

Working with social media data

Quantitative perspectives

Outline of the chapter

In this chapter we cover:

- Choosing how to organise your material.
- Moving beyond raw frequency in calculating results.
- The challenges of concordancing social media texts.
- Annotating social media corpora.
- Analysing social media texts using concordancing software.
- Visualisation tools and social networks.
- Examples of social media text visualisations.

Choosing how to organise your material

Once the materials needed for a research project have been collected, as discussed in [Chapter 8](#), you are ready to organise and then analyse the data. There are some practical concerns that are important to consider at the initial stages of preparing to analyse social media texts from a quantitative perspective. Some of these relate to the size of the datasets that the project includes. Especially if the dataset is large, then it might not be practical for the researcher to sift and search through this material by hand: using some kind of computerised infrastructure can be useful. For example, as we saw in [Chapter 7](#), preparing and storing the collected materials in a package like NVivo or Atlas-ti can enable the researcher to annotate the data with analytical labels (based on the linguistic and contextual, dependent and independent variables). This process can not only be useful for identifying themes that emerge from the data; additionally the software can also count and sort those themes to help the researcher identify the relative prominence of the patterns in the data. Other alternative packages can include using software like a Microsoft Excel spreadsheet to collate and then code examples, or you might choose to prepare material with bespoke statistical packages like SPSS or the open source package R. Once the researcher has stored their

material in a suitable format (for example, choosing the file format required by the software), then they can begin to analyse the materials. The first step is to annotate the data. This might include labelling parts of the data to indicate that a particular linguistic feature is present (or not), or indicating some of the features of the contextual variable (such as the demographic characteristics of the participant, the type of text, number of people involved in an interaction and so on).

A specialised form of analysis which uses particular forms of computing infrastructure belongs to the methods of corpus linguistics. Most of this chapter will consider the practical and analytical concerns that are involved in preparing material for use in conjunction with concordancing software. Those practical concerns include factors that need to be considered when you are compiling and annotating social media materials. After this we describe some basic steps that can be used in corpus linguistics to examine social media texts, moving from simple searches which identify the frequency of individual words in a text, to more complex forms of analysis which compare frequencies across different corpora. We end the chapter by giving some examples of how quantitative forms of analysis might lend themselves to automated forms of representation as visualisation. Before we move on to these more specialised concerns, we begin with some general principles about quantitative analysis that also apply to working with smaller sets of data.

Quantifying features: beyond raw frequency

In [Chapter 8](#), we discussed the importance of collecting data to answer quantitative research questions, by making sure that the variables in the question matched the materials that were collected in type and proportion. However, even if the researcher has specified certain elements of the data collection with quantitative analysis in mind, there can be further factors about the relative size and balance of the data to take into account once analysis begins. This is true whether the scale of the quantitative analysis is relatively small, or is large enough that it requires analysis with the help of automated software like concordancing software. Imagine that the researcher was interested in comparing the intensifiers that were used by women and men when they posted Facebook updates (this was part of a project that I, Ruth, carried out). The participants for the project were selected based on parameters of gender and age, using a snowballing technique to recruit participants who were outside my immediate set of contacts on Facebook. Informed consent was negotiated with each of the participants, until I had recruited ten women and ten men in each of five age groups (15–18 years of age, 19–21 years of age, 22–29 years of age, 30–39 years of age and 40–49 years of age). I then examined the ten most recent posts that the person had published on their timeline at a given point in time (July 2008

and then again in October 2010). But what I had not accounted for in my data collection was that the women and men in different age groups would write posts that varied in word length. In fact, in 2008, the total number of words in the posts written by men was 3,627 while the total number of words written by women in their updates was 3,220. In 2010, a different pattern occurred, the total number of words written by men in their updates was 4,146 and the total number of words written by women in their updates was 4,407.

We could make some simple calculations from quantifying the size of the data samples, such as generating the average number of words per update, and we could go on to separate out the total word length so that the pattern according to age and gender was clearer. It might also be important to examine the word length of posts written by individual updaters so that the variation in the length of the post within each category becomes clearer (rather than just looking at aggregated scores). But just comparing the total or average word length for the posts does not indicate how important the difference in word length of posts is or whether it could be treated as significant: to assess that, statistical tests would need to be applied. And finally, just looking at the word length does not tell us anything about the linguistic feature that the project set out to examine. However, establishing the size of the dataset in terms of word length is important, especially if the linguistic variable you intend to examine matches the grammatical unit of a word (as opposed to a phrase, clause or other kind of unit).

Knowing the size of your dataset is important, because the raw frequencies of particular items can be a misleading comparison if the different sets of your data are different sizes. Imagine a more polarised (and fabricated) example. Say we wanted to examine the intensifier *so* and found eight examples in ten posts written by women and three examples in ten posts written by men, but that the total word length of the women's posts was 120 and the word length of the men's posts was 45. The figures from this imaginary sample could be summarised as in [Table 9.1](#).

If we took the raw frequency of the intensifier (three and eight), or the frequency of the intensifier per post (0.3 and 0.8 per post respectively for men and women), then it would seem that women use *so* more often than do men. But this comparison is misleading as the word length for the posts written by women and men are different. If the frequencies of the intensifiers are calculated as a percentage of the words in each part of the dataset, it

Table 9.1 An example of raw frequency in relation to word length

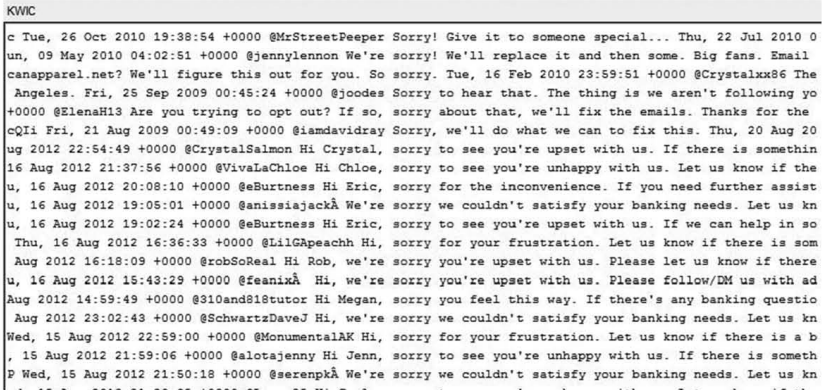
| Gender | Frequency of <i>So</i> | Number of posts | Total word length |
|--------|------------------------|-----------------|-------------------|
| Male | 3 | 10 | 45 |
| Female | 8 | 10 | 120 |

turns out that women and men (in this example) use the same proportion of intensifiers. You can do this calculation for yourself to check: $(8/120 \times 100)$ and $(3/45 \times 100)$ both result in 6.67 per cent. In this case, the difference in word length across elements of the dataset could be offset by converting the raw frequencies to a percentage of the occurrences per number of words. But other kinds of calculations might be appropriate if your linguistic variable exists in units above the level of the word, or if your dataset is so large that percentages are not the most sensible way of making unevenly sized data samples comparable. The possibilities and problems that arise when you might want to use a very large set of materials (like a corpus) are dealt with in the remainder of this chapter.

