5 Population samples

Isabelle Buchstaller and Ghada Khattab

Data without generalisation is just gossip.

1 Introduction

So you want to investigate the language used by a group of people. One of the first questions you might ask yourself is: Who do I collect these data from? A crucial element of empirical linguistic work is to choose not only what type of data to collect (e.g., naturally occurring data, interview data, questionnaire data, experimental data; see Part I of this volume), but also which people to target for data collection. The most reliable method for finding out about the language use of a particular group of people would be to collect linguistic information from every single person in the population, which in the social sciences refers to all members of the community. Obviously, except for very small populations, this method is rather impractical, expensive, and time-consuming. Hence, most researchers only target “some people in the group in such a way that their responses and characteristics reflect those of the group from which they are drawn . . . This is the principle of sampling” (De Vaus 2001: 60). The subgroup of people that reflects the population as a whole (in terms of their social and linguistic characteristics), and therefore lends itself to generalizations above and beyond the scope of the study, is called a representative sample. The question we need to ask as linguists is: To what extent are the findings reported on the basis of a subsample representative of the linguistic habits of a certain population or group?

Many social scientists would argue that representativeness can only be assumed if the characteristics of the sampled group match those of the population at large. This effectively means that our sample must not favor some sectors of the population over others (so that no sectors of the population are excluded or under- or over-represented). For example, in the past 50 years (except for 1981) the census of the United Kingdom was conducted at the end of April, which for many English universities fell during the Easter break. This meant that a large section of the student population were not at their regular place of study, but rather visiting their family back home (sometimes abroad) or on holiday. As a result, the
census did not accurately reflect the populations that live in these areas; the data were biased toward non-student, non-foreign populations. This sampling bias was avoided in 2011, when the census was conducted during term time. Representativeness thus implies the avoidance of biases in the data that would make generalization impossible.¹

Note, however, that in the field of linguistics, social representativeness is not easily achieved, since language varies across a wide range of social dimensions within a population, such as speakers’ age, gender, sexuality, ethnic identity, regional background, educational level, and many others. Also to be taken into account are situational and conversational factors, such as the level of formality, the speaking style, accommodation to interlocutors, conversational topic, and ideological factors, among many others, each of which potentially introduces a bias into the sample. All of this leads to a considerable challenge for linguists: how do we sample in order to avoid biases in our data?

Sankoff (1974: 21) points out that every researcher must make a decision about their “sampling universe,” namely, the groups or communities they want to investigate (e.g., the residents of a particular city or neighborhood, the members of a reading group or a garage band). Having made a decision about whom to investigate, “good data is [then] defined as language materials of sufficient type and quantity, as well as materials which take into account the social context in which the language data is gathered” (Sankoff 1974: 21–2; see also Milroy 1987: 18). As our initial sampling choices establish what type of population our data are representative of and about whom we can make generalizations, those sampling decisions fundamentally constrain the types of questions we can answer. For example, data from teenagers in an affluent suburb in the San Francisco Bay area (Buchstaller et al. 2010) cannot make any generalizations about people or of the state of California as a whole. If the study had intended to make more general claims it would have had to sample across a broader range of social groups (e.g., younger and older speakers, or a wide variety of social backgrounds across the whole state). Tagliamonte (2006) emphasizes the intimate connection between sampling method and research question: “At the outset, a (socio)linguistic project must have (at least) two parts 1) a (socio)linguistic problem and 2) appropriate data to address it.”

Linguistic researchers have been using a wealth of different types of sampling methods — mostly adapted from sociology, developmental psychology, anthropology, or (economic) geography — which vary vastly with time and linguistic subdiscipline. Generally, these sampling methods fall into two basic groups: those that strive for representativeness, also known as probability methods, and those that do not, also called non-probability methods. “Non-probability methods cannot be used to make statistical inferences about the population from which

¹ A sample is hardly ever a perfect replication of the statistical distribution of all subgroups in the population; differences between sample and population are often due to biases called sampling error.
they are drawn. In choosing to adopt non-probability methods [such as single case studies] one must therefore accept that statistically rigorous representativeness is not a primary issue in the research design” (Rice 2010: 232). The main aim of this chapter is to describe the types of sampling that are commonly used in linguistic research, namely convenience sampling, random sampling, stratified sampling, ethnographic sampling, and network sampling. A secondary goal is to describe further issues that arise in specific linguistic subdisciplines.

