Identifying Relevant Change in Diachronic Corpora: 
a perspective from research into formulaic language

Andreas Buerki
University of Basel and Humboldt Universität zu Berlin

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 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 short presentation, we will be focussing on identifying relevant frequency changes, but, as will be demonstrated, changes in frequency often also provide leads to others kinds of change such as change of linguistic form or semantic change. After outlining a number of methods for identifying relevant change, the implementation of one particular approach, based on a simple chi-square test for goodness of fit, is discussed in more detail. Since in this context chi-square assigns high scores also ‘if only one of the frequencies is significantly different to the others’ (Baker 2011: 70), which may be undesirable, a simple coefficient of variance (i.e. the standard deviation as a percentage of the mean) can be used to rank (or filter) results further. Practical challenges that arise in the application of this approach and how they might be overcome will be discussed, as well as sample results.

The implementation presented is concerned with the identification of relevant change in units of formulaic language (also referred to as multi-word units or multi-word expressions) which adds some extra challenges of its own. Other types of linguistic items of which the frequency can be measured should, however, be equally amendable to this method. Examples are based on data taken from the Swiss Text Corpus (Bickel et al 2009), a 20-million word reference corpus of written German spanning the period from 1900 to 2000.

References:

