# Methods for Detection of Word Usage over Time

Ondřej Herman and Vojtěch Kovář

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{xherman1,xkovar3}@fi.muni.cz

**Abstract.** From a natural language corpus, word usage data over time can be extracted. To detect and quantify change in this data, automatic procedures can be employed.
In this work, I describe the application of ordinary and robust regression methods to time series extracted from natural language corpora.

**Key words:** word usage, time series, regression methods, Theil-Sen estimator, Mann-Kendall test

## 1 Introduction

Historically, linguists used to characterize languages based on their own experience and introspection. This methodology can only reflect the nature of an idealized, subjective model, which is inherently frozen in time, unlike the empiric reality of an everyday speech act.

The recent development of large corpora allows us to have a convenient and easily quantifiable view of language change based on actual evidence. The amount of this data is too large to sift through manually, so having a way to summarize it and pinpoint interesting behavior is desirable.

## 2 Time series analysis

A time series is a sequence of discretely spaced observations $(x_i, y_i)$, where $y_i$ is the observation for the time period $x_i$. In the following text, $x_i$ represents a period of time and $y_i$ the amount of appearances of a word over $x_i$ and $n$ is the amount of samples.

### 2.1 Linear regression

In simple linear regression, it is assumed that the true relationship between two variables, $x$ and $y$, is linear: $y_i = a + bx_i$. We are trying to estimate the unknown constants $a$, the slope, and $b$, the intercept. The values of $y_i$ are not

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2013, pp. 79–85, 2013. 

exactly known[1]: $y'_i = y_i + \epsilon_i$, where $\epsilon$ is an unpredictable error component and $y'_i$ is the value observed at $x_i$.

To estimate the values of $a$ and $b$ from a set of observations, the method of least squares [1,2] can be employed. That is, $\hat{a}$ and $\hat{b}$ such that the sum of squared errors $e = \sum_{i=1}^{n} \epsilon_i^2$ is minimal are to be found:

$$\hat{b} = \frac{\sum_{i=1}^{n} (y'_i - \bar{y}')(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{1}$$

$$\hat{a} = \bar{y}' - \hat{b}\bar{x} \tag{2}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

### 2.2 F-test

Even though the estimated parameters $\hat{a}$ and $\hat{b}$ are the best ones in the sense that they minimize the sum of squared errors, the chosen model might not actually describe the observations well. Namely, it is desirable to ensure that the slope of the regression line $\hat{b}$ is non-zero, and that its estimated value is significant compared to the random fluctuations present in the data. That is, the hypotheses to be tested are [1]

$$H_0 : \hat{b} = 0 \tag{3}$$

$$H_1 : \hat{b} \neq 0 \tag{4}$$

One way to obtain a test statistic for (3) is the F-test:

$$F_0 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)} \tag{5}$$

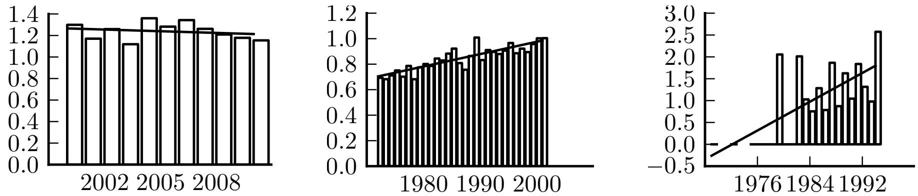
Assuming that the null hypothesis holds, $F_0$ follows the $F$ distribution with 1 and $n-2$ degrees of freedom, therefore the series is considered to exhibit a statistically significant trend when $|F_0| > F_{1-\alpha,1,n-2}$ and the null hypothesis is rejected.

The series shown in Figure 1(a) does not show any evidence of trend. On the other hand, the series in Figure 1(b) shows a very significant trend. According to the result of the F-test in the case of the series shown in Figure 1(c) also exhibit a trend, but its steepnes in this case seems to be caused by the limited volume of text contained in the early years sampled by the corpus and the resulting non-normality of the data.

---

[1] It is assumed that the values of $x_i$ are exactly known. Errors-in-variables models do away with this assumption.

