PRIPA: A Tool for Privacy-Preserving Analytics of Linguistic Data

Jeremie Clos[1], Emma McClaughlin[1], Pepita Barnard[1], Elena Nichele[1], Dawn Knight[2], Derek McAuley[1], Svenja Adolphs[1]
[1] University of Nottingham
{jeremie.clos, emma.mcclaughlin, pepita.barnard, elena.nichele, derek.mccauley, svenja.adolphs}@nottingham.ac.uk
[2] Cardiff University
knightd5@cardiff.ac.uk

Abstract

The days of large amorphous corpora collected with armies of Web crawlers and stored indefinitely are, or should be, coming to an end. There is a wealth of hidden linguistic information which is increasingly difficult to access, hidden in personal and private data that would be unethical and technically challenging to collect using traditional methods such as Web crawling and mass surveillance of online discussion spaces. Advances in privacy regulations such as GDPR and changes in the zeitgeist bring into question the problematic ethical dimension of extracting information from unaware if not unwilling participants. Modern corpora need to adapt, be focused on testing specific hypotheses, and be respectful of the privacy of the people who generated its data. Our work focuses on using a distributed participatory approach and continuous informed consent to solve these issues, by allowing participants to voluntarily contribute their own censored personal data at a granular level. We evaluate our approach in a three-pronged manner, testing the accuracy of measurement of statistical measures of language with respect to standard corpus linguistics tools, evaluating the usability of our application with a participant involvement panel, and using the tool for a case study on health communication.

Keywords: privacy-preserving linguistics, corpus linguistics, software tools

1. Introduction

There is a wealth of hidden linguistic information which is increasingly difficult to access, hidden in personal and private data that would be unethical and technically challenging to collect using traditional methods such as Web crawling and mass surveillance of online discussion spaces. Additionally, advances in privacy regulations and changes in the zeitgeist bring into question the problematic ethical dimension of extracting such information from unaware if not unwilling participants.

Since the generation of knowledge from large amounts of empirical data is at the heart of corpus linguistics, its practitioners have long sought ways to protect the privacy of those who have generated it. However, so far the use of privacy-preserving methods has focused on post hoc processing such as automated anonymisation and de-identification. Those automated methods are severely lacking when faced with modern methods of re-identification and de-anonymisation. Non-automated methods on the other hand are not as scalable.

As a first step towards addressing this issue, we developed PRIPA[1], a software tool using a distributed participatory approach and continuous informed consent by allowing participants to stay in control of their data, and only voluntarily contribute their own censored personal data on their own terms.

We evaluate our prototype by producing a comparison of word frequencies and collocate association scores between two standard state-of-the-art systems and PRIPA, showing that PRIPA is on par with those tools for some of their common features. We produce a small scale quantitative and qualitative evaluation of the tool by users of different levels of expertise, highlighting some key challenges in the production of privacy-preserving linguistic analysis tools.

This paper is structured as follows: In section 2, we discuss the overall methodology of PRIPA: general design for continuous consent, and software architecture. In section 3 we describe our evaluation methodology. We will finally conclude with key challenges and recommendations for further development in section 4.

2. Privacy-Preserving Corpus Linguistics

Being privacy-preserving by design involves the adherence to a set of principles, described in Table 1.

Instead of collecting the data on the online discussion platform, we recruit participants who install a plugin into their Web browser. The PRIPA plugin then allows participants to enrol themselves into different experiments. Those experiments specify multiple things: the websites which will be watched, the words that will be observed, and the statistics that will be collected. In this section we will describe two key aspects of PRIPA for privacy-preserving corpus linguistics: the software architecture allowing data to be collected according to our key principles, and the user interface design allowing for the informed consent of users to be monitored at each key step of the data collection process.

