# Assessing Purpose-Extraction for Automated Corpora Annotations in IoT Privacy Policies

Vincent Miller
*Columbus State University*
*miller_vincent@columbusstate.edu*

Jesus R. Rijo Candelario
*Mercer University*
*Jesus.Rafael.Rijo@live.mercer.edu*

Lydia Ray
*Columbus State University*
*ray_lydia@columbusstate.edu*

Alfredo J. Perez
*University of Nebraska at Omaha*
*alfredoperez@unomaha.edu*

***Abstract*—Privacy policies contain important information regarding the collection and use of user's data. As Internet of Things (IoT) devices have become popular during the last years, these policies have become important to protect IoT users from unwanted use of private data collected through them. However, IoT policies tend to be long thus discouraging users to read them. In this paper, we seek to create an automated and annotated corpus for IoT privacy policies through the use of natural language processing techniques. Our method extracts the purpose from privacy policies and allow users to quickly find the important information relevant to their data collection/use.**

## 1. Introduction

Privacy Policies are legal documents that disclose how a party collects, uses, manages, and shares information on a client or user's data. The type of information collected includes Private Identifiable Information (PII) such as the user's name, date of birth, physical address, email address, telephone number, personal characteristics, metadata, and user-related data (such as sensor data). Many users do not read these policies simply because of their length and complexity. The average user spends 8 minutes reading a single privacy policy of 2,000 words [1]. In addition, the length of privacy policies has dramatically increased in the last years and has now nearly doubled in length [2] which further deters users from reading them. While privacy laws such as the European Union's General Data Protection Regulation (GDPR) require Internet services to provide users with easy-to-read privacy policies, the mandate does not reduce the complexity nor the time it takes to read them [3]. As the Internet-of-Things (IoT) becomes more prevalent in our lives, many privacy policies are not read, making users to agree to use IoT devices which may expose not only their data, but aspects of their lives considered private.

The complexity and time required to read privacy policies have led to research automated privacy policy reading tools to find information relevant to users [4] and minimize the effort to comprehend them. Training many of these automated tools require manual annotated datasets which can take up to 72 minutes per privacy policy [5] to accomplish. Moreover, IoT privacy policies are difficult to find as many times they are embedded into a company's general privacy policy, or they may not be included at all [6]. As more IoT devices are connected to homes or worn by people (and even pets [7]), the understanding of IoT privacy policies has become an important topic to research. In this work, we present a study of IoT privacy policies. Our contributions are as follows:

- We present an algorithm to crawl and find IoT privacy policies in the Internet
- We assess a purpose extraction approach for automatically creating a corpus containing annotations for IoT privacy policies
- We leverage the use of Natural Language Processing (NLP) to find the meaning of sentences in IoT policies and classify them into categories based on the sentence's purpose, allowing users to quickly find relevant information

The rest of this paper is organized as follows. In Section 2, we describe related works. Section 3 presents our methodology to crawl and conduct purpose extraction for IoT privacy policies. Section 4 presents our results. Finally, in section 5 we conclude this work.

## 2. Related Work

The creation of annotated datasets of privacy policies for the Internet has remained a challenge during the last decade. However there have been approaches to solve this problem. For example, the OPP-115 dataset [5], created under the Usable Privacy Policy Project [8], stands as one of the most popular datasets of web privacy policies, as other researchers have used this dataset as a training set for machine learning solutions. This dataset contains 115 privacy policies manually annotated by skilled annotators. A drawback of the OPP-115 dataset is that it has not been updated since 2017, indicating that the dataset may be obsolete with recent legislative changes (e.g., GDPR). The Usable Privacy Policy Project also created the MAPS

