DIACRAN: a framework for diachronic analysis

Adam Kilgarriff
Lexical Computing Ltd., Brighton, UK

Jan Busta
Lexical Computing Ltd., UK and Masaryk Univ, Cz

Pavel Rychlý
Lexical Computing Ltd., UK and Masaryk Univ, Brno, Cz

{adam.kilgarriff,jan.busta,pavel.rychly}@sketchengine.co.uk

Many of the questions that linguists want to explore concern language change, or diachronic analysis. We present Diacran, an implemented system for corpus-based diachronic analysis.

We view diachronic analysis as a special case of keyword-finding. In keyword-finding we want to find the words (or terms, or collocations, or grammatical structures …) that are most characteristic of one text type (or dataset, or corpus) in contrast to another. In diachronic analysis, we usually want to start by finding the words (or terms, etc; hereafter we say just 'word') that have changed most over time. The ingredients for keyword analysis are corpus1, corpus2, and a formula for ranking how interesting each word is. For Diacran, the ingredients are a corpus with at least three 'time-slices' - that is, with documents dated according to at least three different points in time so the corpus can be sliced into three or more subcorpora, each associated with a different time - and, again, a ranking formula.

As in keyword analysis, the challenge for the computational linguist is of getting the 'best' list, where 'best' means the list of (say) the top 500 items, with the largest numbers of items judged interesting (from a text-type, or diachronic, point of view) by a human expert.

The method is this. First we divide the corpus into subcorpora, one for each time slice. Then we normalize the frequency for each word in each time slice, to give frequencies per million words.[1] We then plot a 'best fit' graph, for each word, of change over time, using standard techniques such as linear regression and Theil-Sen gradient estimation.

The 'most interesting' of these graphs have three characteristics:

- high gradient (positive or negative) of the line
  - because we are most interested in words that have changed a lot
- high correlation
  - because we are most interested in words that have changed and stayed changed, not bounced around
- high frequency of the word overall
  - because we are more interested in words where the frequencies in the (say, five) time slices are <100, 200, 300, 400, 500> rather than <1, 2, 3, 4, 5>. The latter is likely just to be noise, whereas for the former, we have ample evidence of a systematic and substantial change.

We then combine the scores on these three factors, to give an overall score for each word. The words with the highest combined scores are the 'most interesting' words, to be shown to a linguist for expert consideration. Diacran is implemented within the Sketch Engine (Kilgarriff et al. 2004) and the expert is supported in their task by 'click through': they can click on an item in the candidate list to see the concordances for the word. They can also see other analyses to show how usage differs between time-slices, within the Sketch Engine, which has a wide range of analysis tools.

The approach should prove useful for various kinds of diachronic analysis. We are using COCA (Davies 2009) and a corpus of blog and newspaper feeds that we have gathered over the last ten years, as test sets.

An ideal test set would give us the 'right answers' so we knew when our system was doing a good job. We are currently searching for datasets that might support threshold-setting and evaluation for Diacran.

Neologisms

The highest-profile kind of diachronic analysis is neologism-finding, particularly by dictionary publishers, where the year's new words are featured in the national press. We are exploring using the set of new words, as added to a dictionary by a dictionary publisher, as the 'ground truth' of the words that our system should put high on the list.

A feature of neologism-finding, particularly for brand-new words (as opposed to new meanings for existing words) is that frequencies, even in very large corpora, will tend to be very low. A sequence of frequencies, over the last five years, of <0, 0, 1, 0, 2> for a word may count as enough to suggest a candidate neologism, that came into existence three years ago. This presents a technical challenge since

[1] Another option is to classify a word as present or absent in a document, and to work with counts for each word per thousand documents. This is often preferable, as we do not wish to give extra weight to a word being used multiple times in a single document. Diacran offers both options.

there are also likely to be many typographical errors and other noise items with profiles like this. It also points to the merits of working with very large corpora, since, the larger the numbers, the better the prospects for using statistics to distinguish signal from noise.

