Velislava Todorova / Maria Chinkina

Significance Filters for N-gram Viewer

Abstract This paper presents a visualization tool for the analysis of tendencies in language use over time. Given a dated and tokenized corpus, it calculates frequencies of selected n-grams and visually presents them as data points on a line chart in a coordinate system, with time on the x axis and relative frequency on the y axis. It provides the option of smoothing the graph in order to make the general tendency more salient. The user can specify an n-gram as a sequence of tokens, lemmas, and/or POS tags, if the corpus provides these annotations. Along with the original text, the tool also accesses the metadata of the corpus, such as dates and authors' names, allowing for a comparison of the use of n-grams by different authors at different time periods in context. The latest version of our tool introduces a filtering mechanism that indicates the periods of time throughout which the observed values within one or more datasets are significantly different. We used Fisher's exact test of independence because it has the advantage of providing reliable results even for sparse data.

1. Introduction

Exploration of the patterns in language use over time is useful for a number of Natural Language Processing (NLP) tasks such as authorship attribution and topic detection. Google Ngram Viewer[1] is one example of such a tool. While providing the functionality of querying the Google Books corpus[2], its current functionality does not include uploading and processing one's own text corpora. This limitation is overcome by our new Ngram Tendency Viewer *Slash/A*[3], which is more suitable for researchers interested in exploring a specific collection of texts.

Given a dated and tokenized corpus, *Slash/A* calculates and visually presents frequencies of selected n-grams as a line chart in a coordinate system, with time

[1] The online Google Ngram Viewer is available here: https://books.google.com/ngrams

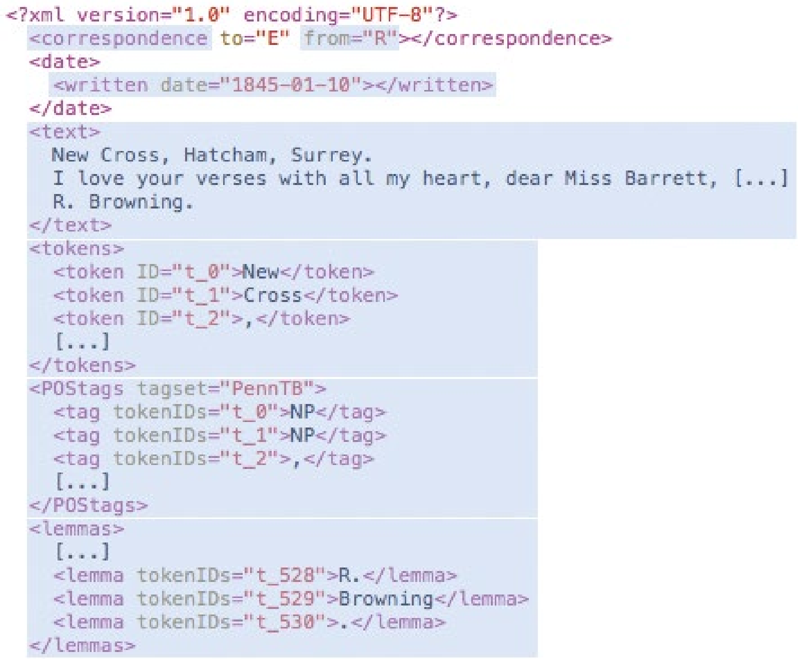
[2] The Google Books corpus can be accessed here: https://books.google.com

[3] The online *Slash/A* tool can be accessed here: https://tinyurl.com/slasha-tool

Published in: Noah Bubenhofer, Marc Kupietz: Visualisierung sprachlicher Daten. Heidelberg: Heidelberg University Publishing, 2018
DOI: https://doi.org/10.17885/heiup.345.474

on the x axis and relative frequency on the y axis and provides the option of smoothing the graph in order to make the general tendency more salient. The user can specify an n-gram as a sequence of tokens, lemmas and/or POS tags, if the corpus provides these annotations. Along with the original text, the tool also accesses the metadata of the corpus, such as dates and authors' names, allowing for a comparison of the use of n-grams by different authors at different time periods in context.