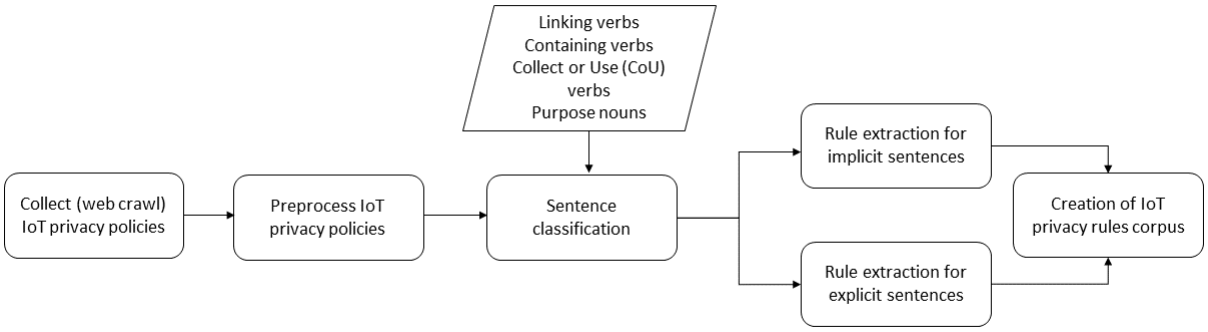


Figure 1: Our proposed approach to create an annotated IoT privacy policy corpus

Policies dataset [9], a dataset of privacy policies of over one million apps from the Google Play Store app. Despite the tremendous number of privacy policies when compared to the OPP-115 dataset, MAPS cannot be applied to supervised machine learning techniques, as this dataset lacks annotations.

Other researchers have focused on evaluating properties of privacy policies. For example, Kuznetsov et al. [10] curated a dataset of IoT privacy policies from several service providers and manufacturers, including well-known companies such as Amazon. Kuznetsov et al. supplemented their dataset with various statistical and semantic analysis. For example, their work highlights the distribution of HTML tags across privacy policies. However, their dataset does not preserve the connection between the privacy policies to their corresponding sources.

Other approaches have experimented with frameworks that step beyond the manual annotation approach of datasets like the OPP-115 take. For instance, PI-Extract [11] is a framework created to extract the action and data object from privacy policies, labeling results into one of four categories: *Collect, Not Collect, Share,* and *Not Share*. However, PI-Extract still relies on manually annotated datasets. Harkous et al. [12] developed an automated question-answering chatbot for privacy policies that uses machine learning models trained on a dataset with over 100 thousand privacy policies, though these privacy policies are not publicly available. Similarly, Yang et al. [13] proposed a framework called *PurExt* to extract the purpose from each sentence in a privacy policy for IoT devices. Yang et al. designed a semantic and syntactic analysis to classify sentences and extract the following elements from each sentence: Actors, actions, data objects, and purpose. Nonetheless, PurExt's model depends on a small dataset made of manual annotations. Recently, the PrivaSeer project has created a search engine for web privacy policies [14].

In this work, we focused on providing an approach to help future researchers assess not only our dataset of over 2000 IoT privacy policies, but any subsets of IoT privacy policies. Moreover, we assessed the readability of IoT privacy policies by devising a readability assessment which could be used to filter IoT privacy policies based parameters outlined by the readability assessments. Finally, our work explores the effectiveness and practicality of the *PurExt* [13] by testing this framework on a large dataset of IoT privacy policies.

## 3. Methods

In this section we describe our approach to automatically annotate IoT privacy policies. We divided this problem into four primary tasks as follows (figure 1):

- Implementing and running a web-crawler scheme to acquire IoT privacy policies.
- Performing a readability assessment across the IoT privacy policies
- Implementing and assessing our implementation of the PurExt approach [13]
- Generating a final corpus of annotated privacy policies

### 3.1. Crawling and Prepropcessing of IoT Privacy Policies

We first compiled a list of IoT company names based on the public database provided by IoT ONE (a U.S.-based company focused on IoT solutions for the Asian market) [15]. IoT ONE's database lists over 3000 unique IoT companies. After removing duplicate entries, we retrieved 3016 unique names. By using a Python library to perform searches through Google [16], we attempted to retrieve the IoT privacy policies as HTML files. The algorithm is shown in figure 2. However, we could not retrieve valid privacy policies from all the companies, as certain companies lacked responsive websites or did not have their policies as HTML files. With these requirements, we obtained 2854 unique HTML files.

