7 Corpora as a source of data

While the face-to-face data collection techniques we've introduced in the previous two chapters are excellent ways to observe language, there are some cases in which they won't be sufficient. Imagine, for instance, that you have developed an interest in a variable that occurs fairly infrequently in speech, such as the alternation between *would have* and *woulda*. After an hour-long sociolinguistic interview, you may only have heard the target construction a handful of times. Or, imagine you wish to track the real-time development of the verb of quotation *be like*, or explore men's and women's language in the Early Modern English period. Without access to historical language data, you won't get very far. This chapter introduces you to data sources known as corpora that you can use to address these and other sorts of research topics.

What is a corpus?

Briefly, a corpus (plural: *corpora*, with stress on the first syllable) is a body of language that someone has collected for a particular purpose. A corpus is designed to be representative of the type of language it has been drawn from, despite constituting only a tiny slice of the whole. Corpora are also typically large: some well-known corpora number in the hundreds of millions of words. This means that corpora cannot usually be analysed by hand, but there are a range of computational tools available to allow you to work with them, which we'll discuss later on.

If you can think of a scenario in which language is used, there's probably a corpus to represent it. Corpora can be based on spoken language or written language (and some corpora provide some of each). Where spoken language is concerned, some corpora comprise only transcripts, while others contain both transcripts and audio, a goldmine for researchers interested in phonetic and phonological variation. Corpora of written language can contain newspaper and academic writings, correspondence, literature and online communications. Video corpora of signed languages are available too, as are many other non-English

corpora, including multilingual corpora containing parallel translations of the same text. Even YouTube has been used as a corpus (e.g. Wrobel 2012, and the study of Victoria Beckham's language in Chapter 14).

Corpora can be used to investigate variation in register, both in speech and writing, as you can find corpora that represent formal language (like MiCASE, the Michigan Corpus of Academic Spoken English; or the academic writing subset of COCA, the Corpus of Contemporary American English) and corpora that represent informal language (like the Buckeye Speech Corpus, a collection of 40 sociolinguistic interviews; or the Enron Email Dataset, a corpus of public-domain emails sent between employees of the now-defunct American energy company Enron). Cheng (2012: 33–34) provides a list of major English-language corpora; you can find a few easy-to-use corpora listed below, and there are plenty of further lists online.

GET UP AND RUNNING WITH SOME ONLINE CORPORA

You can get started with corpora by exploring the following (search for them online):

Corpora of written language:

- The Corpus of Contemporary American English (COCA)
- The Corpus of Global Web-based English (GloWbE)
- The Corpus of Historical American English (COHA)
- Google Books Ngram Viewer

Corpora of spoken language:

- The Switchboard and Fisher corpora via LDC Online (NB: full access requires a subscription through your university)
- The Michigan Corpus of Academic Spoken English (MiCASE)

Why would I want to use a corpus?

By using a pre-existing corpus, it could be argued that you're surrendering control of the nature of your data. After all, you don't get any say in the topics that speakers discuss, or the demographics of your speaker sample, in the way that you do when you design an interview-based study. For many studies, though, the benefits of working with corpora can outweigh these drawbacks.

One such benefit is size: the Corpus of Contemporary American English contains 450 million words, and the Linguistic Data Consortium's Fisher corpus

consists of 2,700 hours of transcribed audio. This can allow you to study lexical or syntactic variables that occur only infrequently in speech. For instance, Rickford et al. (1995) carried out a study of variation in *as far as* constructions, having noticed that people can alternate between, say, 'as far as the government is concerned', 'as far as the government goes' and 'as far as the government'. The authors and their colleagues jotted down each *as far as* construction they noticed in everyday use, and managed to come up with 650 examples – but this required the efforts of several researchers over the course of eight years. By contrast, a collection of spoken and written corpora gave them 500 tokens. If time is of the essence, corpora are your best bet with an infrequent variable.

Working with a corpus can also allow you to generalise over individuals to uncover broader trends about language. Data from a single speaker can tell us a lot about that one individual, but she may not be representative of the larger population of women, teenagers or Philadelphians. The more data we collect, though, the more certain we can be that the patterns we're observing are not a fluke. Working with corpus data has allowed researchers to determine that female users of Twitter are more likely to use emoticons than male users (Bamman, Eisenstein and Schnoebelen 2014), and to track changes in the Philadelphia vowel system over 110 years of apparent time (Labov, Rosenfelder and Fruehwald 2013). A quick check in a diachronic corpus can allow you to confirm whether a case of variation in the present day represents a change in progress. For instance, Laurel noticed people alternating between constructions like 'out the window' and 'out of the window'. Checking the ratio of these phrases in the Google Books Ngram Viewer corpus of digitised books (Michel et al. 2011) resulted in the graph in Figure 7.1, and a study of language change was born.

How do I use a corpus?

You'll first want to figure out what sort of corpus is appropriate to what you want to study. Are you working with a phonological variable, for which having audio is going to be essential? Or a syntactic variable, for which you're fine working just with written data? (If the latter, are you sure you can get away with not analysing sound? What if intonation might be relevant to the variation?) Also consider whether your corpus provides information on factors that might be conditioning the variation. If you're working with a corpus of child language data, for example, information on parents' level of education or occupational status could be useful.