2 Types of sampling

2.1 Convenience sampling

Some researchers recruit subjects bearing not only representativeness in mind, but also convenient accessibility. It is thus not surprising that the most frequent subject pool in convenience sampling is student volunteers. Obviously, convenience sampling excludes a great proportion of the total population, resulting in an unknown amount of systematic biases. This effectively means that there is a fair chance that a study based on convenience sampling reports skewed results, and we have to be careful about any inferences made. Research based on a convenience sampling is thus rather limited in its generalizability. Why do researchers rely on this sampling method? Primarily because it is quick and easy. Convenience sampling is often used in pilot studies since it allows the researcher to survey the field before setting up a more elaborate sample. It is also regularly used in experiments conducted in linguistic paradigms such as theoretical syntax/semantics/phonology, which assume that there is little interpersonal variation (or that such variation is inconsequential for the theoretical model), due to a stable underlying representation across the population. Note that in some instances, convenience sampling is theory independent and therefore more justified – for example, when the speech community is so restricted that the researcher has to sample everyone they can get their hands on. This is particularly the case with fieldwork on endangered languages and/or very small speech communities.

2.2 Random sampling

De Vaus (2001: 60) argues that “the surest way of providing equal probability of selection is to use the principle of random selection. This involves listing all members of the population (this list is called a sampling frame) and then . . . ‘pulling their names out of a hat.’” What this effectively means is that in a random sample of a group or community (i.e., a city such as London, a country such as Brazil), every member of that community has an equal chance of being chosen for participating in the research. Early sampling strategies used in linguistics included choosing people randomly out of telephone books or electoral registers. So in a population of a hundred people, for example, we might select
twenty members based on a set of twenty randomly generated numbers between one and a hundred. The problem with this type of random sampling is that it “requires a good sampling frame. While these may be available in some populations (e.g., organisations such as schools, churches, unions), adequate lists are often not available for larger population surveys of a city, state or country” (De Vaus 2001: 64).

A more tractable way of sampling representatively is to construct a systematic sample. We do this by dividing the population size by the intended sample size and then sampling a representative fraction. For example, if the population is 10,000, but we only have time or money to sample 200, we interview one person out of every 50 (200/10,000) people (see Rice 2010). This can be more feasible than genuinely random selection, and can lead to a more even sample, but assumes a reliably homogeneous population.

The main asset of random sampling is that it can lay claim to representativeness in a statistical sense, which permits extrapolation from the sample studied to the larger population. It also allows the researcher to examine the full spectrum of the target population sampled. This is particularly the case when investigating a large complex community, “especially if it has a high degree of randomness, as in an urban setting in which the neighborhoods are not preselected, [which] requires some kind of indexing procedure in order to cluster the subjects into appropriate social groups” (Chambers 2003: 45).

But random sampling is not without its problems. As De Vaus (2001: 64) points out, the cost and effort involved are often “prohibitive. It would probably involve interviewers travelling long distances just for one interview . . . [Hence random sampling] is most appropriate when . . . the population is geographically concentrated or the data collection technique does not involve travelling.” Furthermore, even supposedly random sampling methods tend to introduce biases into the sample, so that, for example, electoral registers are biased toward the adult native population and telephone books tend to be cut up by geographical areas such as wards, regions, or postal/telecommunication boundaries. Indeed, the random samples used in actual linguistic research hardly ever live up to the stringent exigencies of random sampling under a strict sociological definition: once chosen by the sampling frame, some people move away, refuse to participate, fall ill or die, turn out to be uncooperative, or simply cannot find the time. These individuals either leave gaps in the original survey design or have to be replaced by speakers that have the exact same social characteristics, reducing randomness in selection. Thus, in actual practice, random sampling is hardly ever completely random.

The uneven distribution of populations across space adds a further bias to the data: subgroups tend to be geographically or socially distributed in non-random ways, so the assumption of homogeneity that is necessary for random sampling is often invalid. Random sampling also virtually guarantees that interviewee and interviewer are complete strangers, which tends to result in rather formal speech styles. This is a problem for those linguistic subdisciplines that aim to investigate a range of different stylistic levels or tap into the interviewees’ most casual speech
behavior (see the Observer’s Paradox, Chapter 6). Finally, in some linguistic subdisciplines, participants need to be recorded in labs, which is unfeasible if they are expected to travel from remote, randomly selected locations.

