Chapter 8

Collecting social media materials for quantitative projects

Outline of the chapter

In this chapter we cover:

- Quantitative research questions.
- Types of variables.
- Samples and representativeness.
- Quantitative tools for eliciting data.
- Collecting data for a corpus linguistics project.
- Understanding metadata.
- The features of social streaming data.
- How to scrape social streaming data.

Quantitative research questions

As we've stressed throughout this book, we believe in the value of multi-method approaches for analysing the language used in social media contexts. In Chapter 3, we presented a range of methods that can be used either separately or in combination when planning and executing a research project. While the discourse analytic and ethnographic approaches covered in Chapters 5, 6 and 7 most readily align with qualitative aspects of the research design, in the present chapter, we focus on the principles and practices for collecting data that would be used when a research project includes a quantitative element. In simple terms, quantitative research methods are those that involve counting or measuring particular phenomena (such as people, texts or language features). Providing quantitative accounts of the material you are studying might be important, even if a research project is primarily qualitative in nature. Quantitative information can be useful to contextualise close analysis of individual texts in relation to factors, such as how many people were involved in the interaction, how much material was included in the data collection, whether the data was collected at a single point in time, or collected periodically. But for research with a strongly quantitative focus in terms of the analysis required, including the work in linguistic subfields such as corpus linguistics, computational linguistics or certain types of sociolinguistic projects (especially those within the tradition of what is called variationist sociolinguistics), planning how to collect the data requires particular consideration of the quantity and types of material required for the research.

The kind of data needed for a research project depends a great deal on the aims of the project and the kinds of questions the researcher is trying to answer. Questions which are framed from a quantitative perspective usually involve some kind of measurement and often imply some kind of comparison. Some questions might seek to describe the frequency of linguistic features as they appear in social media forms. For example:

How often do viewers include different types of initialisms in their messages when they live-tweet their responses to given television programmes?

Are temporal adverbs associated with the present moment (like *now*) more likely to occur in blogs or social network sites?

Or they might compare the frequency with which particular groups of people use particular linguistic forms:

How does the use of intensifiers (*so*, *very*, *really*) vary in frequency between the Facebook wall posts written by men and written by women?

Other quantitative questions might compare attitudes towards language use or perceptions of social media use, for example:

What are the most frequent reasons people give for joining a social network site?

Some of these questions may investigate the influence of particular factors on the use of language in social media contexts. For example:

How does the length of time a person spends interacting on a social network site relate to their use of site-specific jargon?

Are women or men more likely to be the target of insults in YouTube comments?

As these examples suggest, and just like qualitative examples, quantitative research questions can vary from being very broad to very specific. Compare the relative breadth of the following set of questions:

Do women blog more often than do men?

How do women and men blog about their experiences of being diagnosed with cancer?

How do Latino-American women and men use evaluative language when they blog about being diagnosed with lung cancer?

The first question sets up a comparison based on the participants' gender (a binary comparison that might be criticised for all kinds of reasons). The second question narrows the type of blog that the research considers, while the third question narrows the groups of women and men who are compared, and introduces the linguistic feature under scrutiny (though "evaluative language" is still a somewhat vague term which could include a range of phenomena), and a more specific type of illness.

Formulating a research question requires an in-depth understanding of the properties and scale of the domain in which you might want to collect data. There is no definitive set of steps for formulating a quantitative research question in linguistics given the multitude of possible areas of inquiry, but some general principles apply. At the most basic level, the nature of the research question determines the kind of data that will then be collected. This is sometimes described in terms of research validity. Dörnyei (2007) describes the two aspects of research validity as the extent to which the results of a project can be generalised to a wider group (external validity) and the extent to which the research design is coherent (internal validity). The question of internal validity can be thought of as making sure you collect data that is consistent with the parameters (and only the parameters) set out in a research question. For example, if a project asked, "To what extent do teenagers use the same slang terms in Twitter and Facebook?" and only collected material from one of those sites (only Twitter or only Facebook), then the data would not help the researcher to carry out the comparative analysis the question required. Likewise, if researchers wanted to find out the most frequent reason that caused people to leave a particular social media site, it would not be very helpful to survey people who were still members of the site. The relative breadth or narrowness of the research question also determines the amount and kinds of data that a researcher needs to collect and in turn the kinds of analysis and claims that they are able to make later. The question of how representative a set of data might be has a specific meaning in statistics (and which we touch on below in the section on constructing a sample), but for now, it is most important to note the relationship between the parameters that a researcher sets in framing a research question and the kind of data they will then need to collect.

The internal validity of a research project also depends on the researcher making sure that the results are not influenced by additional factors outside the parameters of the research question. Controlling the influence of one factor rather than another can sometimes be complicated because of the multifaceted nature of people's identity and interactions. Participants have different aspects to their identities (age, gender, nationality, political affiliation, occupation, membership of different sites), interactions take place in different sites, at different times, accessed from different platforms and for different purposes, and language use can draw on multiple semiotic resources (words, sound, image, layout, gesture), in varying combinations and draw on multiple elements of the language systems at the same time. Isolating which of those multiple factors and features are being investigated is an important step in controlling the data collection process. For example, if you wanted to examine the language used on a social media site by a particular group (say, men), you might want to refine the selection criteria so that you gather or elicit material from only a certain group of men (of a certain age and nationality, or who speak a particular language, or have used the site for a particular length of time). Of course, a person's demographic or site-specific characteristics are not always clear from the profile information they publish on a social media site, and they might not wish for a researcher to use certain categories to describe them: those choices about categorising participants can sometimes be important depending on the kind of research project and approach. As we saw in Chapters 6 and 7, participants' perspectives on their identities are crucial to how the collection and analysis of data might proceed, and might change over time and across different sites. Nonetheless, for a quantitative approach that begins with pre-defined categories, controlling the variables in your project might be an important step in collecting data that enables you to answer your research questions satisfactorily.

A more subtle distinction can arise between collecting data about people's perceived use of language and their actual language use on social media sites. If the project aimed to compare how often different kinds of naming practices were used in Twitter by a group of people, asking people to answer a questionnaire about naming practices will not show the researcher how often people actually use different naming options in Twitter, it will collect data about how people self-report their language use. To find out about actual naming practices, you would need to observe a set of Twitter posts. Of course, self-reports of language practices are interesting in their own right (especially in the context of techno-autobiographies with a linguistic focus as described in Chapter 7), but they may not reveal the same patterns of behaviour that a researcher would observe when they examine actual language practices from the same people. In sociolinguistics, there have been important studies which documented how different groups of people tended to either over-exaggerate or downplay their tendency to use particular linguistic forms when interviewed, depending on the relative meanings associated with the form (Labov 1966 and Trudgill 1974). The same principles can also be at work when we investigate how people use language in social media contexts.

