

Significance Filters for N-gram Viewer

Velislava Todorova and Maria Chinkina
International Studies in Computational Linguistics, University of Tübingen

Slash/A n-gram tendency viewer

Slash/A is a web-based tool that visualizes frequencies of n-grams extracted from a user provided dated text collection. It can access token level annotations and thus allows searches by e.g. lemmas or POS tags.

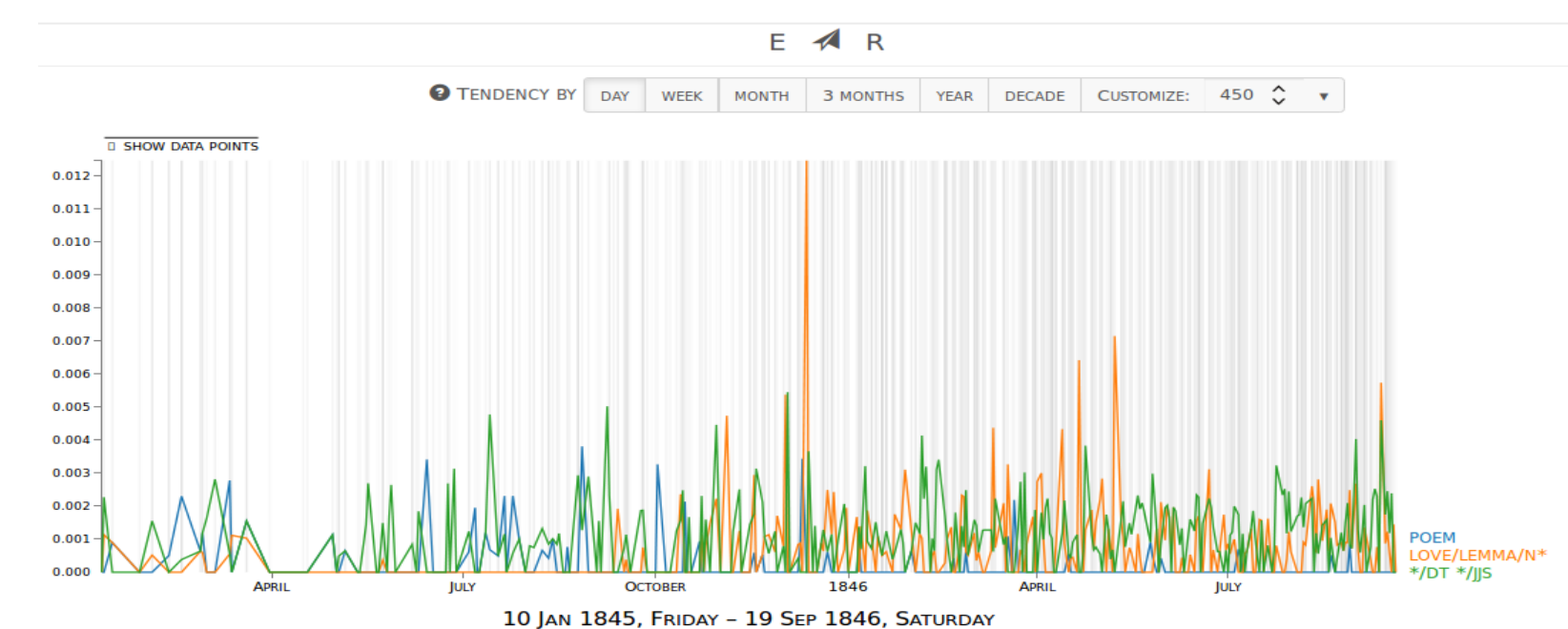


Figure 1: Slash/A query in the Brownings' corpus* for: **poem** (“poem”); **love/lemma/N*** (noun forms with lemma “love”); and ***/DT */JJS** (determiners followed by superlative adjective forms).

The viewer provides the option to smooth the results in order to make the general tendency clearer to see.

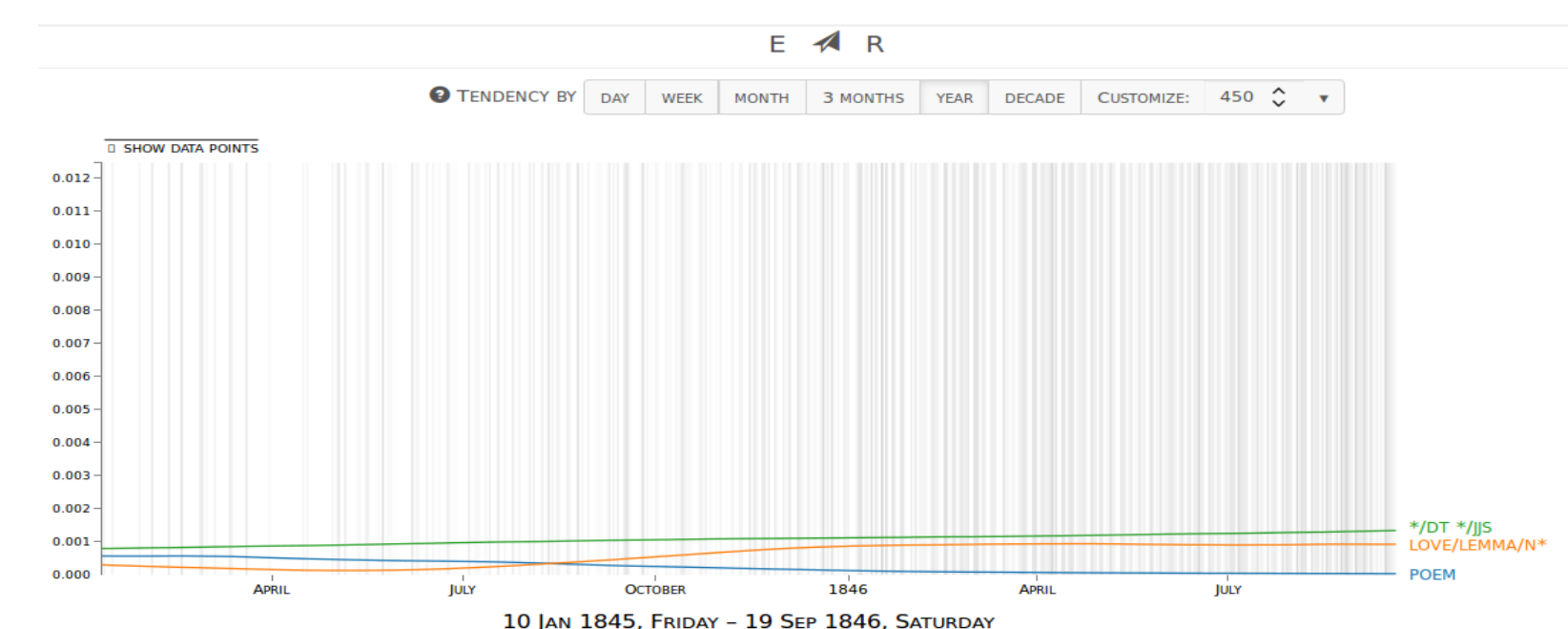


Figure 2: The frequencies from Fig.1, but smoothed.

Slash/A also allows comparing two or more subsets of the corpus (the user can specify the criterion for the division, it can be author of the text or genre or any other piece of document level annotation).