2.1. Data collection process

PRIPA collects 3 types of linguistic information:

[1] https://c19comms.wp.horizon.ac.uk/pripa

P1	Participants are aware of the purpose of the experiment.
P2	Participants are aware of the parameters (web sites, words, time scale) of the data collection.
P3	The features of interest (words, statistical measurements, excerpts) are described in an intelligible way for the participants.
P4	Participants are aware of their right to anonymity.
P5	Participants can consult their data before it is shared with the researchers.
P6	Participants can decide to exclude selected results from the data that is shared with the researchers.
P7	Participants can decide to withdraw completely from a study at any time.
P8	If participants omit to remove personally identifiable information, the researchers should remove it before long-term storage of the data.

Table 1: Key design principles of PRIPA

Word frequencies Word frequencies are the raw number of occurrences for words in a specific word list, defined as part of the experiment. The word list is specific to the experiment and as such a participant that does not want to share a specific word frequency needs to withdraw from the experiment in order to preserve the integrity of the data without violating their privacy.

Collocates Collocates are pairs of words of interest (defined in a word list as part of the experiment) along with their strength of association, given a pre-specified window of words. The list of word pairs is specific to the experiment, and, like word frequencies, a participant that does not want to share a specific word pair needs to withdraw from the experiment.

Concordance lines Concordance lines are lines of text showing the context for a particular word, along with the source of that line. The size of the context is specified in the experiment, and the participant can review the list of concordance lines and exclude the ones they do not want to share.

2.2. Architecture and design

PRIPA is built in a client-server architecture, where the server hosts experiments which are defined in a specific format using JSON syntax[2]. The format is described in Figure 1.

[2]a lightweight data-interchange format documented at `https://www.json.org`

Client-side data collection The client of the application sits in a plug-in for Chromium-based Web browsers (e.g., Google Chrome, Microsoft Edge). We make use of the JavaScript regular expression engine in order to process word lists which are downloaded from the experiment server. Once the user selects an experiment they would like to take part in and accept the disclaimers regarding the way their data will be processed and how they can access/modify/remove it, the PRIPA extension downloads an experiment specification file and watches for the opening or closing of specific websites (depending on the specification of the experiment). When such action (open/close) is triggered, PRIPA attempts to extract the core of the webpage by ignoring banner ads and other informational noise, and runs the analysis based on the word lists provided in the experiment file. The data is stored in the Web browser itself, never leaving the participant's device until they have decided to share their data with the researcher.

Server-side aggregation The statistical measures collected by PRIPA can be aggregated after the fact. Word frequency can be aggregated with a simple sum, and collocate strength is measured using pointwise mutual information (Bouma, 2009) which can be aggregated using simple frequency measures and information about document length. Considering that the Pointwise Mutual Information of two words w_1 and w_2 in a document d can be computed as $PMI(w_1, w_2, d) = log(\frac{P_d(w_1,w_2)}{P_d(w_1)\cdot P_d(w_2)})$ and that $P_d(w) = \frac{\text{freq}(w)}{|d|}$ where $|d|$ is the length of document d, we only need to communicate individual and joint word frequencies as well as length of the web pages in order to aggregate that measure over all participants.

Consent monitoring In order for PRIPA to adhere to the principles laid out in the beginning of the project, consent of the participants needs to be monitored at regular intervals when user data is manipulated. This is done at the following stages:

1. When enrolling in an experiment.
2. When activating the application, which will watch the browser for a pre-defined set of websites.
3. When reviewing concordance lines, where participants can choose to exclude specific data points they deem personal.
4. When submitting results to the researchers, users can choose to instead stop the experiment and delete their data.

3. Evaluation and results

We evaluated our system in a three-pronged approach:

Accuracy of word counting As pointed out by Anthony (Anthony, 2013), corpus linguistics applications often differs in their measurements due to having different standards in the way they process text. For example, some software would break "We'll" into two word