The size of most corpora means that listening or reading through them in real time is an unrealistic means of gathering data. Instead, linguists use computational tools to help them get what they want. Depending on the nature of your linguistic variable and the corpus you select, collecting your data may be as easy as typing out some search queries, or it may require you to learn some basic

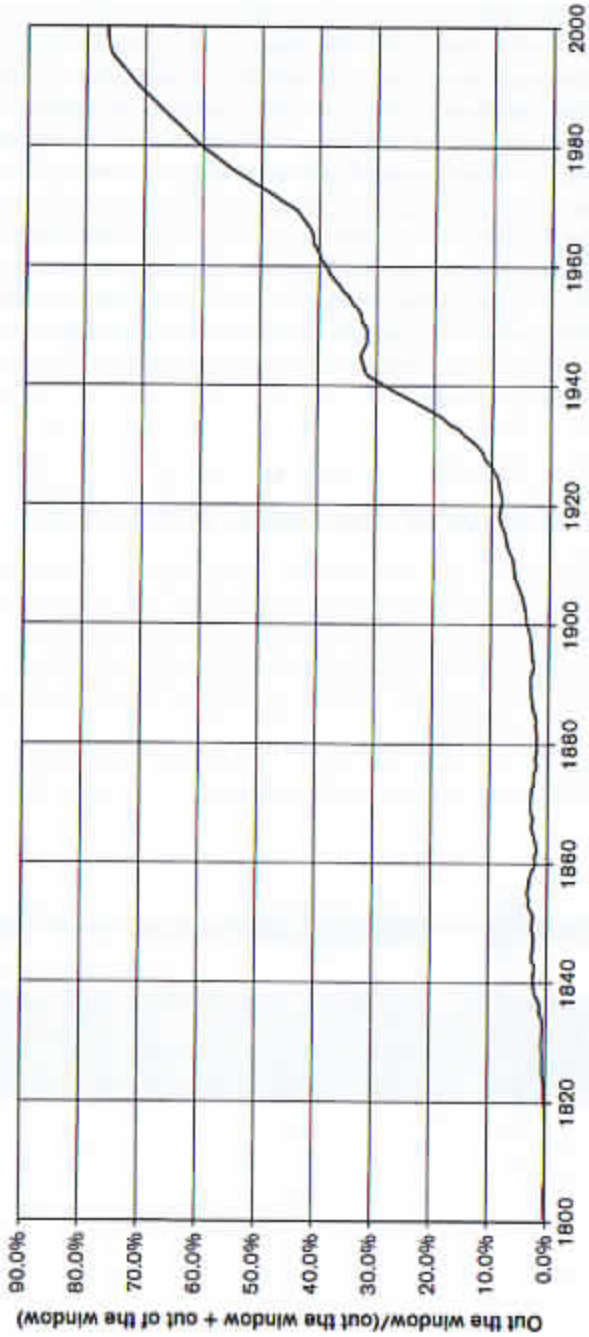

Figure 7.1 The rise of *out the window* in Google Books Ngram Viewer

programming. One of the most common applications of corpora is to generate a concordance – a compilation of all the instances of a given word or phrase in order to examine, say, how it is used, or whether it is used more or differently in one text or register than another – and there is plenty of software available that can do this for you (consult the references in the Further Reading list at the end of this chapter). If your research requires listening to words in context, you will benefit from the existence of web interfaces for audio corpora like BNCWeb, which allows you to search for and hear words and phrases in the British National Corpus as easily as using the Find command in a text document. Learning to write computer code will give you the freedom to do the same even in corpora that don't have existing web interfaces, and has the added bonus of 'teaching you to fish' – that is, of giving you the ability to branch out beyond any limitations you encounter in existing software.

FORCED ALIGNMENT AND VOWEL EXTRACTION

A recent advance in corpus research is the development of software for the automatic time-alignment of speech. Starting with an orthographic transcription, these tools create a phonetic transcription, and then automatically match every word and phoneme in these transcriptions to their precise point of occurrence in the audio. Figure 7.2 demonstrates the output of forced alignment.

Not only can this allow you to jump quickly and easily through your sound files to particular words and phonemes, but it can allow you to use

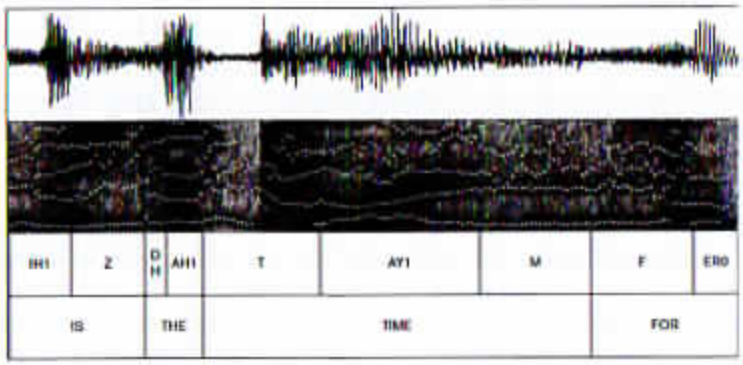

Figure 7.2 A segment of speech with accompanying forced alignment of phonemes and words

tools for vowel extraction: automatic measuring of the formant values of every vowel in a recording. Measuring vowels by hand is arduous, but, as long as you have a transcribed recording, these new techniques can provide thousands of vowel measurements in minutes. Search the web for *forced alignment* to see the kinds of software available.

