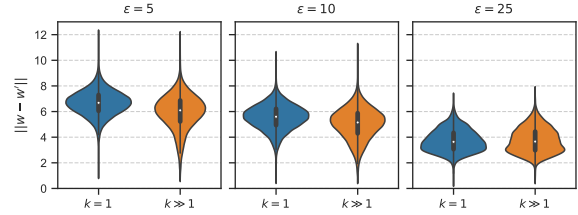
Figure 4 depicts the calculated distances for querying words from our subset 100 times. The distance approximation was carried out at a strategically chosen discrete set of values of $\varepsilon = \{5, 10, 25\}$. Since the distance is calculated as the difference between words and their substitutes, lower values indicate better substitutions. The distances depend on the amount of noise injected into the randomized mechanisms. The more noise, the larger the distances. Apparent across all privacy budgets, the distances between words and their substitutions are slightly shifted towards smaller distances. Since the distributions of distances are almost identical, we can take a principled guess that substitution in both mechanisms generally occurs within a similar region of the embedding space.

### 4.3 Privacy Analysis

Confronted with a non-zero probability that the candidate pool contains the sensitive word and no other word exists in the candidate pool with matching grammatical properties, it could be argued that the privacy guarantees suffer from the increased risk of self-substitution. By calculating the plausible deniability (Bindschaedler et al., 2017), we evaluate the risk of self-substitution arising from our grammatically constrained candidate selection.

In line with previous studies on text-to-text privatization (Feyisetan et al., 2019, 2020; Xu et al., 2021; Qu et al., 2021), we record the following statistics as proxies for plausible deniability.

- $N_w = \mathbb{P}\{M(w) = w\}$ measures the probability that a word is not substituted by the mechanism. This is approximated by counting the number of times a word $w$ is substituted by the same word after running the mechanism 100 times.

- $S_w = |\mathbb{P}\{M(w) = w'\}|$ measures the effective support in terms of the number of distinct

156

substitutions produced for a word from the mechanism. This is approximated by the cardinality of the set of words $w^{`}$ after running the mechanism 100 times.

Since the noise is scaled by $1/\varepsilon$, we can make a connection between the proxy statistics and the privacy budget $\varepsilon$. A smaller $\varepsilon$ corresponds to a more stringent privacy guarantee. Adding more noise to the vector representation of a word results in fewer self-substituted words (lower $N_w$) and a more diverse set of distinct substitutions (higher $S_w$). A higher $\varepsilon$ corresponds to a less stringent privacy guarantee. This translates into less substitutions (higher $N_w$) and a narrow set of distinct substitutions (lower $S_w$). From a distributional perspective, it follows that $N_w$ ($S_w$ should be positively (negatively) skewed to provide reasonable privacy guarantees.

For privacy budgets of $\varepsilon = \{5, 10, 25\}$, we present the distribution of $N_w$ and $S_w$ over 100 independent queries Figure 5. While lower values of $\varepsilon$ are desirable from a privacy perspective, it is widely known that text-to-text privatization requires slightly larger privacy budgets to provide reasonable utility in practice. Values of $\varepsilon$ up to 20 and 30 have been reported in related mechanisms (Feyisetan et al., 2020). The histograms serve as visual guidance for comparing (and selecting) the required privacy budget $\varepsilon$. As both mechanisms build upon the Euclidean distance as a metric, their privacy guarantees should match by using the same privacy budget $\varepsilon$. Directing the substitution to words with a matching grammatical category result in marginal changes to the plausible deniability. This is visually recognizable by the distribution shift. The grammatical constraint risks slightly more self-substitutions and reduced effective support. This is because words are substituted (almost) only by words from the same grammatical category, reducing the pool of unique words that are appropriate for substitution and thus reducing the effective support of the multivariate mechanism. Out of 100 words queried given a fixed privacy budget of $\varepsilon = 10$, self-substitution increases on average from about 29 to 32, while effective support decreases on average from about 66 to 61. The fact that both changes in $N_w$ and $S_w$ do not exceed or fall below 50 indicates that plausible deniability is assured for the average-case scenario. We conclude that the grammatically constrained candidate selection does not come at the expense of privacy and
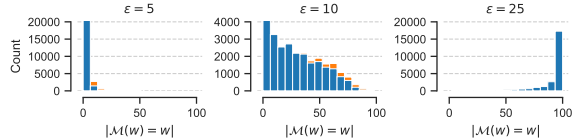