The challenges of concordancing social media texts: making your materials searchable

Chapter 8 introduced the concept of a *corpus* and considered the key issues that arise when building social media corpora. Here we move the focus to using a corpus to undertake linguistic analyses. Corpora are usually analysed using software packages to carry out the counting, sorting and presentation of language features. Some social media corpora contain bespoke search interfaces (as does the Birmingham Blog Corpus and the Twitter Stratified Random Sample). But other software can be used with a range of corpora, which can be more useful if you are using specialised corpora that you have developed for your own project. Commonly used concordancing systems include the propriety software, Wordsmith Tools (hereafter Wordsmith) Scott (2008), Wmatrix (Rayson 2009) and the freeware program, Antconc (Anthony 2005). Each of the software can be used to create a concordance, that is, an indexed list which collates all the instances of a searched for word or phrase within the chosen materials. Of course, concordances are nothing new: concordances have been created in previous centuries to enable searches across lengthy print texts, such as the Bible. The concordances created using software similarly collate (and quantify) words and phrases, often presenting the results in search windows that contain the concordanced lines in a vertical list with the search term (the keyword) positioned centrally in its textual context. Figure 9.1 shows a concordance for the search term *sorry* within a selection of Twitter posts, produced by using Antconc (Anthony 2005).

Before being able to search a social media corpus using this kind of software, you need to ensure that the materials in the corpus are in a format that means they can be analysed. Some problems arise when we need to process large volumes of these social media texts since they have some peculiarities that make them challenging to process for quantitative analysis. Sometimes there are textual features that need to be accounted for or manually cleaned-up before the data can be analysed. Researchers



KWIC

c Tue, 26 Oct 2010 19:38:54 +0000 @MrStreetPeeper Sorry! Give it to someone special... Thu, 22 Jul 2010 0
un, 09 May 2010 04:02:51 +0000 @jennylennon We're sorry! We'll replace it and then some. Big fans. Email
canaparel.net? We'll figure this out for you. So sorry. Tue, 16 Feb 2010 23:59:51 +0000 @Crystalxx86 The
Angeles. Fri, 25 Sep 2009 00:45:24 +0000 @joodes Sorry to hear that. The thing is we aren't following yo
+0000 @ElenaH13 Are you trying to opt out? If so, sorry about that, we'll fix the emails. Thanks for the
cQIi Fri, 21 Aug 2009 00:49:09 +0000 @iamdavidray Sorry, we'll do what we can to fix this. Thu, 20 Aug 20
ug 2012 22:54:49 +0000 @CrystalSalmon Hi Crystal, sorry to see you're upset with us. If there is somethin
16 Aug 2012 21:37:56 +0000 @VivaLaChloe Hi Chloe, sorry to see you're unhappy with us. Let us know if the
u, 16 Aug 2012 20:08:10 +0000 @eBurtness Hi Eric, sorry for the inconvenience. If you need further assist
u, 16 Aug 2012 19:05:01 +0000 @anissiajackÅ We're sorry we couldn't satisfy your banking needs. Let us kn
u, 16 Aug 2012 19:02:24 +0000 @eBurtness Hi Eric, sorry to see you're upset with us. If we can help in so
Thu, 16 Aug 2012 16:36:33 +0000 @LilGapeachh Hi, sorry for your frustration. Let us know if there is som
Aug 2012 16:18:09 +0000 @robSoReal Hi Rob, we're sorry you're upset with us. Please let us know if there
u, 16 Aug 2012 15:43:29 +0000 @feanixÅ Hi, we're sorry you're upset with us. Please follow/DM us with ad
Aug 2012 14:59:49 +0000 @310and818tutor Hi Megan, sorry you feel this way. If there's any banking questio
Aug 2012 23:02:43 +0000 @SchwartzDaveJ Hi, we're sorry we couldn't satisfy your banking needs. Let us kn
Wed, 15 Aug 2012 22:59:00 +0000 @MonumentalÅ Hi, sorry for your frustration. Let us know if there is a b
, 15 Aug 2012 21:59:06 +0000 @alotajenny Hi Jenn, sorry to see you're unhappy with us. If there is someth
P Wed, 15 Aug 2012 21:50:18 +0000 @serenpkÅ We're sorry we couldn't satisfy your banking needs. Let us kn

Figure 9.1 Screenshot of a concordance of the search term *sorry* using Antconc (Anthony 2005).

compiling a social media corpus might need to be particularly aware of the following factors.

Non-standard orthography

Microposts typically contain non-standard orthography of different kinds. These variously creative uses of spelling, punctuation and other typographic resources can cause problems for the tools for corpus search and annotation, which were developed with standard orthography and “traditional” text genres in mind (Beißwenger and Storrer 2008, p.303). Typical problems can include the unconventional use of punctuation which compresses lexical words into a single graphological unit, as shown in bold font in the following example.

I'm glad 2 read **this.I** haven't got many things 2 show in **youtube.I** put ur name just 4 curiosity and u **appeared!I** love twitter

Because the software will read *this.I*, *youtube.I* and *appeared!I* each as single words, this can cause problems for counting the lexical words in a corpus accurately.

Other examples of non-standard orthography include the many spelling variations that occur in social media sites. The spelling variations can make it difficult to search for all the examples of a particular word you might be interested in tracing through a particular corpus. Some variations have become conventionalised over time (such as *ur* for *your* in the example above), but other variations can be more idiosyncratic and therefore more difficult to identify. Some more recent concordancing software has been designed to take account of these spelling variations (see Tagg *et al.* 2013), but not all packages do.

Emoticons and hashtags

The non-verbal resources (like emoticons and hashtags as shown in the examples below) which have emerged from computer-mediated communication and social media sites in particular, can pose further problems for concordancing software. Depending on the settings used by your concordancing software, some of the characters used in emoticons will not be considered 'valid' letters for that system and so will be filtered out of the concordance. They may also be interpreted by the software as marking word breaks or may have other special meanings to the system.

Oh no! Be good to see you back properly next week :-))) I'm great thanks, very H-A-P-P-Y!! Glad it's nearly the weekend too Xxx

I miss kindergarten when the only drama there was, was losing your crayon. #OhJustLikeMe

For instance, Wordsmith uses the hash character # to represent, as a group, all numbers that occur in the corpus. This is to avoid polluting the word frequency list with phenomena that interfere with interpretation of lexis. In addition, Wordsmith does not treat the hash character itself as a valid letter and responds as if it were punctuation. Similar effects may be seen with other kinds of characters and symbols found in social media texts. The researcher may need to adjust the settings of the software in order to be able to search and sift through their corpus to identify and analyse these non-verbal resources used in social media communication.

Abridged posts

Because of the character-constraints imposed in micro-blogging, some users will attempt to circumvent these limitations by using a web service to extend their message (such as longertweets.com). These services allow the user to present an abridged version of the tweet in their stream, usually with a link to the full, longer post. The result is that some posts will appear in an abridged form in the corpus. Alternatively, the post may contain punctuation indicating that it continues in a subsequent post in the stream. Depending on the strategy used to extend the tweet, the elaborating tweet may not be captured when the corpus is constructed. However, these types of tweets are relatively uncommon and may be filtered out of the corpus if the analyst can identify regular syntactic patterns with which they can be identified and removed.

