ProtAnt

A tool for analysing the prototypicality of texts

Laurence Anthony and Paul Baker
Waseda University / Lancaster University

Corpus-based researchers and traditional qualitative researchers, such as those interested in critical discourse analysis, are often required to select prototypical texts for close reading that include the language features of interest that are present in a much larger corpus. Traditional approaches to this selection procedure have been largely ad hoc. In this paper, we offer a more principled way of selecting texts for close reading based on a ranking of texts in terms of the number of keywords they contain. To facilitate this analysis, we have developed a multiplatform, freeware software tool called *ProtAnt* that analyses the texts, generates a ranked list of keywords based on statistical significance and effect size, and then orders the texts by the number of keywords in them. We describe various experiments that demonstrate the *ProtAnt* analysis is effective not only at identifying prototypical texts, but also identifying outlier texts that may need to be removed from a target corpus.

Keywords: *ProtAnt*, critical discourse analysis, prototypicality, keywords, qualitative research

1. Introduction

Corpus-based researchers and traditional qualitative researchers are often required to select texts for close reading that include the language features of interest present in a much larger corpus. Ideally, the texts selected for this close reading should be prototypical, i.e. the clearest, best, most typical, most representative examples (Labov 1973, Rosch 1975, Gries 2003). However, in reality, such selections are difficult to make and thus researchers can sometimes be criticised for "cherry-picking" texts from the larger body of work that illustrate a preconceived idea or point. Critical discourse analysts are particularly targeted for this type of criticism (Widdowson 2004) and even corpus-based discourse analysts can be criticized

International Journal of Corpus Linguistics 20:3 (2015), **273–292**. DOI 10.1075/ijcl.20.3.01ant
ISSN 1384–6655 / E-ISSN 1569–9811 © John Benjamins Publishing Company

for “cherry-picking” as the hypotheses they develop from a top-down analysis of a corpus are usually validated using a bottom-up close reading of a few carefully selected texts from the corpus. To avoid such criticisms, researchers could choose to randomly select texts from a corpus or dataset (e.g. Sajjid 2013), but such a selection procedure may not reveal infrequent but nevertheless important features identified at the top-down level. Alternatively, researchers may attempt a close reading of the corpus texts from the first text to the last. However, this approach is likely to distort perceptions and/or take an unduly long time.

In this paper, we propose a more principled way of selecting texts for close reading based on a ranking of texts in terms of the number of keywords (unusually frequent words in the target corpus compared with a reference corpus) they contain. To facilitate this analysis, we have developed a multiplatform, freeware software tool called *ProtAnt* (Anthony & Baker 2015) that analyses the texts, generates a ranked list of keywords based on statistical significance and effect size, and then orders the texts by the number of keywords in them.

In the next section, we summarize earlier work on the identification of prototypical texts. In Section 3, we explain the design and functions of the *ProtAnt* tool. In Section 4, we describe various experiments that show how the *ProtAnt* analysis is able to correctly rank both short and long texts with known levels of prototypicality. In this section, we also describe an experiment that shows how the *ProtAnt* analysis can be used to identify outlier or miscategorised texts that may need to be removed from a target sample prior to the close reading process. In Section 5, we finish the paper by suggesting potential applications of the software in research, teaching, and learning and possible directions for future development of the *ProtAnt* tool.

2. Text prototype selection methods

To date, language and discourse researchers have mainly adopted two approaches to select texts that exemplify target features of the language variety they are interested in and thus serve as candidates for close inspection. The first and most common approach may be described as ‘opportunistic selection’ and involves making an arbitrary yet hopefully informed choice of texts to analyse. In Wodak’s (2013) edited collection of critical discourse studies, this approach can be illustrated in the work of Caldas-Coulthard et al. (2003/2013: 40), Chouliaraki (2000/2013: 100), and Machin & Suilman (2006/2013: 229). Caldas-Coulthard et al. (2003/2013) describe their selection of target texts on teddy bear discourse as follows: “[…] we purchased all the 15 bear books available in a local children’s book store in London.” Chouliaraki (2000/2013) is interested in the way news broadcast practices implicitly produce hegemonic political positions. To investigate this topic, they select a single news broadcast from August 16, 2000 that reports on the killing of a demonstrator during a riot. Machin & Suilman (2006/2013), on the other hand, are interested in the effect of games on political discourse. Although their target discourse is not textual, they adopt the same selection procedure when they choose just two games for their study, the American Delta Force game and the Hizbollah Special Force game, despite a wide range of similar games existing in the market.

When the researcher has an extensive experience and knowledge of the target area and language, this 'opportunistic selection' is no doubt effective. However, the procedure can in no way be described as principled, which leads to questions concerning the bias of the researcher, the implications of the findings, and the ability to replicate the study.

The second and far less common approach to text selection among language and discourse researchers may be described as 'selective downsizing' and is illustrated in the work of Khosravinik (2010). As part of a joint ESRC-funded project on the representation of immigration in the British press, Khosravinik (2010) takes a 140 million word corpus of 170,000 articles, which was earlier analysed by Gabrielatos & Baker (2008) using corpus-based approaches, and selects sets of articles from five one-week periods where the number of articles about immigration peaked. For each period, all the articles pertinent to immigration were taken from one liberal quality newspaper, one conservative quality newspaper, and one tabloid newspaper. This resulted in 439 articles which Khosravinik (2010) then closely reads and examines in terms of topoi (argumentative strategies) as well as some of van Leeuwen's (1996) socio-semantic categories like aggregation, collectivisation, functionalization and humanisation and individualisation.

A number of studies adopting a similar 'selective downsizing' approach appear in Wodak's (2013) edited collection, such as the study of argumentative discourse by Ehrlich & Blum-Kulka (2010/2013) and the study of Israeli peace discourse by Gavriely-Nuri (2010/2013). Unfortunately, no details are given about the size of the "large corpus" downsized in Ehrlich & Blum-Kulka's (2010/2013: 149) work, or the precise number of metaphor uses in the "hundreds of mentions of metaphor" from which downsizing occurs in Gavriely-Nuri's (2010/2013: 226) work.