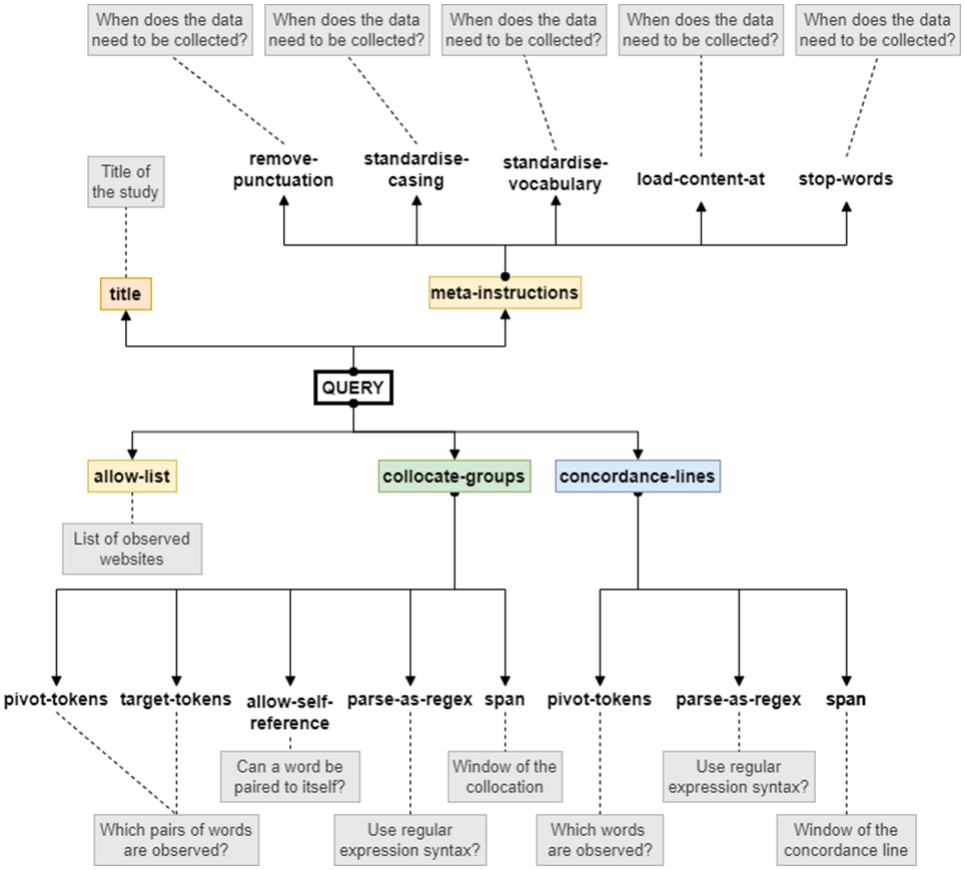

Figure 1: A graphical representation of the experiment file format

tokens, while some would keep it as a singular word token. Small variations, repeated over large corpora, can lead to vastly different linguistic measurements and affect interpretation. As such, we calibrated our measurement so that it is close to standard tools such as AntConc (Anthony, 2005) and LancsBox (Brezina et al., 2018). We designed a set of test web pages with minimal noise and hosted them on a university website, analysing them both offline with AntConc and LancsBox and online through PRIPA.

In Table 2 we show a comparison of frequencies of single words when running a study on modal verbs on a pre-selected corpus. We can observe that counts mostly match. A visual inspection determined that readings which were not matching were due to tokenisation differences when handling punctuation and apostrophes.

In Figure 2 we show a comparative histogram of the differences between measurements of collocation strength between PRIPA, AntConc, and LancsBox on an experiment measuring collocation strength between modal verbs and pronouns. We can see from this graph that out of our samples, most measurements fell within $[0, 0.2[$ of LancsBox and $[0, 0.3[$ of AntConc. A visual inspection showed that the readings that did not match were due to tokenisation differences, like with standard term frequencies.

	PRIPA	**AntConc**	**LancsBox**
may	33	34	33
might	16	16	15
must	15	15	15
should	29	29	29
would	39	39	39
could	30	30	30
can	93	98	91
will	126	126	125
shall	0	0	0
ought to	0	0	0
total	**381**	**353**	**377**

Table 2: Comparative analysis of PRIPA, AntConc and LancsBox on term frequency of modal verbs on a selected corpus (coloured cells indicate identical counts).

Usability study of the software Since the participants are rarely researchers themselves, it is important that the software produced is adapted for laypeople and general non-experts. To test this, we ran a usability study with a small participant involvement panel of 6 people. The quantitative results of the study are summarised in Table 3.