Corpora differ in how much linguistic annotation they provide, meaning that your task of coding a variable may be easier with some corpora than others. The Buckeye Speech Corpus is unusual among audio corpora in having alongside its audio files not only orthographic transcriptions, but also narrow phonetic transcriptions. This takes all the work out of coding a phonological variable like *t/d*-deletion. If you wanted to study *t/d*-deletion in the Switchboard corpus, by contrast, you'd be signing yourself up for the somewhat tedious task of listening to hundreds of *t/d*-final words to collect your data. On the other hand, the transcriptions in Switchboard have been annotated for part of speech, and some of them have even been syntactically parsed. This comes in handy if you need to differentiate in your study between, say, *ring* as a noun versus *ring* as a verb, or *have* as an auxiliary verb versus *have* as a main verb.

Like any other source of data, corpora have advantages and disadvantages. Though we hope the benefits of having a large body of language at your fingertips are apparent, remember that the findings of a corpus-based study will only be as representative as the corpus they come from. As always, you will help yourself out if you clearly articulate the specific research question you are asking before you choose your source of data.

EXERCISES

Exercise 1

Consider the linguistic variables below. What sort of corpus would you need in order to study them? What sort of annotation would be useful?

1 The change in English third singular present marking from *-th* to *-s*
2 The genitive alternation (e.g. *the tree's roots* ~ *the roots of the tree*)
3 The fronting of the GOOSE vowel in American or British English
4 The alternation between *would have*, *would've* and *woulda*

Exercise 2

Contextual style is well known to affect sociolinguistic variation. How could we study contextual style in present-day corpora? How about historical corpora?

References

Bamman, David, Jacob Eisenstein and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18: 135–160.

Cheng, Winnie. 2012. *Exploring Corpus Linguistics: Language in Action*. London: Routledge.

Google Books Ngram Viewer. http://ngrams.googlelabs.com (last accessed 7 November 2014).

Labov, William, Ingrid Rosenfelder and Josef Fruehwald. 2013. One hundred years of sound change in Philadelphia: linear incrementation, reversal, and reanalysis. *Language* 89: 30–65.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331: 176–182.

Rickford, John R., Thomas A. Wasow, Norma Mendoza-Denton and Juli Espinoza. 1995. Syntactic variation and change in progress: loss of the verbal coda in topic-restricting *as far as* constructions. *Language* 71: 102–131.

Wrobel, Emilia. 2012. What can you find on YouTube, that's sociolinguistically interesting? *Journal of Pidgin and Creole Languages* 27: 343–350.

Further reading

Baker, Paul. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Hoffmann, Sebastian, Stefan Evert, Nicholas Smith, David Lee and Ylva Berglund Prytz. 2008. *Corpus Linguistics with BNCweb–a Practical Guide*. Oxford: Peter Lang.

McEnery, Tony, Richard Xiao and Yukio Tono. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.

8 Written surveys and questionnaires as a source of data

Sometimes it's just not possible to access a corpus or collect new, written or spoken language data. Surveys or questionnaires can be a big help here. They produce highly reliable data if used appropriately, and they may be especially good for accessing information on linguistic behaviour that is difficult to observe or record. Writing a good questionnaire is not easy. You have to put a lot of thought into its construction before you can use it for data collection, but one of the pay-offs is that the results can be quick to analyse. In this chapter, we outline (a) the different uses that questionnaires can be put to, (b) their limitations and advantages and (c) how to write good questions and develop, structure, test and administer a questionnaire.

Questionnaires in sociolinguistics

Written surveys and questionnaires allow you to explore how people behave in certain situations, but you can also find out a lot about their beliefs, knowledge, attitudes and their social characteristics. Questionnaires have been used to:

- find out who uses what language, who they use it with and when (e.g. Choi 2005).
- investigate the sociolinguistic profile of ethnic minorities in an urban area (e.g. Extra and Yagmur 2004).
- explore who uses different words and phrases in different contexts, e.g. swear words, loanwords or sexist vocabulary (e.g. Fuller 2005).
- explore attitudes towards dialects, accents, certain lexical items or pronunciations (see Chapter 9). Attitudes are also an important component of ethnolinguistic vitality (Bourhis, Giles and Rosenthal 1981).
- find out about discourse pragmatics in one or more cultures using discourse completion tests (see textbox on DCTs).