The latest version of our tool introduces a filtering mechanism that indicates whether the observed values in a specified time period are significantly different in terms of (i) lower and higher extremes of n-gram frequencies and (ii) use of the same n-gram by different authors. In these cases, a significance filter can facilitate scientific hypothesis testing. The statistical test that we decided to use is Fisher's exact test of independence (Fisher, 1950). It is very similar to the $\chi 2$ test of independence but has the advantage of providing reliable results even in cases with very little data, e.g., if an n-gram only occurs five times in the whole corpus.



```
<?xml version="1.0" encoding="UTF-8"?>
  <correspondence to="E" from="R"></correspondence>
  <date>
    <written date="1845-01-10"></written>
  </date>
  <text>
    New Cross, Hatcham, Surrey.
    I love your verses with all my heart, dear Miss Barrett, [...]
    R. Browning.
  </text>
  <tokens>
    <token ID="t_0">New</token>
    <token ID="t_1">Cross</token>
    <token ID="t_2">,</token>
    [...]
  </tokens>
  <POStags tagset="PennTB">
    <tag tokenIDs="t_0">NP</tag>
    <tag tokenIDs="t_1">NP</tag>
    <tag tokenIDs="t_2">,</tag>
    [...]
  </POStags>
  <lemmas>
    [...]
    <lemma tokenIDs="t_528">R.</lemma>
    <lemma tokenIDs="t_529">Browning</lemma>
    <lemma tokenIDs="t_530">.</lemma>
  </lemmas>
```

Figure 2.1: A simplified sample file in XML format. Required elements are highlighted blue

2. *Slash/A* N-gram Tendency Viewer

Data retrieval

Slash/A is designed to process corpora in XML format, one text per file.[4] To make use of the full functionality of the tool, each XML file should contain the original text, the date, the author's name, and the annotations for tokens, lemmas and POS tags in the order they appear in the text (see Figure 2.1).

All of the token annotations can then be used to compose a corpus query. The following are examples of valid queries using the Penn Tree Bank POS tag set:[5]

our present — a query for the bi-gram *our present*

/VBP presents/NNS — a query for all bi-grams with a non-third person singular verb in present tense as the first token and the plural form of the noun *presents* as the second one

present/lemma/V* — a query for all uni-grams with any form of the verb *present*

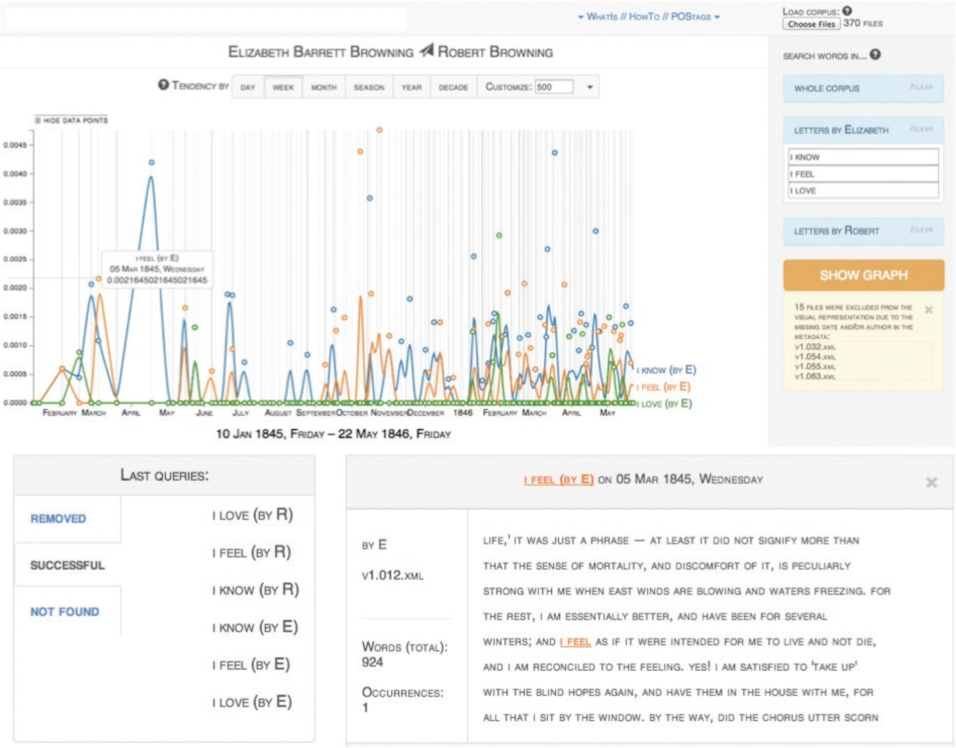
The second example shows that an omitted token in a query leads to matching any token with the specified POS tag. The last example illustrates the use of a combination of all three types of annotations along with the wildcard character (*). The search hits are counted in every document of the corpus and a relative frequency is calculated for each day of the whole time period the corpus covers. These daily relative frequencies can then be smoothed for a more abstract view. We use moving average as a smoothing technique, and the length of the sliding window can be determined by the user. The five preset levels of smoothing correspond to common time intervals: day (no smoothing), week (smoothing parameter p = 3), month (p = 15), three months (p = 45), and year (p = 182). For more detailed information on smoothing as well as the requirements for the corpus format, see Todorova and Chinkina (2014).