A second basic issue to consider when formulating a quantitative research method is whether or not the feature you are interested in can be counted or measured. How you count and compare particular phenomena as part of a quantitative analysis is a more complex question, and is considered in Chapter 9. But at the outset, it can be useful to think about the nature of the feature you want to analyse (your unit of observation) and how you might quantify it (unit of analysis). For example, some features seem relatively easy to quantify, such as how often a particular choice of word occurs. Other linguistic phenomena are harder to pin down, for example evaluative meanings are often constructed in a variety of means, some of which might appear as phrases (e.g. "I can't believe how difficult the task was"), some of which are words (e.g. the intensifer *really*), others are related to punctuation (like exclamation marks or non-standard spellings used for expression, such as *yayyyyy*) or paralinguistic features such as prosody or gesture. But these different aspects of language use often work in combination and the overall combined meaning is hard to reduce to single items that can be counted. The interpretation of evaluative meaning is also highly subjective and dependent on context: one person's opinion of evaluation can be different to someone else's, and determined by the surrounding text and additional factors like cultural values. Finally, some of the multimodal features that are so important in social media interactions (image, layout, icons, audio-visual resources) do not always have clearly defined units of analysis or map on to existing linguistic units.

Questions and variables

Research questions with a quantitative element often investigate variation that can be measured. The underpinning interest in variation necessarily implies a point of comparison and an attempt to map the relationship between one feature or factor and another. These kinds of relationships can be a simple comparison (do emoticons or exclamation marks occur more often in Facebook wall posts?), or they can trace correlations (does the gender of the blog post author correlate with the gender of the blog post commenters?), or they can imply a cause and effect (are apologies posted in Twitter perceived as more or less sincere than apologies posted in mainstream news?), or they might investigate change over time (how did the hashtag *#barackobama* get used before, during and after the American elections in 2008?). Both within sociolinguistics and within statistics, the features and factors that are compared are referred to as different types of variables.

Within the tradition of variationist sociolinguistics, variables are distinguished according to whether they refer to the linguistic features that are being investigated (the linguistic variable) or the factors which relate to the contexts in which the linguistic feature might be used (for example, who, where, when and for what purpose the language is used). These contextual factors are sometimes called *speaker variables* (if they describe the characteristics of the people) or *social variables* or *contextual variables*. Some of the examples of research questions given earlier illustrate the difference between the linguistic and social variables (see Table 8.1).

In experimental studies, the relationship between factors can also be described in terms of cause and effect and described as dependent and independent variables. Imagine a project where the researcher wanted to

Table 8.1 Examples of linguistic and social variables

Question	*Linguistic variables*	*Social/contextual variables*
Are temporal adverbs associated with the present moment more likely to occur in blogs or social network sites?	Temporal adverbs	Type of social media site (blogs or social network sites)
How does the length of time a person spends interacting on a social network site relate to their use of site-specific jargon?	Site-specific jargon	Length of time spent interacting on a social network site
What are the differences in how women of different age groups interpret the use of emoticons in online chat?	Meaning of emoticons	Age

find out whether a person's online anonymity influenced the amount of insults they posted to a social network site. The research question could be posed as, "How does the frequency of insults vary according to the anonymity of site members?" There are two factors (or variables) being compared in this question: the amount of insults generated and whether the site members choose an anonymous form of identity or not. The way data might be collected to answer the question might involve comparing the behaviour of site members who use anonymous representation and those site members who represent their identity through other choices (such as using a name or photograph that identified them). The independent variable is the factor which the researcher manipulates (here the anonymity of the site members) as a means of testing the possible causes of the language use. The dependent variable is the outcome or the effect (here, the frequency of the insults). The quantity of the dependent variable depends on the influence of the independent variable. One way to express this relation is to rephrase the question in a way that makes clear the dependency relationship, such as: "To what extent does the frequency of insults *depend on* the anonymity of the site members?" The difference between the dependent and independent variables can be illustrated with examples from the research questions given earlier (see Table 8.2).

Within statistics, there is a further set of subcategories applied to variables which distinguish between the kinds of measurements a researcher might want to make in their analyses. The different types of categorical variables include dichotomous variables, where a participant in a survey or questionnaire has only two choices of response. For example, if you asked a person if they owned a mobile phone, they could only answer "yes" or

Table 8.2 Examples of independent and dependent variables

Question	*Independent variable(s)*	*Dependent variable*
How does the length of time a person spends interacting on a social network site influence their use of site-specific jargon?	Length of time spent interacting on a social network site	Use of site-specific jargon
Does gender or anonymity most affect the amount of insults posted by a site member to YouTube?	Anonymity of site member Gender of site member	Amount of insults
Are apologies perceived as more sincere when posted to Twitter than when published in the mainstream news?	Medium of publication (Twitter or mainstream news)	Perceived sincerity of the apology

"no". Nominal variables describe responses which are presented as a series of non-numerical categories. For example, a question such as, "What devices do you use to access social media content?" might offer a number of responses, such as a smart phone, a PC, a laptop, a tablet, a game console and so on. Lastly, categorical variables can be ordinal, meaning that they are factors that can be scaled. A well known example of an ordinal variable is a Likert scale. A Likert scale (named after its creator, Rensis Likert) can be used to grade responses to questions on a numerical scale that indicates intensity of response, where the end points of the scale (usually of five or seven points) indicate the least and the most points of intensity. Likert scales are useful because they allow responses to be neutral, as well as in agreement or disagreement with a statement.

There has been a growing body of research in the arts, humanities and social sciences that have employed quantitative methods to explore a range of contextual variables and their role as dependent variables in influencing the creation and interactions found on social media sites. Given that social media sites very often automatically capture contextual information such as the time and the place of the interaction, these variables have been given fresh emphasis along with well recognised speaker variables such as a person's age or gender as indicated through the information stored in their profiles on particular sites (though as we noted in Chapter 3, this kind of information is not always the same as the characteristics the same person might claim for his or her identity in other online or offline contexts).

- Geographic location – often geo-tagged social media texts are used in studies that wish to make claims about the location-based properties of these texts. Sometimes the focus will be on location-based prediction (e.g. Kinsella *et al.* 2011).
- Time – temporal data is readily available in social streaming texts and is used to complement other forms of data such as location with a variety of aims (e.g. Hargittai and Litt 2011 provide a longitudinal analysis using survey data; Altman and Portilla 2012 examine the evolution of language in Twitter).
- Gender – identifying and predicting the gender of social media participants is a growing area of interest (Burger *et al.* 2011; Deitrick *et al.* 2012), and there are studies of the gender of @mentions in Twitter in relation to news stories (Armstrong and Gao 2010) and in relation to hashtags (Cunha *et al.* 2012). Multimodal studies are also beginning to appear looking at how gender is performed in social media images (Rose *et al.* 2012).
- Age – Studies of this variable may consider, for example, the effect of age and gender on blogging (Argamon *et al.* 2007; Rustagi *et al.* 2009; Schler *et al.* 2006).
- Community – since social media services function to establish networked relationships, many studies seek to explore the properties of the social network. For example they may explore the way in which information spreads through a network (Galuba *et al.* 2010; Kwak *et al.* 2010) and how key users 'influence' others within the network (Cha *et al.* 2010).

Most of this work has been undertaken outside the realm of linguistics. However, some studies do consider complementary linguistic variables such as:

- Language variety – this usually is inferred from the content of the social media text, although metadata about language is sometimes available. Some work is beginning to appear on dialectal variation (e.g. study of regional variation in slang (Eisenstein *et al.* 2010)).
- Community (based on language features not just links between users) – some studies attempt to automate classifying users into communities based on specific language features (e.g. based on noun phrases (Haythornthwaite and Gruzd 2007)).