(a) $N_w$ refers to the number of substitute words that are *identical* to a queried sensitive word.
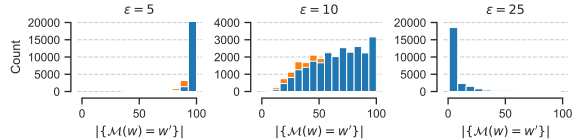


(b) $S_w$ refers to the number of substitute words that are *unique* from a queried sensitive word.

Figure 5: Plausible deniability statistics approximated for a carefully compiled sub-vocabulary of $24,525$ words of varying lexical categories, with each word independently privatized over a total number of 100 queries. We present the baseline in blue and highlight the distribution shift induced by the grammatical constraint in orange.

can therefore be incorporated into the privatization step without the need to recalibrate the proxies for plausible deniability.

Rather than compromising privacy, our constrained candidate selection can be alternatively viewed as a barrier against reconstruction attacks. Recall that the nearest neighbor search is generalized from $k = 1$ to $k \gg 1$. This generalization may impede naïve inversion attacks such as the one proposed in Song and Raghunathan (2020), in which an adversary attempts to recover a word by finding the nearest neighbor to the substitute word. Although this inversion attack is not comprehensive, it can be used as a reference point for investigations regarding the robustness of privacy attacks. We include the setup and the results of a membership inference attack in the Appendix B.

## 4.4 Utility Analysis

To evaluate whether the preservation of syntactic roles translates to better utility in downstream tasks, we conduct experiments with BERT (Devlin et al., 2018) on a subset of GLUE (Wang et al., 2019).

Once for each mechanism under comparison, we privatize the training corpus of each dataset. Since the privacy guarantees do not exactly match, we calculate the available privacy budget for each mechanism such that the .90 quantile of words is plausible deniable. This resembles a practical scenario where we allow a negligible subset of words

| | Level of Privacy | Classification | | Textual Similarity | | | Textual Entailment | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| | | CoLA (MCC) | SST2 (ACC) | QQP (ACC) | MRPC (ACC) | STSB (SCC) | MNLI (ACC) | QNLI (ACC) | RTE (ACC) | - |
| **BERT** | - | 0.5792 | 0.9243 | 0.8879 | 0.8329 | 0.8854 | 0.8229 | 0.8912 | 0.6927 | 0.8146 |
| $k = 1$ | $p = 0.9$ | 0.0248 | 0.8127 | 0.6940 | 0.5603 | 0.6153 | 0.5304 | 0.6327 | **0.5663** | 0.5545 |
| | $p = 0.5$ | 0.2303 | 0.8848 | 0.8181 | 0.6242 | 0.7951 | 0.7114 | 0.8339 | 0.6027 | 0.6875 |
| $k = 20$ | $p = 0.9$ | **0.0928** | **0.8510** | **0.7519** | **0.5946** | **0.6988** | **0.6251** | **0.7423** | 0.4525 | **0.6011** |
| | $p = 0.5$ | 0.3493 | 0.9035 | 0.8397 | 0.6333 | 0.8011 | 0.7301 | 0.8627 | 0.5420 | 0.7077 |

Table 1: Results on a subset of GLUE (Wang et al., 2019). We report Matthews correlation for the CoLA dataset, Spearman correlation for the STSB dataset, and the accuracy score for all remaining datasets. The level of privacy increases with the quantile of words that are provable plausible deniable. $p = .90$ denotes an (almost) worst-case scenario. $p = .50$ denotes an average-case scenario. We fixed the candidate pool to $k = 20$. A candidate pool of $k = 1$ reduces to the randomized mechanism of Feyisetan et al. (2020). Bold font indicates the best result from three independent trials of the worst-case scenario.

without provable privacy guarantees.