[4] The Brownings' corpus, our development corpus, the examples from which are used in this paper, can be downloaded here: http://linguistics.chrisculy.net/lx/resources/. A detailed description of the format of the corpus can be found here: http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format

[5] The Penn Tree Bank POS tag set can be found here: http://www.cis.upenn.edu/~treebank/

Visualization

The basis of our visualization is a simple graph. Relative frequencies of n-grams are plotted as data points, and the selected level of smoothing is represented by a continuous line fitted to the data points (see Figure 2.2). Following Schneiderman's (1996) taxonomy, we have included functionality that permits the high level tasks of overview, zoom, filter, details-on-demand and history. The interactive visualization allows for non-linear exploration of a dataset: adding and deleting word lines, keeping track of successful and unsuccessful queries within the current session and getting access to original text.



15 files were excluded from the visual representation due to the missing date and/or author in the metadata:
v1.032.xml
v1.054.xml
v1.055.xml
v1.063.xml

life,' it was just a phrase — at least it did not signify more than that the sense of mortality, and discomfort of it, is peculiarly strong with me when east winds are blowing and waters freezing. For the rest, I am essentially better, and have been for several winters; and I feel as if it were intended for me to live and not die, and I am reconciled to the feeling. Yes! I am satisfied to 'take up' with the blind hopes again, and have them in the house with me, for all that I sit by the window. By the way, did the Chorus utter scorn

Figure 2.2: *Slash/A* interface: a top panel containing a link to the information about the tool and a button for loading a corpus, a graph showing smoothing by week, a query panel, a history element, and a reading element.

Motivation for Significance Filters

The visualization of relative n-gram frequencies manages to convey a lot of information quickly and efficiently by providing an overview of the general tendency the data follows. However, it can be a source of confusion. Sometimes extreme ups and downs in the use of an n-gram can be perceived as significant even though they are not representing enough data with these high (or low) values. Or, when comparing the frequencies of an n-gram in two subsets of a corpus, some minimal difference might seem insignificant, while it can actually be of significance. To our knowledge, there are currently no n-gram viewers that eliminate the possibility of such confusions.

We try to overcome the problem by suggesting that the inclination of the user to see significance in the visualization should be taken into account, and that various kinds of statistical analysis should be introduced in *Slash/A* for the different tasks that it can be applied to. The following section presents our work on two kinds of statistical analysis and their visualization. We call the technique statistical filtering as its goal is to extract and present portions of data (defined by time intervals) that have the property of obtaining a significant result when subjected to a certain statistical text. The first filter presented here indicates the intervals of time within which the occurrences of an n-gram are significantly higher or lower than in the rest of the period over which the corpus spans. The second filter is to be applied when comparing different subsets of the corpus, and it indicates the time intervals in which the occurrences of an n-gram in the two subsets are significantly different. In what follows, we will present the technical side of the analysis as well as the visual representation of the results and will use various specific cases as illustrations.

3. Filters for significant fluctuations in one data set

Motivation

The first filter that we present filters out the time intervals throughout which a specific n-gram occurs about as much as is expected to occur by chance, given its occurrences in the rest of the documents. This way, the user can focus on the periods with significantly higher (or lower) frequencies than observed in the rest of the corpus. For example, if one is interested in how much a formality marker appears in the first letters of Elizabeth Barrett and Robert Browning's correspondence, one should look for periods at the beginning of their exchange that contain significantly more occurrences of this marker than the subsequent letters.[6]

Mechanism to detect significant time intervals

It is not possible to automatically identify the interval that is of interest for the user in their current task and only test it for significance. It is also not possible to identify the size or the boundaries of this interval. One possible solution could be to ask the user to provide this information for every query. Alternatively, one might just adopt a strategy to select intervals to test. We decided to go for a compromise: we let the user determine the size of the interval and let an algorithm decide the boundaries.

We have taken advantage of the possibility to customize the length of the tested intervals without complicating the use of the program. The different levels of smoothing are related to time intervals of different length, and the same approach can also be used for the statistical tests. Thus, the user is specifying two things at the same time – the size of the sliding window for the moving average smoothing, and the size of the intervals to be tested.

After the size of the intervals of interest is determined, it is checked if the size is meaningful. Having intervals longer than the half of the whole time period tested is not allowed because it can be confusing when presented visually: The tested interval should appear as an object against the rest of the period as background, but if what has to be perceived as background were smaller, the exact opposite visual effect would be produced. Following the principle of a sliding window, all sub-intervals of the specified length are tested for significant difference from the rest of the data. We have decided to apply the standard one-tailed χ^2 test and one-tailed Fisher exact test of independence (Fisher, 1950), each of which determines if two variables are connected, i.e., if one can be used as a predictor for the other. The two variables tested for independence here possess a certain linguistic feature (e.g., being a noun or containing the consonant complex “np”) and belonging to a certain time interval.

In the general case, we use χ^2, which is faster to compute. But whenever there are too few observations, which is usually the case with longer n-grams and with some rare n-grams, we apply the Fisher test of independence as it also provides reliable results in these cases.

[6] See the “Examples” section for clarification.

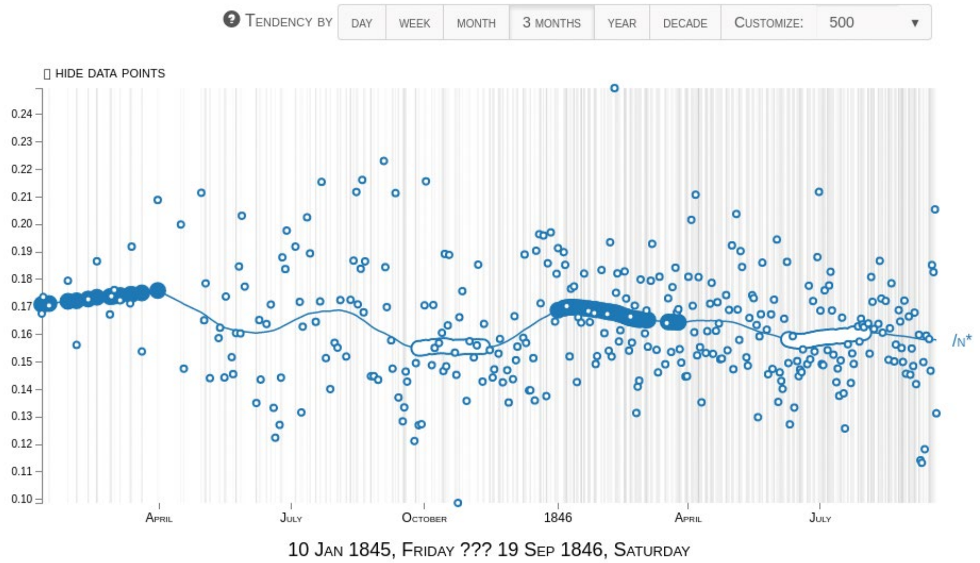


Figure 3.1: Frequency of nouns in the Brownings' corpus (level of smoothing: three months, smoothing parameter p = 45).

Visualization

Visualizing potentially overlapping intervals with either high or low frequency is challenging. First, one needs to make sure that the difference between the indications for significantly higher and significantly lower values is clear. Second, it is important to deal with cases of overlaps without introducing confusion.

The solution we will present here, which is also illustrated in Figure 3.1, is to indicate only the centers of the significant intervals. This way, we avoid the problem of overlaps in the visual presentation. Furthermore, since centers of intervals are points, one can use special marks to indicate them (instead of changing the properties of the part of the line that represents the interval, which is presumably a more confusing approach). These special marks can have two sets of features – one that associates them with the graph they belong to, and the other that makes it clear if this is a period with a lower or higher frequency than in the rest of the documents. We decided to include the stroke color of the mark (which should be the same as that of the graph) in the first set of features and the filling color of the mark in the second set. If the mark has a white center, it indicates that it denotes an interval with a low frequency, and if the mark is