Despite all its shortcomings, however, in cases where the researcher does not know the area to be investigated or its salient social distinctions, random sampling might help to explore which social dimensions correlate with or indeed condition language use (see also Milroy and Gordon 2003).

2.3 Stratified random sampling (judgment or quota sampling)

The difficulties associated with random sampling have led many linguists to weigh “the costs of achieving statistical representativeness against the limited additional benefits it might provide” (Milroy and Gordon 2003: 26). This is a particularly pertinent concern since “speech communities tend to consist of many varieties spoken by groups containing very different numbers of individuals, so that uniform sampling leads to redundancies for some groups and risks missing others entirely” (Sankoff 1988: 900). Contemporary linguistic research tends to rely instead on the principle that a sample needs to be representative for the purposes of the study. Linguists who use this reasoning decide “on the basis of prior experience” (Rice 2010: 240) which stratifying variable(s) matter in a population. They then identify in advance the types of speaker groups they want to investigate – those hypothesized to correlate with linguistic variability – and sample systematically from these groups. For example, a study that wants to investigate the language use of a certain area (such as a particular barrio of Buenos Aires) or of a certain ethnic group (such as Pakistani immigrants in Saudi Arabia) would divide the population into mutually exclusive subgroups, called strata, and sample within these subgroups, making sure that all the subgroups of the population are represented proportionately within the sampling frame. This technique, which has been widely adopted for sampling in linguistics, is called stratified random or judgment sampling.

Although the original conception of judgment sampling was based on sampling within each stratum (using, e.g., a fraction of 1/50th of every social grouping), later studies have suggested that the aim of a sample is not to be “a miniature version of the population but only that we have the possibility of making inferences about the population based on the sample” (Sankoff 1988: 900). For example, Gordon’s (2001) study of phonological changes in two small towns in Michigan aimed at investigating the Northern Cities Vowel Shift. He collected data from sixteen speakers, equally stratified by age and gender in two towns, one relatively close to Detroit and another approximately halfway between Detroit and Chicago. Structuring his sample in such a way, Gordon was able to “examine the interactions of three important social variables [location, age and gender] . . . using a relatively small number of speakers . . . [Importantly], the choice of social variables to investigate was guided by the objectives of the study” (Milroy and Gordon 2003: 34).
Tagliamonte (2006) suggests that a stratified sample should be representative, at the minimum, with respect to age, sex, social class, and educational level. While these social categories have proven important for numerous large-scale studies, Chambers (2003) has pointed out that many artist and student communities tend not to be differentiated by factors such as gender or class. Furthermore, in many communities, particularly less well-studied, non-Western contexts, other factors — such as kinship, experience, urbanness, or religion — underlie the creation and perception of social divisions and linguistic usage. The very relativity of such criteria suggests that we need to consider the local context of the community when making decisions about our data collection strategy. Crucially, in the absence of prior experience with the people we aim to investigate, a sampling strategy that relies on predetermined categories might miss important local social contrasts, or might end up being governed by the prejudices or preconceptions of the researcher rather than orienting to local categories.

2.4 Ethnographic approaches

Ethnographic data collection is the antithesis of random sampling: notions such as randomness, representativeness, or indeed statistically generated generalizability of results are not relevant to this empirical methodology. Instead, ethnographic research aims at the discovery of emic categories, the social, cognitive, cultural, and linguistic contrasts that are salient in a particular community (as opposed to etic criteria, namely, extrinsic concepts and categories imposed by the researcher; see Chapter 10). Eckert (2000: 69) describes this approach succinctly: “while survey fieldwork focuses on filling the sample, ethnographic fieldwork focuses on finding out what is worth sampling.”

Crucially, the local sociocultural distinctions acquired via ethnographic fieldwork expand the researcher’s explanatory possibilities, allowing them to move beyond standardly assumed macrosocial categories, such as age, sex, and gender, and toward participant-designed categories. Indeed, the criteria for stratification that fall out of ethnographic research are generally not objective, global categories, but rather contrasts that reflect the procedures local participants employ in constructing and recognizing social worlds. This can result in groupings as unpredictable as, among the Wishram Chinook tribe (Hymes 1972), adults and children past babyhood as a first community, babies, dogs, coyotes, and guardian spirits as a second community, and those “whose guardian spirit experience had granted them the power of being able to interpret the language of the spirits” (p. 28) as a third.