After collecting the IoT privacy policies in HTML format, our preliminary preprocessing step was implemented to convert the collected HTML privacy policies into usable plaintext files. Recognizing that IoT privacy policies do not

**Algorithm** Automated extraction of IoT privacy policies

**Input:** A list *lst* with one or more names of IoT companies.
**Output:** Returns at most $|lst|$ raw and cleaned HTML files.

```
for name in lst do
    links.append(get-link-web(name))
end for
i ← 0
        ▷ For every 100 requests, the algorithm waits for one minute.
for link in links do
    get-file-web(link)
    name-file(name.txt)
    i ← i + 1
    if i (mod 100) ≡ 0 then
        sleep(60.0)
    end if
end for
```

Figure 2: Algorithm to crawl IoT privacy policies

**Algorithm** Prepocessing of raw HTML IoT privacy policies

**Input:** A IoT privacy policy as an raw HTML.
**Output:** The same IoT privacy policy as a plaintext file.

```
        ▷ Defining HTML files to extract their content.
tags ← {'h1','h2','h3','h4','h5','h6','h7','p','li'}
policy ← open-file(path)
new − file ← make-file()
                ▷ Preliminary preprocessing.
for element in find-all(tags) do
    if element.text ≠ '' ∨ element.text ≠ '\n' then
        x ← element.text
        x ← x.strip()
        x ← x.replace('\n',' ')
        new − file.append(x)
    end if
end for
```

Figure 3: Preprocessing of raw HTML privacy policies

share a similar HTML structure, we chose to keep text contained within common HTML elements, such as the paragraph tag and list element tag a shown in figure 3.

### 3.2. Readability Assessments

Given the nature of our methodology, we considered various automated assessments to measure the quality of each privacy policy in our dataset. We concluded that devising a readability profile for each privacy policy could provide enough insight. For each privacy policy, we created a readability profile containing 12 grammatical features (i.e., number of words, number of conjunctions). The remaining parameters outline scores for various common readability tests which are formulae that numerically estimate the readability of a given text [17].

| Parameters | Maximum | Average | Median |
|---|---|---|---|
| Kincaid Grade Level | 32.180531 | 10.2575044 | 10.4169761 |
| Automated Readability Index (ARI) | 37.6339495 | 10.1365877 | 10.4270026 |
| Coleman-Liau Index | 47.0409318 | 13.0387196 | 13.6803752 |
| Flesch Reading Ease Test | 199.340401 | 43.8386042 | 40.952753 |
| Gunning Fog Index | 28.4712708 | 14.8585273 | 14.8806059 |
| Lix Readability Formula | 74.4731724 | 47.1658984 | 47.3566076 |
| SMOG Readability Formula | 17.8513764 | 12.1044289 | 12.0782209 |
| Rix Readability Test | 10.3157895 | 3.86 | 3.97477395 |
| Dale-Chall Index | 22.203195 | 11.3882384 | 11.7492542 |
| Characters per Word | 12.1828571 | 5.35029532 | 5.53741193 |
| Syllables per Word | 3.76106195 | 1.76921819 | 1.82219135 |
| Words per Sentence | 69.9347826 | 11.2757009 | 11.5515849 |
| Number of Characters | 424556 | 3844 | 5821.53208 |
| Number of Syllables | 222458 | 2152 | 3256.60468 |
| Number of Words | 17857 | 595 | 681.085246 |
| Number of Sentences | 11797 | 187 | 267.587354 |
| Number of Be Verbs | 1040 | 38 | 54.9081967 |
| Number of Auxilliary Verbs | 762 | 36 | 50.0276347 |
| Number of Conjunctions | 2234 | 96 | 139.798595 |
| Number of Pronouns | 4436 | 201 | 274.118501 |
| Number of Prepositions | 5268 | 244 | 342.333021 |

Figure 4: Readability assessments of collected IoT privacy policies

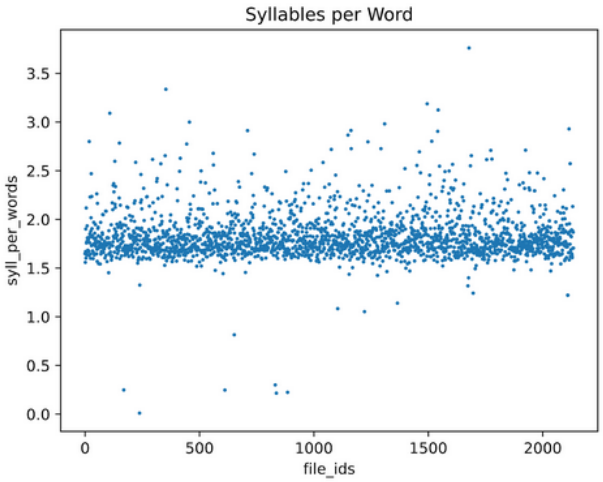


Figure 5: Scatterplot for syllables per word across the dataset. X-axis represents the policy/file id