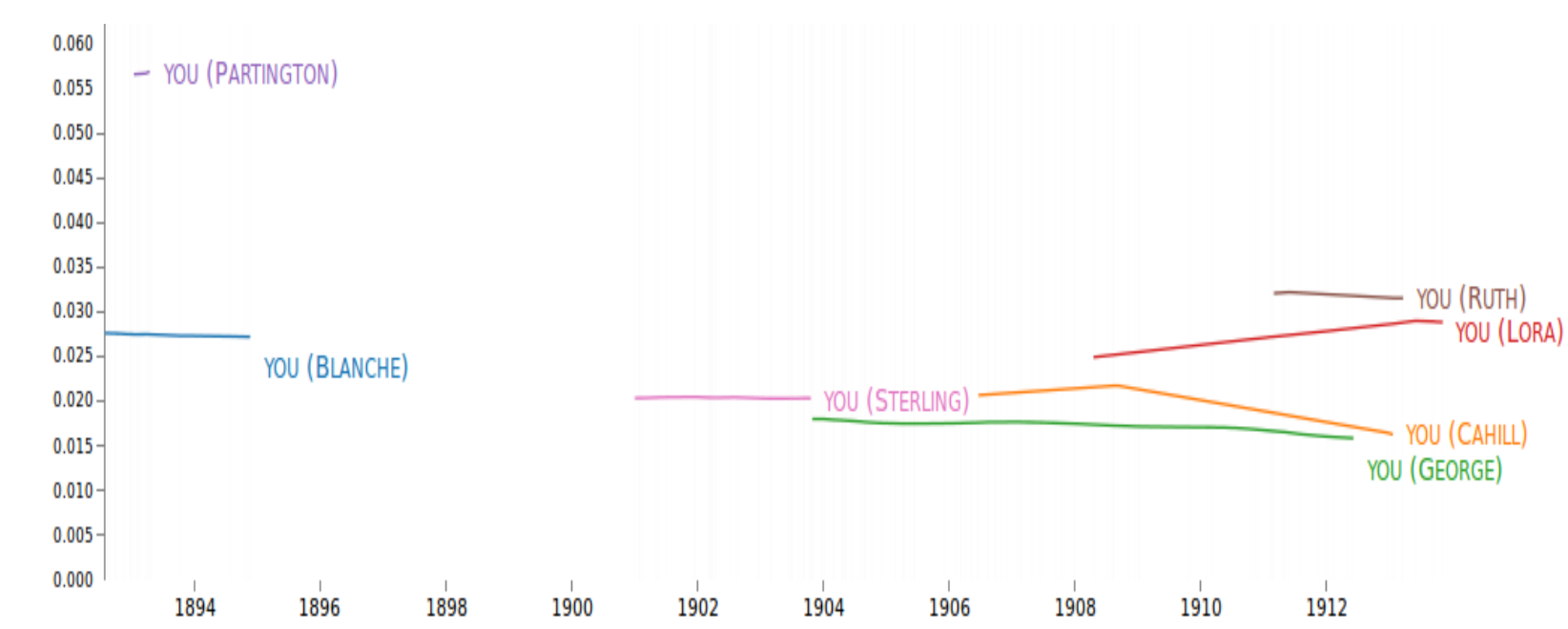


Figure 3: The use of “you” in the Ambrose Bierce corpus,* comparison by recipient.

*The corpus is available from <http://linguistics.chrisculy.net/vistola/index.html#resources>.

Challenge: visible vs. significant

Patterns in one data set

When exploring one set of data, the user is interested in the tendencies. Sometimes the graph is influenced by isolated outliers that are resistant to smoothing and create the wrong impression that the patterns observed are significant.

