AutoRPT: A Tool for Bootstrapping Prosodic Annotation

Seth Heiney, Thomas Hicks, Sally Little, Fernanda Lourenco, Kai Retana, Eliana Stevens, Jonathan Howell

Montclair State University

1 Normal Ave, Montclair, New Jersey, USA

sheiney00@gmail.com, hicksthomas465@gmail.com

littlek6@montclair.edu, lourencodasf1@montclair.edu, retanamontek1@montclair.edu, stevense4@montclair.edu, howellj@montclair.edu

Abstract

Automated Rapid Prosody Transcription (AutoRPT) is a tool for bootstrapping manual annotation of prosodic events in either corpora or standalone audio files using the Rapid Prosody Transcription (RPT) scheme. It functions by utilizing two Long-Short Term Memory (LSTM) models, trained on measures of pitch/F0 and intensity. In addition to discrete, slightly over-generated predictions of prominence and boundary, AutoRPT produces continuous predictions between 0 and 1, similar to crowd-sourced RPT annotations averaged over listeners. Marginal predictions above a given threshold are also indicated discretely by question marks, as in the PoLaR Annotation Guidelines. Annotators achieved a statistically significant increase in annotation speed by modifying AutoRPT-generated annotations compared to creating annotations without assistance. In contrast with older tools such as AuToBI (Rosenberg, 2010), AutoRPT generates more theory-agnostic annotations which can support the work of non-expert annotators, and which we expect will offer greater flexibility in the prosodic annotation of other English language varieties. The source code, documentation, and installation instructions for the tool can be found at https://github.com/Howell-Prosody-Lab/AutoRPT.

Keywords: RPT, Annotation, Speech

1. Introduction

We present a method for automated annotation of prosodic prominence and boundary. While attention to prosody continues to increase in speech sciences, the relative paucity of and lack of diversity among prosodically annotated corpora remains a challenge (Rosenberg, 2018).

Our approach takes inspiration from the AuToBI tool (Rosenberg, 2010) for classification of prosodic events in Mainstream US English (MUSE) using the Tones and Breaks Indices (ToBI) standard (Silverman et al., 1992). Rather than committing to the ToBI standard and to a specific variety of English, however, we make use of the coarser, more theory- and language variety-agnostic Rapid Prosody Transcription (RPT) method (Cole et al., 2019; see also the PoLaR framework of Ahn, Veilleux, and Shattuck-Hufnagel 2019). Building upon a previously developed Recurrent Neural Network-based AutoRPT, development of the tool is part of a larger project involving collection and annotation of African American English and Latine English conversational speech. RPT is a more appropriate standard for uncovering the prosodic inventory of these under-documented varieties and requires little training for native speaker annotators. While the ToBI remains an important and useful annotation scheme, it is usually impractical for annotation of large datasets or by untrained annotators. RPT is a simpler scheme, compatible with ToBI, which has been used successfully for crowdsourcing. Our tool facilitates and speeds the use of RPT by suggesting annotations.

Other related work includes Biron et al. (2021), who primarily used speech rate and silence to detect boundary. We do not look at speech rate in our research, though silence is being investigated to enhance the tool in the future.

The tool is written in Python and takes as input pairs of WAV audio files and Praat TextGrid files. The TextGrid file should be aligned with the audio and contain at least two tiers: a word-level interval tier and a phoneme-level interval tier (using, for example, the Montreal Forced Aligner; see McAuliffe et al., 2017). TextGrids in the training data include word-level intervals annotated for prominence and boundary. Acoustic measures–currently F0 and intensity, with plans to integrate energy and spectral tilt–are extracted using Praat, via the Parselmouth Python library (Jadoul, Thompson, and de Boer, 2018). F0 and intensity measures include maximum, mean, duration, minimum, standard deviation from the mean, and speaker-normalized standard deviation. Part-of-Speech is also included as a feature. These values are fed into a Long-Short Term Memory (LSTM) Multi-Class Classifier model for prediction and output a prosodic structure TextGrid tier as well as a CSV file.

