References

Aijmer, Karin 2009. The pragmatics of adverbs. In Günter Rohdenburg & Julia Schlüter (eds.), *One language, two grammars? Differences between British and American English.* Cambridge: Cambridge University Press. 324-340.

Amador-Moreno, Carolina P. 2010. *An Introduction to Irish English*. London: Equinox.

Barron, A. And Schneider, K. 2005. *The Pragmatics of Irish English.* Berlin: Mouton de Gruyter.

Corrigan, K. P. 2010. *Irish English, Volume 1 – Northern Ireland*. Edinburgh: Edinburgh University Press.

Nevalainen, T. & H. Raumolin-Brunberg 2003. *Historical sociolinguistics. Language change in Tudor and Stuart England.* London: Longman.

Schneider, E.W. 2002. Investigating variation and change in written documents. In P. Trudgill, J.K. Chambers & N. Schilling-Estes (eds.), *The handbook of language variation and change*. Oxford: Blackwell. 67-96.

Tagliamonte, Sali A. 2012. *Roots of English. Exploring the history of dialects.* Cambridge: Cambridge University Press.

Developing *AntConc* for a new generation of corpus linguists

Laurence Anthony
Waseda University
anthony@waseda.jp

1 Introduction

Concordance software is one of the most important but often forgotten components of any corpus study. Over the years, various concordance software tools have been released in what McEnery and Hardie (2012) have described as four generations of tool development. Today, both 3rd-generation software tools, such as *AntConc* (Anthony, 2012), *WordSmith Tools* (Scott, 2012), and *MonoConc Pro* (Barlow, 2000), and 4th-generation web-based tools, such as those hosted at *byu.corpus.edu* (Davies, 2012) are popular choices for most corpus research work.

In recent years, *AntConc* (Anthony, 2012) has seen a rapid growth in popularity among researchers, teachers, and language learners due to its rich set of features, freeware license, multiplatform support, and easy-to-use interface. For researchers, *AntConc* performs speedily and accurately on a wide-range of small and mid-sized corpora. It also offers flexible handling of tags, metadata, and language encodings, and provides a wealth of functions and features. In 2012 alone, the software was downloaded over 120,000 times by users in over 80 countries, and it has become one of the software tools of choice in many corpus linguistics departments looking to introduce students to corpus linguistics through a free and easy tool (Anthony, 2012). For teachers and learners, *AntConc* can perform basic operations, such as producing KWIC concordance lines and keyword lists, in a quick and easy way. Also, it can be used both inside the classroom and as part of student homework projects on Windows, Macintosh OS X, and Linux computers. Finally, to motivate learners to use corpora in their learning, it offers a modern and attractive-looking interface.

Although *AntConc* has many strong features, it also has a number of weaknesses when compared to the most popular web-based and commercial tools. To address these issues, various design and performance improvements have been introduced in the latest version of the software. In this paper, I first review the current status of corpus analysis tools discussing their respective strengths and weaknesses, and explaining the motivation to introduce changes to *AntConc*. Next, I describe the changes introduced in the latest version of *AntConc*. As part of the discussion, I explain the choice of programming

language, the importance of including a flexible and explicit token definition, and the approach used by *AntConc* to handle whitespace issues that provides far greater transparency and flexibility over other tools. Many of these changes are directly relevant not only to corpus linguists but also teachers and learners who use corpus tools as part of a Data-Driven Learning (DDL) approach.

2 Current status of corpus analysis tools

Recently, Tribble (2012) conducted a major survey of the most popular tools used by corpus linguists around the world. Based on responses from 891 linguists, he showed that three tools are predominantly used today: *corpus.byu.edu*, *WordSmith* Tools and *AntConc* (see Figure 1). Viewing these results, it is apparent that the most popular tools are fast, easy-to-use, and feature-rich.