No doubt more studies in the variation of social media language will continue to emerge long after this book is published, which address a range of research questions. You might like to evaluate the research questions and methods for gathering data in these (and other studies) along with your own project design by using the reflective questions in the list below.

Points for reflection

What kinds of variables are compared in your research question?
Can you count the linguistic features or language practices you want to investigate?
How broad or narrow is your research question?
What kind of data will you need to collect to answer it?
How might you need to control the variables in your project?

Sampling and representativeness

Having established your research question and identified the variables you want to investigate and how you want to measure them, you are some way to deciding what kinds of material you need to gather for analysis. These choices will determine whether you can gather existing materials together for content analysis (and so compile either a dataset or a large-scale collection of materials in a corpus) or whether you will need to elicit responses from participants (for example through a survey or in an experiment). Regardless of the choice of tool you use, there will be criteria you need to apply when selecting your material in terms of the choice of texts, contexts and participants, and in terms of how much material you might want to collect. It is worth saying again at this point that although the focus of this chapter is on the principles and issues relating to quantitative approaches, very often quantitative approaches can be combined with other kinds of analysis, and that, just as in the qualitative approaches covered in Chapters 5, 6 and 7, documenting the criteria used to collect your data for a quantitative project is an important part of the research process to be included in writing up.

When approaching the process of selecting data, it may be useful to think about the types and amount of texts we are collecting along a "continuum" of discourse data (Bednarek 2009). We may be interested in rich analysis of only a handful of texts or statistical analyses of large volumes of data. In addition we may wish to combine both of these approaches in a single study. For example, Baker (2006) advocates a "two-pronged" approach to analysis, combining the qualitative insights of close discourse of small volumes of texts with the quantitative analysis made possible by using a corpus. Extending this kind of approach, Bednarek (2009) suggests a "three-pronged" method to corpus-based discourse analysis. This approach incorporates close manual analysis of single texts with manual, or partially-automated, small-scale corpus-based analysis that might complement quantitative work, often highly automated, using large-scale million word corpora (Figure 8.1). For example, I might be interested in certain kinds of evaluative language that are used in blogging. In order to study evaluative meanings in blog posts I might begin with close discourse analysis of a single post, considering

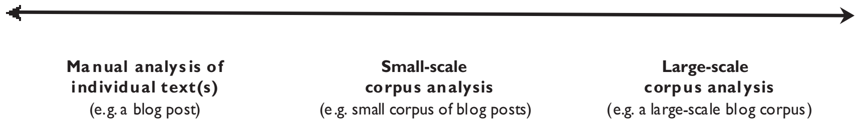
Manual analysis of individual text(s)
(e.g. a blog post)
Small-scale corpus analysis
(e.g. small corpus of blog posts)
Large-scale corpus analysis
(e.g. a large-scale blog corpus)

Figure 8.1 A continuum of discourse data. Adapted from Bednarek (2009, p.19).

how meanings are being made in a particular text type, for instance, analysing how evaluation functions in a post in a food review blog. I might complement this work by considering a range of blog posts collected as a small corpus of food review blog posts (assembled according to particular selection criteria). In addition I might like to compare how the language of food review blogging compares with general blogging by comparing them with a much larger set containing a representative sample of posts from a large volume of different kinds of blogs.

An important methodological concern is sampling, that is, determining the scope and criteria for collecting texts. Because it is impossible to examine all the possible language uses of all possible kinds of social media sites by all possible people, the researcher has to limit their choice of material in some way. In statistics, the ideal, generalised data a researcher might want to investigate (such as blogs written by men about mental health) is called the population and the actual set of materials that the researcher selects for analysis (the particular set of blog posts written by a particular set of blog authors) is called the sample. In the case of social media interactions, the sample can consist of texts, or of participants (e.g. survey respondents) or both, depending on the contextual variables being examined.

If a project is working within strictly quantitative methods with the aim of generalising the results of the research, then it may be important that the sample of selected material is representative, that is, that it matches the general characteristics of the population of interest. Imagine that a researcher wanted to examine the evaluative language used by women and men when they talk about a critical illness, like cancer. The researcher might start by finding out what kinds of cancers get blogged about most frequently and then select their data accordingly (for example, if more blogs were written about brain cancer than lung cancer, then the researcher might want to account for this difference in how many blogs of which type they collect). In other quantitative approaches, it might be more important to balance the size of each data sample proportionately, so for example if you were comparing the frequency of naming choices in Facebook and Twitter, you would want to collect equal numbers of posts from equal numbers of participants from each site.

The desire to produce quantitative findings that can be generalised can also determine the criteria used to decide how much data should be collected (how many participants, posts or survey responses and so on). The practicalities of carrying out a small-scale student term paper often mean that the researcher is working within constraints of time, budget and availability for recruiting participants. But this does not mean that sample size should be disregarded as a luxury available only to more advanced researchers. If the analysis will involve statistical tests, then it is important to make sure you collect enough data required for the particular test. It is easier to collect too much data and then select a subset for analysis later than not collect enough and have to go back to the participants or texts you are sampling from to gather more material afterwards. You should build in a 'safety margin' for collecting enough data to allow for the parts of the sample which might not be used in the final analysis (such as participants who drop out of a study, or who only partially complete a questionnaire; blogs or Twitter accounts that cease to be maintained actively over time, and so on).

The question of how much to collect depends not only on the extent to which you want to generalise from your findings (more is better) but also on the number of variables you are testing (the more variables you are testing, the more material you need). Dörnyei (2007, p.99) provides some approximate 'rule of thumb' measures, suggesting that for research that is attempting to correlate one factor with another, at least 30 participants or examples should be included, and for experiments, at least 15 participants per group. There are also mathematical formulae that will calculate the precise quantities that are suggested for projects that involve multiple variables, and we refer you to specialist quantitative methods handbooks for more details on how to apply these to your project design (see Rasinger 2008).

Points for reflection

What is the population and the sample that you intend to study in your project?

What factors might require you to build in a 'safety margin' for your data collection?

How will you balance the number of participants and texts in your sample?

Quantitative tools for eliciting data

Once you have identified the variables in your project and formulated your research question, and decided how many participants or examples you need to gather you then need to consider which tools might be best to collect the material you will analyse. There is a broad distinction between quantitative tools that can be used to elicit information (that is, to generate new material for analysis) and the processes used to gather existing material together (in a dataset or corpus). In the remainder of this chapter we cover two well-known tools for eliciting quantitative data (surveys and experiments) and close by considering the factors which are important for designing and gathering material for a social media corpus. The factors which underpin the design of surveys, experiments and corpora were well-established prior to the advent of social media genres, so the outlines presented here will refer you to existing, more detailed textbooks on the subject at various points. However, the characteristics of social media which enable rapid, collaborative interactions with potential large audiences mean that these tools can be adapted and adopted with new emphasis for research about the language used on social media sites.

Surveys

Using a survey in the form of a questionnaire to be completed by the participants can be an efficient way of gathering large amounts of data in a relatively short amount of time. This is especially true when the process of gathering the survey responses can be handled through an online mechanism and distributed to a large audience through social media sites (such as a Facebook Friend list or a Twitter Follower list), and then re-distributed to further audiences as the survey is posted on again by the respondents to their own Friend or Follower list in a snowball effect. However, being able to elicit large amounts of material is only useful if the material can be used to answer the research questions set by a project. The design of a questionnaire is thus crucial: you need to make sure that you ask questions that will prompt responses useful to your project. It might seem very simple to make this observation, but working out how to frame and phrase survey questions effectively takes some thought and it is always essential to do some sort of piloting.