Clearly, the 'selective downsizing' approach is more principled than the 'opportunistic selection' approach or an even more simplistic approach that relies on choosing texts based on their order in a pre-compiled corpus or the degree to which a text might "look interesting" to the analyst. Despite this, the approach can still result in a large number of texts that require close-reading, as in the case of Khosravinik's (2010) approach, or a small number of texts that may appear to be “cherry-picked” to some extent, as in the case of Ehrlich & Blum-Kulka’s (2010/2013) and Gavriely-Nuri’s (2010/2013) work.

The issue of appropriate text selection is an important consideration in organizations that have vast repositories of textual data and need to quickly find texts for management purposes, decision making, or customer profiling (Durfee et al. 2007). As a result, the selection of prototypical texts has become an active area of research in natural language processing (NLP). Visa et al. (2001) and Kloptchenko et al. (2002, 2004) adopt an approach that initially requires human judgment to pick prototypical texts from a larger corpus that represent different text types (or classes). Next, they analyse the prototype texts in terms of word and sentence structure, i.e. word order and paragraph structure, to create a profile of the prototype texts and the remaining corpus texts. Next, they classify all the corpus texts into the different classes based on their distances from the initial prototype. Many other machine learning algorithms also employ a similar concept of prototypically. The nearest neighbour classifier (e.g. Manning et al. 2008), for example, requires a set of known samples (prototypes) from particular classes from which it builds a set of features (e.g. words for text classification). Then, new documents are assigned to a class dependant on how “near” they are to the training documents in a particular class. Chen et al. (2011) also use a prototype method to classify newsgroup posts into particular categories. A useful review of prototype-based classifiers can be found in Fayed et al. (2007).

Fayed et al. (2007) propose a classification method that still requires a set of training samples of a particular class but also utilizes an algorithm to auto-select the most prototypical sample from that set. These self-generated prototypes are then utilized in the classification of future unseen samples. Bahrololoum et al. (2012) propose one such prototype detector that is based on a so-called gravitational search algorithm, which models features in the training samples (e.g. words in a text dataset) as small masses that interact with each other and cluster together.

The NLP methods described above are dependent on the number of classes perceived to exist and in the case of Visa et al.’s (2001) and Kloptchenko et al.’s (2002, 2004) work, they rely on prototypical text (or texts) selected at the initial stage. Despite these limitations, NLP approaches offer a researcher a way to detect many texts that are similar in nature to the prototype(s) and hence partially overcome the initial selection bias. However, in the task of determining prototypical texts for a detailed qualitative analysis, these approaches are perhaps less attractive. First, the researcher either needs to select a set of one or more prototypical texts (the motivation for our current research) or at least define the possible classes or groups that a text may be classified into. The mathematics behind these classification approaches can also be extremely complex rendering the methods as “black-boxes” to subsequent users. A further problem is that these methods group texts based on their general profile rather a particular set of unique features that might be of interest to qualitative researchers (e.g. in critical discourse analysis).

It should be noted that there are also NLP techniques that can cluster texts into groups in an unsupervised way (i.e. without relying on any pre-defined categories), but again, these techniques rely on the general profiles of texts. They can also be highly sensitive to the clustering approach adopted (e.g. suggesting very different groups if clustered from top down or bottom up), pose problems for the user when attempting to label these groups, and again, may appear to the user as "black-boxes" (Manning et al. 2008). Also, while it is possible to identify prototypical texts within these groups, the methods of identification often rely on simple approaches such as treating all samples in a group as prototypes or picking a sample that represents the mean or "centre" of the group when plotted in a geometric space (for a discussion, see Fayed et al. 2007).

In the next section, we propose a novel approach to prototype text selection that does not require prerequisite groups or an advanced knowledge of mathematics to understand the methodology. Our approach is also tailored to qualitative studies that target unique text features and is implemented as a single-file, portable software tool that requires no advanced training to use.

3. *ProtAnt* analytic tool

ProtAnt is a freeware, portable software tool designed to profile corpus texts and rank the texts by the degree to which they are prototypical of the corpus as a whole using the concept of keywords. We hypothesise that a text which contains a greater number of keywords from the corpus as a whole is also likely to be a more central or typical text in that corpus.

Keywords are generally conceived as words that occur statistically significantly more frequently in one corpus than in a reference corpus. In this sense, the keywords themselves can be considered as distinctive of the target corpus. Various statistical tests have been used to determine keywords, with log-likelihood becoming a de facto standard in much of corpus work due to its presence in popular concordance tools, such as *AntConc* (Anthony 2014) and *WordSmith Tools* (Scott 2014). Once determined, keywords can also be ranked.

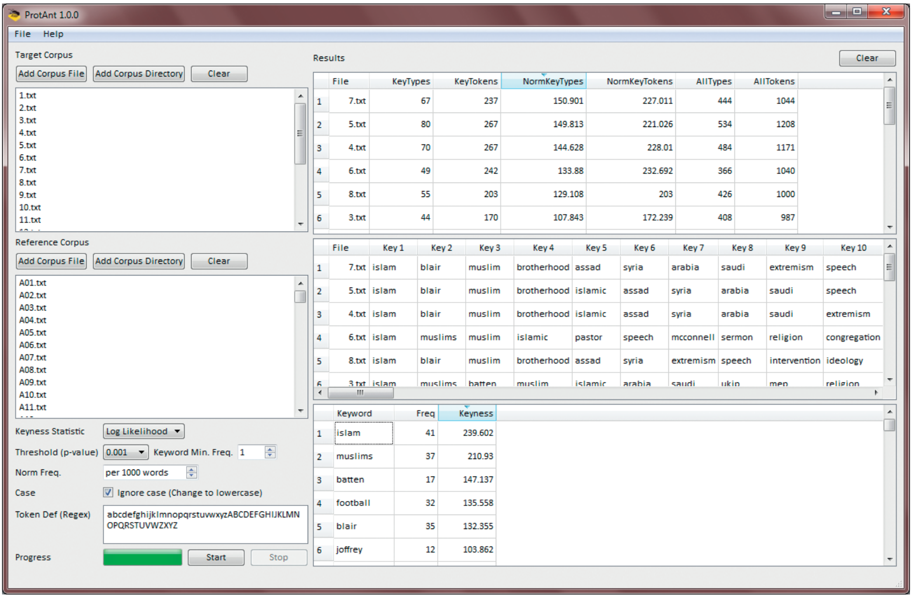
Traditionally, keywords have been ranked using the 'keyness' values from the statistical measure directly (effectively a ranking by *p* value). However, it is also possible to rank keywords in different ways. Scott (2014), for example, implements a key-key-words approach in which keywords are generated for each text in a corpus separately, and then these are ranked according to their distribution across texts in the corpus as a whole, with keywords appearing in a large number of texts ranked highly. The problem with this approach, however, is that it can only work with relatively long individual texts. Otherwise, the assumptions of the keyness statistic will be invalidated for many of the candidate words, i.e. those appearing less than 5 times in each individual text. More recently, effect size measures such as relative frequency (Demarau 1993) or a log of relative frequency (e.g. Hardie 2014), have been used to rank keywords. Effect measures are statistically rigorous, well understood, and produce ranking values that can be easily compared across studies.