Automated and rebroadcast posts

Spam has unfortunately infiltrated social media meaning that unwanted posts can be present in a corpus. These, and other various kinds of non-human, automatically-generated posts generated may be present in the corpus

materials that may not conform to the researcher's selection criteria. For instance some services will generate automatic Twitter updates such as:

I favorited a YouTube video -- Vocal training <http://youtu.be/S9hruS0ET18?a>

These kinds of posts are relatively easy to identify in a corpus since they will occur multiple times with unexpected similarity. For example all instances of the post above contained the non-standard punctuation "--", giving us a clue that they were not manually produced.

If your corpus contains micro-blogging posts then it will also incorporate rebroadcast material such as retweets. The status of these tweets in your study is an important theoretical question. For example, in my (Ruth's) study of celebrities and their use of Twitter, I excluded retweets as I was most interested in the language being used by the celebrities (not in the people whose tweets were being forwarded by the celebrities). However, if I had wanted to study what kinds of topics were typically in forwarded messages, then including the retweets would have been vital. Depending on the intended use of the corpus, a researcher may consider some or all of these tweets as 'noise' and decide to:

- filter out all instances of repetition;
- filter out any tweets that seem to be spam;
- filter out retweeted tweets;
- filter out any automated, non-human tweets.

The decisions that are made in this respect will clearly have an impact upon the size of the corpus, and consequently on the quantitative analyses that involved calculations based on frequency.

Questions to ask yourself if you want to use linguistic concordances in your project are summarised in the following points for reflection.

Points for reflection

Can the social media data you intend to analyse be compiled as a searchable concordance (e.g. if the data is multimodal how will images and layout be archived and annotated)?

Which concordance software will be used (either proprietary or open source)?

How will you get the social media data into a format the concordance software will accept in light of features such as non-standard orthography, emoticons, hashtags, abridged posts, automated and rebroadcast material?

Annotating social media corpora

Having built a corpus and refined the format of the materials, what kinds of linguistic features and patterns might we search for? An important factor determining the kinds of features that can be found in a corpus is how it has been annotated. Corpus annotation involves adding a layer of linguistic information to a corpus so that the researcher can search for recurring patterns. There are many different kinds of annotation that can be undertaken depending on the research question. A very common form of annotation is part of speech (POS) tagging. POS tagging can be employed to aid word sense disambiguation tasks that may be helpful for some kinds of discourse analysis. For example, perhaps the researcher might be interested in exploring emotional language in a micro-blogging corpus and wish to know how *like* functions in its evaluative sense. An unannotated corpus such as HERMES (Zappavigna 2012) will return results such as those in the following examples.

@User.the whole concave look is gone cause its **like** almost the same length._. hahahaha

RT @IDoThat2: RT if you sit there smiling **like** an idiot when you think of a happy memory. #idothat2

@User I **like** the Ditty Bops! Never heard of them before, very cute x why does it have to be **like** this coba?

ALWAYS ON THE RADIO RANTING, HE DON'T UNDERSTAND OUR PRESIDENT,. NO A PERSON **LIKE** U NEVER CAN. EVER. U SEE, ASS KISSER joe scarboro-

Its seems **like** forever since they dated. lol

As these examples suggest, in order to isolate the instances where *like* is functioning as a verb (e.g. "I really **like** it!") we would need to disambiguate these from instances where it is used as a softener (e.g. "It was, **like**, so great!") or as a comparative (e.g. "It looks more **like** a dog than a cat."). In order to automatically retrieve only the desired instances, a POS tagger can be used so that the search function will only return particular kinds of items (e.g. all instances where *like* functions as a verb). The tags shown in bold in following examples would distinguish instances where *like* functions as a verb (VBP) from instances where it functions as preposition (IN), for example:

@User I *like* **/VBP** the Ditty Bops! Never heard of them before, very cute x

RT @IDoThat2: RT if you sit there smiling *like* **/IN** an idiot when you think of a happy memory. #idothat2

Unfortunately, many POS taggers are not trained to work with social media texts and so do not cope well with properties of social media texts such as

non-standard orthography. However, an example of a POS tagger that has recently been developed to work with Twitter data is the Twitter POS tagger¹ (Gimpel *et al.* 2011). This POS tagger was trained on manually annotated tweets. Features² that are specific to Twitter (and online discourse in general) which this tagger aims to annotate automatically include:

- Hashtags (#): indicates topic/category for tweet.
- At-mentions (@): indicates another user as a recipient of a tweet.
- Discourse markers (~): indications of continuation of a message across multiple tweets.
- Links (U): URL or email addresses.
- Emoticons (E): typographic conventions indicating emotion or involvement.

Other concordancing programmes which include POS taggers include WMatrix (Rayson 2009), and there are a range of tools available to support corpus annotation. Some tools focus on supporting relatively low volume manual annotation, while other tools focus on higher volume, partially automated analyses. An example of a tool supporting corpus-based discourse analysis of small corpora is the UAM Corpus Tool (www.wagsoft.com/CorpusTool/) (O'Donnell 2008). This tool allows the researcher to annotate a text using a schema which they define in the form of a network of choices. As we mentioned in Chapter 5, tools for multimodal annotation are also beginning to be developed such as UAM Image Tool (www.wagsoft.com/ImageTool/) and the relatively established, ELAN, used for video annotation (<http://tla.mpi.nl/tools/tla-tools/elan/>) (Wittenburg *et al.* 2006). These tools can be combined with quantitative as well as qualitative approaches. For a useful overview of the current state of corpus annotation tools see O'Donnell (in press).

In summary, annotating social media data involves:

- Deciding what features and patterns are important as the search terms in your corpus and whether part of speech tagging might be needed.
- Considering what kinds of manual and automatic annotation are possible for the corpus.
- Evaluating what you might do with the product of the annotation (e.g. statistical processing) and weighing up if it is worth the intensive effort.

Analysing social media texts using concordancing software

Having made your corpus searchable, you are now ready to begin analysing the materials. Many different kinds of analysis can be done using concordancing software. Here we introduce three frequently used kinds of analysis: frequency, keyness and collocations.

Frequency lists

The most basic analysis that can be carried out using concordancing software is to establish the frequency of words within the collected materials: corpus linguists are often interested in establishing the regular patterns or norms within a particular set of materials. To start with, the researcher might be interested in the frequency of individual words, or may want to create a list of the words contained in the materials (for example, in decreasing order of frequency). Table 9.2 shows us the ten most frequent words in the Birmingham Blog Corpus and in the HERMES Twitter Corpus.

There are certain features we might notice from these frequency lists. For example, we might notice that there is quite a lot of overlap between the words which occur on both lists: (*the, to, a, of, and, in, I*) all occur in both lists, and the top two words (*the* and *to*) are exactly the same. We might conclude from this that perhaps the language of blogs and micro-blogs is broadly similar to each other. However, we might also question how useful this information actually is. For example, the Birmingham Blog Corpus and HERMES are different sized corpora (628,558,282 words and 100,281,967 words, respectively), so the raw frequencies in the table are not comparable for individual items. Instead, the frequencies will either need to be expressed as a percentage of the dataset as a whole or to be normalised (which in corpus linguistics often adjusts these to a relative frequency of occurrences per million words (see McEnery and Hardie 2012)). Some software will make this calculation for you; other software may require you to make this calculation for yourself (which can be done relatively easily in a package like Microsoft Excel).