There are several factors that are important when preparing survey questions. First, different types of questions can be selected depending on how you want to quantify your results. Closed questions can be used if you want to measure dichotomous responses, such as "Are you a member of Twitter?" which can be answered with either an affirmative or negative response. Other questions can be used to gather graduated responses, which might be more useful if you are trying to elicit self-reported accounts of behaviour or opinions about a topic. With these questions, multiple choice or rank ordered responses might be useful which employ ordinal or non-numerical options. For example, you might ask participants to rank which kinds of social media sites they find most useful for accessing information about current events, or to indicate from a list of options which tools they use for accessing social media sites. Second, just as in face-to-face contexts, the way in which a survey question is phrased can make a great deal of difference to the way in which people answer it. Questions need to be worded in simple, unambiguous language, and to avoid jargon that may not be understood by the people answering the questions. The wording of questions should try to avoid presupposing a particular answer or using language that might be clearly evaluative (for example, compare the difference in asking, "What do you hate about the Facebook timeline?" and a more neutrally worded, "How would you describe the effects of the Facebook timeline?") Questions should only require one response: complex or multiple questions with only one response option make it difficult to process the responses. Finally, open-ended questions can be included in a survey (though strictly speaking they generate material that is more suited to qualitative analysis rather than a purely quantitative approach). These open-ended questions can be of different kinds: used to elicit responses to specific questions, or as a follow-up clarification. For example, it might provide a space for respondents to explain why they had ranked their choices in a particular order, or to provide further context about an answer given previously.

Surveys can be administered in a variety of formats: participants can be given paper copies of the survey and asked to complete them in writing, researchers can ask the participants questions and then fill in the answers on their behalf (either face-to-face or by telephone), or online survey templates can be used, which enable the participants to type in their answers to an online form set up by the researcher (and which then allow the researcher to export the responses in an electronic format of some kind). Different social media contexts can expand the range of survey tools too: the rating tools available (such as the 'like' button in sites like Facebook) can be used to create simple polls, and more bespoke polling tools have been created as apps that work with other social media sites like Twitter to allow participants to respond to questions or presentations in real time. Other kinds of contextual data automatically generated from devices like smart phones and streamed into social media services (such as location data in the form of GPS information, live-streamed video content from mobile camcorders or details generated from check-ins) can also be used in quantitative surveys that might wish to correlate participants' language or responses with their geographical location. However, as with surveys which are carried out in offline contexts, the wording of the question and the formats provided for responses will determine the data that is generated (even if this is in a form that is readily automated).

Experiments and quasi-experiments

Unlike surveys where the researcher observes relations between linguistic and contextual variables, in experiments or quasi-experiments, the researcher alters one (or more) independent variables in order to test its effect on the dependent variable. It is often quite difficult to create experiments where the influence of additional variables can be eliminated from the research adequately (see Dörnyei 2007; Rasinger 2008). For example, imagine that you wanted to test whether using a social network site influenced the slang terms that were adopted when students joined a college course. Before they joined the college course, you might ask them to list all the slang terms they knew and then once they had joined the course ask them to repeat the test. But how could the researcher be sure that it was using the social network site that caused the change in slang terms reported by the students after they had joined the college course? It might be possible that students would just acquire new slang terms anyway, without using the site to interact with their peers. One way in which researchers can try to ensure that the influence of a particular variable on a given group is the cause of the observed effect is to compare the group who has been manipulated in some way with another comparable group who has not been manipulated: a control group. In the (fabricated) example given here, a control group of students (matched in terms of number and demographic character with the test group) who did not use the social network should also be tested to see whether their knowledge of slang terms also changed as they entered college, and if so, whether they changed to the same extent as their site-using counterparts.

Group 1: students who do not use the social network site tested for their knowledge of slang terms before joining college.
Group 2: students who use the social network site tested for their knowledge of slang terms before joining college.
Group 1: students who do not use the social network site tested for their knowledge of slang terms after joining college.
Group 2: students who use the social network site tested for their knowledge of slang terms after joining college.

Quasi-experimental research can use simulated scenarios to test participants' perceptions of language use, or their anticipated responses. For example, Schultz *et al.* (2011) presented respondents with fictional apologies made by a company but with one version posted to Twitter whilst another was posted to a blog. The research then examined the effect of the social media context (Twitter or blog) on the participants' perceptions of the apology's sincerity. The ability to create simulations in certain social media contexts such as virtual worlds would seem to lend itself to quasi-experimental research, but there are also many unknown and contentious issues related to research design in these contexts (such as being sure about the demographic characteristics of the participants, and how far the observer's paradox might operate in these contexts). In the last part of this chapter, we shift our attention from the tools used to elicit data for a quantitative project, to some of the technical processes involved in gathering existing texts that can be used for a particular subfield of language study: corpus linguistics.

Collecting data for a quantitative project using corpus linguistics

A *corpus* is a technical term used within the field known as Corpus Linguistics to refer to "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" (Sinclair 2005). Corpus Linguistics was made possible as an academic discipline with the advent of computer processing technology[1] that allowed researchers to search for words and textual patterns in large volumes of digitised text. A social media corpus is a corpus that includes texts collected from social media services such as WeChat, Twitter and Facebook. The included texts may be verbal content encoded in plain text format (e.g. the written content of a micro-blog post). The texts may also incorporate multimodal resources such as images and video which require metadata in order to be processed and analysed in a quantitative project.

Due to the careful design principles that need to be applied when building a corpus, it is "not simply a collection of texts" (Biber *et al*. 1998, p.246). A corpus is only as useful as the rigour of its selectional constraints (the parameters used to decide which material to include) and the relevance of these selection criteria to the research question at hand. When building a corpus a researcher will take into account issues of representativeness (the extent to which the corpus can be said to generalise a particular variety or dimension of language given the variation possible in a linguistic sample) and balance (the range of genres, text types or linguistic features included in the corpus) (Sinclair 2005). For example, is a corpus of blog posts representative of the communicative patterns found in journalism blogging, or does the sampling of the texts (how many and what kind of text) mean that some text types are more frequent than others in the corpus? If you want to make claims about blogging within journalism, how many texts would you need to analyse? What range of blogs would you need to gather these texts from and how many different electronic sources would you need to consider? Applying Sinclair's (2005) suggestion of some important corpus design factors (shown in italics below) to the domain of social media corpora, we might consider:

- *Mode* – social media involve a range of semiotic modes (e.g. spoken and written modes in YouTube videos and their attendant comments/video responses).
- *Text type* – social media encompass a variety of text types and genres (e.g. blog posts, comments which include genres such as observations, anecdotes, etc.).
- *Domain* – social media are used across multiple domains, for example specialist or academic domains (e.g. live-tweeting[2] a conference) and popular domains (e.g. live-tweeting TV).