Keywords have been commonly used in corpus research as a way of identifying a salient set of lexis in one or more corpora, which can then be subject to more qualitative, interpretative analyses of collocates and concordance lines. Leńko-Szymańska (2006), for example, compares corpora consisting of American and Polish students writing essays on the same topic, finding keywords relating to writing style (e.g. Polish students used more linking expressions such as *moreover*, *however* and *thus*), and focus (American students used fewer generalising words and their keywords tended to relate to specific situations). Keywords have also been used to indicate ideologies in texts. For example, Baker et al. (2013) compare keywords in corpora consisting of newspaper articles about the topic of Islam. They find that tabloid newspapers tend to use keywords that focus more on terrorism and extremism (e.g. *fanatics*, *terror*, *bomber*) whereas broadsheet newspapers focus more on conflict in general through keywords such as *conflict*, *military*, *revolution*, and *occupation*.

The *ProtAnt* software tool we have developed utilizes keywords in a novel way to identify prototypical texts. Our keyword-based prototype detection approach works in the following way. Researchers wishing to find prototypical texts in their corpus first load their target corpus into *ProtAnt* as individual UTF-8 encoded plain text files. Next, a suitable UTF-8 encoded reference corpus must be loaded. This can be a single file or a set of separate files. Finally, the researcher must specify the keyness statistic (e.g. log likelihood), a statistical threshold value (e.g. $p < 0.05$), a ranking statistic (e.g. log of relative frequency) and several other settings (see below). Once these are made, the *ProtAnt* tool compares the frequencies of words in the target corpus with those in the reference corpus and calculates the complete set of keywords for the *entire target corpus*. Based on this list, it next calculates how many keywords from the entire corpus are in each target corpus text and then ranks the texts by the number of keywords in them. Finally, *ProtAnt* displays the corpus keywords, the corpus keywords appearing in each individual corpus text, and the overall rankings of the texts in a tabular form.

In this way, *ProtAnt* works in a kind of reverse way to Scott's (2014) key-keyword approach, by first finding keywords that are distinctive of the target corpus as a whole, and then counting how many of these keywords appear at the individual text level. What is not immediately apparent, however, is why our approach might select prototypical (i.e. central, typical) texts of a target corpus over perhaps distinct, unique texts in the corpus. To understand our reasoning, it is important to remember that the keywords selected are distinctive of the corpus as a whole along many possible different dimensions. If a particular text in the target corpus has many of these keywords, it does not follow that it is especially distinctive. Rather, it suggests that it represents many of these different facets. In other words, the text can be considered prototypical of the corpus as a whole. Indeed, an extremely distinctive text (e.g. one containing a few unusual words repeated many times) would only produce a small number of unique keywords and thus be ranked low by *ProtAnt*. We test our reasoning through a series of experiments in the following section.

Figure 1 shows a screenshot of *ProtAnt* during the analysis of 20 newspaper articles of which 10 are related to the topic of Islam. The results of this analysis are discussed in detail in the following section.



1.txt
2.txt
3.txt
4.txt
5.txt
6.txt
7.txt
8.txt
9.txt
10.txt
11.txt

A01.txt
A02.txt
A03.txt
A04.txt
A05.txt
A06.txt
A07.txt
A08.txt
A09.txt
A10.txt
A11.txt

Keyness Statistic: Log Likelihood
Threshold (p-value): 0.001
Keyword Min. Freq.: 1
Norm Freq.: per 1000 words
Case: ☑ Ignore case (Change to lowercase)
Token Def (Regex): abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWZXYZ
Progress

	File	KeyTypes	KeyTokens	NormKeyTypes	NormKeyTokens	AllTypes	AllTokens
1	7.txt	67	237	150.901	227.011	444	1044
2	5.txt	80	267	149.813	221.026	534	1208
3	4.txt	70	267	144.628	228.01	484	1171
4	6.txt	49	242	133.88	232.692	366	1040
5	8.txt	55	203	129.108	203	426	1000
6	3.txt	44	170	107.843	172.239	408	987

	File	Key 1	Key 2	Key 3	Key 4	Key 5	Key 6	Key 7	Key 8	Key 9	Key 10
1	7.txt	islam	blair	muslim	brotherhood	assad	syria	arabia	saudi	extremism	speech
2	5.txt	islam	blair	muslim	brotherhood	islamic	assad	syria	arabia	saudi	speech
3	4.txt	islam	blair	muslim	brotherhood	islamic	assad	syria	arabia	saudi	extremism
4	6.txt	islam	muslims	muslim	islamic	pastor	speech	mcconnell	sermon	religion	congregation
5	8.txt	islam	blair	muslim	brotherhood	assad	syria	extremism	speech	intervention	ideology
6	3.txt	islam	muslims	batten	muslim	islamic	arabia	saudi	ukip	mep	religion

	Keyword	Freq	Keyness
1	islam	41	239.602
2	muslims	37	210.93
3	batten	17	147.137
4	football	32	135.558
5	blair	35	132.355
6	joffrey	12	103.862

Figure 1. Main display of *ProtAnt*

In Figure 1, the top left pane lists the target files that the researcher is interested in analysing. The middle left pane lists the reference corpus files. The bottom left pane offers the researcher various options to change the way that keywords are generated and ranked, including the selection of different keyword statistics and different keyword cut-off values. There are also options that affect the way in which tokens in the target and reference corpora are counted. Here, the token definition is specified as a regular expression with case being ignored if necessary (the default). The token definition can also include Unicode character classes, such as \p{L}, which represents any Unicode 'letter' class character, or \p{N}, which represents any Unicode 'number' class character. As a result, *ProtAnt* can work not only with texts in English but also those in any other language specified in the Unicode standard, including Japanese, Chinese, Korean, Arabic, and Hebrew. Finally, there is also an option to change the constant used to calculate normalized word frequencies (e.g. frequency per thousand words or frequency per million words).