We trained the tool on a prosodically annotated subset of the Boston University Radio News Corpus (BURNC) (Ostendorf et al., 1996), featuring roughly 2 hours of annotated audio data from 6 speakers and divided into 80-20 training/test sets. Annotations for pitch accent and

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5640–5648
11–16 May 2026.

phrase/boundary tone are collapsed into prominence and boundary, respectively. Preliminary models, based solely on pitch measures or on intensity measures, alongside Part-of-Speech, achieve lower accuracy when compared with AuToBI (83% and 93% for prominence and boundary), although accuracy of AutoRPT is already sufficient to facilitate manual annotation (cf. Escudero et al., 2014) and simpler to learn. We anticipate improved performance on future iterations, with the inclusion of additional acoustic measures (e.g. duration, energy and spectral tilt) sensitive to the nuclei of stressed syllables (already identified in the phoneme-level transcription).

2. Data and Pre-Processing

2.1 Data

The data used to train the models comes from the Boston University Radio News Corpus. This corpus was chosen as it is one of the few prosodically annotated corpora available. Annotations of this corpus were made using the ToBI annotation scheme. We utilized the subset of data that was annotated using ToBI, consisting of six speakers and roughly two hours of audio. The corpus consists of WAV files and TXT files containing the transcripts. Annotations are included as TON files.

2.2 Pre-Processing

In-house tools were created to facilitate the transfer of ToBI annotations into RPT annotations. The Montreal Forced Aligner (MFA) was used to ensure alignment of the words and phones in the transcript to their audio equivalent in the WAV file. The ToBI scheme for Mainstream American English (MAE_ToBI) (Beckman and Elam, 1994) provides for pitch accents and boundary tones, which map neatly to prominence and boundary in the RPT annotation scheme. Unlike MAE_ToBI, however, the RPT scheme does not distinguish different classes of pitch accent and boundary tones, although a researcher using MAE_ToBI would nonetheless be able to use the broad categories of prominence and boundary to do so. Once the TON file annotations were converted into RPT labels, the timestamp included was then compared to each interval in the MFA-produced TextGrid file. This associated each prosodic marking with the word in which it appeared. This then allowed for a proper file setup to train our model using the overall structure outlined in the next section.

3. Structure

3.1 Extraction

Classes named *extraction* are utilized for both F0 and intensity respectively. These classes contain the needed functions to collect and return the desired features from each interval.

3.2 File Processing

This class contains the functions needed to utilize TextGrid Tools (Buschmeier and Włodarczak, 2013) to select only intervals that contain text from the input TextGrid file, ensuring that blank space is avoided. Once these timestamps are gathered, the *start* and *end* time stamps are used to grab an interval of the WAV file using Parselmouth before being saved to the dictionary themselves. The functions defined and stored inside the *extraction* classes are also called and utilized within this class. The final product of the main function of this class is a dictionary containing interval numbers that have corresponding values for *text, minimum, maximum, mean,* and *duration*.

3.3 Normalizing

In an effort to normalize the data the features *Standard Deviation* and *Z-Score* are taken into account. There is a small class with a series of functions that calculate these metrics using up to three intervals behind and ahead of the target interval. In the case where an interval does not have three intervals ahead or behind it, only the existing ones are calculated without padding. This function uses the output of the File Processing function as input and in turn outputs the input dictionary with added values for standard deviation and z-score.

3.4 Part-of-Speech

Once the final dictionary is created the *text* of each interval is run through the SpaCy Part-of-Speech (POS) tagger (Montani et al., 2023). To do this the output dictionary from Normalizing is used as input to this function. This tagger contains 18 different numerical values that all correlate with a unique POS. These results are then appended to the dictionary in a new list named *POS* before the dictionary is returned from the function.

3.5 CSV Outputs

This model is designed to take input as CSV files, and therefore functions are needed to translate the output dictionary into a CSV file. These files are also kept on the chance that they may be repurposed for retraining the model later on after they have been reviewed by humans.

3.6 TextGrid Outputs

To allow for ease of annotation and evaluation, the outputs are printed into a TextGrid file. This file contains a total of four tiers: Words (an interval tier at word-level), Prominence (an interval tier showing the percentage of a prominence event), Boundary (an interval tier showing the percentage of phrase boundary), and RPT (a point tier with RPT protocol markings). It is worth noting that these are generated using both TextGrid Tool and the textgrid package from praatio (Mahrt, 2019).