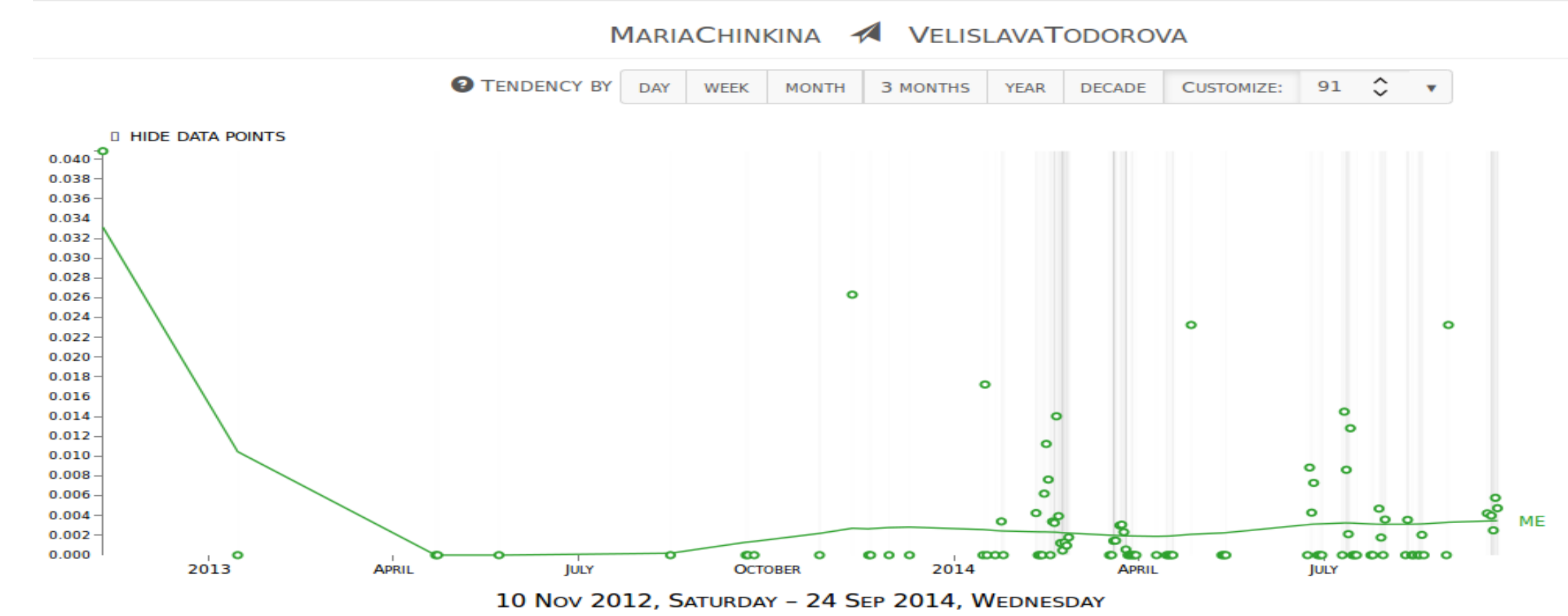


Figure 4: The use of “me” in the personal correspondence between the two authors of this poster.** There is a single data point (the first one) with extremely high value that affects visibly the shape of the graph.

Solution: filtering

Periods with significantly high/low values

We conduct statistical tests (χ^2 and Fisher) to determine if the frequencies in one period are significantly higher (or lower) than in the rest of the corpus and we include the results in the visualization.

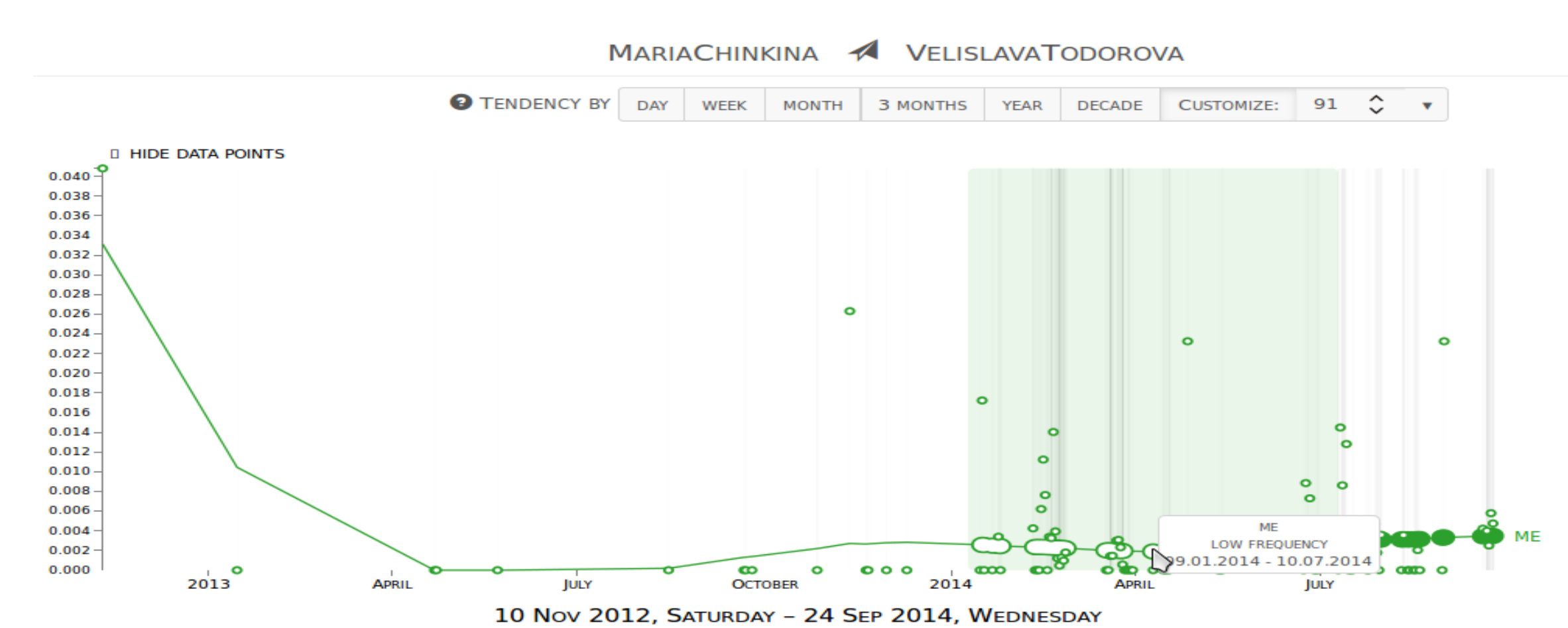


Figure 6: The same data as in Fig. 4. The centers of the periods with significantly high and low frequencies (compared to the rest of the data) are marked with ● and ○ respectively. When mousing over a mark, the whole period is highlighted in the color of the graph and details are shown.