The right-side of the *ProtAnt* main window, shown in Figure 1, displays the results of the analysis. In the top right pane, the frequencies and normed-frequencies of key types and key tokens can be viewed for each target corpus file, together with the frequencies all of types and tokens in the file. The values in the table can be sorted on any column simply by clicking on the column label. The default setting is to sort by normalized key types. Hence, this table provides an immediate view of the ranking of texts according to the number of key types contained in them. In Figure 1, the file 7.txt has been revealed as the most prototypical for this corpus. Clicking on any filename in this table causes the program to open up the original file for close reading.

The middle right pane of *ProtAnt* shows all the keywords that appear in each target corpus file ranked by their keyness values. In Figure 1, it can be seen that the most salient keyword *islam* is appearing in the first five prototypical texts (as expected), but the second most salient keyword *muslims* only appears in one of them (i.e. 6.txt). By viewing the ranked list of keywords in each file, a researcher can get an immediate sense of the text as a whole. Again, clicking on any filename in this table causes the program to open up the original file for close reading.

The bottom right pane of *ProtAnt* shows the complete list of keywords for the target corpus and their associated raw frequencies and keyness values. As with all other tables in the *ProtAnt* tool, clicking on the column name sorts the table by that column.

4. Experiments using *ProtAnt*

In order to test the usefulness of the *ProtAnt* tool for identifying useful prototypical texts, several experiments were conducted. These are described in Sections 4.1 to 4.5.

4.1 Experiment 1: Identification of prototypical newspaper articles

The first experiment was designed to see if *ProtAnt* was able to correctly rank a set of known texts in terms of their general topic focus. For this first experiment, a corpus of 20 files (Corpus 1) was created by using the searchable online news database *Nexis UK* to collect national newspaper articles from March 2014. Corpus 1 comprised 10 articles containing the term *Islam*, 5 articles containing the term *football*, and an additional 5 articles that were randomly selected by searching on 5 frequent words derived from the top 100 most frequent words in the British National Corpus (*time*, *people*, *other*, *know*, *see*). This produced news articles about tennis, art, and science, as well as an obituary and a television review. The corpus size was 22,353 tokens with a mean file size of 1,118 tokens. For a reference corpus, the 1 million word BE06 (Baker 2009) corpus was used. This contains 200 samples of British writing from 15 genres published in 2006.

Table 1 shows the ranking of the Corpus 1 texts by normalized key type and normalized key token value when using the log-likelihood statistical measure of keyness. In the table, four different cut-off *p* values were used, i.e. 0.05, 0.01, 0.001, and 0.0001. The values in parentheses in the column headers indicate the number of keywords that were elicited for each cut-off point, and the numbers in parentheses in the cells indicate the file number. All the files are given a suitable short name, and the files related to Islam are shaded so they can be more easily identified. Considering that half of the corpus texts contain articles about Islam, we would expect these articles to be viewed as most typical of the corpus as a whole and thus would be ranked higher than the others. We would then expect the 5 football articles to appear next in the rankings, with the 5 randomly chosen articles appearing as least typical.

Table 1. Rank ordering of Corpus 1 (newspaper articles) texts by normalized key type and normalized key token

	0.05 (1069)		**0.01 (610)**		**0.001 (234)**		**0.0001 (150)**	
	Types	**Tokens**	**Types**	**Tokens**	**Types**	**Tokens**	**Types**	**Tokens**
1	Islam (5)	Islam (6)	Islam (5)	Islam (6)	Islam (7)	Islam (6)	Islam (5)	Islam (6)
2	Islam (2)	Islam (5)	Islam (7)	Islam (5)	Islam (5)	Islam (4)	Islam (4)	Islam (5)
3	Islam (7)	Islam (4)	Islam (6)	Islam (7)	Islam (4)	Islam (7)	Islam (7)	Islam (4)
4	Islam (4)	Islam (7)	Islam (4)	Islam (4)	Islam (6)	Islam (5)	Islam (8)	Islam (7)
5	Islam (6)	Islam (3)	Islam (2)	Islam (3)	Islam (8)	Islam (8)	Islam (6)	Islam (8)
6	Islam (3)	Islam (1)	Islam (3)	Islam (8)	Islam (3)	Islam (3)	Islam (2)	Football (11)

Table 1. (*continued*)

	0.05 (1069)		**0.01 (610)**		**0.001 (234)**		**0.0001 (150)**	
7	Review (19)	Islam (2)	Islam (1)	Islam (2)	Islam (1)	Football (11)	Islam (1)	Islam (3)
8	Islam (1)	Obituary (16)	Islam (8)	Football (11)	Islam (2)	Obituary (16)	Islam (3)	Islam (9)
9	Obituary (16)	Islam (8)	Review (19)	Obituary (16)	Islam (9)	Islam (9)	Islam (9)	Islam (2)
10	Islam (8)	Review (19)	Obituary (16)	Science (17)	Football (11)	Islam (2)	Obituary (16)	Football (14)
11	Science (17)	Football (11)	Football (14)	Islam (1)	Obituary (16)	Islam (1)	Football (11)	Islam (1)
12	Football (11)	Science (17)	Football (11)	Football (14)	Islam (10)	Football (14)	Islam (10)	Review (19)
13	Football (14)	Football (14)	Science (17)	Islam (9)	Review (19)	Science (17)	Review (19)	Obituary (16)
14	Islam (9)	Islam (9)	Islam (9)	Review (19)	Football (14)	Review (19)	Football (14)	Science (17)
15	Islam (10)	Tennis (18)	Islam (10)	Tennis (18)	Science (17)	Islam (10)	Tennis (18)	Football (12)
16	Tennis (18)	Islam (10)	Tennis (18)	Islam (10)	Tennis (18)	Tennis (18)	Science (17)	Tennis (18)
17	Football (12)	Football (12)	Football (13)	Football (12)	Football (12)	Football (12)	Football (13)	Islam (10)
18	Football (13)	Football (13)	Football (12)	Art (20)	Football (13)	Art (20)	Football (12)	Art (20)
19	Art (20)	Art (20)	Art (20)	Football (13)	Football (15)	Football (15)	Art (20)	Football (13)
20	Football (15)	Football (15)	Football (15)	Football (15)	Art (20)	Football (13)	Football (15)	Football (15)

The results in Table 1 show that the *ProtAnt* analysis reliably ranks the Islam files as the most prototypical of the corpus as a whole, regardless of cut-off value. The ordering of files also remains relatively stable whether the type and token ordering is used. It is reassuring that all eight experimental conditions resulted in the top 5 files being about Islam. In the table, it can be seen that files 4, 5, 6 and 7 always rank in the top five, suggesting that these four files are the most prototypical of the corpus as a whole.