- *Language(s)/language varieties* – most social media are cross-cultural with a variety of languages present.
- *Location* – since social media are enacted online the entire concept of place becomes problematic.
- *Date* – time is crucial to social media and social media services usually encompass forms of streaming data (see explanation later in this chapter).

An important distinction that is commonly made when classifying corpora is the difference between a specialised corpus that aims to study a particular type of text (e.g. restaurant reviews in 'food blogs') and a general corpus intended to span a representative sample of language across a range of text types (e.g. a reference corpus such as The British National Corpus or the Corpus of Contemporary American English). There are many other ways we might classify corpora, but given the significant role of time in social media, issues surrounding the development of traditional diachronic corpora (corpora that include texts produced over a particular time period used to study language change) are of particular relevance. Some examples of the types of corpora which include social media texts are listed in Table 8.3, along with examples of published papers (where available) which show how these corpora have been used.

Is the web a corpus?

Just as the World Wide Web (which is effectively an unprocessed collection of hypertext documents linked together in various ways) should not be thought of as a corpus[3] "because its dimensions are unknown and constantly changing, and because it has not been designed from a linguistic perspective" (Sinclair 2005) we should not think of the data made available by social media services

Table 8.3 Examples of different sorts of social media corpora

Type of corpus	*Example*	*Studies using the corpus*
Specialised	Various hashtag corpora	Page (2012)
Reference	HERMES Twitter Corpus	Zappavigna (2012)
	Birmingham Blog Corpus	Kehoe and Gee (2012)
	Cambridge and Nottingham e-Language Corpus	Knight *et al.* (2014)
	Corpus of Global Web-based English	
Diachronic	Twitter Stratified Random Sample SRS	None available at time of writing

as corpora. Collections of social media texts evolve as people add, modify and delete content and the linguist needs to approach this naturally-occurring text with the same careful gaze used in other data collection contexts. Searching with a search engine provided by a social networking service (e.g. Twitter's search interface) is not the same thing as searching a corpus. The algorithms used by these search functions are generally not available for public inspection, and thus we do not have adequate information about their parameters or which texts have been included to determine whether the search results map to what might have been returned using a corpus designed according to particular selection criteria. Nevertheless, just as online discourse has proven to be useful to linguists working in many areas, including lexicography, syntax, semantics and translation (Volk 2002) and corpora derived from Computer Mediated Communication (CMC) begin to appear, social media corpora will eventually gain traction.

Researchers usually build their own social media corpora since CMC is a semiotic mode that remains under-represented in most available traditional corpora. However, very large web-based billion-word corpora are beginning to emerge, such as the 25 billion word USENET corpus (2005–9) (Shaoul and Westbury 2010), a corpus consisting of public USENET postings collected between October 2005 and January 2010, and the Birmingham Blog Corpus (http://wse1.webcorp.org.uk/blogs/), consisting of approximately 629 million words of blog texts extracted from the web. In addition the field of Computational Linguistics has generated many very large Twitter corpora, particularly in the area of sentiment analysis (e.g. Bollen *et al.* 2011). An example of a corpus incorporating data from a range of social media types is the Cambridge and Nottingham eLanguage Corpus (CANELC), designed to form part of the Cambridge International Corpus. This corpus is a one million word collection of electronic communication including emails, tweets, blogs, text messages, post in electronic fora and chat room communication. Depending on the nature of the corpus and the institution within which it was created, access will either be free or via paid subscription (e.g. the USENET corpus (2005–9) is currently available as a Public Data Set on Amazon Web Services). The page hosting a corpus will usually provide links to researcher projects that have used the corpus and this can be a useful way to find relevant literature and people who are likely to be interested in the work that you are doing.

Understanding metadata

Metadata is information about information. It usually describes something about the content of digital media, such as how it was created, its display, or its context. There are different kinds of metadata that can be created by an individual or automatically generated by software. For instance, a digital image might incorporate three forms of metadata: *technical metadata*, including information about the technical specification of the image captured by a camera, such as the ISO speed at which the image was taken, *descriptive metadata* such as keywords about the image assigned by the content-creator, and *administrative metadata* such as licensing information (Photometadata. org 2011). Social media services make use of a large range of metadata that can be mined by third parties. For example a service may attach information about the time and location at which particular content was created. An example of a linguistic study that has made use of such metadata is research into lexical variation using a geo-tagged Twitter corpus[4] that found that many slang terms such as *af* (as fuck) and *hella* (very) "have strong regional biases, suggesting that slang may depend on geography more than standard English does" (Eisenstein *et al.* 2010, p.1285). Other work adopting this kind of approach includes Brice Russ' study of dialect variation. In his case study, he describes the processes he used to exploit the geo-tagging metadata in a social media study of dialectology.

Box 8.1 Twitalectology: examining large-scale regional variation through online geo-tagged corpora

Brice Russ

Dialectology is a key component of sociolinguistic study; studying regional variation can often provide valuable insights into other social aspects of language, and discussing dialects can be a useful tool in building general interest in linguistics. Mapping dialects on a broad scale, however, has traditionally required surveying large, geographically scattered populations in a process that can take years of time and significant effort.

Twitalectology seeks to examine whether regional variation in some linguistic items can be studied and analysed using data from Twitter. If so, data from these variables could be collected in a matter of weeks or months without elicitation or supervision, making it significantly easier to conduct dialectological studies. Selecting linguistic variables which were suitable for Twitalectology was a challenging process. Twitter, being a textual medium, cannot accurately represent phonetic or phonological variation in most instances, so I limited my first round of experimentation to lexical and morphosyntactic variables. I also wanted to use variables which exhibited variation across a significant part of the continental United States, and to incorporate both well-studied regional variables and variables which had not yet been thoroughly nationally surveyed.

After examining data from the Harvard Survey of North American Dialects and other linguistic surveys, I decided on three variables: soft drink terms (specifically, *soda* versus *pop* versus *coke*), the usage of *hella* as an intensifier (in contrast with, for example, *very*), and the *needs X-ed* construction, as seen in phrases such as *The car needs washed*. Using a Python script which sent requests to the Twitter API (application programming interface), I constructed a corpus of tweets which contained a token of at least one of the desired linguistic variants, as well as the self-reported location of the account posting each tweet.

Certain terms in the corpus, however, may not refer uniquely to the desired variant; a tweet about *pop*, for example, could be talking about *Pop Tarts*, *pop music*, or other non-soda-related terms. To disambiguate these word-senses, I used collocations: series of words which, when juxtaposed, can refer to a unique concept. For soft drinks, for example, I removed all tweets that did not contain the collocations *drink X* or *drinking X* (e.g. *drinking pop*); those which remained were almost categorical examples of the desired linguistic variable. I then calculated the frequency of each linguistic variant on a city-by-city level and plotted the results on Google Maps; an example of this work can be found at www.briceruss.com and in Chapter 9 of this book.

Because corpus-based approaches to computer-mediated discourse require insights from a variety of linguistic fields, many of which are new and cutting-edge, I recommend that any prospective researchers not be afraid to ask others for help. Twitalectology would not be possible without the contributions and advice of many colleagues. I'd also note that a little programming knowledge can go a long way; I'm definitely not a computational linguist by nature, but I found that the scripting skills I did know went a long way towards helping me explore and analyse the large datasets that this project entailed.

#advice: Corpora created using data from social media can be useful and easy to collect, but don't trust them blindly.