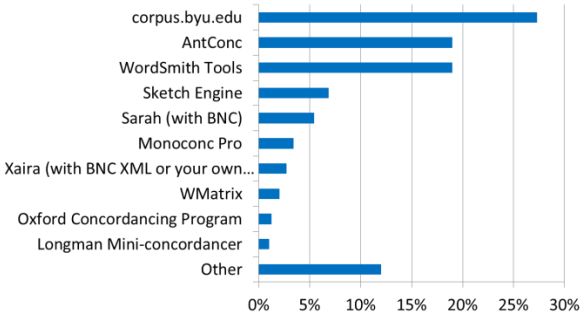

Figure 1. Survey results in response to the question: "Which computer programs do you use for analysing corpora?" Reponses: 891. (Tribble, 2012)

On the other hand, Tribble also reports that more advanced tools are being increasingly desired by corpus researchers. As a result, there is a growing interest in software development and coding, for example, using the R statistical package.

From the results of Tribble's survey, it is clear that for many corpus linguists a number of areas need addressing. Firstly, most web-based tools (e.g. corpus.byu.edu) provide only a window to a general English language corpus and offer no access to the raw corpus data. This makes it difficult to use them in many situations, for example, when the researcher is interested in the specialized language use in English for Specific Purposes (ESP) research. In this context, general corpus tools, such as *WordSmith Tools* (Scott, 2012) and *MonoConc Pro* (Barlow, 2000) can be used successfully as they can process almost any corpus for which the raw data is available. On the other hand, these tools come with a restrictive commercial license that often prevents their use in countries and regions where budgets are limited. In particular, they are difficult to use outside of the classroom, for example, as part of student homework projects. Also, statistical information is becoming increasingly important in corpus work, but the current tools do not easily generate the latest statistical results needed by researchers.

AntConc is a freeware tool that is able to process raw corpus data of various kinds. As a result, it can be useful in many contexts. However, in recent years, *AntConc* has begun to fall behind other tools in terms of speed, mainly due to its database architecture. *AntConc* processes all data in active memory and uses only a primitive indexing system. In addition, some common computer operations, such as drag-and-drop, are not available due to aging nature of the programming languages used to develop the software, i.e., *Perl* and *TclTk*. Various other limitations have recently become apparent, such as its limited support for handling annotated files and its limited statistical functions. In view of these limitations, an effort to update *AntConc* has been undertaken over the past three years. The results of this work are described in the following section.

3 New Features in *AntConc*

Programming language: This biggest change to *AntConc* has been to recode the software in the *Python* programming language together with the *Qt* graphical user interface package. The use of Python allows more advanced statistics modules to be included directly in the software thus addressing a major weakness in previous versions. Also, the use of *Qt* allows *AntConc* to adopt a more modern appearance and enables standard computer operations, such as drag-and-drop, to be incorporated. The use of *Qt* also allows rich-text tables to be utilized, leading to fast rendering of color-highlighted results.

Database architecture: The new version of *AntConc* incorporates a SQLite backend database that can operate in an indexed mode or on plain text files. This gives the program the ability to search for results on much larger, multi-level annotated files, while also being able to search in plain text files using regular expressions.

Multi-language support: *AntConc* has always offered multi-language support. However, the new version extends this to include flexible definitions of words that can extend or replace various Unicode character classes. It can also handle cross-platform line breaks and various forms of whitespace in and between text lines.

Performance improvements: Various other coding changes have led to the latest version of *AntConc* performing at greatly increased speeds over previous versions. This has led to some operations that took minutes to perform in the past completing in a matter of seconds or fractions of a second. The

software no longer has a sluggish feel and can handle very large corpora of 100s of millions of words without problem. A screenshot of the new software is shown in Figure 2.