However, the results also reveal that files 9 and 10 are perhaps less typical than the other Islam files. File 9 mentions a speech by Tony Blair on Islam, which is also referred to in files 4, 5, 7, 8. On the other hand, file 9 only refers to the speech tangentially, and the majority of this article is about a potential threat to the West from China and Russia rather than being about Islam. File 10 is quite different from the other nine Islam files in that it is not discussing politicians' views on Islam, but rather, it is a story about a school which told parents that children had to attend a workshop on Islam or be called racist.

Looking at the results in Table 1, one may question why file 19 (a review article) appears around the middle of the table under most of the experimental conditions, even though it was the only file of its type. An answer is perhaps found when we consider the 20 strongest keywords across the 20 files in order of strength: *Islam*, *Muslims*, *Batten*, *Blair*, *football*, *Joffrey*, *Muslim*, *Mara*, *Brotherhood*, *Islamic*, *Assad*, *Kundnani*, *Syria*, *Arabia*, *Saudi*, *manager*, *Sansa*, *UKIP*, *pastor*. While 14 of these keywords relate to the Islam news stories, 2 of them are used in the file 19 review (*Joffrey* and *Sansa*, occurring 12 and 7 times respectively). These occurrences are enough to propel the review to the middle of the table.

The bottom five rankings in Table 1 almost always include files 12, 13, 15 and 20. The first three of these files are about football and would perhaps be expected to appear towards the middle of the list as there were five football files in the corpus. However, although all these files are about football, each deals with a different aspect of football; file 12 is an autobiographical article about footballers switching teams, file 13 refers to an adjudication about misconduct in football, and file 15 relates to the start of the new football league season. Unlike the Islam files which mainly contain a set of shared keywords related to Tony Blair, the Muslim Brotherhood, Syria, and Assad, the football files have fewer keywords and fewer shared keywords too, indicating that they are more lexically diverse than perhaps expected.

4.2 Experiment 2: Identification of prototypical novels

A second experiment was designed to see if *ProtAnt* was able to correctly rank a set of longer texts from a completely different genre (fiction rather than news). Corpus 2 comprised 10 excerpts from the novel *Dracula* (1897), five from *Frankenstein* (1818) and five more each from individual novels (two taken from the 19th century, three written in the late 20th or early 21st century). The overall corpus size was 40,759 tokens with a mean length of 2,038 tokens (about the length of a file in the Brown family). Again, the BE06 (Baker 2009) corpus was used as a reference corpus to elicit keywords.

Table 2 shows the ranking of the Corpus 2 texts by normalized key type and normalized key token value, again using the log-likelihood statistical measure of keyness and four different cut-off *p* values, i.e. 0.05, 0.01, 0.001, and 0.0001. As in

Table 1, the values in parenthesis in the column headers indicate the number of keywords that were elicited for each cut-off point, and the numbers in parentheses in the cells indicate the file number. All the files are given a suitable short name, and the files related to *Dracula* are shaded so they can be more easily identified. Again, we would expect the 10 Dracula files to be viewed as most typical, followed by the five Frankenstein files, with the 5 other novels appearing at the bottom of the table.

Table 2. Rank ordering of Corpus 2 (novels) texts by normalized key type and normalized key token*

	0.05 (1794)		**0.01 (1175)**		**0.001 (442)**		**0.0001 (274)**	
	Types	**Tokens**	**Types**	**Tokens**	**Types**	**Tokens**	**Types**	**Tokens**
1	D. (10)	D. (8)	D. (9)	D. (8)	D. (8)	D. (8)	D. (8)	D. (8)
2	D. (3)	D. (3)	D. (7)	D. (6)	D. (6)	D. (6)	D. (7)	D. (2)
3	D. (7)	D. (6)	D. (8)	D. (5)	D. (7)	D. (5)	D. (6)	D.(5)
4	D. (8)	D. (10)	D. (10)	D. (9)	D. (9)	D. (7)	D. (9)	F. (13)
5	D. (9)	F. (13)	D. (6)	F. (13)	D. (10)	D. (2)	D. (10)	D. (6)
6	F. (12)	D. (9)	D. (3)	D. (2)	D. (2)	D. (4)	D. (5)	D. (7)
7	D. (6)	D. (5)	D. (2)	D. (1)	D. (5)	D. (9)	D. (2)	D. (3)
8	F. (13)	D. (2)	D. (5)	D. (3)	F. (15)	F. (15)	D. (4)	D. (1)
9	D. (2)	D. (1)	F. (13)	D. (7)	D. (4)	F. (13)	D. (3)	F. (15)
10	F. (11)	D. (7)	F. (12)	D. (10)	D. (3)	D. (3)	F. (15)	D. (4)
11	D. (5)	F. (12)	D. (4)	D. (4)	F. (14)	D. (1)	F. (13)	D. (9)
12	D.(4)	F. (15)	F. (14)	F. (15)	F. (13)	D. (10)	D. (1)	F. (14)
13	F. (14)	D. (4)	F. (15)	F. (14)	D. (1)	F. (14)	F. (14)	D. (10)
14	D. (1)	F. (14)	D. (1)	F. (12)	F. (12)	F. (12)	M. (17)	F. (12)
15	F. (15)	F. (11)	F. (11)	F. (11)	M. (17)	F. (11)	F. (12)	F. (11)
16	J. (16)	J. (16)	J. (16)	J. (16)	F. (11)	T. (19)	F. (11)	T. (19)
17	M. (17)	M. (17)	M. (17)	M. (17)	J. (16)	M. (17)	J. (16)	J. (16)
18	H. (20)	H. (20)	H. (20)	T. (19)	H. (20)	J. (16)	H. (20)	M. (17)
19	T. (19)	T. (19)	I. (18)	H. (20)	T. (19)	H. (20)	T. (19)	H. (20)
20	I. (18)	I. (18)	T. (19)	I. (18)	I. (18)	I. (18)	I. (18)	I. (18)