We might also notice that there are some differences between the words included on each list, such as the inclusion of the pronoun *you* and the abbreviation *http* in the top ten items for the HERMES corpus, but not in

Table 9.2 Top ten word frequencies in Birmingham Blog Corpus and HERMES Twitter Corpus

| Rank | Birmingham Blog Corpus | | | HERMES Twitter Corpus | | |
|------|------------------------|---------------|------|-----------------------|---------------|------|
| | Word | Raw frequency | % | Word | Raw frequency | % |
| 1 | THE | 24,986,273 | 3.98 | THE | 3,358,659 | 3.15 |
| 2 | TO | 15,523,666 | 2.47 | TO | 2,379,223 | 2.23 |
| 3 | AND | 13,226,449 | 2.10 | I | 2,236,470 | 2.10 |
| 4 | A | 13,169,504 | 2.10 | A | 1,674,654 | 1.57 |
| 5 | OF | 11,541,388 | 1.84 | HTTP | 1,631,187 | 1.53 |
| 6 | I | 9,736,364 | 1.55 | AND | 1,545,943 | 1.45 |
| 7 | IN | 8,442,980 | 1.34 | OF | 1,217,398 | 1.14 |
| 8 | IS | 6,700,706 | 1.07 | YOU | 1,194,631 | 1.12 |
| 9 | THAT | 6,635,605 | 1.06 | IS | 1,120,058 | 1.05 |
| 10 | FOR | 5,829,087 | 0.93 | IN | 1,118,227 | 1.05 |

the Birmingham Blog Corpus, but, on the whole, this does not give a very full picture of the variation between the language which occurs commonly in these two corpora from social media. Nor do these frequency lists on their own allow the researcher to see if these frequencies are typical of blogs and micro-blogs as distinct genres, or if the frequencies are simply common to English language use more generally. To get a clearer picture of how to interpret these frequencies, the researcher needs to search further. One option is to compare the frequency of the words with existing corpora drawn from offline examples of language such as the British National Corpus or the Concordance of Contemporary American English (COCA). Compared with offline language corpora, the researcher might see that a similar range of grammatical words are amongst the most frequently occurring items. (The most frequent items on the COCA word list are *the*, *be*, *and*, *of*, *a*, *in*, *to*, *have*, *to* and *it*.) This comparison tells us that the frequency of the definite article, *the*, is not particularly distinctive as a characteristic of blogging or micro-blogging, but the occurrence of the personal pronouns *I* and *you*, and the URL prefix *http* are more so. Perhaps this might suggest the interactional dimension of Twitter posts and its importance as a site for sharing linked material. However, once again, the frequency list alone is only the first step in the analysis: the researcher might want to find out more accurately whether or not the variation between one set of frequencies is meaningful or not.

Keyword lists

If the researcher wants to find out which features are statistically significant when comparing corpora, then sometimes a small corpus will be compared with a second corpus to determine which items in the smaller corpus are determined to be ‘key’, that is, which features occur at higher frequency than the pattern represented in the reference corpus. This kind of work is referred to as keyword analysis. Baker (2010, p.104) defines a keyword as “a word which occurs statistically more frequently in one file or corpus, when compared against another comparable or reference corpus”. For instance I might have built a specialised corpus of blog posts about climate change and wish to compare how climate change is represented in this corpus with a second, specialised corpus of traditional news media texts about climate change. I could then compare the differences between the blogs and traditional news media by comparing the two sets. Or, if the researcher wanted to see if the frequency of a particular word is typical of their corpus in particular, or of language use more widely, they could compare a specialised corpus with a larger, reference corpus. An example of this would be comparing the frequency of a specialised set of Twitter posts (say of posts made by a set of celebrities) with a corpus of Twitter language more generally (such as the HERMES Twitter corpus). As Julia Gillen’s case study shows, a comparison

of keyness can help the researcher see what was genuinely distinctive about the language use in the context she was studying: Second Life.

Box 9.1 Investigating language use in a virtual world

Julia Gillen

It is sometimes thought that the language used in new media, especially by young people, is of impoverished quality and that topics are trivial. It may be assumed that only everyday language is used, or, conversely, that dialogues are full of jargon. Furthermore, the quantity of written language used in many virtual worlds is sometimes also overlooked.

I investigated the language used in the Scheme Park Programme pilot, led by Peter Twining at the Open University. This pilot, designed to investigate radically different models of education, received funding in 2007 from the National Academy for Gifted and Talented Youth and used a protected environment within Teen Second Life.³

In order to investigate the students' use of written language in the project, I conducted a corpus linguistic analysis of a large, randomised sample of the students' turns in chat logs, collected with fully informed consent. I used WordSmith as the concordancing software, and compared my specialised corpus of Second Life Student Chat with the reference corpus of BNC Baby, a four-million word cut-down version of the British National Corpus. This samples diverse genres, including newspaper articles, correspondence and everyday conversation by adults. My aim was to find out which lexical items were distinctive to the interactions "in-world", as we referred to the 3D simulated environment of the Scheme Park Programme.

Significant findings included the following:

- Students' turns were characterised by considerable interrogation and inquiry, with a preponderance of question words such as *how* and *what*.
- Orienting in space and time was evidenced through the high frequency of words such as *time*, *here*, *there*, *now*, etc.
- There are various indications of positive relationship building and collaborative activities – *yes* was a keyword but *no* was not. *Haha* and *LOL* indicated shared humour. *Help* and *thanks* reflected an environment where assistance was asked for and readily given. Indeed politeness is prevalent with *please* also common.

- A few genre-specific terms – such as *schomer*, *RL* and *IM* – featured with high keyness, indicating the role of these specialised, ‘in group’ to suggest familiarity within the discourse community.
- *Thing*, *things*, *make* and *stuff* indicated activities around the construction of ‘objects’ and ‘scripts’. These words appear often in the frequency list and in comparison with language overall. The use of such simple terms in complex and abstract domains of communication, such as laboratories, has been found in other studies.
- *Meeting* and *library* appear more often than they do in the overall language corpus, which is not unremarkable given that the reference corpus is adult and contains a considerable amount of text in formal genres.

#advice: It is not difficult to learn how to carry out a simple lexical analysis with corpus linguistics tools. It can be fruitful in investigating a specific discourse community, probing behind common assumptions.

Sometimes the keyword lists can show the researcher findings that they did not anticipate at the outset. For example, I (Ruth) built a specialised Twitter corpus gathered between 2010 and 2012 which gathered publicly available posts from celebrity, ‘ordinary’ and corporate accounts. By running a keyword test on different sections of the corpus, I was able to see that the addressed messages posted by corporate accounts were quite different in the relative frequency of the vocabulary choices. Words like *hi*, *thanks*, *sorry*, *please* occurred in the top ten keywords for addressed messages, suggesting that these interactions (as compared with other kinds of posts like general broadcasts or retweets) favoured a kind of institutionalised customer care talk not used in other kinds of Twitter interactions or by other groups of Twitter users. The data was not collected with the intention to explore customer care talk (such as greeting or apologising): it was only by using the search tools to examine frequency patterns that this distinctive use of language was brought to light. Even once the keywords had been isolated, the research was incomplete. It did not tell me how those customer care terms were used in context. To examine the function of those apologies, I needed to combine concordance searches with other, manual, qualitative analysis of particular posts.