[3] The unit of $y$ is the logarithm of the relative frequency per million words, for the reasons explained in 2.3




(a) Oxford English Corpus, $p = 0.514$[3]

(b) Google ngrams, $p = 4.3 \times 10^{-10}$

(c) British National Corpus, $p = 0.009$

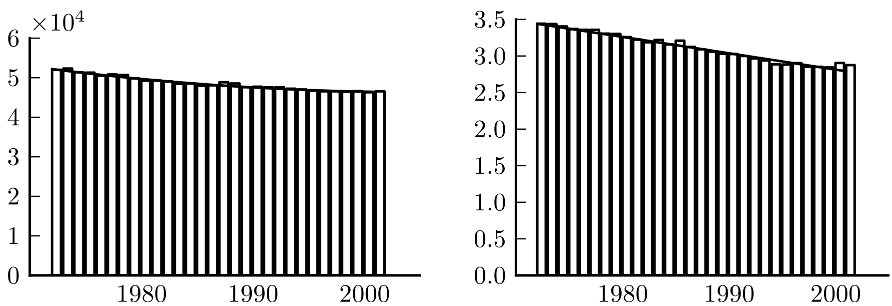
Fig. 1: Linear regression models and the respective p-values obtained using the F-test calculated for the word 'carrot'

### 2.3 Weighted linear regression

It is possible to extend the least squares method to fit a higher degree polynomial to the data, and also to weight the samples to account for heteroscedasticity[4]. As discussed in [3], this does not provide a significant improvement over the ordinary least squares.

The adjusted coefficient of determination $R^2_{adj}$ [1] can be used to find a suitable degree of the polynomial to fit to the time series. For most of the series examined, the value of $R^2_{adj}$ reaches the maximum for quadratic polynomials. Applying a logarithmic transformation linearizes the regression line, as can be seen in Figure 2. Treating the models as multiplicative therefore yields better results.

In almost all other cases, the higher-order models do not accurately describe the time series.




(a) the original series

(b) log-transformed series

Fig. 2: 'the' from Google Ngrams

[4] In heteroscedastic data, the sample variances are not equal.

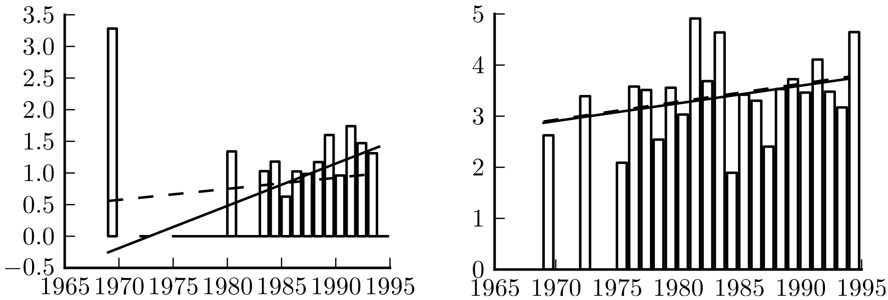
### 2.4 Theil-Sen estimator

The least squares methods are based on some assumptions that cannot always be met in practice. Namely that the error terms are normally distributed with known variances and mean zero. Rank-based robust methods do away with these requirements and are also less sensitive to the presence of outliers.

The Theil-Sen estimator [4,5,6,7] is a statistic used to estimate the slope of the regression line. It is model-free and non-parametric. The resulting estimate is a linear approximation of the trend line.

The Theil-Sen estimator is defined as the median of the pairwise slopes of the samples[8]:

$$\hat{\beta}_{ts} = \text{med}\, \frac{y_i - y_j}{x_i - x_j}, \quad i \neq j \tag{6}$$




(a) 'spice' (b) 'snow'

Fig. 3: Behavior of the Theil-Sen estimator for words encountered in the British National Corpus

As shown in Figure 3(a), outliers can easily confuse the ordinary least squares estimator represented by the dashed line, while the Theil-Sen estimator is able to ignore them and estimate the trend better.

On lower quality data, this estimator provides superior estimates of the slope compared to standard regression models. Another benefit is that it does away with the assumption that the data follows a predetermined model, so any monotonic trend can be estimated, therefore it works just as well even on non log-transformed data.

### 2.5 Mann-Kendall test

To test the significance of a model obtained using the Theil-Sen estimator, the Mann-Kendall test statistic [2,9] can be used:

$$S = \sum_{i=1}^{n} \sum_{j=1}^{i} \text{sgn}(x_i - x_j)\, \text{sgn}(y_i - y_j) \tag{7}$$

A top score of $S = \binom{n}{2}$ indicates that the series is increasing everywhere while $S = -\binom{n}{2}$ means that the series is decreasing.

Under the null hypothesis $S$ has the following properties[9]:

$$E[S] = 0 \tag{8}$$

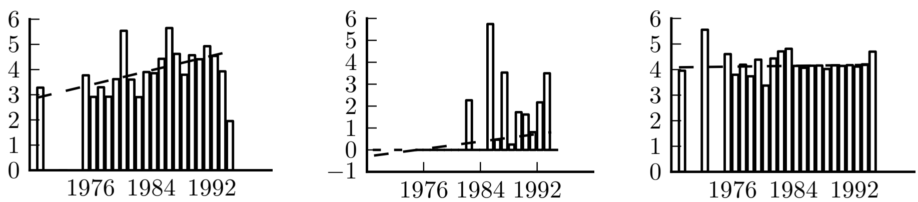
$$V[S] = \frac{n(n-1)(2n+5) - \sum_{i=1}^{n} t_i(i)(i-1)(2i+5)}{18} \tag{9}$$

where $t_i$ is the number of tied values in the $i$-th group[5]

The standardized[6] $Z$ statistic is computed as

$$Z = \begin{cases} \frac{S-1}{\sqrt{V[S]}} & S > 0 \\ 0 & S = 0 \\ \frac{S+1}{\sqrt{V[S]}} & S < 0 \end{cases}$$

The null hypothesis is to be rejected if $|Z| \geq u_{1-\frac{\alpha}{2}}$ at the significance level of $\alpha$, where $u_a$ is the quantile function of the standard normal distribution.




(a) 'oil', $p = 0.021$ (b) 'disk', $p = 0.009$ (c) 'slow', $p = 0.821$

Fig. 4: Words from the British National Corpus tested using the Mann-Kendall test with the trend line fitted using the Theil-Sen estimator

While a weighted linear model does not fit the series in the Figure 4(a) well (F-test $p = 0.24$), the $p$-value obtained using the Mann-Kendall test is considerably more significant. For the series in the Figure 4(b), the situation is similar: the linear model tests at $p = 0.67$. Interestingly, the slope calculated using the Theil-Sen estimator is, in this case, zero. No trend is found in the series in 4(c) by any of the methods. On well-behaved series the behavior of this test is comparable to the standard linear F-test.

The significance test based on Spearman's $\rho$ was also examined [3]. It behaves very similarly as the Mann-Kendall test [9,8,10].

[5] For example, the sequence [1,2,2,1,3,4,5,1,5,5] has 3 tied groups of lengths 3, 2 and 3.

[6] This statistic is only approximately normal.

## 3 Future work

The relationship between the word usage frequency and time is not inherently polynomial and would probably be better modeled as a sequence of possibly discontinuous linear segments.

Determining if and at which points the behavior of a time series changes is a well studied problem with a large body of research results available, such as [11], [12], [13] or [14]. These methods build on the framework described in this document and are likely to model the time series extracted from natural language corpora better than a single linear function.

## 4 Conclusion

The methods contained in this text were described with the potential application to the data from the Oxford English Corpus and the British National Corpus in mind.

Even though the ordinary regression models applied to log-transformed series work quite well, their use has more drawbacks than the robust methods have.

The most suitable method seems to be the Theil-Sen slope estimator, along with the Mann-Kendall or Spearman's $\rho$ tests to investigate a possible trend present in the word usage data.

**Acknowledgements** This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013.

## References

1. Montgomery, D., Johnson, L., Gardiner, J.: Forecasting and time series analysis. 2nd edition. McGraw-Hill (1990)
2. Forbelská, M.: Stochastické modelování jednorozměrných časových řad (Stochastic Modelling of one-dimensional time series, in Czech). Masarykova univerzita (2009)
3. Herman, O.: Automatic methods for detection of word usage in time (2013)
4. Rousseeuw, P.J., Leroy, A.M.: Robust regression and outlier detection. John Wiley & Sons, Inc., New York, NY, USA (1987)
5. Matoušek, J., Mount, D.M., Netanyahu, N.S.: Efficient randomized algorithms for the repeated median line estimator. Algorithmica **20** (1998) 136–150
6. Wilcox, R.R.: Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy. 2nd ed. Springer New York (2010)
7. Wilcox, R.: A note on the Theil-Sen regression estimator when the regressor is random and the error term is heteroscedastic. Biometrical Journal **40**(3) (1998) 261–268
8. Onoz, B., Bayazit, M.: The power of statistical tests for trend detection. Turkish Journal of Engineering and Environmental Sciences (27) (2003)
9. Neeti, N., Eastman, J.R.: A contextual Mann-Kendall approach for the assessment of trend significance in image time series. Transactions in GIS **15**(5) (2011) 599–611