* D. = *Dracula*; F. = *Frankenstein*; T. = *The Intimate Adventures of a London Call Girl*; H. = *Harry Potter and the Deathly Hallows*; I. = *It*; M = *Moonstone*, J. = *Jane Eyre*

As found in the first experiment, the results in Table 2 show that changing the cut-off *p* value has little effect on the ranking of the files. Similarly, there was little difference in the ranking when ordering the files by the number of included key types or key tokens. Under six of the eight conditions file 8 (*Dracula*) is viewed as the most typical. It is interesting to observe that file 1 is placed as the least typical *Dracula* file in the corpus under four of the conditions (sometimes as low as the 14th position) indicating that it was not a very typical file. This file is written at an earlier time narratively speaking, and begins with a description of Jonathan Harker's stay at Dracula's castle in Transylvania before many of the other characters have been introduced. In contrast, the other files are written at a later time narratively speaking, and they are largely set in locations in the UK.

In this experiment, the *ProtAnt* analysis also appears to be good at identifying atypical files regardless of the cut-off value, with the three most recent novels (*The Intimate Adventures of a London Call Girl* (2007), *Harry Potter and the Deathly Hallows* (2007) and *It* (1986)) almost always appearing in the bottom three ranking positions. *It* is the only American novel in the corpus, with the rest being written by British authors. As some keywords contained UK spellings like *sympathised* or references to British concepts like *solicitor*, this may help to explain why this file was usually ranked as least typical.

4.3 Experiment 3: Identification of prototypical texts in a large corpus

The third experiment we conducted was designed to see if *ProtAnt* could identify prototypical texts in a larger corpus, where it would not be possible to credibly point at the most typical files independently of the automated analysis. For this experiment, we used the 1 million word AmE06 Corpus (Potts & Baker 2012), keeping the BE06 (Baker 2009) corpus as the reference. The AmE06 corpus is identical in structure to BE06 (described above) except it contains American, rather than British texts. Using *ProtAnt*, the analysis should identify the files that are the most "American" in nature (compared to the British reference corpus). By ranking the files in the AmE06 corpus according to the number of keywords they have when compared against BE06, those files which are most distinctive when this comparison is made can be deemed the most typical of AmE06.

Keeping the same log-likelihood statistical measure of keyness with a cut-off p value of 0.0001 and looking at normed key types, the three files which emerged as having the most keywords were H24, H17 and H13 (H files contain texts in the register "Miscellaneous: Government documents, industrial reports etc."). File H24 is from the Department of Treasury and relates to State and Local General Taxes. A possible reason why this file was chosen as being most typical of AmE06 (in relation to BE06) was due to a high number of references to American states and cities which would be much less likely to appear in BE06. H17 is a government document describing the appointment of a director of the Administrative Office of the US Courts, and also contains references to American states as well as concepts distinct to the US like *federal* and *congress*, as well as American spellings of frequently cited words like *program*, *toward* and *center*. H13 is a Congressional Record and contains similar keywords to H17.

The three AmE06 files which were shown to be least typical of American English (in relation to BE06) were N06, P27 and P19. The N and P categories are from "Adventure" and "Romantic fiction" respectively. Text N06 is from an adventure novel set in Vietnam in 1975, and the excerpt it is taken from describes the main character jumping out of an aeroplane. Text P27 is from an historical romance novel set in France in 1885, and as such does not contain much language which marks it as "American". P19 contains a description of a sexual encounter between two characters, and its keywords are related to this encounter (e.g. *kissed*, *hear*, *cried*, *finger*, *willing*) rather than referring to anything of specifically American context.

In this experiment, a close reading of the texts identified by *ProtAnt* as typical and atypical texts clearly revealed why the tool might categorize the texts in such a way. This gives the tool a high degree of face validity (in addition to its innate reliability).

4.4 Experiment 4: Impact of reference corpus on the identification of prototypical texts

In the *ProtAnt* analysis, the choice of reference corpus plays an important role in determining which texts are selected as typical or not. For example, a comparison of AmE06 against BE06 will focus on typicality in terms of the "Americaness" of files in AmE06. To investigate the impact of the reference corpus on the selection of texts, we carried out a fourth experiment where instead of BE06, the Brown Corpus was used as the reference against AmE06. In this experiment, while both corpora were from American writing, the salient dimension was time as the Brown corpus contains text samples from 1961 and AmE06 was collected 45 years later. Therefore, the typical texts should be those that were most typical of 2006.

Interestingly, H24 appears as the most typical text again, followed by H21 and G40. Although H24 is a fairly dry government text about tax, it is written with a direct address to the reader and makes high use of the second person pronoun keywords *you* and *your* (a feature of personalising language that appears to have become more popular since 1961). H24 also refers to the 2005 event hurricane Katrina, which was discussed in 15 different files across the AmE06. H21 contains era-specific keywords that were relevant to the political mood at the time: *terrorism*, *Bush*, *preparedness* and *Palestinian*, while G40 is a first person autobiography from a woman who grew up in Mississippi, and as well as the personalising language, it contains references to issues around racial segregation and gender which suggest a more 21st century perspective on such events.

The three files least typical of AmE06 when compared to Brown are N02, G21 and G71. N02 is an adventure story that is set in a jungle. Considering it occurs out of an American context and has no reference to modern technology or current events it could have just as easily been written in 1961 as 2006. G21 is from a text about the American civil war, so again contains no reference to 1961 or 2006, while G71 is a text about a 20th century artist called Nozkowski which also has no reference to time period. As with Experiment 3, the typical and atypical AmE06 files when compared to Brown look credible.

4.5 Experiment 5: Identification of atypical texts in a large corpus

A fifth experiment was designed to see if the *ProtAnt* analysis was able to find outliers in a corpus. To test this usage of the tool, we again used AmE06 (with BE06 as the reference), but this time selected all the files from one of its fifteen registers plus a file from a different register (using a random number generator). We then repeated this experiment for each register. Ideally *ProtAnt* should rate the "outlier" file towards the bottom of the list of typical files. Table 3 shows the results based on a ranking by normed key types using the log-likelihood statistical measure of keyness with a cut-off *p* value of 0.0001.