Metadata can often appear in 'unfriendly' forms that appear difficult to make sense of. Some of the metadata that were collected in the process of building the HERMES corpus are summarised in Table 8.4. Although difficult to read, these automated examples of metadata can be very useful in a research project, for example the geo-tag, as we have seen in the studies on dialect variation mentioned earlier in this chapter, can be used to filter posts based on the location where they were produced. Location information of this kind will not always be available as this kind of feature is typically 'opt-in' due to the privacy concerns of social media users. This may also be the case with other forms of metadata such as information about gender. When designing your study you will need to consider whether the dimensions that you are interested in exploring correspond to the kinds of features that are tracked in a particular social media service's metadata. For example, some

Table 8.4 Examples of metadata about a post and a user account

Type of metadata	*Tag*	*Description*
About the post	created_at	Coordinated Universal Time (UTC) timestamp for tweet creation
	in_reply_to_screen_name	Display name for the user that replied
	Geo	Geotag (location information)
	Source	Application used to tweet
	in_reply_to_status_id	Unique id for the user that replied
	Lang	language
	time_zone	The user's time zone
About the user's account	statuses_count	Number of posts the user has made
	profile_image_url	The URL of the user's avatar file
	utc_offset	Time between user's time zone and UTC time

proprietary archiving systems will provide this kind of metadata in formats that can be exported for further analysis. If you are not using one of these systems to archive material, then you may need to consider whether you have the resources to create bespoke archiving tools to extract the information you need.

Aside from the metadata recorded by social media services about the context of a social media text, another important kind of social media metadata is created when users themselves produce tags to classify their own content. These might be labels assigned to blog posts or hashtags used in micro-blog posts. A broader definition of what we might term 'social metadata' includes "information about a resource resulting from user contributions and online activity—such as tagging, comments, reviews, images, videos, ratings, recommendations—that helps people find, understand, or evaluate the content" (Smith-Yoshimura and Shein 2011, p.10). The kind of collaborative tagging evolving with community use in social media is often referred to as a practice of folksonomy (Vander Wal 2007), or social tagging. This community-based metadata is very different to the top-down hierarchical approaches developed by subject classification in libraries. Whereas document classification involves experts, social tagging engages communities of general users. For example, it is used heavily on photo-sharing sites, such as Flickr, where it often functions as a cooperative form of verbal indexing involving a bottom-up approach to the kind of classification previously achieved by reference librarians (see Barton and Lee (2012) for an exploration of Flickr and the changing literacy practices involved in social tagging). An example of a linguistic study that leverages the insight into the ‘topic’ of texts afforded by collaborative metadata is work by Kehoe and Gee (2012), who in their work used Delicious tags and corpus linguistic methods to aid how researchers might go about determining textual ‘aboutness’. Unfortunately since different social media services will store and present this kind of information in different ways there generally is not one elegant solution for isolating this phenomena so that it can be used in a research project.

Social-streaming data

In order to build a social media corpus we need to understand the particular properties of the data we are collecting. Social media aims to allow people to connect with each other seamlessly during their daily lives. Thus most services will incorporate some form of ‘streaming’ capability whereby a user can broadcast a chronologically organised ‘feed’ of information (e.g. status updates, blog posts and vlogs) that can be delivered to other users in near real-time, depending on how other parties choose to consume their social media feeds.[5] This capability is sometimes referred to more generally as the ‘real-time web’. Users are able to subscribe to feeds of their associates’ status updates and multimedia content (e.g. photos and video). Often these updates are shared via a mobile device at the time an event occurs or an observation is made. The advantage that social streams of this kind offer to the researcher is the hope of automated data collection using technologies to access and store the data (which, as we will see in the following section of this chapter, can be achieved if the site makes its Application Programming Interface (API) available, as do Twitter and Wikipedia).

Streaming data does, however, pose some challenges for researchers, particularly those interested in exploring structural patterns, such as the ‘conversational’ interactions between users that can occur on social media. These include considerations such as:

- What is a useful time frame for a ‘snapshot’ of the unfolding social stream that you will capture? What will this time frame reveal? What are its limitations? How will the time frame skew the kinds of language you might find in your sample (e.g. increased frequency of sport-related lexis during international sporting events like the Olympics)?
- What kind of connections between user accounts or between unfolding instances of language use (e.g. a micro-blog post or a wiki edit) do you want to capture, keeping in mind that you may end up needing to track many-to-many relationships?

- How will you keep track of relationships between embedded and externally referenced multimedia (e.g. pictures referenced in a micro-blog post)?
- How will you cope with the changing nature of the technology underlying streaming data when it changes and causes collection tools to become redundant? (E.g. updates to the Twitter Application Programming Interface (API) have 'broken' many tools developed to work with the first version.)

As we suggested earlier in this chapter, social-streaming discourse is highly temporally bound, since its real-time production has increased the capacity of contextual variables to skew the data: people will often post about events as they happen and about topics that are on their mind at a particular time, often in reaction to shared situations. The time and duration of sampling from a social stream impacts on the representativeness and balance of language retrieved.

Alongside these abstract concerns is the practical problem of sampling in a principled way the vast amount of streaming data generated by social media services and the technical skills required to do this. Generally some basic programming and text processing skills are required to work with streaming data since most social media services will not release a tool with an interface that allows people to easily detect, sample and extract their data. Those services are instead likely to expect developers and researchers to interact with an Application Programming Interface (API). This is because the people working with social-streaming data are usually third-party developers with high-level programming skills who are using it as input into software applications that they are building. An API is the language that software tools use to communicate with a social media service's back-end database. If the API is public, developers can use it to write custom applications that interface with the service's data feeds, allowing them to 'scrape' its data feeds, in other words to download selected types of data. We deal with this process in the next section. If you do not possess basic programming skills it will usually be necessary to collaborate with someone with the appropriate technical background or investigate the range of paid data collection services that have emerged to meet demand for social media data in areas such as marketing. As Brice Russ points out, analysing the language of social media can often turn out to be a collaborative endeavour across the boundaries of different disciplines.

Scraping social-streaming data

Social media scraping is a form of the more general technique known as web scraping, the process by which data is automatically collected from the websites using custom software. Collecting data from social media services usually involves working with the particular service's Application

Programming Interface (API). For example, the Twitter API is used by many third-party developers to create applications that deploy this streaming data in a range of ways, from simply allowing a user to track different kinds of phenomena that they may be interested in, to complex data visualisation and social media analytics. Social media data of this kind is often used in web 'mash-ups', combining the functionality from two or more sources (such as data made available by an open API) to create a new service that meets some particular, novel need. For example, Twittervision (Twittervision.com) is a web mash-up that combines Twitter feeds with Google Maps to create a display of tweets unfolding in real time on a map.

There is a range of custom software tools available to assist with scraping data from different resources, the details of which are not important here as they rapidly become obsolete. The best way to find information about where current resources might be found is to search developer message boards where you will find developers and research working on social media scraping software (e.g. the Twitter Development Talk Google Group). We should keep in mind that there is no universal solution due to the frequency with which social data changes. These tools need to keep up with the evolving nature of web services and the particularities of different APIs. What will remain more stable, however, is the general principle that, however streaming data is collected, it must meet the kind of selection criteria for building corpora that were introduced earlier in this chapter, inflected as they are by some difficult questions regarding how time units affect dimensions such as representativeness and balance.