Since a deep sense of the locally salient social groupings and values can only be achieved via sustained presence in the community, researchers doing ethnographic fieldwork typically “hang . . . out” (Giddens 2006: 85) or live with the group whose practices are of interest, becoming a participant observer rather than merely an outsider/interviewer, and sampling data according to ethnographic relevance. Ethnographic research thus has the crucial advantage that it allows us
not only to collect reports of the cultural context that might impinge on language use, but also to observe these practices first hand. The researcher’s access goes beyond linguistic behavior to include other behavioral practices, attitudes, ideologies, and information on how the people understand their own and others’ behavior, all of which can help us interpret linguistic practices. As such, the data produced by ethnographic fieldwork is, by its very nature, much *richer* than data resulting from other data collection techniques. It often documents informants’ behavior in a variety of situations, and thus a range of speaking styles, including the vernacular sought after in variationist sociolinguistic research (see Chapter 6).

While unparalleled in terms of the depth and quality of the data produced, the ethnographic approach to sampling is also the most cost-intensive in terms of time and effort committed per researcher. As a rule of thumb, most researchers estimate an outcome ratio of at least 10:1 (ten hours spent in the field yield roughly one hour of recorded data). In addition, since the data are so highly specific to the local setting, the resulting findings are inherently difficult to compare, contrast, or collate with other datasets. Indeed, testing the reliability of findings culled from ethnographic data would mean spending an equal amount of time in the same community. Hence, ethnographic data is not representative in a statistical sense (see Eckert 2000; Tagliamonte 2006: 27); any generalized claims beyond the confines of this community have to be treated with care. Note, however, that the categorizations that emerge from ethnographic fieldwork are not merely subjective, but rather intersubjective since they (ideally) converge with the community’s assessment.

### 2.5 Social network or snowball sampling

The technique known as the social network or snowball sampling technique also aims to investigate locally specific, participant-designed groups. Unlike ethnographic sampling, this approach does aim to examine quantitative variation across the group, but uses networks for the recruitment and sampling of participants. The term *friend-of-a-friend approach* in network sampling was coined by Lesley and James Milroy (1992), who contacted their participants by being referred from friend to friend, neighbor, or acquaintance in working-class neighborhoods in Belfast. In conjunction with this recruitment approach, they measured individuals’ network status and found that linguistic variation did not necessarily correlate with the etic categories used in stratified sampling, but did correlate with certain types of networks.

While this approach shares some of the limitations of ethnographic fieldwork in terms of the time investment needed in order to enter the network, and the tact and emotional involvement required to operate within it, one of the inherent advantages of this technique is that informants are less likely to decline a request for an interview if the researcher has been referred to them by a friend. Another positive side effect of network sampling is that the researcher encounters a great amount of social information that flows in networks.
The Community of Practice approach is similarly concerned with the social practices of a subsection of the speech community, whereby membership is established by the participants themselves, rather than by the researcher (Holmes and Meyerhoff 1999: 175–6; Meyerhoff 2002: 527–8). A Community of Practice framework assumes that participants negotiate the meaning of linguistic variables and therefore may not always behave similarly with respect to their use of these resources (Eckert 1989, 2000). This sampling method can lead to more ethnographic analysis, with full immersion into a network, or it can be combined with random sampling, only using snowball sampling for recruitment purposes and to establish participant-designed categories, but aiming at a more stratified sample overall.

Social network sampling thus combines elements of random sampling and ethnography: whereas random and judgment sampling see the individual speaker as a representative of abstract, predefined social categories, such as age, gender, and class, social network sampling focuses on voluntary membership in participant-designed networks or social groupings. Unlike ethnographic methods, however, social network sampling usually does not entail the researcher’s complete immersion into the local community.

The preceding subsections have shown that linguists draw on a wealth of strategies to collect samples from a given population or group of people. These sampling techniques can be, and often are, combined, so that, for example, networks might be recruited to fill in the sample cells for a judgment sample or, alternatively, an ethnographic project might attempt to contrast participant-designed categories against macro-social categories (see Labov et al.’s 1968 studies in South Harlem). Generally, though, the choices we make in terms of sampling strategies tend to be motivated by our theoretical persuasions about the fabric of social structure, our research questions, and, at times, our wider epistemological commitments. These choices have crucial consequences for the interpretation of data, since the type and amount of data we collect fundamentally determines the results we get, the statistical models we can use (see Chapters 14–16), and, ultimately, the questions we can answer.