The results in Table 3 show that for 10 out of 15 cases the outlier file is correctly identified as being at the bottom or very close to the bottom of the list, indicating excellent success in two thirds of cases. However, in the five other cases, i.e. rows F, G, J, M and R, the tool was less effective at identifying the outlier file. One explanation for the poor performance with F ("Popular lore"), G ("Belles lettres, biographies, essays") and R ("Humour") is that the terms of these registers are somewhat difficult to define, and the language within them appears to be more varied from one file to another than in, for example, P ("Romance and love story") or A ("Press: Reportage"). This may make identification of an outlier more difficult as there are fewer similarities between the files within the register. The academic register (J), on the other hand, could be said to be more clearly defined. Here, the outlier was from R ("Humour"). The file R8 is from a novel which is voiced by a nine year old genius who self-identifies as an inventor, amateur entomologist, origamist, and amateur archaeologist. The text contains keywords like *learning*, *dictionary*, *humans* and *age*. The excerpt from this file quickly switches between the following topics: entomology, microphones, jujitsu, childbirth, music, the magazine *National Geographic*, skyscrapers and limousines. This is a somewhat unusual text (even for the R category), which may explain why it was not seen as an outlier for the J register. Finally, the poor performance for register M ("Science fiction") may be influenced by the fact that the outlier file was from a similar register N ("Adventure and western), i.e. both are from fiction.

Table 3. Ranking of outlier files based on a *ProtAnt* analysis using log-likelihood normalized key types with a cut-off *p* value of 0.0001

Category	Register	Outlier file	Ranking of outlier file
A	Press: Reportage	K12	40/40
B	Press: Editorial	L9	28/28
C	Press: Reviews	P13	18/18
D	Religion	C8	18/18
E	Skills, trades and hobbies	N7	34/37
F	Popular lore	A3	28/49
G	Belles lettres, biographies, essays	M6	40/76
H	Miscellaneous: Government documents, industrial reports etc.	L13	30/31
J	Academic prose in various disciplines	R8	8/80
K	General fiction	E15	28/30
L	Mystery and detective fiction	C6	25/25
M	Science fiction	N8	4/7
N	Adventure and western	A7	29/30
P	Romance and love story	A5	30/30
R	Humour	L2	2/10

From this experiment, we can conclude that *ProtAnt* is good at identifying unknown miscategorised files in cases where the corpus comprises files from a single easily-defined register and the outlier file is taken from a clearly different register.

5. Discussion and conclusion

Our experiments with *ProtAnt* confirm the utility of identifying prototypical texts using the concept of keywords, with results largely matching predicted outcomes. *ProtAnt* clearly does a useful job generally, but we should note that it does not perform perfectly. On the other hand, in cases where the rankings of texts did not match the expected values, a closer look at the texts offered reasonable explanations, such as idiosyncrasies inherent to the texts themselves.

As *ProtAnt* bases the notion of typicality of texts on keywords generated using a reference corpus, the reference corpus itself can be adjusted to suit particular research questions. In the experiments with AmE06, the choice of reference corpus allowed us to search for "Americaness", by comparison with BE06, or "2006ness", by comparison with the Brown corpus. This feature of *ProtAnt*, i.e. its ability to judge the prototypicality of individual texts in relation to some kind of corpus-level distinctiveness category, can be considered as an advantage, but may also be viewed as a limitation if no obvious reference corpus is available. In these latter cases, a possible solution would be to treat each file as a pseudo target corpus, and the corpus as a whole as the reference corpus. This method would require each file (and the corpus as a whole) to be sufficiently large for the keyword statistic to be valid, but interpreting the results would be simple (although perhaps counter intuitive); files with the least number of keywords would be the most prototypical of the corpus as a whole. This is because keywords produced by this method would identify more atypical uses of language in individual texts of the corpus. Thus, the files with the most keywords would be least typical and those with the fewest keywords would be most typical. Of course, a disadvantage of this approach would be that the prototypicality could then not be tailored in any way to a global distinctiveness category, such as "Americaness" above

We hope to investigate this alternative prototype detection method and other possibilities in future work. We also hope to further investigate the interaction between the rankings of texts and the cut-off points for significance, the average file size, the number of files in a corpus, and whether key types or tokens are counted.

In this paper, we have discussed the possibilities of *ProtAnt* in relation to finding prototypical texts for a closer, more qualitative analysis at a later stage. However, other applications of the tool can also be imagined. For example, corpus linguists might use the tool to quickly identify a typical text for use as an illustrative example or to analyse as part of a supplement to their corpus analyses. In language teaching, educators and examiners might find the tool useful for identifying and scrutinising a small number of student essays written at different ability levels in order to get a feel for how the levels are distinct from one another. The prototypical texts identified by *ProtAnt* could also be used as models for students in teaching activities. Another application of the *ProtAnt* tool may be in training forensic specialists to make decisions about whether a text should be categorised as say, extremist literature, or to identify typical author style. A further application is in corpus building, where the researcher may use the tool to identify texts which are atypical of a particular genre and thus should be categorised as such or selected for exclusion. The identification of atypical texts may also help discourse analysis spot "resistant" or "minority" discourses that go against the grain.

In making the *ProtAnt* tool free and publicly available, we hope both researchers and educators can use it to quickly identify prototypical and atypical texts across a large data set and use these texts as part of more refined language analyses and applications.

Acknowledgment

The research presented in this paper was supported by the ESRC Centre for Corpus Approaches to Social Science, ESRC grant reference ES/K002155/1.

References

Anthony, L. (2014). *AntConc (Version 3.4.3)* [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from http://www.laurenceanthony.net/software/antconc/ (last accessed May 2015).

Anthony, L., & Baker, P. (2015). *ProtAnt (Version 1.0)* [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from http://www.laurenceanthony.net/software/protant/ (last accessed May 2015).