Collocation

One of the advantages of using concordancing software is that it allows researchers to examine large stretches of text. This can be combined productively with a more qualitative kind of textual analysis where the researcher can bring close focus to particular parts of the corpus and begin to examine the patterns that emerge in more detail. One way to do this is to

trace the collocational patterns that are found in the data. Collocation refers to the co-occurrence of words in patterns that are regularly repeated and statistically significant. This principle of co-occurrence is familiar from language use in both offline and online contexts: the adjective *blonde* can be said to collocate with hair in a way that the synonym *yellow* does not (you might like to think of other examples of hair colours that have strong collocational patterns). Corpus analysis software can be a helpful way to approach collocations, because the software can sort selected words not only in terms of their frequency, but also organise examples where the selected words occur according to their textual contexts (the words which co-occur to the left or the right of the word). Sometimes this can reveal patterns that might not appear obvious without this large-scale process of sifting and sorting. For example, when I (Ruth) was comparing the temporal adverbs that occurred in the Twitter corpus collected to compare celebrity, corporate and ordinary Twitter use, I found that adverbs which emphasised the present moment (*today*, *tonight*, *now*) occurred more often than references to points of time in the past like (*yesterday*). However, when the keywords were examined in context, the results suggested that the celebrities and ordinary Twitter members used *today*, *tonight* and *now* in different ways. For example, the phrase *the show* collocated with the temporal adverbs for celebrities:

Mannequin's going **into the show tonight** for the first time. I'm really excited for everyone to see it!—Britney

I always love having Kid Inventors **on the show and today** is no exception. <http://su.pr/23ZYZO>

However, it did not collocate with these adverbs for the ordinary Twitter members, as in the next example.

Taking a break from coding **today**, need a break from all those letters and numbers. Instead I am spreadsheeting. Oh, damn #itsjustasbad

By working from the frequency lists to the keywords and then looking at the collocations for those keywords, the promotional strategies used by celebrities in Twitter could be identified, suggesting that their use of recency in tweeting behaviour was put to a promotional use not typical of all Twitter members (Page 2012).

Visualisation tools and social networks

Why text visualisation?

Because social media corpora are typically very large, linguists can often benefit from visualisation tools to aid the interpretation of complicated

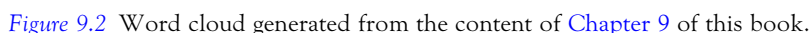
patterns and thematic trends that they might not be able to observe otherwise (Wise *et al.* 1995, pp.51–2). In addition, some of the metadata available in social media materials, such as the timestamp, geo-spatial information and information about connections in the social network, can open up the possibility of visualising patterns of language use in multidimensional forms. Given that sometimes researchers might be interested in patterns that involve relationships between many textual variables across many contextual dimensions of meaning (such as time, space or network), visualisations can be very useful. Existing techniques commonly used in linguistics, such as statistical analyses of corpus-based data, essentially flatten the text into a countable product, without showing other relationships between contextual features, such as how variation in the frequency of particular words might fluctuate over time, or how particular memes become popular and then disappear from use. Fortunately, advances in computer technology afford us the possibility of annotating, managing and visualising highly complex data. We can now track multiple relationships between variables unfolding in time or along other dimensions. As a result we have the potential to model the unfolding of meaning in text. However, visualisation tools and techniques are not without their problems. As with all forms of computing, ‘bad data in equals bad data out’. We have to be careful that we use visualisation strategies that illuminate the kinds of linguistic relationships that we want to explore, which can be based on features that can be clearly identified and which are readily ‘countable’. If we do not do this we risk creating a representation that does not accurately reflect the patterns in the texts that we want to understand and which takes on ‘a life of its own’.

What is text visualisation?

The field of visualisation has the potential to provide linguists with some much-needed help by supporting their analytical gaze when they work with texts and large corpora. Visualisation, in general, is concerned with finding methods of representation that best leverage the characteristics of human visual perception to make complex data meaningful. The term “information visualisation” (often abbreviated to “InfoVis”) refers to “the use of computer supported, interactive, visual representations of abstract data to amplify cognition” (Card and Mackinlay 1997, p.7). While most visualisation techniques share the general aim of amplifying and enhancing human cognition, they vary considerably in the type of data they seek to represent (e.g. financial data, scientific data, medical data, etc.). There is a large and growing body of visualisations available worldwide and a number of taxonomies are proposed to classify visualisation tools (e.g. Chi 2000; Tory and Moller 2004).

Those interested in visualising text often have a background in both computer science and digital art, bringing both technical and aesthetic skills

Some concordancing software integrates visualisations like word clouds to help the researcher in their process of analysis. For example, WMatrix provides word clouds in conjunction with the verbal frequency and keyword lists. In this case, the differences in the font size can help the researcher see more immediately the more or less frequent items in the list: a visual form of differentiation which eases the burden of searching through and sorting a tabulated list of numerical data. Of course, this interpretive strategy might tend to privilege the larger sized font and therefore the higher frequency items. The researcher has to look more closely to find the smaller items which occur less frequently. So while visualisations can be helpful, we should be aware that, like all parts of the research process, they are forms of interpretation which are not neutral but present information in a particular and selective way.



Social media text visualisation

A common form of social media visualisation can be found in techniques for representing relationships between users (e.g. Heer and boyd 2005). These approaches represent non-linguistic links (e.g. 'friendship' relationships on Facebook or 'following/follower' relationships on Twitter) between users in a social media network. Sometimes they are used as part of Social Network Analysis (SNA), a method for using network theory to analyse social relationships (e.g. Ugander *et al.* 2011). While this form of visualisation can usefully supplement linguistic analysis, linguists are also interested in the more difficult problem of representing 'who is saying what to whom' and the even more challenging problem of representing the meanings being negotiated when someone says something to someone in the putative social network. However, as John Caulfield points out in his case study in this chapter, modelling network connections can be a useful starting point in helping the researcher identify which members of a network might be core participants (and so worth following up with additional forms of analysis) and those who are peripheral (and so might not form the main focus for the analysis).

Box 9.2 Imagining the Irish language blogosphere: a social network analysis using comments and link data

John Caulfield

Irish is a minority language and its survival as an everyday community language is under threat. My research explores how some Irish speakers worldwide use social media to create new forms of language communities online. I aim to describe who's participating, the social processes taking place, and how users have adapted the language to computer-mediated communication. In seeking to visualise the invisible ties formed through interacting online, I turned to social network analysis.

I knew that the cluster of active Irish language bloggers would be small, and I hoped to take a whole network approach. The first challenge was defining what exactly being 'active' and 'Irish language' entailed. Each study sets its own parameters for inclusion. In this case, I included any blogger who had made one post over a three-month period in which Irish was the primary language of communication, and the commenters that responded to these posts. This was admittedly a very low threshold, but it enabled me to expand the network quickly (and later focus analysis on the core of prolific users).