**The corpus is compiled by Maria Chinkina and consists of 4455 Facebook messages.

Differences between two data sets

When exploring two sets of data, the user is interested in the differences between them. The distance between the graphs can sometimes be taken to mean significance level even though it is not meant to represent that.

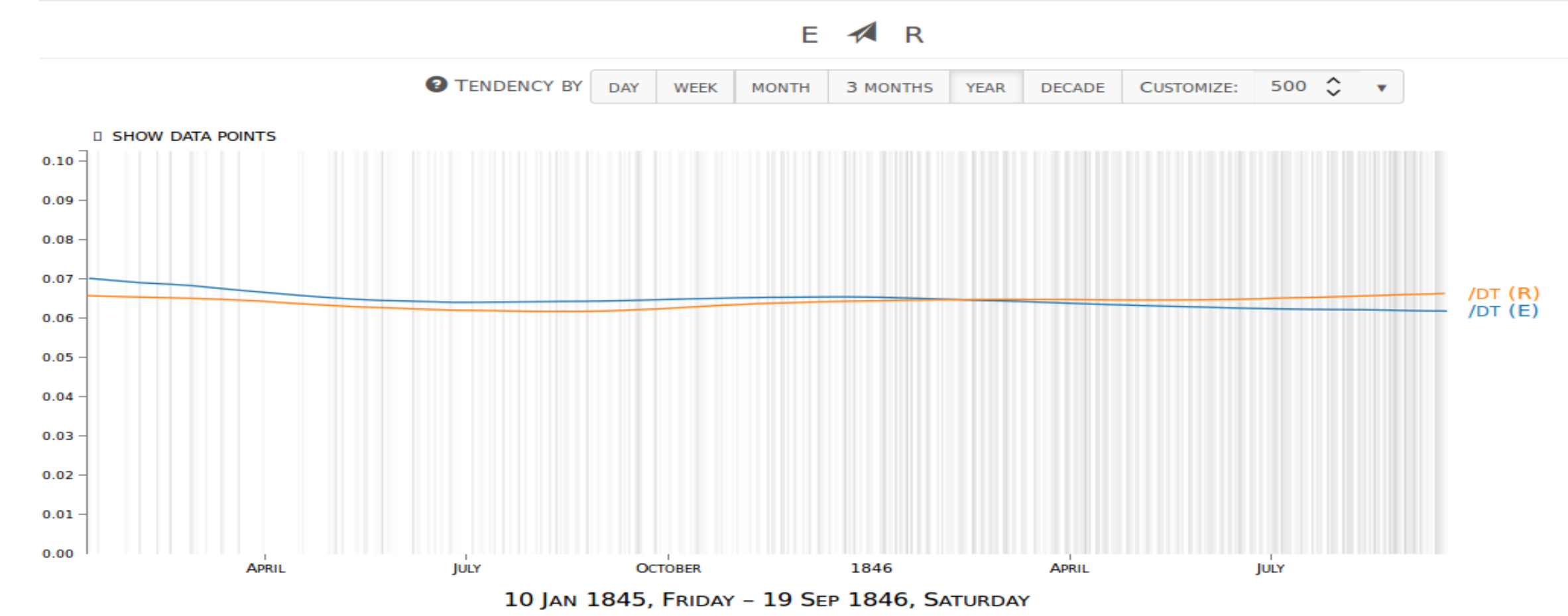


Figure 5: The use of determiners in the letters by Elizabeth Barrett and by Robert Browning (from the Brownings' corpus), with visible differences in the very beginning and in the very end of the correspondence.

Periods with significantly different values

We conduct statistical tests (χ^2 and Fisher) to determine if the frequencies observed in two subsets of the corpus for a given period, are significantly different from each other and we include the results in the visualization.