With the help of an external Python library for readability scores and syntactical elements, we produced a readability profile for our collected privacy policies, storing these profiles as plain text files [18]. After measuring these parameters, we calculated the average, median, and maximum of each parameter as presented in figure 4. Additionally, we visualized our results using scatterplots for each parameter, as illustrated in figure 5. Based on our results, we decided to exclude any privacy policies that had less than 100 words for the purpose extraction phase, resulting in a shrunk dataset of 2134 unique IoT privacy policies.

$$Sentence_{explicit} = Noun_{purpose} + Verb_{link} \vee Verb_{contain} + Purpose$$

$$Sentence_{implicit} = Subject + Verb_{CoU} + Data\ Object + Purpose$$

$$Purpose\text{-}Aware\ Rule = \{Action, Data\ Object, Purpose\}$$

Figure 6: Syntactic structures of purpose-aware rules, explicit sentences, and implicit sentences

### 3.3. Sentence Classification

We decided create our own implementation in Python of the *PurExt* framework as described by Yang et al. [13]. In our Python implementation, we used the spaCy library. The spaCy library is an open source, industrial strength Natural Language Processing library (NLP) [19] which allowed us to easily create a pipeline to tokenize, create part-of-speech (POS) tags, dependency parse, and named entity recognition (NER) labels for each word in the dataset. Tokenization splits a text into its smallest units for processing, and stores the POS, dependency and NER. The POS tag provides lexical information that is required for dependency parsing and the NER. Dependency parsing looks at phrases to assign the relationship between words. It assigns labels to each token; the action verb (root), noun subject (nsubj), data object (dobj), etc. labels to each token and builds a tree that can be traversed from the root. Lastly, the NER tags relationships to real world objects. For example, the phrase "United States of America" would be labeled as a geopolitical location. Moreover, the NER is capable of tagging tokens entities as names, organizations, and dates.

*PurExt* classifies sentences into three categories including *explicit sentences*, *implicit sentences*, and *other sentences*. Explicit and implicit sentences involve sentences related to the data collection or use in privacy policies. Explicit sentences are syntactically-based, whereas implicit sentences are semantically-based. Nevertheless, both categories have a syntactic structure (shown in figure 6) that *PurExt* considers for rule extraction. The *other sentences* category includes sentences related contact information, terms of service, and policy updates, among others.

The sentence classification begins at the root node of the dependency tree. The explicit sentence starts by checking if the root is a linking verb or a containing verb. A linking verb connects the subject to the predicate of a sentence. Common linking verbs are forms of "to be" and the five senses, "smell, taste, touch, sight, hearing." A containing verb links the subject and predicate through the meaning of containing, such as "containing" and "including." Next, the noun subject is checked against a list of purpose nouns provided by the *PurExt* authors. Afterwards, the framework checks if the noun subject is modified by a complement verb involving data collection or use (CoU).

The *PurExt* authors also provided a list of CoU verbs. If the sentence meets these conditions, the framework will classify it as an explicit sentence. Now if the root of the sentence is a CoU verb, it has a chance of being implicit. The last condition for implicit is to check if the sentence contains at least one Data Object. These two conditions will classify sentences as implicit. Otherwise sentences are classified as other sentences. Our list of words used in our implementation is shown in table 1.