Sourcing the sample was time-consuming and involved six months of participant-observation: writing, reading and searching for Irish

language blogs. I expanded the sample by tracing links between blogs and hand-coding them for Irish content. The online researcher is always faced with the niggling doubt that some activity will remain undetected. To overcome this, I continued visiting blogs after my data capture period, and added the small number of newly discovered blogs and commenters that met the criteria to the sample. I could confidently say that the resulting 73 blogs and 68 commenters comprised most, if not all, of the Irish language blogosphere in early 2011.

The next challenge was establishing what would constitute a connection between network members. Again, each study defines ties differently. I set an admittedly low threshold of just one comment, link, blog roll entry (or a few lesser-used functions, like 'likes' and 'notes') between members. In a spreadsheet I compiled an edgelist listing all the individual members, the others they interacted with, and the number of times they interacted. This data was analysed using the *igraph* package in the statistical software R. I chose the Fruchterman-Reingold layout for my visualisations as it pulls together nodes that are well connected and pushes less well-connected nodes to the periphery. Combining this with colour-coding nodes according to their language use, geographic locations, longevity and subject matter helped reveal patterns of social structure that would have remained hidden in the raw data.

The above approach to social network analysis helped reveal a well-connected core of prolific bloggers who had adapted the features of computer-mediated communication to maintain as monolingual an Irish space as possible. These would later become the focus of discourse analysis.

#advice: Set strict criteria for who is included in your study. It will help you present your findings within very clear parameters.

Figure 9.3 is an example of a networked visualisation taken from Caulfield's study. The square nodes represent blogs and circular nodes represent individual commenters, with edges representing those who interacted across the blogosphere through comments and linking during the data capture period. Arrows show the direction of the interaction. Isolated blogs (those with no interactions) appear unconnected at the edge of the network. In the colour version of this plot, nodes are colour-coded according to their language use, with monolingual blogs/commenters appearing in dark green, and other nodes appearing in other colours according to how much/little Irish they used. The plot indicated that language use affected network position, with monolingual nodes being the most well-connected in the network. Caulfield went on to carry out discourse analysis of core nodes and so identified a number of innovative ways in which core users maintained as monolingual a space as

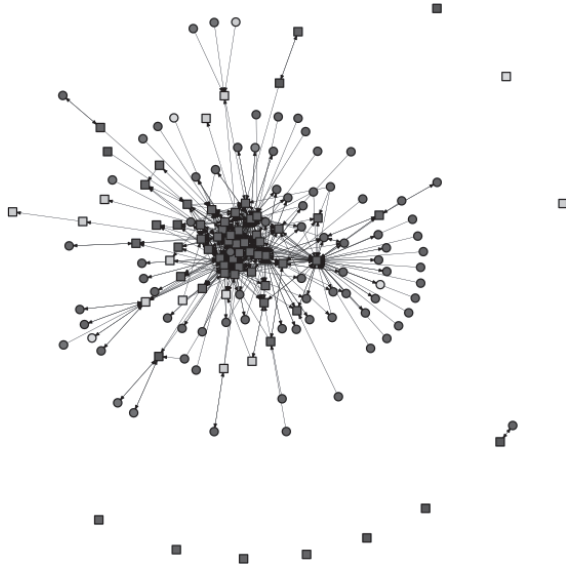


Figure 9.3 Visualisation of the Irish language blogosphere created by John Caulfield using Gephi software.

possible on their blogs, including hyperlinking from unusual or difficult Irish words to their English language translations. In this way, social network analysis helped identify nodes and themes worthy of further focused analysis, and showed how the partly automated process of quantitative analysis can be complemented with a qualitative approach to other parts of the research project.

Social media text visualisation usually focuses on visualising social text stream data. Because of the limitations of automatic language processing, language-focused social media visualisation tends to concentrate on what is loosely defined as the ‘topic’ of the communication, in part, because of visualisation’s reliance on identifying linguistic features at the level of the word. However, in these social media visualisations, the ‘flattened out’ dimensions of word frequency are counter-balanced to include visual representation of how the topics relate to other contextual data provided in corpus, such as the geographical or temporal location of the post. Examples include:

- ThemeCrowds, a visualisation that was applied to a micro-blogging corpus “with the goal of identifying groups of users within a large geographical area, who discuss similar topics over time” (Archambault *et al.* 2011, p.81);
- visualisations of the temporal evolution of topics (Kraker *et al.* 2011);

- visualising social media ‘events’, defined as “a set of relations between *social actors* on a specific *topic* over a certain *time period*” (Zhao and Mitra 2007, p.1);
- Twitinfo, a system design to allow users to visually explore Twitter events base on keyword queries (Marcus *et al.* 2011).

Topic-based techniques have also been used to create social media analytics to support journalistic inquiry (Diakopoulos *et al.* 2010), to represent discourse about climate change (Scharl *et al.* 2013), and explore information propagation (Chien-Tung *et al.* 2011). In addition social media services such as Facebook have begun developing their own in-house visualisation techniques (e.g. Facebook graph search, www.facebook.com/about/graphsearch) to allow users to visually explore their own networks. Visualisations drawn from social media can also have very practical uses in areas such as improving emergency response (Mazumdar *et al.* 2012), and have been used to track how social media sites disseminate knowledge during crises (Procter *et al.* 2013).

TwitterStreamgraph: an example of social media text visualisation

A visualisation technique that has been used with streaming data is the streamgraph (Byron and Wattenberg 2008). This is an example of a text visualisation technique that allows the researcher to represent visually “usage over time for the words most highly associated with ... [a] search word” (Clark 2008). The streamgraph builds on visualisation formats that are familiar from more traditional types of graphs and charts. For example, area graphs represent the frequency of a particular kind of data graphically by blocking out (usually in colour) the portion of a graph that falls beneath the plot line (the larger the blocked out portion of the shape, the greater the quantity of the item being quantified at that point). In packages like Microsoft Excel, for example, it is possible to create simple area graphs by entering numerical data into the spreadsheet and selecting the “Area Chart” format from the menu. The area chart in [Figure 9.4](#) is a visual representation of the number of times that messages were retweeted from a selection of named accounts in the hours following the news reporting the death of former British Prime Minister, Margaret Thatcher, in April 2013.

From [Figure 9.4](#), you can see that some users posted messages that were redistributed as retweets more frequently than others (in this case, the celebrity figures such as the comedian Frankie Boyle, boy band member Harry Styles and television host Piers Morgan were retweeted more often than mainstream news accounts like the BBC or CNN, for example). But while an area graph usually shows a single data series (in the example above, the number of retweets), stacked area graphs can represent multiple data series by stacking one on top of the other. This can be useful if you wanted

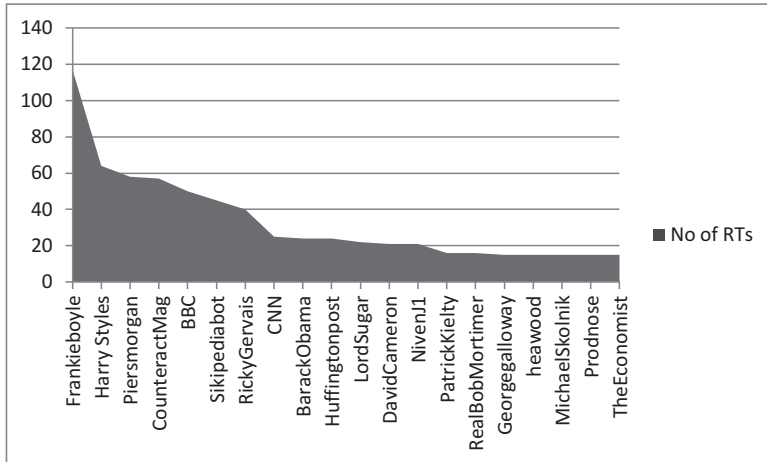


Figure 9.4 An area graph for rebroadcast tweets following the death of Margaret Thatcher.