Often linguists will need to create their own software (or hire a developer) in order to undertake the kind of scraping required for a particular research project. For example, in order to build the Hermes corpus, I (Michele) used a simple script that repetitively downloaded tweets using the Twitter API as this was the simplest solution at the time given the limited custom software available. The general method used to build the Hermes corpus (Zappavigna 2012, p.24) was:

1. Use a script to interact with the Twitter API and download all the unfiltered tweets from Twitter across a particular time-window or until a certain quantity is captured (keeping in mind that, in step 4, non-English tweets will be removed, reducing the overall number of tweets).
2. Separate the content of the tweets from other metadata, depending on how they will be imported/processed by the particular concordance software to be used with the corpus.
3. Convert any entity sequence, such as escaped characters,[6] into their native form.
4. Filter the text so that it contains only English tweets (or tweets from the particular language of interest). This step is not an exact science as current language filtering technologies are not 100 per cent accurate.

Another significant issue is the form that the scraped data will take once it has been captured. This again depends on the research design. If variables found in the metadata as well as the content of the social feed will be analysed (e.g. location data plus the content of a micro-blog post) then the material may need to be stored in a database, however often this solution does not scale well for large amounts of data which may require custom-designed scripts for processing. The reason why these technologies may be required is that calculating multiple relationships between large numbers of variables in large volumes of data can require a lot of computing power. If, however, more traditional text analysis using, for example, a concordancing system will be undertaken on only the textual content of a social media text (i.e. the linguistic patterns in, for example, a micro-blog post without the attendant metadata) then it may best be stored as plain text. However, this is not necessarily a simple option since getting the data into plain text format may require stripping it of superfluous information as well as processing the kinds of peculiarities that arise from differences in encoding formats (which we will discuss later in Chapter 9). In summary, a linguist thus has the following options in terms of data collection when using a social media corpus: use an existing corpus, modify that corpus in some way, or build a new corpus. Some questions that you might like to ask yourself when planning a quantitative research project are summarised in the following list.

Points for reflection

Is there an existing corpus of that type of text which is already available and accessible that you might use for your analysis?

If you need to build a specialised corpus of some kind, is this restricted to material from a particular site?

Does the site you want to study have an API which allows automated text extraction?

What automated tools or services are available that can be used to gather material from the site you are interested in?

What contextual factors are important in your study and are they available as a form of metadata that can be collected automatically?

What format will the ‘scraped’ data be provided in? Will you need to adapt this to accommodate the size of the files, or to enable analysis using other automated tools (such as concordancing software or visualisation models)?

The process of selecting a particular text type and compiling a corpus for analysis is described in Andrew Kehoe’s case study of building the Birmingham Blog Corpus, with which we conclude this chapter.

Box 8.2 Using reader comments to help determine the topic of blog posts

Andrew Kehoe

In order to index the web and make it easier for people to find the information they need, it is essential to determine what online texts are about. In the early days, web texts were often classified manually in topic hierarchies like Yahoo. With the growth of the web and recent explosion of social media, it has become increasingly important to develop effective techniques for determining topics automatically. Our focus was on blogs. We wanted to determine whether reader comments on a blog post would provide us with information about the topic of that post which could not be determined by looking at the post alone.

Our approach was a corpus linguistic one, compiling large amounts of data and using statistical analyses to extract the 'key' topic-related words from each post and set of comments. To achieve this, we needed a large collection of blog data, and this case study focuses on how we compiled the *Birmingham Blog Corpus (BBC)*. In doing this, we faced two main challenges. The first was in finding a source of blog data large enough to contain posts on a wide range of topics and to allow us to draw meaningful conclusions from our analyses. The second was in separating the reader comments from the blog posts which, given the required corpus size, would need to be achieved with minimal human input.

With these requirements in mind, we turned to the blog-hosting sites Blogger and WordPress. These sites are vast – Google returns more hits from Blogger (over 840 million) than it does for all *.edu* and *.gov* sites combined – and cover a diverse range of topics, from technology to tennis and politics to parenting. However, despite this apparent heterogeneity, we found that each hosting platform has a limited number of text formatting conventions (e.g. Blogger comments begin with the tag < *dd class* = "*comment-body*" > or similar). This is ideal in corpus compilation as it makes the removal of 'boilerplate' (advertisements, headers, menus) and the separation of comments from posts much more straightforward. Another advantage of blogs over general web data is that accurate publication dates are recorded. We were careful to preserve these in our corpus to assist users interested in language change across time.

To collect our data, we started with the lists of 'trending' blogs made available on each hosting site: 'Blogs of Note' on Blogger and 'Freshly Pressed' on WordPress. We wrote Perl scripts to download each post from each featured blog, together with its associated comments. From this initial dataset we extracted all links to other blogs, whether these were in a post, in a comment, or in a 'blogroll' (list of recommended blogs). New blogs were then added to the crawling queue, widening our coverage beyond the initial 'trending' list. This automated process continued for a month, during which time we processed 222,245 posts, totalling 95 million running words of text. In addition, we downloaded 86 million words of associated comments, revealing a wealth of linguistic knowledge which we could utilise to improve document indexing.

The Birmingham Blog Corpus is available to search at www.webcorp.org.uk/blogs.

#advice: Have clear requirements, understand the data, preserve info that may be useful later (to you or others), share your results and your data.

Notes

[1] Corpora, however, have been used to study language as early as the thirteenth century when monks painstakingly compiled manual concordances of the Christian Bible, and non-digital corpora were used during the 1950s in work on English grammar (O'Donnell, in press).

[2] Live-tweeting refers to the practice where users will post real-time reactions to or descriptions of events or media such as public gatherings or television programmes.

[3] Nevertheless, there is a body of research interested in finding ways to process web data so that it may be made useful to corpus linguists and more closely approximate the rigours of traditional corpora (Baroni and Bernardini 2006; Hundt *et al.* 2007; Kilgarriff and Grefenstette 2003).

[4] There are also studies that use geo-tagged corpora to create models used to predict the location of tweets based on the language patterns in the post (Kinsella *et al.* 2011).

[5] Many users adopt tools, such as a feed reader, to aggregate multiple web feeds into a single view, meaning that they do not have to visit sources individually for current information.

[6] An escape character e.g. \ (a single backslash) signals that the character (or sometimes the sequence) following it is not an operator (a symbol interpreted by the computer as 'syntax' for a program) or some other special case.

References

Altman, E. and Portilla, Y. (2012) *Geo-linguistic Fingerprint and the Evolution of Languages in Twitter*, paper presented at the Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) 2012, 26–29 August 2012, Kadir Has University, Istanbul, Turkey, available: http://hal.inria.fr/docs/00/69/06/08/PDF/c6.pdf [accessed 12 November 2013].

Argamon, S., Koppel, M., Pennebaker, J. and Schler, J. (2007) 'Mining the blogosphere: age, gender, and the varieties of self-expression', *First Monday*, 12 (9), available: http://firstmonday.org/ojs/index.php/fm/article/view/2003/1878 [accessed 26 November 2013].

Armstrong, C.L. and Gao, F. (2010) 'Gender, Twitter and news content', *Journalism Studies*, 12 (4), 490–505.