DISCOURSE COMPLETION TESTS (DCTs)

In DCTs, you give participants scenarios, which include the setting, the status of participants and the social distance between them, as well as incomplete scripted dialogue in a variety of different social situations. Then you ask respondents what they would say there. Here's an example from Blum-Kulka, House and Kasper (1989: 14), constructed to elicit a request (we've removed the questionnaire instructions here):

> Ann missed a lecture yesterday and would like to borrow Judith's notes.
>
> *Ann*: _____
>
> *Judith*: Sure, but let me have them back before the lecture next week.

Sometimes, multiple-choice options are given for the missing piece of dialogue, and sometimes respondents are only given a situation and some space for their response. DCTs are a frequently used method in cross-cultural pragmatics, and some interesting findings have been made. For example, how would you ask your roommate to clean up the kitchen, if they had left it in a mess the night before? Turns out that 74 per cent of Argentineans prefer direct bald-on strategies, as in "Hey, Fernanda, you have to clean the kitchen before you leave", in contrast to only 12 per cent of Australians (Blum-Kulka 1997: 56). Who knew?! But remember that what people tell you they do, may tell you what they would like to *think* they do, rather than what they actually *do* do (see Beebe and Clark Cummings 1996).

Questionnaires can also find out whether particular constructions are considered grammatical or not. Traditionally, linguists did this by asking people to report judgements in terms of categories such as *acceptable*, *marginally acceptable*, *unacceptable*, *good*, *terrible*, etc. This isn't ideal. Bard, Robertson and Sorace (1996) outline advantages and shortcomings of the traditional format, and they propose the method of magnitude estimation as a way to overcome these issues. Magnitude estimation allows graded acceptability to be measured, e.g. people can tell you that they think a sentence is four times more acceptable than another.

Written surveys are frequently used to investigate regional and social variation. Many traditional dialect surveys are based on a questionnaire, e.g. the *Linguistic Atlas of England* (Orton, Sanderson and Widdowson 1978). Search for Laurel MacKenzie's *Dialect Variation Maps* online for a modern example. You can also combine questionnaires with other data elicitation methods. Llamas (2007) developed a set of questionnaires, which she used in combination with an interview, to collect data on Teeside English.

There are some clever question formats that you can use when you investigate regional and social variation with a written survey.

1. You can elicit forms and features. You can find out whether someone is familiar with a particular form by asking directly, or indirectly, by prompting them to use a particular item. You could ask, "Are you familiar with the word *bairn*? If so, how often would you use it?" (the direct approach). Or you might ask, "What do they call a person who's not an adult around here?" (more indirect). Or you might give respondents a fill-in-the-blanks task.
2. You can collect judgements about structural constraints on a feature. This can be helpful if you want to find out if a particular variant can or cannot be used in certain tenses or with certain subjects, etc. Judgement or permutation tasks are outlined in Schilling (2013: 71–75).
3. And you may also want to find out whether a feature is the same or different from another one (see box below on phonological variables).

Some variables are more suited to use in written surveys than others. Many morphological, syntactic, lexical and semantic variables are easy to represent in writing and quite suited to use in a questionnaire if they are not affected by strong, overt social evaluation. Variable forms are often given in a list of options, asking participants to select a form that they would use in daily conversations, possibly with some space for alternatives not listed.

PHONOLOGICAL VARIABLES IN WRITTEN SURVEYS

It's difficult to design a written survey on phonetic variation, e.g. vowel shifts, as phonetic details cannot easily be transmitted to lay-people in writing. Variation at the phonemic level, on the other hand, can often be represented in conventional orthography, so mergers of phonemes can be studied relatively well with a questionnaire. You can give participants in your study minimal pairs and ask them whether the words sound the same or different or whether they rhyme. For example, Dollinger (2012: 93) asked respondents questions such as:

> Do the words ***cot*** and ***caught*** sound the same to you? Yes No
>
> Does the ending of AVENUE sound like ***you*** or ***oo***?

Alternatively, give respondents a pre-recorded list of word pairs and ask whether each of the pairs is the same or different words. A more indirect version of this task is the "communication test" (Labov 1994: 356ff). Whether or not someone can distinguish between members of minimal

pairs is an indication of whether they merge the sounds under investigation or not. This method can reveal *near mergers*: sometimes people claim not to hear a difference but spectrographic analysis of their speech reveals that they produce the sounds differently.

But you may not be interested in mergers at all. Similar methods can be used if you just want to find out which phonemes occur in particular words. Boberg (2010: 140) asked respondents to classify words such as *soprano*, *Iraq*, *pasta* and *plaza* into one of two categories: whether their vowel is pronounced like the <a> sound as in *cat* or the <ah> sound as in *father*.

Limitations and opportunities

Questionnaires are a great resource; however, we must be aware of their limitations. They are not particularly suited to delve deeply into an issue, and what respondents claim they do may not match up with what they actually do, partly because they may find it difficult to assess their language behaviour. In addition, they may understand and talk about non-standard features in a way that limits what we can find out from them. People tend to equate grammaticality with standard language and although we can frame our questions avoiding terms such as *correct* and *grammatical* by asking "How would you say . . .?", "Can you say . . .?", "Which is more natural to you . . .?", standard forms often continue to influence assessment.