3.7 Models

Each of these models (Pitch and Intensity separately) are fed their corresponding CSV file

where they will output raw predictions (values between 0 and 1). From here the outputs are added into the dictionary that was created during the Normalizing functions. This allows for storage of predictions so that the predictions from Pitch and Intensity can be averaged for a final decision. For example, if Pitch predicts a value of 0.65 and Intensity predicts a value of 0.75, the final value would be 0.70.

3.8 Final Outputs

3.8.1 Prominence and Boundary

After averaging the two models, the data is written to a TextGrid file. The two tiers, Prominence and Boundary, are created as interval tiers. The intervals are aligned with each word that was provided in the input TextGrid. The text within these intervals is the averaged score from each of the models. We did this in an attempt to mimic how RPT annotation would be represented from a crowdsourcing approach.

3.8.2 Rapid Prosody Transcription

During this process, specific thresholds are set. These thresholds determine what scores are needed to output a marking. Due to the low number of examples of boundary, the thresholds are lower than those for prominence. The exact thresholds were calculated via ROC analysis against human annotated data that did not appear in prior training or testing iterations. ROC curves can be found in Appendix C.

Type	Empty	?	Marker * or]
Prominence	0-0.4	0.41-0.98	0.99-1
Boundary	0-0.16	0.17-0.8	0.81-1

Table 1: Marking Thresholds

Following the RPT protocol, the *Marker* item refers to "*" for prominence and "]" for boundary. A question mark is appended to these in order to designate lower confidence: "*?" and "]?".

A middle range was also identified in order to mitigate over-generation of prosodic markings, and to identify borderline markings for human annotators. The "?" was chosen as an uncertainty marker following the PoLaR framework (Ahn et al., 2019).

The marker tier is created as a point tier. By using an aligner (MFA in our case), vowels can be annotated for stress at the phoneme level. We utilize this to create discrete points at which we place our point annotations. If a word is marked with boundary by the model, the mark is by default output in line with the end of the word interval. If a word is marked with prominence, the mark is then placed at the center of the interval for the vowel containing the highest stress.

4. Features

4.1 Pitch and Intensity

The tool collects five features from each interval with regard to pitch and intensity: (1) the minimum value, (2) the maximum value, (3) the mean/average value, (4) the standard deviation of maximums when compared to surrounding intervals, and (5) the corresponding z-score of maximums when compared to surrounding intervals.

These features are informed by Rosenberg (2010) to be relevant and important to determining the presence of prosodic events.

4.2 Part-of-Speech

Part of Speech (POS) has been shown to be a relevant indicator of prosody (Rosenberg, 2010) and so was used in our model. Its relevance lies in how certain POS classes of words are much more likely to show prominence than others. Content word POSs are more likely to contain prominent events than their function word POS counterparts (Ananthakrishnan and Narayanan, 2008). This indicates that there is a structural consistency in how prominence appears as related to Part of Speech and highlights its value as a considered feature.

For implementation, a SpaCy tagger is adequate, as this gives detailed tags while limiting possible tags to a range of 18. From these tags, there is an associated number ID for each as designated by SpaCy. These numerical values are what the model reads as input, allowing them to be used with minimal processing.

4.3 Duration

We utilized duration as a feature due to it being indicative of prosodic events. Longer words tend to have more prominence and may also indicate a prosodic boundary (Klatt, 1975; Turk and Sawusch, 1996; Turk and White, 1999; Turk and Sawusch, 1997).

5. Results

Scores for Precision, Recall, and F1 were generated using the scikit-learn metrics (Pedregosa et al., 2011). These models perform strongly in prominence and with acceptable boundary scores. Recall scores across both features are high, which was an intentional decision to facilitate bootstrapping for human annotations. The combined scores reference the overall performance of the models when predicting both Prominence and Boundary together, giving a more holistic view of their performance.

Class	Precision	Recall	F1
Prominence	0.74	0.94	0.83
Boundary	0.40	0.79	0.53
Combined	0.62	0.91	0.74

Table 2: Pitch model scores

Class	Precision	Recall	F1
Prominence	0.76	0.94	0.84
Boundary	0.40	0.79	0.53
Combined	0.64	0.90	0.75

Table 3: Intensity model scores

6. Utility

To evaluate the utility of the tool we tested the speed of annotators working with and without the tool.