Baker, P. (2006) *Using Corpora in Discourse Analysis*, London and New York: Continuum.

Baroni, M. and Bernardini, S., eds (2006) *Wacky! Working Papers on the Web as Corpus*, Bologna: GEDIT. Available: http://wackybook.sslmit.unibo.it/ [accessed 12 November 2013].

Barton, D. and Lee, C.K. (2012) 'Redefining vernacular literacies in the age of web 2.0', *Applied Linguistics*, 33 (3), 282–98.

Bednarek, M. (2009) 'Corpora and discourse: a three-pronged approach to analyzing linguistic data', in Haugh, M., Burridge, K., Mulder, J. and Peters, P., eds, *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian*, available: www.lingref.com/cpp/ausnc/2008/paper2283.pdf [accessed 26 November 2013].

Biber, D., Conrad, S. and Randi, R. (1998) *Corpus Linguistics: Investigating Language Structure and Use*, London: Cambridge University Press.

Bollen, J., Mao, H. and Xiao-Jun, Z. (2011) 'Twitter mood predicts the stock market', *Journal of Computational Science*, 2 (1), 1–8.

Burger, J.D., Henderson, J., Kim, G. and Zarrella, G. (2011) 'Discriminating gender on Twitter', *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Edinburgh, United Kingdom.

Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, K.P. (2010) 'Measuring user influence in Twitter: the million follower fallacy', Washington, DC, 23–26 May 2010. California: The AAAI Press. Available: www.aaai.org/Library/ICWSM/icwsm10contents.php [accessed 21 May 2011].

Cunha, E., Magno, G., Almeida, V., Andr, M. and Benevenuto, F. (2012) 'A gender based study of tagging behavior in Twitter', *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, Milwaukee, Wisconsin, USA: ACM, 323–24.

Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T., Hu, W. and Cusani, R. (2012) 'Gender identification on Twitter using the modified balanced winnow', *Communications and Network*, 4 (3), 189–95.

Dörnyei, Z. (2007) *Research Methods in Applied Linguistics*, Oxford: Oxford University Press.

Eisenstein, J., O'Connor, B., Smith, N.A. and Xing, E.P. (2010) 'A latent variable model for geographic lexical variation', *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1277–87, available: www.cs.cmu.edu/~nasmith/papers/eisenstein+oconnor+smith+xing.emnlp10.pdf [accessed 26 November 2013].

Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z. and Kellerer, W. (2010) 'Outtweeting the twitterers-predicting information cascades in microblogs', paper presented at the 3rd Workshop on Online Social Networks (WOSN 2010).

Hargittai, E. and Litt, E. (2011) 'The tweet smell of celebrity success: explaining Twitter adoption among a diverse group of young adults', *New Media and Society*, 13 (5), 824–42.

The Harvard Survey of North American Dialects, available: www4.uwm.edu/FLL/linguistics/dialect/ [accessed 12 November 2013].

Haythornthwaite, C. and Gruzd, A. (2007) 'A noun phrase analysis tool for mining online community conversations', in Steinfield, C., Pentland, B., Ackerman, M. and Contractor, N., eds, *Communities and Technologies*, London: Springer, 67–86.

Hundt, M., Nesselhauf, N. and Biewer, C. (2007) *Corpus Linguistics and the Web*, Amsterdam: Rodopi.

Kehoe, A. and Gee, M. (2012) 'Reader comments as an aboutness indicator in online texts: introducing the Birmingham Blog Corpus', in Oksefjell Ebeling, S., Ebeling, J. and Hasselgård, H., eds, *Studies in Variation, Contacts and Change in English Volume 12: Aspects of Corpus Linguistics: Compilation, Annotation, Analysis*, University of Helsinki e-journal.

Kilgarriff, A. and Grefenstette, G. (2003) 'Introduction to the special issue on the web as corpus', *Computational Linguistics*, 29 (3), 333–47.

Kinsella, S., Murdock, V. and O'Hare, N. (2011) '"I'm eating a sandwich in Glasgow": modeling locations with tweets', *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, Glasgow, Scotland, UK.

Knight, D., Adolphs, S. and Carter, R. (2014) 'CANELC – The Cambridge and Nottingham eLanguage Corpus', *Corpora* 9 (1). In Press.

Kwak, H., Lee, C., Park, H. and Moon, S. (2010) 'What is Twitter, a social network or a news media?' *The 19th World-Wide Web (WWW) Conference*, 26–30 April, Raleigh, North Carolina.

Labov, W. (1966) *The Social Stratification of English in New York City*, Washington, DC: Center for Applied Linguistics.

Lexicalist, available: www.lexicalist.com/ [accessed 12 November 2013].

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.

Photometadata.org (2011) Classes of metadata, available: www.photometadata.org/node/46 [accessed 12 February 2014].

Rasinger, S. (2008) *Quantitative Research in Linguistics: An Introduction*, London: Continuum.

Rose, J., Mackey-Kallis, S., Shyles, L., Barry, K., Biagini, D., Hart, C. and Jack, L. (2012) 'Face it: the impact of gender on social media images', *Communication Quarterly*, 60 (5), 588–607.

Rustagi, M., Prasath, R., Goswami, S. and Sarkar, S. (2009) 'Learning age and gender of blogger from stylistic variation', *Pattern Recognition and Machine Intelligence*, 205–12.

Schler, J., Koppel, M., Argamon, S. and Pennebaker, J. (2006) 'Effects of age and gender on blogging', paper presented at the AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs.

Schultz, F., Utz, S. and Göritz, A. (2011) 'Is the medium the message? Perceptions of and reactions to crisis communciation via Twitter, blogs and traditional media', *Public Relations*, 37, 20–7.

Shaoul, C. and Westbury, C. (2010) A USENET corpus (2005–9), available: www.psych.ualberta.ca/~westburylab/downloads/usenetcorpus.download.html [accessed 17 March 2011].

Sinclair, J. (2005) 'Corpus and text – basic principles', in Wynne, M., ed., *Developing Linguistic Corpora: A Guide to Good Practice*, Oxford: Oxbow Books, 1–16, available: http://ahds.ac.uk/linguistic-corpora/ [accessed 12 November 2013].

Smith-Yoshimura, K. and Shein, C. (2011) *Social Metadata for Libraries, Archives and Museums Part 1: Site Reviews*, Dublin, Ohio: OCLC. Available: www.oclc.org/research/publications/library/2011/2011–02.pdf [accessed 12 November 2013].

Trudgill, P.J. (1974) *The Social Differentiation of English in Norwich*, Cambridge: Cambridge University Press.

Vander Wal, T. (2007) Folksonomy Coinage and Definition, available: http://vanderwal.net/folksonomy.html [accessed 12 November 2013].

Volk, M. (2002) 'Using the web as corpus for linguistic research', *Tähendusepüüdja. Catcher of the Meaning. A Festschrift for Professor Haldur Õim*, Publications of the Department of General Linguistics 3, University of Tartu, available: www.halskov.net/files/Volk_Web_as_Corpus.pdf [accessed 26 November 2013].

Zappavigna, M. (2012) *Discourse of Twitter and Social Media: How we use Language to Create Affiliation on the Web*, London: Continuum.