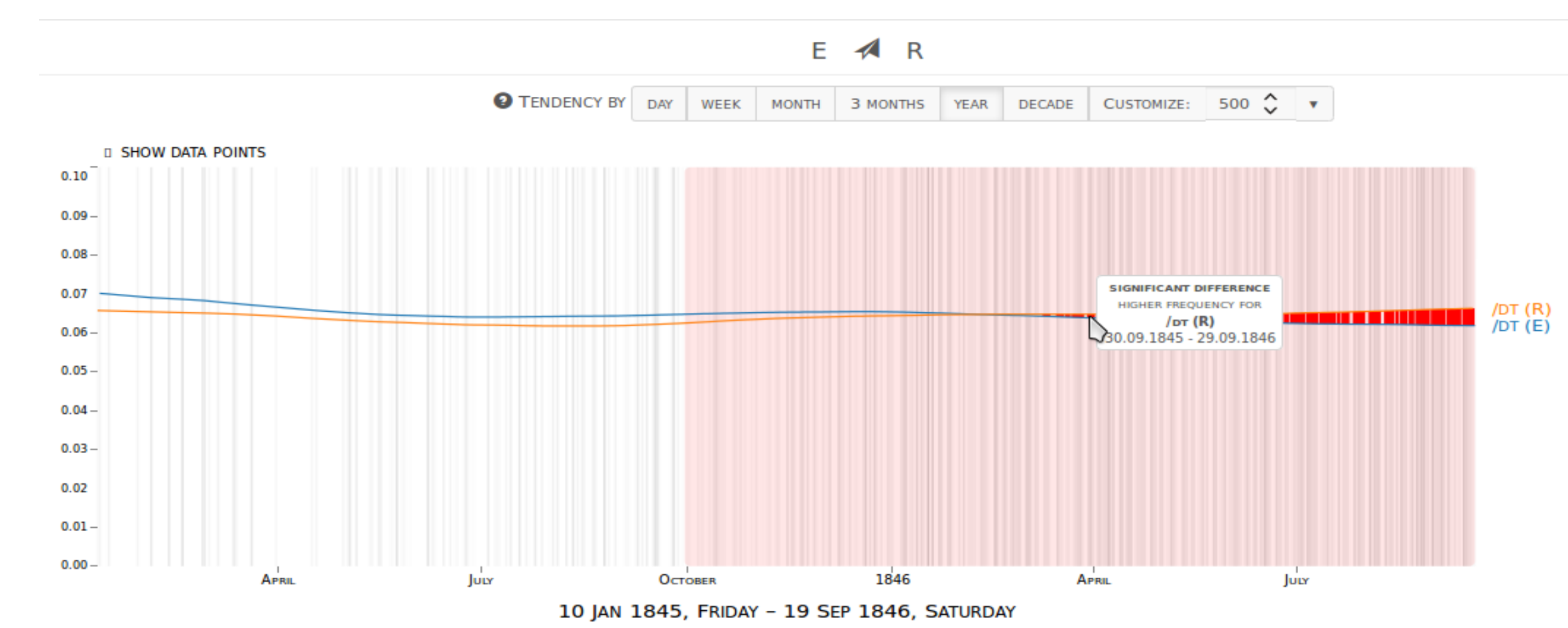


Figure 7: The same data as in Fig. 5. The centers of the periods when the frequencies in the two subsets were significantly different from each other are indicated by a line (|), connecting the corresponding points on the graphs. On mouse-over, an area appears in the background, representing the length of the period and details are shown in a tooltip.

Conclusions

Sometimes visualizations can mislead or hide information. We try to reduce this risk to a minimum for our n-gram viewer Slash/A by introducing a mechanism to detect potentially interesting periods and draw the user’s attention to them.

Future work

- We have to evaluate how much the additional filtering functionality is facilitating the linguistic research.
- Now the two filters can be applied one at a time. We need to test how useful it is to look at their results simultaneously, and eventually allow for this.
- Other filters might be introduced too, for example to show periods in which two or more tokens occur together more often than expected by chance.

Acknowledgements

We are grateful to Dr. Christopher Culy, who supervised the project, for the advice and support.

Reference

V. Todorova and M. Chinkina.
Slash/a n-gram tendency viewer.
ESSLLI 2014 Student Session Proceedings,
pages 229–239.

Download Slash/A

<http://linguistics.chrisculy.net/vistola/tools/slasha.html>

Contact us

todorova.slava@gmail.com
maria.chinkina@gmail.com