If our question is very direct, we effectively ask respondents to pay attention to how they speak, which we don't really want because it often results in irregular linguistic patterns and hypercorrection (Labov 1972: 134; see discussion of the observer's paradox in Chapter 5). Though direct questions often tell us more about people's opinions than how they use language, this is much truer for highly stereotyped, non-standard grammatical forms and obsolescent forms than less socially sensitive variables (Boberg 2013).

A further issue with questionnaires concerns ordering effects. This refers to the way a response may be influenced by previous ones. People can also misread questions or interpret questions differently to what you'd intended, or they may not be able to read and write at all. Dörnyei (2003) notes that people tend to agree rather than disagree, and they show a tendency to overgeneralise. So, when they like one aspect of a person or event, they overestimate *all* characteristics associated with that person or event. Long questionnaires may simply tire people out so they give incorrect answers or leave questions blank.

Many of these limitations can be minimised: you can randomise items, use indirect elicitations if possible, avoid the term *grammatical* and design the survey to be more like a game than a test. Most importantly, you can keep its length

manageable: four pages, 30 to 50 items, or 30 minutes of completion time are the maximum.

Having said all that, written questionnaires also have many advantages; they are cost-effective and collect a lot of data quickly. That's good news because confidence in the results of a study improves as the number of respondents increases. If the questionnaire was designed well, preparing the data for quantitative analysis will also be efficient.

Developing questionnaire items

Thorough preparation will make your questionnaire infinitely more useable. It should be as clear as possible and one way to achieve this is to develop a precise research question. You can then scaffold your knowledge or theory on this to build survey questions. Focus groups, interviews, a short open-ended pilot survey and a trawl of the existing literature may help you decide what further questions to include.

Questions in a questionnaire are sometimes called *items*. That's because they often don't look like a canonical question. There are two main types of questions/items: closed questions and open questions. Closed questions provide a closed set of possible answers, and because the answers are pre-defined, this type of question can be analysed quickly. A closed question typically has three clearly marked parts: (a) instructions, (b) a question or statement and (c) possible answers, as in the following example:

> Below are listed five questions about swearing. We would like to ask you to answer each one by circling the number that most closely matches your opinion.
>
> 1 How acceptable do you think swearing is?
>
> Not at all ... Very
>
> 1 2 3 4 5 6

Be mindful of the hazards associated with biasing question options. If there is a neutral position, there should be the same number of response options on either side. There is no consensus as to whether the points on scales should be an even or an odd number. Sometimes researchers worry that people may drift to the mid-point in a series to weasel out of committing firmly to something, but sometimes the middle is precisely what respondents feel if they are undecided.

There are many different closed question formats. Sociolinguists most frequently use checklists, rankings, rating scales (as in the example above), true–false questions, multiple-choice questions and semantic differential questions (where

people get polar adjectives and rank a speaker along a scale: e.g. rich←→poor). More detailed examples are in De Vaus (2005) and Schleef (2013).

Open-ended questions do not provide pre-formulated answers. You can ask specific open-ended questions, questions of clarification, sentence completion items and short-answer questions, which ask for one concept or one idea only. Open-ended questions give respondents some space to provide an answer themselves, and, thus, assume they can express themselves in writing (but remember if they are completing the questionnaire by hand, you may have to decipher what they express!). The idea here is to get more personal responses, but the downside is that you often get lots of blank spaces and a large number of individual answers that may be difficult to categorise and quantify (assuming that's the goal). Answering open-ended questions is also quite time-consuming, so they should be used sparingly and towards the end of the questionnaire.

SOME TIPS FOR WRITING GOOD ITEMS (BASED ON DÖRNYEI 2003)

- If you ask questions about behaviour that respondents may disapprove of, mitigate questions, e.g. by assuming it occurs and asking about detail, or by casualising it.
- Write short, simple and natural-sounding items.
- Do not put all negative terms on one side of the questionnaire and all positive ones on the other. Alternate them.
- *Don't know* or *Other* options may be appropriate when none of your options may apply, otherwise you are better not to include these.
- A well-tested method that reduces idiosyncratic interpretation of some questions is multi-item scaling: you use differently worded items that all focus on the same target, for example the perceived acceptability of swearing, and then you average out responses.

Avoid

- acronyms, abbreviations, technical terms, colloquialisms;
- questions that ask about two different things at once, while expecting only one answer;
- items containing negatives;
- unclear, unspecific terms, such as *frequently, sometimes, good*;
- potentially loaded or leading questions;
- sensitive questions – if that's not possible, renew the promise of confidentiality when you do;
- words that do not allow exceptions, such as all-inclusive or all-exclusive words (e.g. *all the time, nobody*), as they may result in a lack of variability in answers. They are fine as options on a continuum.

Questionnaire structure

Your questionnaire should consist of more than just a list of questions. A good questionnaire starts with an informative, reassuring and polite introduction, and it ends with a conclusion. The introduction should include: (1) the questionnaire title; (2) a very brief outline of what the research is about and who is responsible for conducting the study; (3) a friendly request to fill in the questionnaire fully and honestly; (4) a brief overview of what will be asked in the questionnaire and, roughly, how long its completion will take; (5) a clear statement saying that responses will be treated with absolute confidentiality and that respondents will remain anonymous; (6) the researcher's name, institution and (7) a thank you for participating in the study.