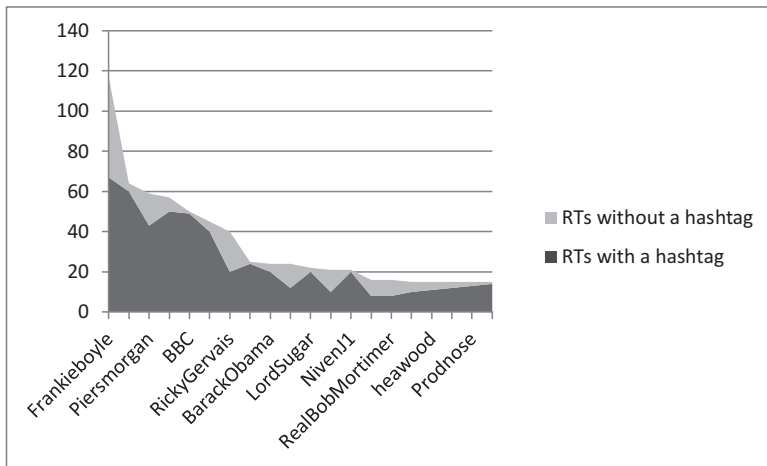


Figure 9.5 Stacked area chart showing retweeted messages with and without a hashtag.

to visualise the relationship between two or more different variables. For example, if you wanted to represent how many of those retweeted posts contained a hashtag and how many did not, you could use coloured areas to visualise that difference, as in Figure 9.5. In the chart in Figure 9.5, the proportions of retweeted messages with and without a hashtag are indicated through the relative size of the coloured blocks for each of the different Twitter members. Again, the correlation between size and quantity allows the viewer to see fairly quickly certain correlations (for example the extent

to which retweeting and hashtags were used in combination as a means of making a particular tweet more visible).

Streamgraphs build on this established model of stacked area graphs but generate smooth curves for the different data streams by interpolating between points to produce a flowing river of data. An example of a web-based tool for doing limited interactive streamgraph visualisations of lexis occurring in tweets is TwitterStreamgraph⁴ (Clark 2008). In the TwitterStreamgraphs, the distribution of the most ‘interesting’ capitalised words that occur in a database of Twitter messages for either a single Twitter account or a group of Twitter accounts can be represented by combining the size of font (for individual words) with the size of the stacked areas of the chart. For example, Figure 9.6 shows a streamgraph using *linguistics* as a search word.

Other visualisation tools can adapt models of representation that move beyond mathematical visualisations derived from graphs and charts. For example, Brice Russ’ study of American dialects in Twitter (see his case study in Chapter 8) used Google Maps to plot the geographical distribution of different lexical variables (such as *soda*, *coke* and *pop* for soft drinks). You can see an example of one of the dialect maps in Figure 9.7, and more on Brice Russ’ web pages: www.briceruss.com.

Rather than representing the relationship between the different variables in Russ’ study (the lexical variables, *soda*, *pop* and *coke*), and their relationship to geographical location (different towns in America) in a chart, visualising the distribution of the terms across the map more powerfully suggested the clustering of the terms across physical locations.

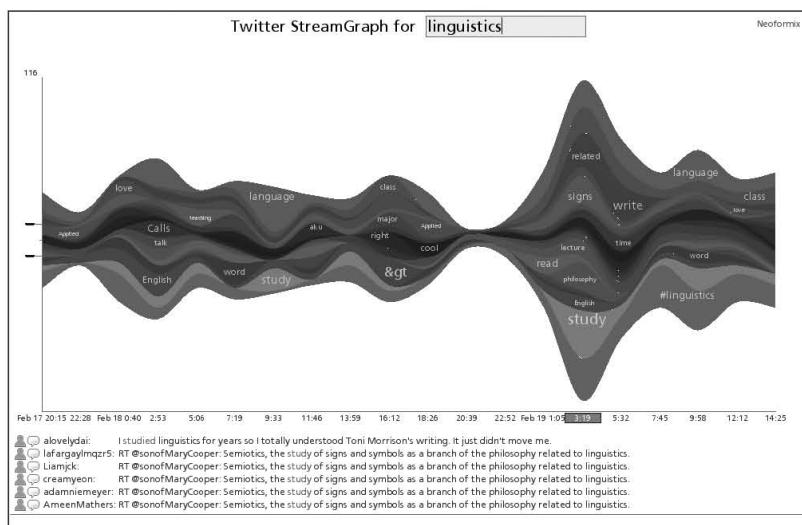


Figure 9.6 A Twitter StreamGraph generated with the search word *linguistics*.

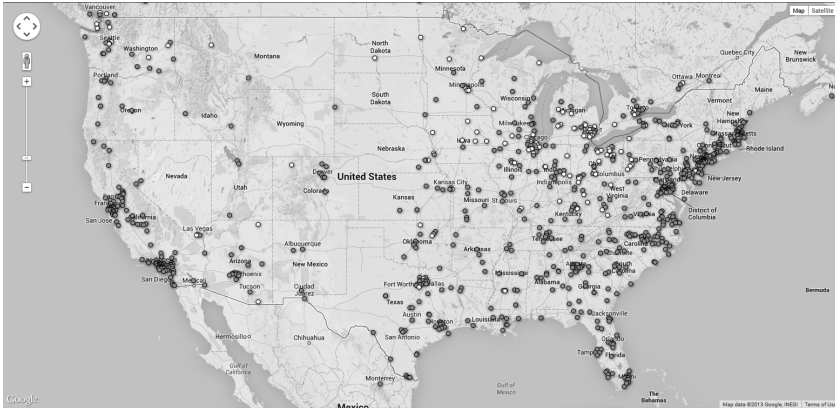


Figure 9.7 Dialect map showing the distribution of *soda*, *coke* and *pop* (Russ 2012).

As even the brief range of visualisations suggests, there are now a variety of tools and models for making a picture ‘tell a thousand words’ – or at least represent the frequency of thousands of words! No doubt as time goes by, additional tools for visualisation will emerge. Whether your research project is relatively simple or involves more complex, large-scale analysis of data, it is worth exploring how interpreting your results can be modelled in different ways. All of these visualisations, like any representation of analysis (verbal or visual), will be partial and selective, so just as with all your choices in the research design, from formulating your question through to gathering and analysing your data, make sure that your choice of visualisation is thought through carefully with a clear rationale.