### 2.6 Size matters

Once we have chosen our sampling method, the next question that arises is how much data is enough? Social scientists tend to believe in the mantra that big numbers are beautiful. This is because large samples allow us to draw more reliable inferences about the behavior of the whole population. Crucially, however, statisticians tell us that the choice of sample size essentially depends on the degree of accuracy we are aiming for: we need to choose the degree of error we are prepared to tolerate in our sample. Table 5.1 (from De Vaus 2001: 71) depicts the sample sizes needed in order to be 95 percent confident that the behavior of the sample chosen from the population is the same as the behavior of the population at large plus or minus the sampling error (for confidence limits, see also Moser and Kalton 1971; Woods, Fletcher, and Hughes 1986; also Chapter 15).
Thus, for example, if we find that 34 percent in our sample of 816 people merge their low back vowels [ɒ] and [ɔ:], as is common in many American dialects, we can be 95 percent confident that between 37.5 and 30.5 percent (34 percent plus/minus 3.5 percent) of the population at large do indeed merge their vowels. Table 5.1 reveals that increasing the sample size with smaller numbers has a disproportionately large effect on improving the sampling error. In fact, De Vaus points out that many survey companies restrict their samples to 2,000, because the extra cost involved in increasing the sample does not have enough payout in terms of increased accuracy. Giddens (2006: 88) suggests a similar figure for research in sociology, arguing that “studies of only two or three thousand voters, for instance, can give a very accurate indication of the attitudes and voting intentions of the entire population.” These are obviously very large numbers that are unattainable for most linguistic research programs. We might want to ask what is the absolute minimum sample size that allows us to generalize from our data with reasonable confidence. Generally speaking, Neumann (2007: 222) gives as a rule of thumb for the social sciences that a small population (< 1,000) would be accurately represented by a sample of 300 (hence 3 percent), whereas a larger population (> 150,000) would require a sample size of at least 1,500.

Linguistic studies tend to be based on much smaller samples of informants than research in other areas in the social sciences (but see below). The reason for this is first and foremost a practical one: The lions’ share of quantitative research in fields

---

Table 5.1  *Relationship between sample size and sampling error: De Vaus 2001: 71*

<table>
<thead>
<tr>
<th>Sampling errora</th>
<th>Sample size</th>
<th>Sampling error</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>10000</td>
<td>5.5</td>
<td>330</td>
</tr>
<tr>
<td>1.5</td>
<td>4500</td>
<td>6.0</td>
<td>277</td>
</tr>
<tr>
<td>2.0</td>
<td>2500</td>
<td>6.5</td>
<td>237</td>
</tr>
<tr>
<td>2.5</td>
<td>1600</td>
<td>7.0</td>
<td>204</td>
</tr>
<tr>
<td>3.0</td>
<td>1100</td>
<td>7.5</td>
<td>178</td>
</tr>
<tr>
<td>3.5</td>
<td>816</td>
<td>8.0</td>
<td>156</td>
</tr>
<tr>
<td>4.0</td>
<td>625</td>
<td>8.5</td>
<td>138</td>
</tr>
<tr>
<td>4.5</td>
<td>494</td>
<td>9.0</td>
<td>123</td>
</tr>
<tr>
<td>5.0</td>
<td>400</td>
<td>9.5</td>
<td>110</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>10</td>
<td>100</td>
</tr>
</tbody>
</table>

*Notes:* a This is in fact two standard errors.

---

2 The numbers in Table 5.1 are given for very large populations. Indeed, De Vaus (2001: 71–2) points out that “the size of the population from which we draw the sample is largely irrelevant for the accuracy of the sample. It is the absolute size of the sample that is important. The only exception to this is when the sample size represents a sizable proportion of the population (e.g. 10 percent). In such cases a slightly smaller sample is equally accurate.”
such as geography, demography, or political science is based on secondary data, which means that the cost of data collection and handling are borne by large agencies like the national census or commercial institutions. Also, those social scientists who do not work with pre-collected data tend to rely on surveys, questionnaires, GPS tracking, or notes from focus groups, none of which requires the time-intensive linguistic recording and transcription work which explodes data-handling time by a subdiscipline-dependent ratio (ten hours of word-for-word transcription per recorded hour; 1:100 for close phonetic transcription, and potentially even more when transcribing children’s speech; see Chapter 12). Consequently, even linguistic projects that initially aimed for big numbers have not been able to process all the data gathered. For example, the Shuy, Wolfram, and Riley study (1967) randomly sampled and interviewed 254 families (702 subjects in total) in Detroit. Exigencies of time and data handling reduced the number of speech samples they could analyze to only sixty speakers, and the principal linguistic analysis was limited to thirty-six of those (Wolfram 1969). Thus, the great majority of linguistic studies tend to rely on a fraction of the sample size commonly used in other social sciences.

Note, however, that the ratio of observations to informant numbers in linguistics is in an inverse relationship to most other social science research. While, for example, a study on voting behavior or religious affiliation tends to collect one or very few observations from many informants, linguistic research typically relies on fewer informants, but collects many observations from every single one of them. For example, Khattab and Al-Tamimi (in press) collected longitudinal data from ten toddlers for a phonological acquisition study, but the token number in their statistical design was 5,697 words. Hence, linguistic research can attain a relatively large number of observations by relying on a smaller number of informants. What this effectively means is that in the field of linguistics, the notion of sample size needs to be further specified: Are we referring to the number of observations or the number of informants? Note that, ideally, linguistic research that relies on a large number of observations per speaker ought to treat speaker as a random variable in a mixed-effects model (see Roberts 2012; Johnson in press; also Chapter 16).

In fact, we might want to argue that large informant numbers are not even necessary for the purposes of language research. It is well known that “the larger the size and the lower the population heterogeneity, the more precise sample estimates will be” (Rice 2010: 230). Hence, the precision of a sample is a function not only of the number of observations, but also of the amount of variability within the population as regards the feature of interest. The fact that language is based on mutual intelligibility “places a limit on the extent of possible variation, and imposes a regularity (necessary for effective communication) not found to the same extent in other kinds of social behaviour” (Sankoff 1980: 51). Indeed, the kind of speech that tends to be investigated in most linguistic subdisciplines is more homogeneous than other types of variable behavior, since it is “not subject to the informants’ control in the way that answers on voting choices [or other forms
of social behavior] would be” (Labov 1966: 180). Consequently, the recommendation in the literature is that “even for quite complex communities samples of more than about 150 individuals tend to be redundant, bringing increasing data-handling problems with diminishing analytical returns” (Sankoff 1980: 51–2).

In terms of minimum numbers, Meyerhoff and Schleef (2010) argue that five or six speakers per cell suffice in order to make statistically sound generalizations about the data collected. For example, the Spanish learner database created by Mitchell et al. (2008) consists of twenty learners of Spanish at each of three levels, illustrated in Table 5.2.

Making the sample more robust by adding just five speakers per cell would require fifteen more speakers overall. Including another factor, such as the speaker’s motivation to learn Spanish (divided, say, into two levels, higher or lower), would require sixty more speakers overall, an increase of 100 percent. This example demonstrates the extent to which sampling design is fundamentally determined by practical considerations such as financial and temporal resources. Needless to say, if a researcher decides to exclude a factor, variation along this dimension must be controlled for rather than ignored in the sample, which effectively shrinks the size of the population under investigation.

### 2.7 Applications and developments

If we divide the data collection methods used across linguistic subfields into “armchair,” “field,” and “laboratory” (Clark and Bangerter 2004: 25), it is not surprising that the interest in sampling and stratification issues has primarily grown out of those subfields using empirical methods in which there is a perceived need to justify one’s database. Since sociolinguistics, developmental linguistics, and cognitive linguistics place more weight on the representativeness of their data, these subdisciplines have historically contributed most to our knowledge base on sampling strategies. However, as linguists’ thinking about the impact of social, cognitive, and idiolectal effects on language use and of the scientists’ role in

---

**Table 5.2 The database for Spanish second language acquisition. Mitchell et al. 2008: 293**

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 9 (13–14)</td>
<td>20</td>
</tr>
<tr>
<td>A2 students (17–18)</td>
<td>20</td>
</tr>
<tr>
<td>Undergraduate (19 plus)</td>
<td>20</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>60</strong></td>
</tr>
</tbody>
</table>

---

3 Age can be sliced up in a number of ways. In Table 5.2, it is treated as a categorical variable, but numerical age could also have been treated as a continuous variable. The decision about whether to treat age as continuous or categorical has important ramifications for the statistical model we choose for the analysis of our data (see Chapter 14).
defining and ratifying “social categories” has evolved, sampling methodology has become more precise throughout the field. In the remainder of this chapter, we provide a brief overview of how sampling has evolved in a selection of linguistic subdisciplines.

3 Sampling in subdisciplines of linguistics

3.1 Theoretical linguistics

Theoretical research in the generative tradition has often been associated with so-called “armchair” methods: “you imagine examples of language used in this or that situation and ask yourself whether they are grammatical or ungrammatical, natural or unnatural, appropriate or inappropriate” (Clark and Bangerter 2004: 25). This practice may have arisen out of (early) generativist claims that the intuitions of every native speaker fully represent linguistic competence (see Chomsky’s 1965 concept of the “ideal speaker”). Indeed, if every native speaker has the same hard-wired language faculty – I-Language, bioprogram, or Universal Grammar, depending on the flavor of generative theory – consulting a range of speakers about the same phenomenon would only lead to replications of information and amount to a waste of time.

In recent years, theoretical linguistics has adopted increasingly sophisticated methods of sampling data, often continuing to rely on small numbers of respondents, but using more finely calibrated elicitation procedures (see Chapter 3). Usage-based theoretical research is also engaged in more robust empirical inquiry, often taking advantage of the ever-increasing number of pre-collected, pre-transcribed, even syntactically parsed corpora (see Chapters 11 and 13). Concerns with empirical methods, including questions of sampling, have consequently become more widespread in these subfields. Most recently, theoretical research, in particular situated at the interface with cognitive and/or sociolinguistics, has started to rely on very large samples (in terms of informants as well as observations). Wolk et al. (in press), for example, extracted 3,824 tokens of ‘s-genitive and of-dative constructions from the Archer Corpus (which spans the period 1650–1990), in order to demonstrate how short-term distributional fluctuations can trigger long-term changes in probabilistic grammars. A number of linguists have also used the World Wide Web to sample even larger numbers of tokens (e.g., Walter and Jaeger [2008], who extracted 260 million instances of that is and over 37 million instances of this is). Such large samples allow testing for a large amount of interacting effects. Research based on questionnaires, such as magnitude estimation, and collected via web-based platforms also tends to rely on larger numbers (see, e.g., Bard, Robertson, and Sorace 1996).

Note, however, that these large-scale projects are for the most part based on convenience samples and often give no information about the number or social profile of their informants. As such, they do not claim to sample randomly or to fill
The sampling strategy of early dialectological projects was driven by the perceived need to document a “genuine” or “pure” form of the language (see Milroy and Gordon 2003) before it disappeared, resulting in the practice of targeting conservative rural speakers who were assumed to portray the most traditional dialect features (non-mobile older rural males, or NORMs; see Chambers and Trudgill 1998). In the 1970s and 1980s, in line with the positivist epistemologies that pervaded the social sciences at the time, a range of sampling methods were adopted from economic geography. In particular, grid sampling, which involves superimposing a matrix of equivalently sized cells on the area to be sampled and selecting an equivalent number of participants from each, was adapted for quantitative dialectological projects. The rationale for this random spatial sampling was that it avoids the problem of missing small spatial structures, thus avoiding bias. Note, however, that grid sampling is the geographical equivalent to random sampling in a single speech community: it is a priori, static, and entirely independent of human activity (Romaine 1980; Milroy and Gordon 2003).

More recently, critical reflections on sampling in dialectology have problematized the operationalization of space as socially uniform carrier material over which linguistic variability can be superimposed (Britain 2004, 2010). Geographers as well as dialectologists have argued for socially sensitive projects relying on a greater amount of ethnographic fieldwork. As a consequence, sampling methods in contemporary dialectology have become increasingly cognizant of the fundamental role which human agency and social relations play in shaping and construing geography. This has resulted in the development of geo-demographically and socio-geographically sensitive sampling criteria (see, e.g., Cheshire, Edwards, and Whittle 1989, 1993; Buchstaller and Alvanides in press).