We report the performance scores in Table 1. A baseline trained on unprotected data is listed as an upper bound on the performance. All trials mimic the training of the baseline. To privatize the texts in the datasets, we use our modification with a varying candidate pool of size $k \in 1, 20$. Recall that $k = 1$ reduces our modification to the multivariate mechanisms of Feyisetan et al. (2020). Although we focus our analysis on a worst-case scenario in which the .90-quantile of words is plausibly deniable, we included test results for an average-case scenario in which only a .50-quantile of words enjoys plausible deniability.

On average, BERT bounds at $81.46\%$ when trained on sensitive text. Compared to the baseline, BERT trained on surrogate texts attains $55.45\%$ when the candidate pool is $k = 1$. By broadening the candidate pool to $k = 20$ and directing the substitution to words with matching grammatical categories, BERT trained on surrogate texts ranks at $60.11\%$. This corresponds to narrowing down the performance loss by $4.66\%$.

Contrary to our initial assumption that preserving the syntactic role of words is particularly relevant to sentiment analysis, we find evidence that accounting for syntactic information during privatization benefits a variety of downstream tasks. We conclude that linguistic guidance is a legitimate alternative perspective to previous extensions that focus on the geometric position of words in the embedding.

## 5 Conclusion

Privatizing written text is typically achieved through text-to-text privatization over the embedding space. Since text-to-text privatization scales the notion of indistinguishably of differential privacy by a distance in the geometric space of embeddings, prior studies focused on geometric properties (Feyisetan et al., 2019; Xu et al., 2020; Carvalho et al., 2021b).

Unlike prior studies on amplifying text-to-text privatization by accounting for the geometric position of words within the embedding space, we initialized a set of strategies for amplification from the perspective of grammatical properties, such as *category*, *number*, or *tense*.

By incorporating grammatical properties in the form of part-of-speech tags into text-to-text privatization, we direct the privatization step towards preserving the syntactic role of a word in a text. We experimentally demonstrated that that surrogate texts that conform to the structure of the sensitive text outperform surrogate texts that strictly rely on co-occurrences of words in the embedding space.

**Limitations.** We note that directing the substitution to candidates with matching grammatical categories incurs additional information leakage that is not accounted for by our modification. Too remedy the unaccounted information leakage, one could recast the candidate selection through the exponential mechanism (McSherry and Talwar, 2007).

## Acknowledgment

# References

Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato Recupero, and Venkatramana S Subrahmanian. 2007. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. *ICWSM*, 7:203–206.

Vincent Bindschaedler, Reza Shokri, and Carl A Gunter. 2017. Plausible deniability for privacy-preserving data synthesis. *arXiv preprint arXiv:1708.07975*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2020. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*.

Ricardo Silva Carvalho, Theodore Vasiloudis, and Oluwaseyi Feyisetan. 2021a. Brr: Preserving privacy of text data efficiently on device. *arXiv preprint arXiv:2107.07923*.

Ricardo Silva Carvalho, Theodore Vasiloudis, and Oluwaseyi Feyisetan. 2021b. Tem: High utility metric differential privacy on text. *arXiv preprint arXiv:2107.07928*.

Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of differential privacy using metrics. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 82–102. Springer.

Maximin Coavoux, Shashi Narayan, and Shay B Cohen. 2018. Privacy-preserving neural representations of text. *arXiv preprint arXiv:1808.09408*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jinshuo Dong, Aaron Roth, and Weijie J Su. 2019. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006a. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 486–503. Springer.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006b. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2018. Author obfuscation using generalised differential privacy. *arXiv preprint arXiv:1805.08866*.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *International Conference on Principles of Security and Trust*, pages 123–148. Springer, Cham.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 178–186.

Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019. Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 210–219. IEEE.

Oluwaseyi Feyisetan and Shiva Kasiviswanathan. 2021. Private release of text embedding vectors. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, pages 15–27.

Wei Jiang, Mummoorthy Murugesan, Chris Clifton, and Luo Si. 2009. t-plausibility: Semantic preserving text sanitization. In *2009 International Conference on Computational Science and Engineering*, volume 3, pages 68–75. IEEE.

Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2011. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826.

Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022. The limits of word level differential privacy. *arXiv preprint arXiv:2205.02130*.

Frank McSherry and Kunal Talwar. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Ilya Mironov. 2017. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE.

Debra A Myhill, Susan M Jones, Helen Lines, and Annabel Watson. 2012. Re-thinking grammar: The impact of embedded grammar teaching on students' writing and students' metalinguistic understanding. *Research Papers in Education*, 27(2):139–166.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331. IEEE.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Chen Qu, Weize Kong, Liu Yang, Mingyang Zhang, Michael Bendersky, and Marc Najork. 2021. Privacy-adaptive bert for natural language understanding. *arXiv preprint arXiv:2104.07504*.

David Sánchez and Montserrat Batet. 2016. C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1):148–163.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 377–390.

Congzheng Song and Vitaly Shmatikov. 2019. Auditing data provenance in text-generation models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 196–206.

Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.

Aleena Thomas, David Ifeoluwa Adelani, Ali Davody, Aditya Mogadala, and Dietrich Klakow. 2020. Investigating the impact of pre-trained word embeddings on memorization in neural networks. In *International Conference on Text, Speech, and Dialogue*, pages 273–281. Springer.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Xuan-Son Vu, Son N Tran, and Lili Jiang. 2019. dpugc: Learn differentially private representation for user generated contents. *arXiv preprint arXiv:1903.10453*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

Stanley L Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69.

Benjamin Weggenmann and Florian Kerschbaum. 2018. Syntf: Synthetic and differentially private term frequency vectors for privacy-preserving text mining. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 305–314.

Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2020. A differentially private text perturbation method using a regularized mahalanobis metric. *arXiv preprint arXiv:2010.11947*.

Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2021. On a utilitarian approach to privacy preserving text generation. *arXiv preprint arXiv:2104.11838*.

# Appendices

## A Linguistic Evaluation

Covering three levels of privacy budgets $\varepsilon$, we include the detailed linguistics analysis of the multivariate substitutions obtained from MADLIB (Feyisetan et al., 2020) in Figure A.1.

Without a constraint on syntactic roles, we cannot expect the privatization step to yield surrogate texts that conform to the structure of the sensitive texts. From the diagonal, it can be clearly seen that our grammatical constraint retains most grammatical categories across all budget budgets and all types of categories. At a low privacy budget of $\varepsilon = 5$, the preservation capability of grammatical categories is 0.4163. At a moderate privacy budget of $\varepsilon = 10$, the preservation capability bounds at 0.8145. At a high privacy budget of $\varepsilon = 25$, the advantage in the preservation capability diminishes as the perturbation probability in general decreases.
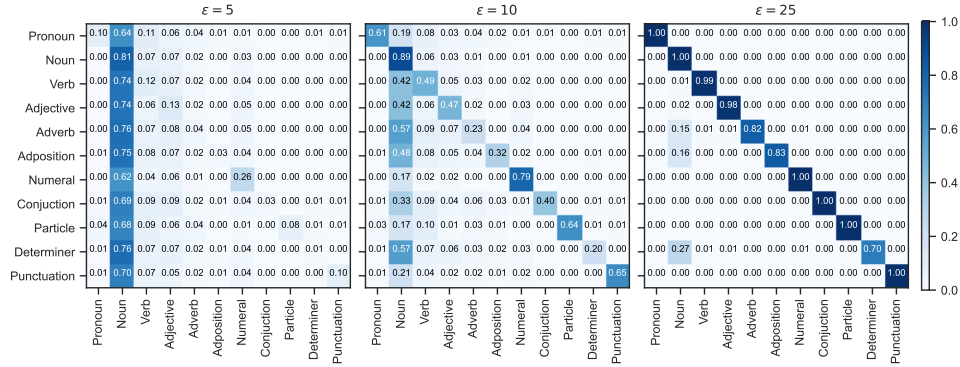