6.1 Materials and Methods

The annotators were 5 graduate and undergraduate students from Montclair State University. 5 were native speakers of English. 1 was a native speaker of Brazilian Portuguese, fluent in English. They ranged in ages from 21-30. 1 identified as Latine, 2 as white, and 2 as both. 1 had extensive experience with the annotation system, 1 was entirely new to prosodic annotation, and the rest had some intermediate degree of experience with the annotation system.

The annotation data were taken from the Montclair Map Task Corpus (Pardo et al., 2019), a set of 48 recorded and transcribed conversations around 45 minutes to an hour and a half in length. From each of these, a two-minute audio clip was randomly selected and extracted alongside the accompanying section of the TextGrid with word and phoneme tiers.

Each file was split by channel, with one conversational participant in channel 1, the left channel, and the other in channel 2, the right channel. Half of the conversations were randomly chosen and fed to the AutoRPT tool, which read the TextGrid file and added the tiers described in 3.6. On a few occasions, the tool failed due to some file naming errors and was rerun only until a viable file was produced.

Each annotator was assigned both channels of each of 4 conversations, 2 with a TextGrid that had gone through the AutoRPT tool ("prepopulated") and 2 with a TextGrid that had not and contained only the transcription and phoneme tiers ("clean"). To mitigate learner bias, each annotator was given the files in a strict order alternating between prepopulated and clean files. Annotators were instructed to annotate each file in a single sitting wherever possible. Annotators recorded the time it took to annotate each file.

All data was analyzed using R Statistical Software (v2025.09.0; R Core Team, 2025). Plot data was obtained via the tidyverse package (Wickham et al., 2019). Three outliers were removed (time > 55 minutes for clean files for annotators with ranges otherwise 10-35). The data was normalized by word count of the clip to account for clips that were mostly confirmations of the other person's statements

6.2 Results

When annotating with the tool we found that the annotators averaged 4.94 more words per minute (M = 4.43, SD = 2.68) than without, $t(5) = 3.70$, $p = .01045$. See Appendix A for more details.

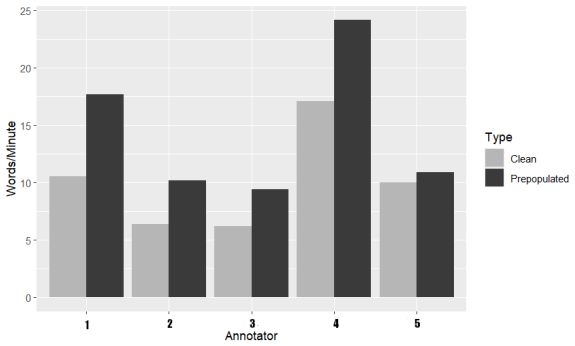


Figure 1: Annotators averaged more words per minute using the tool (avg 15.16) than without (avg 10.22).

We used lme4 (Bates et al., 2015) to perform a linear mixed effects analysis of the relationship between minutes spent annotating and whether the annotation tier was prepopulated with annotations. We included order of presentation and number of words (excluding fillers) as a fixed effect and annotator as random effect. P-values were obtained by likelihood ratio tests of the full model with the effect prepopulating annotations against the model without the effect of prepopulating annotations. The difference between models was significant ($\chi 2 = 11.644$; $p < 0.001$).

7. Conclusion

This approach yields acceptable scores that can be utilized for bootstrapping the human annotation processes. The high recall rates fit our goal of ensuring that minimal resources would be needed to find prosodic features and that time could be spent narrowing down the predicted features as opposed to locating them.

8. Limitations

This tool is built using limited data and from a single variety of English. Therefore, it inherently

contains a bias to perform better on "Standard" American English. As the tool is designed for bootstrapping annotations, overall F1 and precision scores tend to be lower at the boon of higher recall scores.

Boundary scores also underperform due to what is believed to be a lack of data when compared to prominence for an LSTM model type (support for prominence being 1189 and for boundary being 409) and potentially suboptimal acoustic measures.

The thresholds for each feature have not been fully optimized, and the tool may perform better on different datasets when paired with different thresholds. To this point, within the Utilities.py file there exists a loop inside the create_point_tier function, which may be used to adjust thresholds.

9. Ethics

This tool is intended to facilitate the acoustic annotation of speech data, making under-researched varieties of English more easily annotated as well as easing the training process of new annotators. This tool does not aim to replace human annotators, nor does it aim to advocate for a 'standard' of Prosody in English.

10. Future Works

Moving forward, the immediate goal of this project is to retrain the models using more varieties of American English such as Latine and African American varieties. Depending on data collection, a switch from LSTM for boundary detection to either classical machine learning techniques such as Support Vector Machines (SVM) or reverting to the Recurrent Neural Network approach may be made.

As this is ideally a tool to be used on corpora, user interface improvements are being made to facilitate working on large amounts of data at once, as the original tool runs on one file set at a time.

11. Supplementary Materials

The source code, models, documentation, and installation/usage instructions for the tool can be found at https://github.com/Howell-Prosody-Lab/AutoRPT.

12. Acknowledgments

We would like to thank the members of our lab, past and present. Without their assistance and hard work this tool would not be possible.

This work supported by the National Science Foundation under Grant No. FY22-23-3000.

13. References

Ahn, B., Veilleux, N., and Shattuck-Hufnagel, S. (2019). Annotating prosody with polar: Conventions for a decompositional annotation system. In *Proceedings of the 19th International Congress of Phonetic Sciences*, pages 1302–1306, Canberra, Australia. Australasian Speech Science and Technology Association Inc.

Ananthakrishnan, S. and Narayanan., S.S. (2008). Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):216–228.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Beckman, M.E and Elam, G.A. (1994). Guidelines for ToBI labelling. Technical report, Ohio State University.

Biron T, Baum D, Freche D, Matalon N, Ehrmann N, Weinreb E, Biron D, and Moses E. (2021) Automatic detection of prosodic boundaries in spontaneous speech. *PLoS ONE*, 16(5): e0250969. https://doi.org/10.1371/journal.pone.0250969

Buschmeier, H. and Włodarczak, M. (2013). TextGridTools: A TextGrid processing and analysis toolkit for Python. In *Proceedings der 24. Konferenz zur elektronischen Sprachsignalverarbeitung*, pages 152–157, Bielefeld, Germany.

Cole, J., Hualde, J.I., Smith, C.L., Eager, C., Mahrt, T., and Napoleão de Souza, R. (2019). Sound, structure and meaning: The bases of prominence ratings in English, French and Spanish. *Journal of Phonetics*, 75:113–147.

Escudero, D., Aguilar-Cuevas, L., González-Ferreras, C., Gutiérrez-González, Y., and Cardeñoso-Payo, V. (2014). On the use of a fuzzy classifier to speed up the Sp_ToBI labeling of the glissando Spanish corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 1962–1969.

Jadoul, Y., Thompson, B., and de Boer, B. (2018). Introducing Parselmouth: A python interface to Praat. *Journal of Phonetics*, 71:1–15.

Klatt, D.H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, 3(3):129–140.

Mahrt, T. (2016). PraatIO: A python library for working with praat files. https://github.com/timmahrt/praatIO

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M.,

and Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proceedings of Interspeech 2017*, pages 498–502.

Montani, I, Honnibal, M, Honnibal, M., Boyd, A., Van Landeghem,, S., and Henning Peters, H. (2023). explosion/spaCy: v3.7.2: Fixes for APIs and requirements (v3.7.2). Zenodo. https://doi.org/10.5281/zenodo.10009823

Ostendorf, M., Price, P., and Shattuck-Hufnagel, S. (1996). Boston University Radio Speech Corpus. Technical report, Linguistic Data Consortium. Language Resource Association (ELRA).

Pardo, J. S., Adelya, U., Gash, H., Jaclyn, W., Mason, N., Sherilyn, W., Keagan, F., and Decker, A. (2019). The Montclair map task: Balance, efficacy, and efficiency in conversational interaction. *Language and Speech*, 62(2), 378-398. https://doi.org/10.1177/0023830918775435

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., and Vanderplas. J. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning research*, 12:2825–2830.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org./

Rosenberg, A. (2010). Autobi - a tool for automatic ToBI annotation. In *Proceedings of INTERSPEECH*, pages 146–149.

Rosenberg, A. (2018). Speech, prosody, and machines: Nine challenges for prosody research. In *Proceedings of the International Conference on Speech Prosody*, pages 784–793.