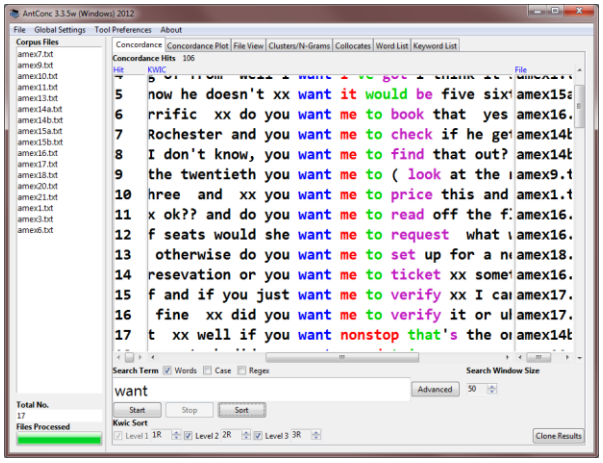

Figure 2. Screenshot of latest version of *AntConc*. (Anthony, 2012)

4 Conclusions

Users of *AntConc* should find the latest version to be much improved over previous versions. It is hoped that these changes will enable to software to continue meeting the needs of the corpus linguistics community.

References

Anthony, L. (2012). *The Past, Present, and Future of Software Tools in Corpus Linguistics*. Presentation given at KACL 2012, Busan, Korea.

Anthony, L. (2012). *AntConc* (Version 3.3.5) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/.

Barlow, M. (2000). *MonoConc Pro* [Computer Software]. Available from http://www.athel.com/mono.html.

Davies, M. (2012). corpus.byu.edu.

McEnery, T and Hardie, A (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Scott, M. (1996). *WordSmith Tools* [Computer Software]. Available from http://www.lexically.net/software/index.htm.

Tribble, C. (2012). T*eaching and Language Corpora: Quo Vadis?* 10th Teaching and Language Corpora Conference (TALC). Warsaw, 11th-14th July 2012.

Bridging lexical and constructional synonymy, and linguistic variants – the Passive and its auxiliary verbs in British and American English

Antti Arppe
University of Alberta
arppe@ualberta.ca

Dagmara Dowbor
University of Alberta
dowbor@ualberta.ca

1 Introduction and Background

In the case of *constructional alternations* – or *synonymous syntactic variants* – studies utilizing multivariate/multicausal predictive models have been recently undertaken, though primarily for binary/dichotomous alternations of only two outcomes, and typically for English, e.g. the possessive alternation: X of Y vs. Y's X (Gries 2002; Rosenbach 2003), verb particle placement (Gries 2003a), and the dative alternation: GIVE NP NP vs. GIVE NP PP (Gries 2003b; Bresnan 2007), Exeeptions scrutinizing polytomous settings with more than two outcomes are various studies on German word order variation and the German active vs. *werden*-passive vs. *bekommen*-passive (Bader & Häussler). Inspired by the latter study, Arppe (2011) has scrutinized the English four-way constructional alternation of active vs. *be*-passive vs. *get*-passive vs. *become*-passive, which can be seen to merge lexical with constructional alternation, using British English data (British National Corpus: BNC) (see examples 1-4):

(1) BECOME: ... how the siege$_{\text{PATIENT}}$ became interpreted by today's protestant loyalists$_{\text{AGENT}}$ ... [original sentence in BNC]
(2) GET: ... how the siege got interpreted by today's protestant loyalists ...
(3) BE: ... how the siege was interpreted by today's protestant loyalists ...
(4) ACTIVE: ... how today's protestant loyalists interpreted the siege ...

Based on a statistical multivariate analysis of the British English corpus data using *polytomous logistic regression* (Arppe 2008, 2012) on a range of contextual predictors, (a) active constructions were observed in the British English corpus data to be significantly associated with having an explicitly expressed AGENT or other argument which could be turned into subject in a corresponding active construction, or a co-ordinated verb, (b) *be*-passive constructions with ABSTRACTIONs, ACTIVITIES/EVENTs, ARTEFACTs, [forms of] COMMUNICATION, HUMAN GROUPs or [forms of] POSSESSION as PATIENTs (i.e. grammatical subjects