Notes

- 1 At the time of writing this POS tagger is available for download here: www.ark.cs.cmu.edu/TweetNLP/.
- 2 At the time of writing the annotation guideline from which the features in this list were extracted is available here: https://github.com/brendano/ark-tweet-nlp/blob/master/docs/annot_guidelines.md.
- 3 Teen Second Life and Second Life are trademarks of Linden Lab.
- 4 The interactive application is available at: www.neoformix.com/Projects/TwitterStreamGraphs/view.php.

References

- Anthony, L. (2005) ‘AntConc: design and development of a freeware corpus analysis toolkit for the technical writing classroom’, in *Professional Communication Conference, 2005. IPCC 2005. Proceedings. International*, USA: IEEE, 729–37.
- Archambault, D., Greene, D., Cunningham, D. and Hurley, N. (2011) ‘Theme-Crowds: multiresolution summaries of Twitter usage’, paper presented at the 3rd

- International Workshop on Search and Mining User-generated Contents, Glasgow, Scotland, UK.
- Baker, P. (2010) 'Corpus methods in linguistics', in Litosseliti, L., ed., *Research Methods in Linguistics*, London: Continuum, 93–116.
- Beißwenger, M. and Storrer, A. (2008) 'Corpora of computer-mediated communication', in Lüdeling, A. and Kytö, M., eds, *Corpus Linguistics: An International Handbook* (Vol. 1), Berlin and New York: Mouton de Gruyter, 292–308.
- Byron, L. and Wattenberg, M. (2008) 'Stacked graphs—geometry and aesthetics', *Visualisation and Computer Graphics, IEEE Transactions on*, 14 (6), 1245–52.
- Card, S. and Mackinlay, J. (1997) 'The structure of the information visualisation design space', in *Proceedings of the 1997 IEEE Symposium on Information Visualisation*, Phoenix: IEEE, 92–9.
- Chi, E.H. (2000) 'A taxonomy of visualisation techniques using the data state reference model', paper presented at the Information Visualisation, 2000. InfoVis 2000.
- Chien-Tung, H., Cheng-Te, L. and Shou-De, L. (2011) 'Modeling and visualizing information propagation in a micro-blogging platform', *2011 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, USA: IEEE, 328–35.
- Clark, J. (2008) Twitter Topic Stream, available: <http://neoformix.com/2008/TwitterTopicStream.html> [accessed 19 February 2012].
- Diakopoulos, N., Naaman, M. and Kivran-Swaine, F. (2010) 'Diamonds in the rough: social media visual analytics for journalistic inquiry', *2010 IEEE Symposium on Visual Analytics Science and Technology (VAST)*, Salt Lake City: IEEE, 115–22.
- Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J. and Smith, N.A. (2011) 'Part-of-speech tagging for Twitter: annotation, features, and experiments', *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, 42–7.
- Heer, J. and boyd, d. (2005) 'Vizster: visualizing online social networks', *2005 IEEE Symposium on Information Visualisation*, Minneapolis: IEEE, 32–9.
- Kraker, P., Wagner, C., Jeanquartier, F. and Lindstaedt, S. (2011) 'On the way to a science intelligence: visualizing TEL tweets for trend detection', in Kloos, C., Gillet, D., Crespo García, R., Wild, F. and Wolpers, M., eds, *Towards Ubiquitous Learning* (Vol. 6964), Berlin Heidelberg: Springer, 220–32.
- McEnery, T. and Hardie, A. (2012) *Corpus Linguistics: Method, Theory, Practice*, Cambridge: Cambridge University Press.
- Marcus, A., Bernstein, M.S., Badar, O., Karger, D.R., Madden, S. and Miller, R.C. (2011) 'Twitinfo: aggregating and visualizing microblogs for event exploration', *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, ACM, 227–36.
- Mazumdar, S., Ciravegna, F., Gentile, A.L. and Lanfranchi, V. (2012) 'Visualising context and hierarchy in social media', paper presented at the International Workshop on Intelligent Exploration of Semantic Data (IESD), Galway City, Ireland.
- O'Donnell, M. (2008) 'Demonstration of the UAM CorpusTool for text and image annotation', *Proceedings of the ACL-08:HLT Demo Session (Companion Volume)*, Columbus, Ohio, June 2008, Association for Computational Linguistics, 13–16.
- O'Donnell, M. (in press) 'Between man and machine: the changing face of corpus annotation software', in Yan, F. and Webster, J.J., eds, *Developing Systemic Functional Linguistics*, London: Equinox.

- Page, R. (2012) 'The linguistics of self-branding and micro-celebrity in Twitter: the role of hashtags', *Discourse and Communication*, 6 (2), 181–201.
- Procter, R., Vis, F. and Voss, A. (2013) 'Reading the riots on Twitter: methodological innovation for the analysis of big data on Twitter', *International Journal of Social Research Methodology*, 16 (2), 197–214.
- Rayson, P. (2009) 'Wmatrix: a web-based corpus processing environment', Computing Department, Lancaster University, available: <http://ucrel.lancs.ac.uk/wmatrix/> [accessed 27 November 2013].
- Russ, B. (2012) 'Examining large-scale regional variation through online geotagged corpora', paper presented at American Dialect Society, Portland, 5–7 January, available: www.briceruss.com/ADStalk.pdf [accessed 27 November 2013].
- Scharl, A., Hubmann-Haidvogel, A., Weichselbraun, A., Lang, H. and Sabou, M. (2013) 'Media Watch on climate change – visual analytics for aggregating and managing environmental knowledge from online sources', *46th Hawaii International Conference on System Sciences (HICSS-46)*, 7–10 January 2013, Maui, Hawaii.
- Scott, M. (2008) *Wordsmith Tools Version 5*, Liverpool.
- Tagg, C., Baron, A. and Rayson, P. (2013) "'I didn't spel that wrong did i. Oops": analysis and standardisation of SMS spelling variation', *Linguisticae Investigationes*, 35 (2), 367–88.
- Tory, M. and Moller, T. (2004) 'Rethinking visualisation: a high-level taxonomy', paper presented at the Information Visualisation, 2004. INFOVIS 2004.
- Ugander, J., Karrer, B., Backstrom, L. and Marlow, C. (2011) 'The anatomy of the Facebook social graph', *arXiv preprint arXiv:1111.4503*, available: <http://arxiv.org/pdf/1111.4503v1.pdf> [accessed 27 November 2013].
- Wise, J., Thomas, J., Pennock, K., Lantrip, D., Pottier, M. and Schur, A. (1995) 'Visualizing the non-visual: spatial analysis and interaction with information from text documents', *Proceedings of the IEEE Information Visualisation Symposium*, Atlanta, Georgia: IEEE, 51–8.
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. and Sloetjes, H. (2006) 'ELAN: a professional framework for multimodality research', paper presented at the LREC, Fifth International Conference on Language Resources and Evaluation, available: www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf [accessed 27 November 2013].
- Zappavigna, M. (2012) *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*, London: Continuum.
- Zhao, Q. and Mitra, P. (2007) 'Event detection and visualisation for social text streams', paper presented at the International Conference on Weblogs and Social Media (ICWSM'2007), Boulder, Colorado, USA.