Next follows the main text, which should be logically ordered. Make the first questions ones that keep the reader interested and involved and vary the question formats. You don't want to bore, or put off people at the start. This is also why you might want to put open-ended, personal or more demanding items towards the end of the questionnaire. If your questionnaire consists of different subsections, you can use headings to help people through. It also helps readers when you emphasise instructions clearly. A good questionnaire is attractive, looks professional and is error-free.

You end the questionnaire by thanking respondents once again and leaving them your contact details. Additionally, you could renew the promise of anonymity and provide further information you consider vital for respondents or your study, e.g. ask respondents to make sure no questions are blank, tell them how questionnaires can be returned and where they can view the survey results.

Testing, administering and processing questionnaires

Before you distribute your questionnaire, make sure to test it. One strategy is to ask a few friends to read it (without completing it) and tell you what they think as they read it. You might also want to ask people to answer the survey questions and think aloud, explaining their reasoning as they go (Willis 2005). This allows you to check whether questions are worded clearly. You may also want to consider piloting your questionnaire in a small group of people and then have a good look at the answers. Are some questions left blank? Are some questions open to misunderstanding? Is the questionnaire too long? Will you be able to process the responses?

Once all this is done, you can finally distribute the questionnaire! For your sample to be roughly representative, you want to collect data from more than 30 people per subgroup (i.e. per category investigated, for example at least 30 per age group, per speaker sex, etc.). That's because your sample will likely have a normal distribution with more than 30 people (Hatch and Lazaraton 1991). This is different from the minimum of five or six speakers per cell we recommend in Chapter 2. It makes sense, though. Think about it. You will get only one questionnaire per person. However, you normally get many, many tokens of a single variable per speaker in a study that's based on speech recordings. Thirty may be more than you can handle in a small student project, so if that's not feasible, try to get as many responses per subgroup as you can – definitely more than ten per group – but keep in mind that results may change radically with more data.

You can administer written surveys and questionnaires face-to-face (with the researcher noting down answers) or long-distance (with the respondent writing all answers without the researcher being present). Both have advantages and disadvantages. Have a look at De Vaus (2005: 126ff) for a comparison of distribution methods, including expected response rates. Dörnyei (2003: 83ff) outlines some tips to ensure many respondents spend sufficient time and effort completing the questionnaire.

LONG-DISTANCE DISTRIBUTION AND CROWD-SOURCING

Long-distance surveys can be sent by e-mail, post, or they can be distributed by hand or on the internet. Telephone, computer or internet surveys make it possible to play recordings even to participants who are miles away. Online surveys also allow you to make use of crowd-sourcing tools such as Amazon Mechanical Turk. At the time of writing, its main limitation is that you need to have an American address to use it. You could also advertise your survey on various social media sites.

Once all questionnaires are collected, you process the data. Assign each questionnaire a number and enter the data into a spreadsheet, using codes that you've made clear notes on for future reference (your coding guide). Now check for potential inaccuracies and mistakes in the database and, if appropriate, reverse the scores of negatively worded items. For example, you may want to make sure all positive items have a high or positive number code and negative items have a low one. If data is missing, you'll have to consider removing participants or certain questions from the survey. Dörnyei (2003) talks about ways to evaluate questionnaires for validity and reliability.

If space allows, all decisions about questionnaire structure, items and data exclusion should briefly be mentioned in the methods section of your report. When presenting your results, consider who may have filled in your questionnaire. Was it based on self-selection? A particular medium? This may bias your findings.

Now you can finally analyse the data. You can jump to Chapters 12 to 14 for help on how to do this or to Chapter 15 for some thoughts about complementing questionnaire data with qualitative analysis. This can be a valuable reminder about the people behind your numbers.

EXERCISES

Exercise 1

What are the advantages and disadvantages of administering questionnaires face-to-face and by distance, especially online?

Exercise 2

Consider Figure 8.1 from Campbell-Kibler (2011: 441). This is *part* of a written survey which aims to find out what social attributes speakers associate with the variants of (ing). Respondents listened to audio recordings of eight speakers. For every one of these speakers, several naturally produced sentences had been digitally altered to create triplets: one had an [ɪŋ], one an [ɪn] and one was a neutral guise with no audible (ing) token. A survey page for one speaker is shown over the page.

As mentioned above, questionnaire length must be limited. Campbell-Kibler has limited the number of recordings each participant would hear to eight and she managed to keep the survey page for each of the eight voices quite short and to the point.

What decisions did she have to make to do this? What kind of question types does she use and which ones could she have used – with and without increasing questionnaire length? How does question type influence the kind of analysis you can conduct? Think about the type of answer you get from different questions and consult Chapter 12, if you need help.