(a) MADLIB with $k = 1$
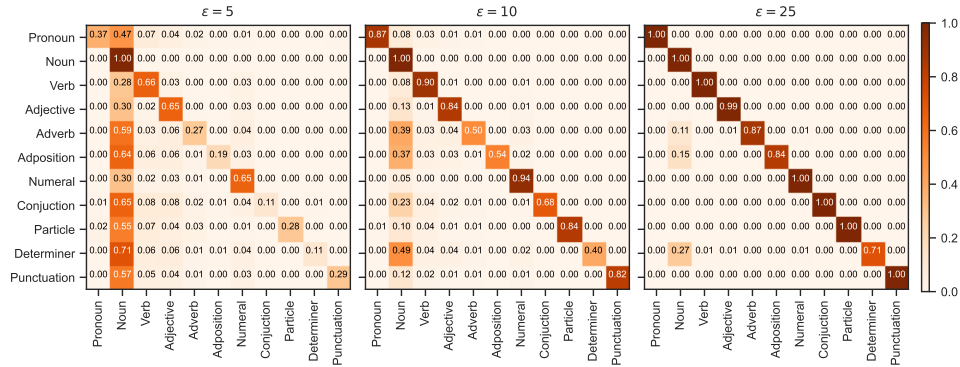


(b) MADLIB with $k = 20$

Figure A.1: Linguistics analysis with respect to the grammatical category of a sub-vocabulary after 100 times of querying a randomized mechanism. Given a candidate pool $k$ of nearest neighbors, $k = 1$ represents substitutions solely based on co-occurrences, whereas $k = 20$ represents grammatically constraint substitutions. The size of the candidate pool has been approximated by the sub-vocabulary's neighborhood.

## B Setup and Results from Membership Inference Attack

To eliminate the possibility that the performance gain is caused by mismatching privacy guarantees, we perform a *Membership Inference Attack* (MIA) introduced by Shokri et al. (2017). Given black-box access to a model, an adversary attempts to infer the presence of records from an inaccessible training corpus. We follow the experimental setup of Carvalho et al. (2021b) for our membership inference attack. To maximize the attack uncertainty, we divide the IMDb dataset into four *disjoint* partitions with an equal number of members and non-members, respectively. The target model is trained on the first partition after privatization by each mechanism, whereas the shadow model is trained on the non-privatized second

partition. The shadow model architecturally mimics the target model. We then build an attack model composed of a two-layer multi-layer perception with a hidden size of $64$ and non-linear activations. To train the attack model, we feed the logits obtained by the second and third partitions given by the shadow model, where logits from the second first partition are labeled as members and logits from the third partition are labeled as non-members. Once the attack model is trained, we feed the logits of the first partition and the fourth partition obtained by the target model, where logits from the first partition are labeled as members and logits from the fourth partition are labeled as non-members.

We measure the success rate of our membership attack using macro-averaged metrics for precision and recall. Precision captures the fraction of records for which the membership was correctly inferred. Recall captures the coverage of the membership attack. Since the baseline accuracy of the membership attack is $0.5$, we consider a randomized mechanism to be provably private if and only if it holds the attack accuracy close to that of random guessing. We report the attack accuracy as the area under the precision-recall curve. We report a non-private membership accuracy of $0.53$. Given a practical privacy budget, both mechanisms fluctuate around the $0.5$ mark averaged across three independent trials. With no clear hint, we thus conclude that the performance gain induced by a grammatical constraint cannot be attributed to a latent privacy loss.