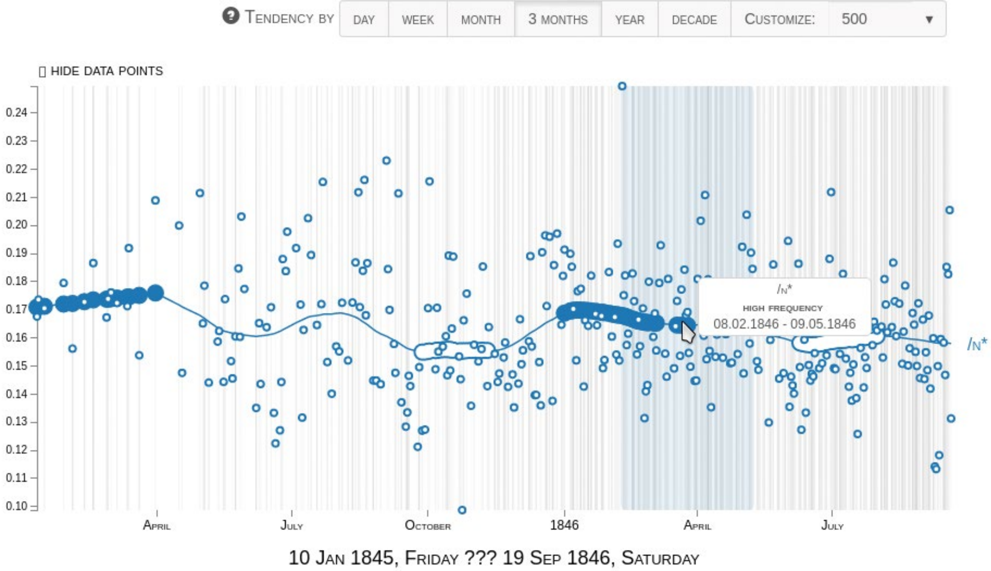


Figure 3.2: Statistical details on demand: Frequency of nouns in the Brownings' corpus (level of smoothing: three months, smoothing parameter p = 45).

filled with a solid color (the color of the graph), it indicates an interval with a high frequency.

We have included functionality of statistical details on demand, which is illustrated in Figure 3.2. To indicate the boundaries of the intervals, we color the background whenever the center of a significant interval is pointed at. For this, we use the color of the selected mark, but with a very low opacity. The details about the significance hypothesis are shown in a tooltip.

4. Filters for significant differences between two data sets

Motivation

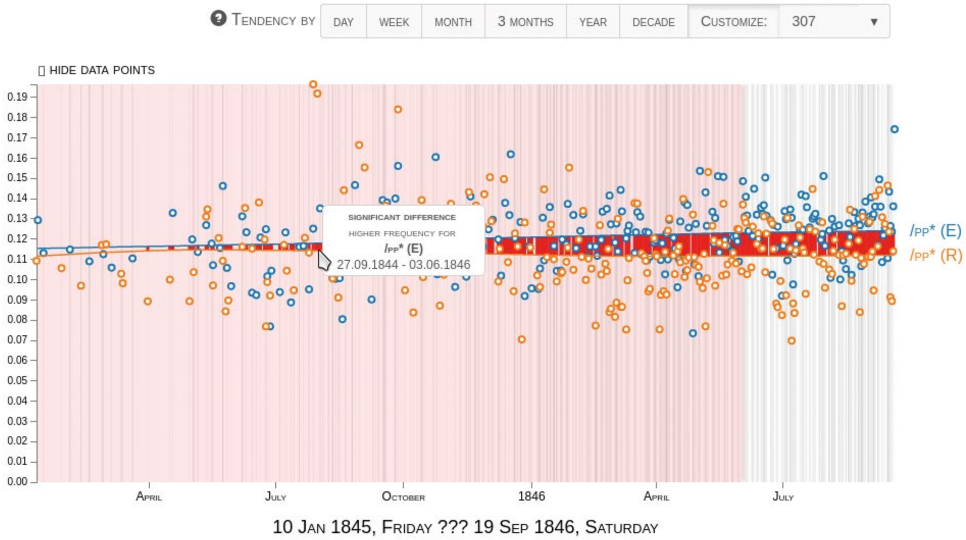
The filter described above is applicable to one time series at a time. However, significance filtering can also be very useful in a task that involves comparison between two time series – that is, to compare the frequencies of a certain n-gram in two distinct groups of texts.