Exercise 3

Dollinger (2012: 95ff) compares data collected using written questionnaires with interview data. Consider the results for the three examples of the low-back vowel merger in Canadian English (Table 8.1). Dollinger analysed interview data to find out whether ten people have the merger (or not) and then compared these results with what they claimed to do in a written questionnaire. The 'Merge/Nonmerge'

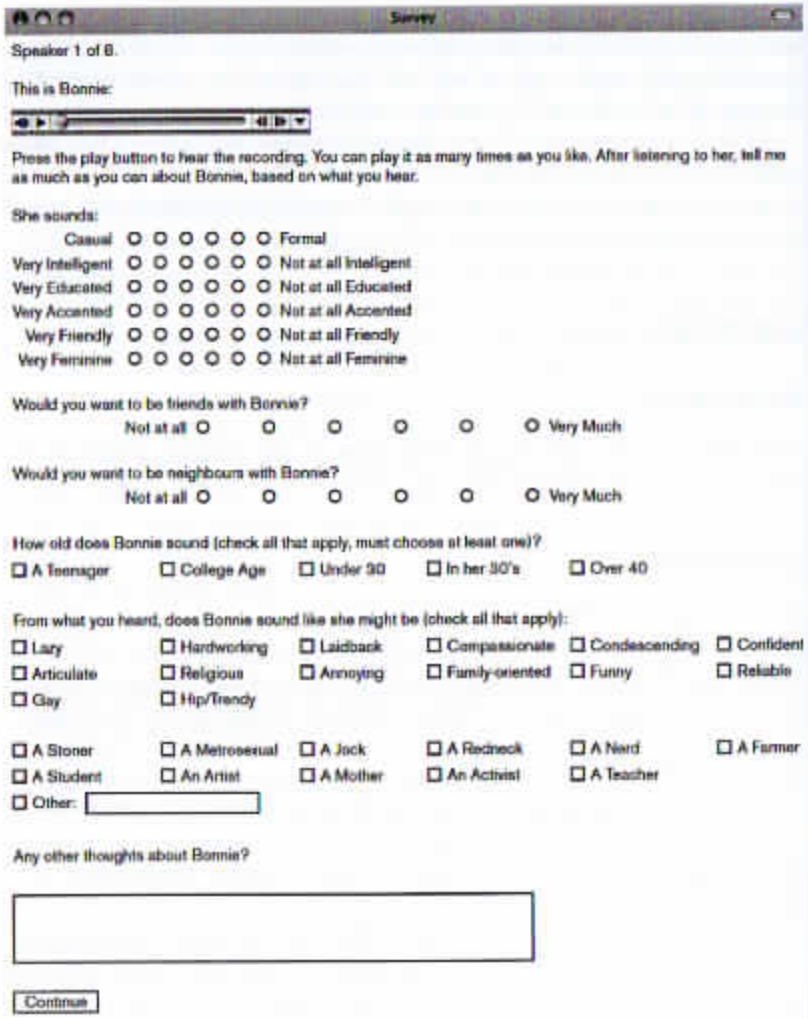

Figure 8.1 Sample from an online survey

Source: Campbell-Kibler (2011: 441)

columns show what the participants claim they do. If the questionnaire data matched the data in the interview there is a 'Yes' in the column 'Match'.

How reliable is the written questionnaire? What factors seem to influence reliability?

Table 8.1 Match between self-reporting and acoustic analysis

	cot/caught			*Don/Dawn*			*sorry/sari*	
	Merge	*Nonmerge*	*Match*	*Merge*	*Nonmerge*	*Match*	*Merge*	*Nonmerge*	*Match*
Anton	X		No	X		Yes		X	No
Mario	X		Yes	X		Yes		X	Yes
Gustave	X		Yes		X	No		X	Yes
Lola	X		Yes	X		Yes		X	Yes
Kelsey	X		Yes	X		Yes		X	Yes
Ella		X	No		X	Yes		X	Yes
Chad	X		Yes	X		Yes		X	Yes
Nancy	X		No?	X		Yes		X	Yes
Carl	X		Yes	X		Yes		X	Yes
Carla	X		Yes	X		Yes		X	Yes
			7/10			9/10			9/10

Source: Dollinger (2012: 95)

References

Bard, Ellen Gurman, Dan Robertson and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72: 32–68.

Beebe, Leslie M. and Martha Clark Cummings. 1996. Natural speech act data versus written questionnaire data: How data collection method affects speech act performance. In Susan Gass and Joyce Neu (eds) *Speech Acts across Cultures: Challenges to Communication in a Second Language*. Berlin and New York: Mouton de Gruyter, 65–86.

Blum-Kulka, Shoshana. 1997. Discourse pragmatics. In Teun A. van Dijk (ed.) *Discourse as Social Interaction*. London: Sage, 38–63.

Blum-Kulka, Shoshana, Juliane House and Gabriele Kasper. 1989. *Cross-Cultural Pragmatics: Requests and Apologies*. Norwood, NJ: Ablex.

Boberg, Charles. 2010. *The English Language in Canada: Status, History and Comparative Analysis*. Cambridge: Cambridge University Press.