Silverman, K.E., Beckman, M.E., Pitrelli, J.F., Ostendorf, M., Wightman, C.W., Price, P., and Hirschberg, J. (1992). Tobi: A standard for labeling English prosody. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, volume 2, pages 867–870.

Turk, A.E. and Sawusch, J.R. (1996). The processing of duration and intensity cues to prominence. *The Journal of the Acoustical Society of America*, 99:3782–3790.

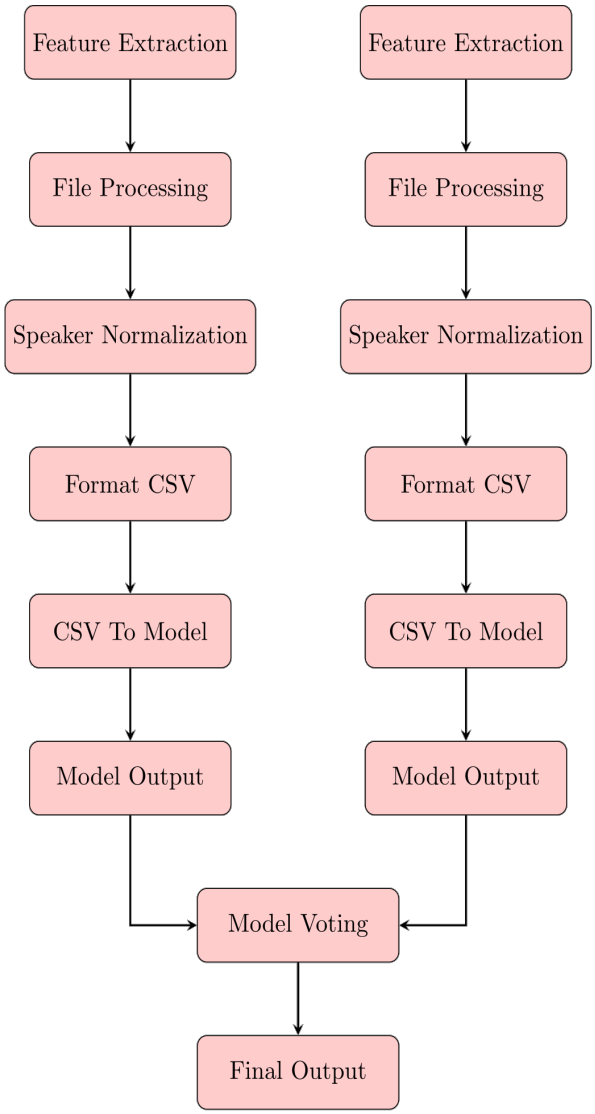
Turk, A.E. and Sawusch, J.R. (1997). The domain of accentual lengthening in American English. *Journal of Phonetics*, 25(1):25–41.

Turk, A.E. and White, L. (1999). Structural influences on accentual lengthening in English. *Journal of Phonetics*, 27(2):171–206.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., ... Yutani, H. (2019). “Welcome to the tidyverse.” *Journal of Open Source Software*, **4**(43), 1686. doi:10.21105/joss.01686.

14. Appendix A

This image serves to illustrate the process of AutoRPT as it pulls features, runs analysis, and outputs results:



15. Appendix B

Improvement in words per minute (outliers removed)

Group	With tool (wpm)	Without tool (wpm)	% difference
High experience (n=1)	24.16	17.06	34.45
Avg. middling experience (n=3)	13.38	8.70	45.20
Low experience (n=1)	10.17	6.38	45.80
Avg. of all	15.16	10.22	38.97

Some reviewers wondered whether experience level correlated with improvement. The data is presented for readers' edification, but it is evident that there are not enough participants in each group to make a strong claim.

Three attempts were removed as outliers (minutes > 55 in a dataset where all other attempts (n=42) were <40 minutes). With those data left in, the chart reads as follows:

Improvement in words per minute (outliers left in)

Group	With tool (wpm)	Without tool (wpm)	% difference
High experience (n=1)	24.16	17.06	34.45
Avg. middling experience (n=3)	13.38	8.36	46.18
Low experience (n=1)	10.17	4.61	75.24
Avg. of all	15.16	9.26	48.35

16. Appendix C

ROC curves for Prominence and Boundary detection against human annotations (n=74,844 intervals) on non BURNC data.