Bahrololoum, A., Nezamabadi-pour, H., Bahrololoum, H., & Saeed, M. (2012). A prototype classifier based on gravitational search algorithm. *Applied Soft Computing*, *12*(2), 819–825. DOI: 10.1016/j.asoc.2011.10.008

Baker, P. (2009). The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics*, *14*(3), 312–337. DOI: 10.1075/ijcl.14.3.02bak

Baker, P., Gabrielatos, C., & McEnery. T. (2013). *Discourse Analysis and Media Attitudes: The Representation of Islam in the British Press*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/CBO9780511920103

Caldas-Coulthard, C. R., & van Leeuwen, T. (2013). Teddy bear stories. In R. Wodak, (Ed.), *Critical Discourse Analysis Volume II: Methodologies* (pp. 35–60). Los Angeles, CA: Sage. (Original work published 2003).

Chen, L., Guo, G., & Wang, K. (2011). Class-dependent projection based method for text categorization. *Pattern Recognition Letters*, *32*(10), 1493–1501. DOI: 10.1016/j.patrec.2011.01.018

Chouliaraki, L. (2013). Political discourse in the news: Democratizing responsibility or aestheticizing politics? In R. Wodak, (Ed.), *Critical Discourse Analysis Volume II: Methodologies* (pp. 97–118). Los Angeles, CA: Sage. (Original work published 2000).

Damerau, F. J. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management*, *29*(4), 433–447. DOI: 10.1016/0306-4573(93)90039-G

Durfee, A., Visa, A., Vanharanta, H., Schneberger, S., & Back, B. (2007). Mining text with the Prototype-matching method. *Information Resources Management Journal*, *20*(3), 19–31. DOI: 10.4018/irmj.2007070102

Ehrlich, S. Z., & Blum-Kulka, S. (2013). Peer talk as a 'double opportunity space': The case of argumentative discourse. In R. Wodak, (Ed.), *Critical Discourse Analysis Volume II: Methodologies* (pp. 145–168). Los Angeles, CA: Sage. (Original work published 2010).

Fayed, H. A., Hashem, S. R., & Atiya, A. F. (2007). Self-generating prototypes for pattern classification. *Pattern Recognition*, *40*(5), 1498–1509. DOI: 10.1016/j.patcog.2006.10.018

Gabrielatos, C., & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press (1996-2005). *Journal of English Linguistics*, *36*(1), 5–38. DOI: 10.1177/0075424207311247

Gavriely-Nuri, D. (2013). If both opponents "extend hands in peace", why don't they meet? Mythic metaphors and cultural codes in the Israeli peace discourse. In R. Wodak, (Ed.). *Critical Discourse Analysis Volume II: Methodologies* (pp. 169–186). Los Angeles, CA: Sage. (Original work published 2010).

Gries, S. Th. (2003). Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics*, *1*, 1–27. DOI: 10.1075/arcl.1.02gri

Hardie, A. (2014). *CQPWeb (Version 3.1.10)* [Computer Software]. Lancaster, UK: Lancaster University. Retrieved from https://cqpweb.lancs.ac.uk/ (last accessed May 2015).

Khosravinik, M. (2010). The representation of refugees, asylum seekers and immigrants in British newspapers: A critical discourse analysis. *Journal of Language and Politics*, *9*(1), 1–28. DOI: 10.1075/jlp.9.1.01kho

Kloptchenko, A., Back, B., Visa, A., Toivonen, J., & Vanharanta, H. (2002). Toward content based retrieval from scientific text corpora. In *Proceedings of the 2002 IEEE International Conference on Artificial Intelligence Systems (ICAIS), Divnomorskoe, Russia*, 5-10 September 2002 (pp. 444–449). Washington, DC, USA: IEEE Computer Society.
DOI: 10.1109/ICAIS.2002.1048170

Kloptchenko, A., Magnusson, C., Back, B., Visa, A., & Vanharanta, H. (2004). Mining textual contents of financial reports. *The International Journal of Digital Accounting Research*, *4*(7), 1–29.

Labov, W. (1973). The boundaries of words and their meanings. In J. Fishman (Ed.), *New Ways of Analyzing Variation in English* (pp. 340–73). Washington, DC: Georgetown University Press.

Leńko-Szymańska, A. (2006). The curse and blessing of mobile phones: A corpus-based study into American and Polish rhetorical conventions. In A. Wilson, D. Archer & P. Rayson (Eds.), *Corpus Linguistics around the World* (pp. 141–151). London, UK: Rodopi.

Machin, D., & Suleman, U. (2013). Arab and American computer war games: The influence of a global technology on discourse. In R. Wodak, (Ed.), *Critical Discourse Analysis Volume II: Methodologies* (pp. 229–252). Los Angeles, CA: Sage. (Original work published 2006)

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *An Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/CBO9780511809071

Potts, A., & Baker. P. (2012). Does semantic tagging identify cultural change in British and American English? *International Journal of Corpus Linguistics*, *17*(3), 295–324.
DOI: 10.1075/ijcl.17.3.01pot

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192–233. DOI: 10.1037/0096-3445.104.3.192

Sajid, F. (2013). Critical discourse analysis of news headline about Imran Khan's peace march towards Wazaristan. *Journal of Humanities and Social Science*, *7*(3), 18–24.
DOI: 10.9790/0837-0731824

Scott, M. (2014). *WordSmith Tools (Version 6)* [Computer Software]. Liverpool, UK: Lexical Analysis Software. Retrieved from http://www.lexically.net/wordsmith/index.html (last accessed May 2015).

van Leeuwen, T. (1996). The representation of social actors. In C. R. Caldas Coulthard & M. Coulthard (Eds.), *Texts and Practices* (pp. 32–70). London, UK: Routledge.

Visa, A., Toivonen, J., Vanharanta, H., & Back, B. (2001). Prototype matching: Finding meaning in the books of the bible. In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34)*, Hawaii, USA, 3-6 January 2001 (pp. 3002). Washington, DC, USA: IEEE Computer Society.

Widdowson, H. G. (2004). *Text, Context, Pretext: Critical Issues in Discourse Analysis*. Oxford, UK: Blackwell. DOI: 10.1002/9780470758427

Wodak, R. (2013). *Critical Discourse Analysis*. Los Angeles, CA: Sage. DOI: 10.4135/9781446286289

Authors' addresses

Laurence Anthony
Faculty of Science and Engineering
Waseda University
3-4-1 Ohkubo, Shinjuku-ku
Tokyo 169-8555
Japan

anthony@waseda.jp

Paul Baker
Department of Linguistics and English Language
Lancaster University
Bailrigg
Lancaster LA1 4YL
UK

j.p.baker@lancaster.ac.uk