Mechanism for detecting significant periods

Detection of significant periods in this filter is similar to that used for the first one, with the important difference that not one interval is compared to the rest of the data, but two subsets of the data are compared to each other at a given time interval. Again, we use a sliding window with the size of the moving average sliding window to go through the whole time series and detect intervals with significant results.

Visualization

The visualization of significant differences between data subsets inherits the idea of only indicating the center of the intervals, and it does so by placing a thin red line connecting the two time series. On mouse-over over a red line, one obtains the indication of the interval length in the background and a tooltip with additional information, as shown in Figure 4.



SIGNIFICANT DIFFERENCE
HIGHER FREQUENCY FOR
lpp* (E)
27.09.1844 - 03.06.1846

10 Jan 1845, Friday ??? 19 Sep 1846, Saturday

Figure 4: Comparison of the frequency of pronouns in the letters by Robert and Elizabeth (smoothing parameter p = 307).

5. Examples

General notes

The statements that we will use here to demonstrate the statistical filtering functionality of *Slash/A* are from the domains of formality theory and of gender linguistics. More specifically, the statement about nouns as an indicator of language formality is taken from Heylighen and Dewaele (1999), and the statement about the gender specific use of pronouns comes from Koppel et al. (2002) and Argamon et al. (2003). The corpus that will be surveyed is the Brownings' corpus mentioned earlier.

Investigating one data set

Let us explore the statement "Formal language is characterized by higher frequency of nouns" using *Slash/A*. Figure 3.1 shows that the beginning of the correspondence is characterized by a significantly higher frequency of nouns than in the rest of the corpus. The end of the correspondence is characterized by a significantly lower frequency of nouns. Thus, the language in the corpus contains more formal markers (nouns in this particular case) in the first quarter of the exchange and less formal markers in the last quarter. It is interesting that the middle part of the time series doesn't follow the expected pattern of a gradual lowering of the formality level. The rare occurrence of nouns in the second quarter of the period is followed by a frequent use of them, which could be indicative of two things: (i) that the formality of the correspondence does not evolve linearly, but rather goes back and forth; or (ii) that the frequency of nouns is also a marker for something else, and this other thing interferes with the formality of the language producing unclear patterns. To check if (i) holds, we can look at a higher level of smoothing. The expectation would be that when longer intervals are taken into account, not quarters but halves of the whole periods will be marked for significance, and the first half will contain more occurrences of nouns than the second one. Figure 5.1 is obtained by applying the largest smoothing parameter that is accessible for fluctuations testing. One can no longer see anything informative about the beginning and the end of the letter exchange as the differences in noun usage are not significantly different. It appears that the presence of this formality marker does not change linearly over time – at least not for the whole collection of letters.

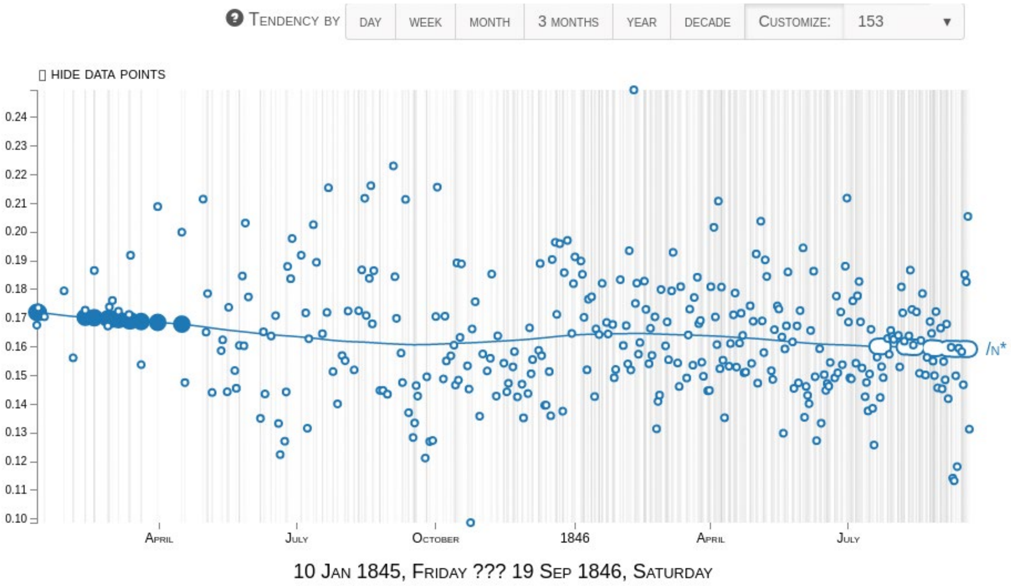


Figure 5.1: Frequency of nouns in the Brownings' corpus (a custom smoothing parameter p = 153)