Table 1: Part of speech words in sentence classification

| Part of speech | Used words |
|---|---|
| Purpose nouns | purpose, reason, intention, goal, motivation, way |
| Collection or use verbs | access, check, collect, disclose, gather, keep, know, obtain, process, provide, receive, request, retain, save, share, store, transfer, update, use, utilize |
| Linking verbs | act, am, appear, are, be, became, become, can be, come, could be, could have come, did, do, does, fall, feel, fell, felt, get, go, got, grew, grow, had, had become, had been, had seemed, has, has appeared, has become, has been, has seemed, have, have appeared, have become, have been, have seemed, indicate, is, is being, is getting, keep, look, may be, might be, might have been, must, prove, remain, seem, shall be, shall have been, should be, should have appeared, should have been, smell, sound, stay, taste, turn, was, was being, went, were, will be, will become, will have become, will have been, will seem, would be |
| Containing verbs | contain, include |

### 3.4. Rule Extraction and Corpus Creation

Lastly, *PurExt* extracts privacy rules from the sentences. The original rule structure involved the extraction of the actor, action, data object, and purpose from the explicit and implicit sentences. We chose to omit the actor element, as *PurExt* found only 172 actors from their 600 sentences, and 141 of those actors were the word "we". The actors are often the company and the user. While performing sentence classification, we realized the rule extraction can be done side by side, as the classification checks are also the required extractions. The action is extracted by examining the subject to identify the token matching CoU verbs. At the same time, Data Objects can be acquired by checking for any tokens annotated as "dobj." The purpose can be extracted by examining the predicate and later acquired from the root of the sentence. Because we noticed we could perform extraction alongside classification, we opted to process our *other sentences* and show their rule extraction.

We organized and created our final corpus using the Pandas open source library which is a high performance tool with convenient data structures and tools for data analysis [20]. Each IoT privacy policy received its own Pandas dataframe (a table). Each entry contains the sentence type, the original sentence, action, data object, and the purpose. This allowed us to easily examine our data to find and correct any issues. We exported each dataframe as a comma separated values (CSV) file. Furthermore, the file's name represents the company the data was extracted from. Our

corpus has a total of 2134 rule extracted files, one for each processed IoT privacy policy.

## 4. Results

We present our results for the explicit and implicit extraction in table 2. We had surprisingly low explicit classification, even though the *PurExt* authors stated that their explicit classification approach is syntactically based. When we tested the database provided by Yang et al. with our implementation, it classified 112 out of the 120 explicit sentences correctly. These eight sentences follow a simple sentence structure and failed the *PurExt's* check to verify if the subject is modified by a complement containing at least one CoU verb. These sentences then get classified as "other" sentences, instead of explicit sentences. When we observed our data and examined potential reasons behind the low explicit classification, we inferred that privacy policies appear to have complex sentence structures. There are 8 types of complete sentences [**?** ]. Therefore, we determined that the *PurExt* framework must perform additional checks regarding the syntactic structure of sentences to handle more complex sentences.

We also had low implicit classification results. We expected low results once we tried to follow PurExt authors' approach of retraining the NER model. When we contacted the *PurExt*, they did not share their annotated dataset for retraining of the NER model, which made replicating their work impossible. The implicit classification, although it has some syntactic structure, is mainly semantically based. Due to the legal definitions of words in privacy policies, the NER labels would need retraining to correctly label them. Words like "advertisers" would need to be labeled as an entity but do not get the correct label without retraining. Because of the lack of access to skilled annotators, we were unable to retrain the NER model to improve these results. We performed purpose extraction in parallel with sentence classification, our action, data object, and purpose results all showing 100% accuracy for the explicit and implicit sentences.

We also observed that library updates may have also played a key role in our results. In their work Yang et al.'s implementation, they used version 2.x of the spaCy library, whereas we used version 3.3 in our implementation. Our dataset is publicly available in GitHub [21].

Table 2: Results from our approach

| | |
|---|---|
| Total explicit sentences | 65 |
| Total implicit Sentences | 9948 |
| Average explicits per policy | 0.03 |
| Average implicits per policy | 4.66 |
| Action extraction accuracy | 100% |
| Data extraction accuracy | 100% |
| Purpose extraction accuracy | 100% |

## 5. Conclusion and Future Work

Privacy policies stand as a core component toward helping users understand the primary implications of using any IoT device. Given the length and complexity of most IoT privacy policies, researchers will continue working towards making privacy policies more readable for most users.