We can see from the data that most participants felt confident in using PRIPA, but had a difficult time un-

	Question	Median
Q1	I think that I would like to use this extension frequently.	3
Q2	I found it difficult to understand what the extension does.	4
Q3	I found it easy to set up and run the project in the extension.	4.5
Q4	I think that I would need the support of a technical person to be able to use this extension	1.5
Q5	I found the analyses and results were clearly explained in the extension	2.5
Q6	I felt very confident using this extension	4
Q7	I would imagine that most people would learn to use this extension very quickly	3.5
Q8	I am concerned about the privacy and security of my personal data (i.e., who may be able to access my personal information and how it is protected) when using the extension	2.5

Table 3: Usability study of 6 participants - Median value of the Likert data (1 = strongly disagree, 5 = strongly agree).

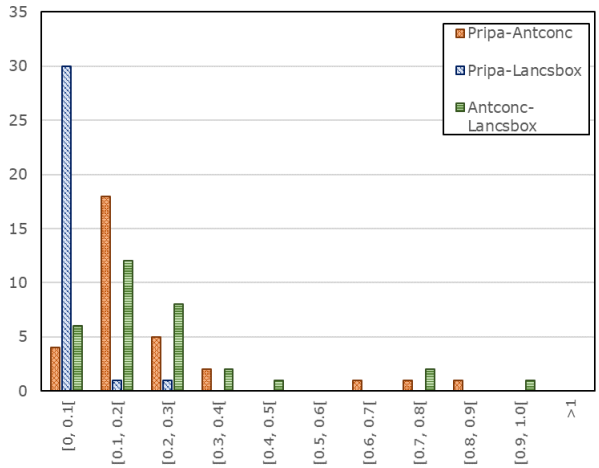

Figure 2: Histogram of differences between PRIPA, AntConc and LancsBox in calculating strength of association between collocates on a sample corpus. Difference between LancsBox and AntConc also provided for baseline.

derstanding the goal of the application. This raises the issue of the importance of a clear user interface and shows that PRIPA can be improved with respect to its first key design principle: participants are aware of the purpose of the experiment. Additionally, we note from the quantitative data reported in Table 3 as well as from qualitative data collected during the same survey that participants were concerned about the privacy of their data. This is partly explained due to the permission model of Chrome-based extensions, which require asking the participants access to their entire browsing experience and them trusting that we will filter only the websites and the data that is stated in the experiment details. Recent updates in the Chrome permission models allow for fine-grained website permissions at runtime and therefore that problem will soon be patched out of PRIPA.

In-depth study of health communication In order to evaluate our tool in the field, we ran a study of health communication from the British government during the COVID-19 pandemic. We defined a list of websites of interest based on an empirical study of the most visited news websites in the UK, on which to carry out a pilot study to examine modality markers surrounding key terms from health messages (e.g., ”mask”, ”vaccine”, ”lockdown”, and more). Results from our study shows that PRIPA allows us to access language data from the perspective of the people consuming it. However, it also highlighted a weakness of PRIPA in that when dealing with communication-oriented web applications such as Twitter direct messages or Facebook Messenger, it cannot differentiate between language being produced by the participant and language being consumed. Such information would be useful from a linguistic perspective and will therefore be added in future versions of PRIPA.

4. Conclusion

In this paper we present PRIPA, an early prototype of a new family of corpus linguistics tools which allow for collecting personal data in a privacy-preserving way. PRIPA is an early prototype and therefore a work in progress, but its development raised a number of questions and helped us uncover a set of research directions and good practices for a more trustworthy privacy-preserving type of linguistic analysis.

5. References

Anthony, L. (2005). Antconc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *IPCC 2005. Proceedings. International Professional Communication Conference, 2005.*, pages 729–737. IEEE.

Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, 30(2):141–161.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

Brezina, V., Timperley, M., and McEnery, A. (2018). # lancsbox v. 4. x.

Rayson, P. (2009). Wmatrix: a web-based corpus processing environment.