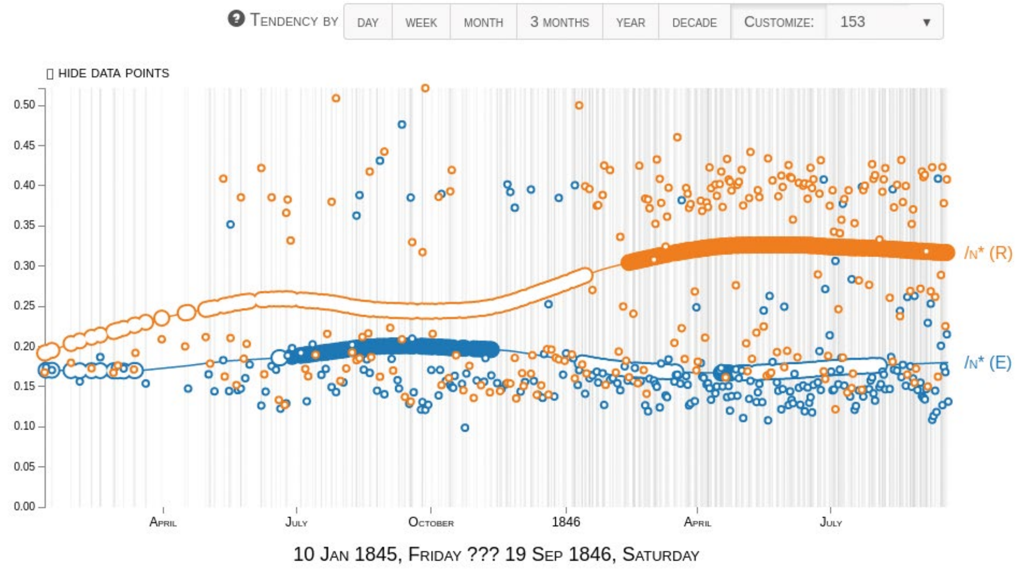


Figure 5.2: Frequency of nouns in the letters by Robert and Elizabeth separately (smoothing parameter p = 153).

To explore the question further, one can look at Elizabeth's and Robert's letters separately.[7] An interesting observation can be made from Figure 5.2: Elizabeth's letters show a linear decrease in the use of nouns. That is, on a larger scale, the formality level of her letters lowers neatly linearly with time. On the other hand, Robert appears to be following a contrary pattern, his use of nouns in the first half of the exchange being lower than in the second one.

Making sense of all these observations would require some exploration of other formality markers. What we can suggest as a possible interpretation is that Robert was quicker in dropping this particular formality marker (there are multiple low frequency marks in the second quarter of Robert's time series in Figure 5.1). Elizabeth held on to her nouns longer and lowered their use more significantly only after the first half of the correspondence period. A process of language adaptation can explain why Robert then raised his use of nouns to a level closer to the one his beloved respondent was sustaining.

Comparing two data sets

For the second type of filtering, we present the results obtained by exploring the statement "Women use more pronouns than men". Figure 4.1 above illustrates the statement – Elizabeth uses more pronouns than Robert in the course of their correspondence.[8] More interesting is what can be seen in a less smoothed view, like the one in Figure 5.3. It again shows that Elizabeth uses pronouns significantly more often, but only in certain periods, and that there are times when the frequencies of pronouns in the letters of the two authors are so close that their differences are insignificant.

6. Conclusion and Future Work

We have presented *Slash/A* N-gram Tendency Viewer that extracts data from a corpus, conducts searches on it, calculates and plots n-gram frequencies and smooths them. We have also discussed the advantages of including a significance filtering functionality and proposed two significance filters to improve the user's

[7] Note that the number and distribution of data points are different when we inspect Elizabeth's and Robert's letters separately compared to inspecting the corpus as a whole. In the latter case, each data point represents one day, and the frequency is calculated based on all the letters written on this day.

[8] The parameter 307 is chosen because it is the largest parameter value that can be applied to this data.