Background and Related Work

The traditional way to find neologisms is 'reading and marking'. Lexicographers and others are instructed to read texts which are likely to contain neologisms – newspapers, magazines, recent novels – and to mark up candidate new words, or new terms, or new meanings of existing words. This is the benchmark against which other methods will be measured. It is a high-precision, low-recall approach, since the readers will rarely be wrong in their judgments, but cannot read everything, so there are many neologisms that will be missed.

For a dictionary publisher, one reading of 'neologism' is 'words which are not in our dictionary (yet)'. Of course words may be missing from dictionaries for many reasons, of which newness is one (and simple oversight is another). On this reading, one kind of neologism-finding program identifies all the words in a corpus (over a frequency threshold) that are not in a particular dictionary. Corpora have been used in this way to mitigate against embarrassing omissions from dictionaries since large, general corpora (for example, for English, the British National Corpus[2]) became available, in the late 1980s and 1990s. Note that this process has complexities of its own, and where the language has complex morphology, identifying the word forms not covered by the lemmas in the dictionary is far from simple.

There are some 'lexical cues' that speakers often use when introducing a word for the first time: "so-called", "defined as", "known as". In writing, the language user might put the new item in single or double quotation marks. One kind of corpus strategy for identifying neologisms looks for items that are marked in these ways. An implemented system for English, which shows these methods to be strikingly useful, is presented by Paryzek (2008).

The approach is extended for Swedish by Stenetorp (2010) who starts from lists of neologisms from the Swedish Academy and Swedish Language Council, and develops a 'supervised' machine learning system which finds features of neologisms *vs.* non-neologisms, and can then classify new items as neologism-like or not. Stenetorp uses a very large corpus of documents each with a time stamp, as do we.

O'Donovan and O'Neil (2008) present the system in use at Chambers Harrap at the time for identifying neologisms to add to the dictionary, so is of particular interest as a system which, in contrast to the academic ones, is used in earnest by a publisher. One component of the software suite builds a large time-stamped corpus; another, the word-tracking component (based on Eiken 2006) identifies items which have recently jumped up in relative frequency; and a third, echoing the third of our criteria above, promotes higher-frequency items so they will appear higher in the lists that lexicographers are asked to monitor.

Gabrielatos et al. (2012) present an approach to diachronic analysis similar to ours, but focusing on one specific sub-issue: what are the most useful time-slices to break the data set up into. There is usually a trade-off between data sparsity, arguing for fewer, fatter time-slices, and delicacy of analysis, which may require thinner ones. We hope to integrate the lessons from their paper into the options available in Diacran.

[2] http://www.natcorp.ox.ac.uk

References

Davies, M. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159-190.

Eiken, U. C., Liseth, A. T., Witschel, H. F., Richter, M., and Biemann, C. 2006. Ord i dag: Mining Norwegian daily newswire. In *Advances in Natural Language Processing* (pp. 512-523). Springer Berlin Heidelberg.

Gabrielatos, C., McEnery, T., Diggle, P. J., & Baker, P. (2012). The peaks and troughs of corpus-based contextual analysis. *International journal of corpus linguistics*, *17*(2), 151-175.

Kilgarriff, A., Rychlý, P., Smrz, P., Tugwell, D. 2004. The Sketch Engine. Proc. EURALEX. pp. 105–116.

O'Donovan, R., & O'Neil, M. 2008. A systematic approach to the selection of neologisms for inclusion in a large monolingual dictionary. In *Proc 13th Euralex International Congress* (pp. 571-579).

Paryzek, P. 2008. Comparison of selected methods for the retrieval of neologisms. Investigationes Linguisticae XVI, Poznan, Poland.

Stenetorp, P. 2010. Automated extraction of swedish neologisms using a temporally annotated corpus. Masters' thesis, KTH (Royal Institute of Technology) Stckholm, Sweden.