Automatically Identifying Instances of Change in Diachronic Corpus Data

Andreas Buerki
Humboldt-Universität zu Berlin
andreas.buerki@hu-berlin.de

1 Introduction

With the increasing availability of diachronic corpora, machine-aided identification of linguistic items that have undergone significant change is set to become an important task. This importance is heightened further if, as Hilpert and Gries (2009:386) have argued, approaching linguistic change in a data-driven manner can reveal otherwise unnoticed phenomena. Key to this endeavour is being able to tell apart relevant change from noise and random or other synchronic variation. This non-trivial task differs in important ways from the much more widely investigated comparison of linguistic features between two (usually contemporary) corpora and has to date not received the attention it should perhaps be afforded.

In this paper, a number of methods for identifying relevant change are reviewed and a procedure suggested which has not so far been documented. This new procedure is based on a simple chi-square test for goodness of fit, combined with additional parameters. Its operation is illustrated using the example of a study conducted to investigate motivation of recent and ongoing change in Multi-word Expressions (MWEs) using data taken from the 20-million word Swiss Text Corpus (STC). The STC is a corpus of 20th century written German as used in Switzerland (Bickel et al 2009). Results of the application of the proposed method indicate that the procedure yields high-quality instances of significant change in the data and is applicable to MWEs as well as a range of other linguistic items. It is able to identify instances of change with fewer arbitrary decisions and able to identify a wider range of different types of change than other suggested methods. Additionally, it shows that both the structure of the data as well as particular research interests will guide the choice of method used to identify relevant change.

2 Goals, data and possible procedures

The method presented is designed to deal with data that consist of a large number of linguistic items and their frequencies in different time periods. An example for this type of data is seen in table 1 where one of nearly 18,000 MWEs extracted from the STC is shown with associated frequencies over five time periods.

Item	Frequencies in 5 periods
Im Laufe der Jahre *'in the course of time'*	23 26 76 35 31

Table1: Data structure

The method presented can be applied to any linguistic item as long as types and tokens can be extracted from corpus materials and counted (e.g., words, constructions, morphological features, etc.).

Before discussing ways to identify instances of change, it is pertinent to consider two preliminary questions. The first concerns what should be considered relevant change. For present purposes, a relevant instance of change is defined as an observable instance of change in the data which is not due to noise or accidental variation, but rather reflects what could reasonably be thought to indicate diachronic change in the language of which the corpus is a sample.

The second question concerns the types of change that should be identified. These are taken to be

1. the appearance of (new) types
2. the disappearance of (old) types
3. semantic shifts (stable form)
4. change in form (stable semantics)
5. notable in- and/or decreases in frequency.

Following other diachronic corpus studies, these five types of change are investigated from the vantage point of changes in frequency which, as will be demonstrated, can be used not only to identify change of types 1, 2 and 5, but also of types 3 and 4.

Only a small number of methods for identifying relevant shifts in frequencies have been suggested to date. These include the use of a coefficient of variance (CV) as applied in Baker (2011) and a rank-order correlation measure where frequencies are correlated with the sequence of time periods in the corpus (i.e. periods 1 to n) suggested in Hilpert and Gries (2009). A further possibility is to use Belica's (1996) coefficient of difference (D), the values of which can be squared for each period and then summed to arrive at an overall measure of change. A fourth option which has not so far been applied is to use a chi-square test for goodness of fit to test if the differences in frequency across the periods is significantly different from what could be expected due to chance.

To find the most useful method for purposes of identifying relevant change in the data outlined, all four methods were applied to MWEs extracted from the STC. In all cases, the following additional parameters were set: only MWEs were considered which showed a frequency of at least four occurrences per a million words in at least two of the five time periods (i.e. they could occur less often or

not at all in maximally three of five time periods). This ensured that items identified occured with notable frequency at one point, but also allowed for patterns where an item might have appeared or disappeared (or both) during the period of investigation. Further, frequencies were capped at three times the number of documents in which they occurred. This was used to prevent burstiness (caused, for example, by a topical concentration of certain items in individual documents) from unduly influencing the tracing of diachronic developments.

For the rank-order correlation method (we used Spearman's rho) and the method using the chi-square test, levels of significance were defined which provided a non-arbitrary divide between significant and non-significant change. For Spearman's rho, which was more restrictive, a significance level of alpha = 0.05 was used, for the chi-square test, a more stringent significance level of alpha = 0.001 was specified. For the other two methods, the highest scoring third of changes was considered to have undergone relevant change (cf. Baker 2011).

3 Results of the evaluation

Instances of relevant change identified by each of the four methods were compared by looking at individual test cases as well as the overall number of items identified. Table 2 shows the number of MWE-types identified as having undergone relevant change in each of the four methods. The total number of MWE-types occurring with minimum frequency in at least two time periods was 17,645.

Method	Types with relevant change
Top third (CV and D)	5,881
Spearman's rho	1,268
Chi-square	7,563

Table 2: Number of types with relevant change

The figures of table 2 show that the correlation method identifies the fewest types as having undergone relevant change. In fact, significance is only reached for perfect correlations (i.e. either a progressive in- or decrease in frequency over the five time periods). This is because the five data points provided by the temporal structuring of the source data provide insufficient statistical power; more data points would make this measure more meaningful, but such are not available in many cases.[1] The correlation based method, requiring a perfect rank-correlation is therefore too narrow to be useful for the data structures described. Comparing the detailed results of the remaining methods shows that in a number of instances the chi-squared based method appears to make more sensible decisions than the other two. Unlike the other two remaining methods, it also provides a non-arbitrary cut-off point for identifying significant change. A cut-off point, furthermore, which is easily interpretable: items whose frequencies differ in ways that are unlikely to be due to chance are identified as having undergone relevant change. This links in well with the definition of relevant change given above.

[1] Neither is it possible to apply the more sophisticated methods suggested by Hilpert and Gries (2009), for the same reason.

4 Conclusions

In the study of MWE-change in which the methods were evaluated, the chi-square-based approach served as the most useful method for identifying relevant change among a vast number of potential changes. It showed important advantages over other possible procedures. The chi-square-based method was shown to be 1) broad enough to include a variety of patterns, rather than only a progressive in- or decrease in frequency, 2) to fit well with the definition of relevant change used, 3) to provide a statistically robust, non-arbitrary cut-off point and 4) it was well suited for application to data that cannot provide a large number of data points. The additional parameters used (an item needing to appear in at least two periods with a minimum frequency of 4/M and the cap on frequencies) added to the robustness of results and therefore their usefulness which was confirmed when a sample of MWEs identified as having undergone significant change was investigated in detail to establish motivations for change. While a comprehensive and broad identification of all significant change in the data was advantageous in the application reported on, for other purposes, a more selective method might be desired. In such cases, a measure such as the coefficient of variance used in Baker (2011) could additionally be used to rank change and limit investigation to top-ranking cases.

References

Baker, P. (2011). Times may change, but we will always have money: Diachronic variation in recent British English. *Journal of English Linguistics*, *39*(1), 65-88.

Belica, C. (1996). Analysis of temporal changes in corpora. *International Journal of Corpus Linguistics*, *1*(1), 61-73.

Bickel, H., Gasser, M., Häcki Buhofer, A., Hofer, L., & Schön, C. (2009). Schweizer Text Korpus. *Linguistik Online*, *39*(3), 5-31.

Hilpert, M., & Gries, S. T. (2009). Assessing frequency changes in multistage diachronic corpora. *Literary and Linguistic Computing*, *24*(4), 385-401.