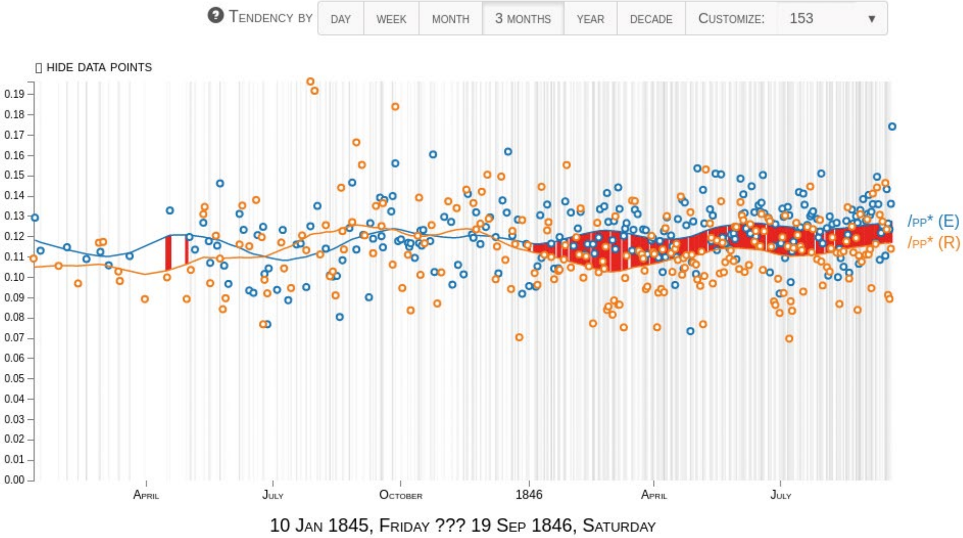


Figure 5.3: Comparison of the frequency of pronouns in the letters by Robert and Elizabeth (level of smoothing: three months, smoothing parameter p = 45).

certainty level in their conclusions. One of the filters shows periods with significantly different values as compared to the rest of the time series, and the other one shows significant differences between two time series.

As future work, other similar filters could be introduced for other kinds of tasks. For example, one that indicates the time intervals throughout which a certain combination of words is used more often than it is expected for these words to appear together by chance. This can be helpful for collocation strength monitoring if collocations are understood simply as words occurring together unexpectedly often (for the definition of collocation, see Dale et al., 2000).

7. Acknowledgements

We would like to thank the supervisor of this project and a former professor at the University of Tübingen, Dr. Christopher Culy, for his continuous support and valuable advice. Velislava Todorova is also grateful to the German Exchange Service (DAAD) for the grant that made it possible for her to study in Tübingen and to work on this project there.

8. References

Argamon, Shlomo, Moshe Koppel, Jonathan Fine and Anat Rachel Shimoni. 2003. “Gender, genre, and writing style in formal written texts.” In *Text – Interdisciplinary Journal for the Study of Discourse,* 23, no. 3: 321–346. https://doi.org/10.1515/text.2003.014.

Dale, Robert, Hermann Moisl, and Harold Somers, 2000. Eds. *Handbook of Natural Language Processing*. CRC Press.

Fisher, Ronald Aylmer. 1950. “Statistical methods for research workers.” 11th ed. Edinburgh: Oliver and Boyd. https://doi.org/10.1038/123866a0.

Heylighen, Francis and Jean-Marc Dewaele. 1999. “Formality of language: definition, measurement and behavioral determinants.” Technical report. Center “Leo Apostel”, Free University of Brussels & Birkbeck College, University of London pespmc1.vub.ac.be/Papers/Formality.pdf.

Koppel, Moshe, Shlomo Argamon, and Anat Rachel Shimoni. 2002. “Automatically categorizing written texts by author gender.” *Literary and Linguistic Computing* 17, no. 4: 401–412. http://dx.doi.org/10.1093/llc/17.4.401.

Shneiderman, Ben. 1996. “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations.” In *Proceedings of IEEE Visual Languages*, 336–343 https://doi.org/10.1109/vl.1996.545307 (accessed 12 January 2018).

Todorova, Velislava, and Maria Chinkina. 2014. “*Slash/A* N-gram Tendency Viewer: Visual Exploration of N-gram Frequencies in Correspondence Corpora.” In *Proceedings of the ESSLLI 2014 Student Session*, 229–239. http://www.kr.tuwien.ac.at/drm/dehaan/stus2014/proceedings.pdf (accessed 12 January 2018).