In this work we created a web crawling framework for IoT privacy policies and we implemented a purpose extraction tool based on NLP. In order to assess the quality of our implemented natural language processing framework, we curated a publicly available dataset of IoT privacy policies using our web crawling and purpose extraction framework. In future works, we plan to use or create a dataset with manual annotations to train the NER model of our purpose extraction implementation. Likewise, we plan to explore the syntactic structure to find ways to handle more complex sentences.

## Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1950416. Vicent Miller and Jesus Rijo would like thank Dr. Lydia Ray, Dr. Alfredo Perez, and Dr. Yesem Peker for their mentorship during the Research Experience for Undergraduates program at Columbus State University.

## References

[1] A. M. McDonald and L. F. Cranor, "The cost of reading privacy policies," *Isjlp*, vol. 4, p. 543, 2008.

[2] M. Lewis, L. He, and L. Zettlemoyer, "Joint a* ccg parsing and semantic role labelling," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1444–1454.

[3] D. Basin, S. Debois, and T. Hildebrandt, "On purpose and by necessity: compliance under the gdpr," in *International Conference on Financial Cryptography and Data Security*. Springer, 2018, pp. 20–37.

[4] F. Liu, S. Wilson, P. Story, S. Zimmeck, and N. Sadeh, "Towards automatic classification of privacy policy text," *School of Computer Science Carnegie Mellon University*, 2018.

[5] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. G. Leon, M. S. Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell *et al.*, "The creation and analysis of a website privacy policy corpus," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1330–1340.

[6] A. J. Perez, S. Zeadally, and J. Cochran, "A review and an empirical analysis of privacy policy and notices for consumer internet of things," *Security and Privacy*, vol. 1, no. 3, p. e15, 2018.

[7] A. J. Perez and S. Zeadally, "Recent advances in wearable sensing technologies," *Sensors*, vol. 21, no. 20, p. 6828, 2021.

[8] N. Sadeh, A. Acquisti, T. D. Breaux, L. F. Cranor, A. M. McDonald, J. R. Reidenberg, N. A. Smith, F. Liu, N. C. Russell, F. Schaub *et al.*, "The usable privacy policy project," in *Technical report, Technical Report, CMU-ISR-13-119*. Carnegie Mellon University, 2013.

[9] S. Zimmeck, P. Story, D. Smullen, A. Ravichander, Z. Wang, J. R. Reidenberg, N. C. Russell, and N. Sadeh, "Maps: Scaling privacy compliance analysis to a million apps," *Proc. Priv. Enhancing Tech.*, vol. 2019, p. 66, 2019.

[10] M. Kuznetsov, E. Novikova, I. Kotenko, and E. Doynikova, "Privacy policies of iot devices: Collection and analysis," *Sensors*, vol. 22, no. 5, p. 1838, 2022.

[11] D. Bui, K. G. Shin, J.-M. Choi, and J. Shin, "Automated extraction and presentation of data practices in privacy policies," *Proceedings on Privacy Enhancing Technologies*, vol. 2021, no. 2, pp. 88–110, 2021.

[12] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer, "Polisis: Automated analysis and presentation of privacy policies using deep learning," in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 531–548.

[13] L. Yang, X. Chen, Y. Luo, X. Lan, and L. Chen, "Purext: Automated extraction of the purpose-aware rule from the natural language privacy policy in iot," *Security and Communication Networks*, vol. 2021, 2021.

[14] M. Srinath, S. N. Sundareswara, C. L. Giles, and S. Wilson, "Privaseer: A privacy policy search engine," in *International Conference on Web Engineering*. Springer, 2021, pp. 286–301.

[15] About iot one - leading digitalization consultancy. [Online]. Available: https://www.iotone.com/suppliers

[16] googlesearch-python 1.1.0. [Online]. Available: https://pypi.org/project/googlesearch-python

[17] S. Tonelli, K. M. Tran, and E. Pianta, "Making readability indices readable," in *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, 2012, pp. 40–48.

[18] readability 0.3.1. [Online]. Available: https://pypi.org/project/readability/

[19] Spacy. industrial-strength natural language processing in python. [Online]. Available: https://spacy.io/

[20] Pandas. [Online]. Available: https://pandas.pydata.org/

[21] 2022 reu@csu: Assessing purpose-extraction for automated corpora annotations. [Online]. Available: https://github.com/jesusrrc/reu_csu_2022