Boberg, Charles. 2013. The use of written questionnaires in sociolinguistics. In Christine Mallinson, Becky Childs and Gerard Van Herk (eds) *Data Collection in Sociolinguistics: Methods and Applications*. New York and London: Routledge, 131–141.

Bourhis, Richard Y., Howard Giles and Doreen Rosenthal. 1981. Notes on the construction of a 'subjective vitality questionnaire' for ethnolinguistic groups. *Journal of Multilingual and Multicultural Development* 2: 145–155.

Campbell-Kibler, Kathryn. 2011. The sociolinguistic variant as a carrier of social meaning. *Language Variation and Change* 22: 423–441.

Choi, Jinny K. 2005. Bilingualism in Paraguay: Forty years after Rubin's study. *Journal of Multilingual and Multicultural Development* 26: 233–248.

De Vaus, David. 2005. *Surveys in Social Research*. 5th edition. London and New York: Routledge.
Dollinger, Stefan. 2012. The written questionnaire as a sociolinguistic data gathering tool: Testing its validity. *Journal of English Linguistics* 40: 74–110.
Dörnyei, Zoltán. 2003. *Questionnaires in Second Language Research: Construction, Administration and Processing*. Mahwah, NJ: Lawrence Erlbaum.
Extra, Guus and Kutlay Yagmur. 2004. *Urban Multilingualism in Europe*. Clevedon: Multilingual Matters.
Fuller, Janet M. 2005. The uses and meanings of the female title Ms. *American Speech* 80: 180–206.
Hatch, Evelyn and Anne Lazaraton. 1991. *The Research Manual*. New York: Newbury House.
Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia, PA: University of Pennsylvania Press.
Labov, William. 1994. *Principles of Linguistic Change Volume I: Internal Factors*. Malden and Oxford: Wiley-Blackwell.
Llamas, Carmen. 2007. A new methodology: Data elicitation for regional and social language variation studies. *York Papers in Linguistics* 8: 138–163.
Schilling, Natalie. 2013. *Sociolinguistic Fieldwork*. Cambridge: Cambridge University Press.
Schleef, Erik. 2013. Written surveys and questionnaires. In Janet Holmes and Kirk Hazen (eds) *Research Methods in Sociolinguistics*. Oxford: Wiley-Blackwell, 42–57.
Orton, Harold, Stewart Sanderson and John Widdowson. 1978. *The Linguistic Atlas of England*. London: Croom Helm.
Willis, Gordon B. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage.

Further reading

Boberg, Charles. 2013. The use of written questionnaires in sociolinguistics. In Christine Mallinson, Becky Childs and Gerard Van Herk (eds) *Data Collection in Sociolinguistics: Methods and Applications*. New York and London: Routledge, 131–141.
Brown, James Dean. 2001. *Using Surveys in Language Programs*. Cambridge: Cambridge University Press.
De Vaus, David. 2005. *Surveys in Social Research*. 5th edition. London and New York: Routledge.
Dollinger, Stefan. 2015. *The Written Questionnaire in Social Dialectology: History, Theory, Practice*. Amsterdam: John Benjamins.
Dörnyei, Zoltán. 2003. *Questionnaires in Second Language Research: Construction, Administration and Processing*. Mahwah, NJ: Lawrence Erlbaum.
Gillham, Bill. 2007. *Developing a Questionnaire*. 2nd edition. London and New York: Continuum.
Schleef, Erik. 2013. Written surveys and questionnaires. In Janet Holmes and Kirk Hazen (eds) *Research Methods in Sociolinguistics*. Oxford: Wiley-Blackwell, 42–57.

9 Studying perceptions and attitudes

Finding out what people think or believe about language is a very exciting area of research, and it is crucial for sociolinguistics because language attitudes influence language use. We'll tell you about three methods of data collection in this chapter: direct methods, indirect methods and the collection of pre-existing speech or text for further analysis of the social meaning of language. In keeping with the spirit of this book, we'll point you to some key readings that will help kick-start your own investigation. Our examples will mostly be about the social evaluation of speech, and Garrett (2010) is a particularly useful resource in this area, but we'll also consider social identification, speech perception and perceptual dialectology.

Direct methods

You might think the easiest way to find out about attitudes is to ask people. But for various reasons, people aren't always good sources of direct information on their attitudes. They may fudge (or even lie) if they think their answer might be interpreted negatively, or they may lack the necessary degree of introspection to answer your questions well. So be careful when analysing data based on this approach.

You can use a variety of different methods to find out about attitudes directly: interviews, rapid and anonymous surveys, questionnaires, etc. In the last few chapters, we offered advice on these methods and you could also look at a few studies that use direct methods: Hickey (2009) who explored language attitudes towards Irish and English in Ireland, or Sharp et al. (1973) who investigated attitudes to Welsh. Sharp et al. studied 12,000 children and, as you can see right at the start of their survey, they made no attempt to hide their interest in language: "The following exercise is designed to find out what kind of idea you have of the Welsh language" (Sharp et al. 1973: 167). Once the data is collected, it is often analysed by identifying the main emerging themes.