On the other hand, large-scale linguistic atlas projects sample vast numbers of informants across a geographical area, but the number of participants included in any one place tends to be relatively restricted. For example, the Linguistic Atlas of North America (Labov, Ash, and Boberg 2006) collected 417 speakers across the territory of English-speaking North America, but only two to six informants per locality. The Syntactic Atlas of the Netherlands Dialect is based on data in 250 cells of variable geographical size for the whole of the Netherlands (Barbiers, Cornips, and Kunst 2007), with two speakers per cell. Obviously, while these
informants are assumed to represent the spatial location for which they stand, they tend not to be sampled on the basis of stratificational criteria, which means that their representativeness in the social scientific sense of the term is not necessarily a given. Indeed, many atlas-style projects (a notable exception is Kurath 1972) have focused on either urban (the Linguistic Atlas of North America) or rural (the Survey of English Dialects, the Syntactic Atlas of the Netherlands Dialect) speech, which reduces the demographic representativeness of the study to just this settlement type.

3.3 Variationist sociolinguistics

The quantitative paradigm established by William Labov in the 1960s and 1970s (Labov 1966) is the main framework within which sociolinguists investigate quantitative linguistic variability in its social setting. Naturally, sampling is inherently important to uncovering socially conditioned patterns of linguistic variation. Early sociolinguistic research drew heavily on random sampling, en vogue in other social sciences at the time (see Massey 1985).

Increasing awareness of and interest in stylistic and situational variability prompted the discipline to construct samples of various kinds to answer questions about the distribution of variability across social space and to acknowledge both inter- as well as intra-speaker variation. Judgment sampling seemed to alleviate many of the problems of random sampling and continues to be favored by many sociolinguistic research projects. Since the advent of network studies in the 1970s, a number of researchers have started to target smaller groups of self-defined communities, often drawing on the methods of participant observation, adapted from anthropological linguistics and ethnography (see Eckert 2000).

The wealth of sampling methodologies currently used in sociolinguistic research suggests that "researchers are now more relaxed than they once were about methodological issues such as whether or not their account ... [is] technically representative or whether strict random sampling procedures should be used. This shift in attitude ... enables researchers to select more freely from a range of methods those which, within a defensible theoretical framework, will best enable them to achieve their goals" (Milroy and Gordon 2003: 46–8). Tagliamonte (2006: 28) proposes that sociolinguistic research is now characterized by a certain mix-and-match attitude: “the critical component of this hybrid methodology for variationist analysis is that the researchers decide which type of representativeness is sufficient – or attainable – depending on the focus of the study.”

3.4 Phonetics and sociophonetics

Traditional experimental phonetic/phonological research places stringent control over the technical methods used in data collection (Foulkes, Scobbie, and Watt 2010; Scobbie and Stuart-Smith 2012). A great deal of attention is paid
to the physical environment in which recordings are made, favoring laboratory over naturalistic settings, and using high-quality recording equipment in order to capture the best sound quality for auditory, acoustic, perceptual, or articulatory analysis. Laboratory phonetic research also aims to control for as many of the potential linguistic confounds as possible when analyzing sounds, yielding elicitation techniques that result in rather stilted speech (e.g., words in isolation or embedded in the infamous carrier sentences which help control for speech rate, rhythm, and neighboring sounds). Due to this laborious and detailed methodology, small numbers of speaker and tokens are justified, and male speakers are often preferred over females due to the relative ease in analyzing their spectrographic outputs. Results regarding the phonetic pattern(s) of interest (e.g., consonants, vowels, or suprasegmental features) are normally taken to be characteristic of the language of the speaker, and not much information is collected about individual sociolinguistic backgrounds, although monolingual speakers who are thought to speak a standard representative variety are often recruited to avoid noise in the data.

More recently, however, advances in theory and technology have seen an unprecedented integration of the fields of phonetics and sociolinguistics, resulting in a revamping of methodologies used (Di Paolo and Yaeger-Dror 2010; Thomas 2011; Scobbie and Stuart-Smith 2012), with positive effects on sampling.

In terms of theoretical advances, a recognition of the place of speech variability and social variation in phonological theory and models of cognition (e.g., Pierrehumbert 2002; Hawkins 2003; Kristiansen 2006; Foulkes 2010) has necessitated the use of larger, socially stratified corpora to allow gradient listener- and speaker-induced sound change to be tracked either longitudinally or in apparent-time studies (e.g., Harrington 2006, 2010; Fromont and Hay 2008). In sociolinguistic research, interest in non-categorical fine-grained variation has risen due to the discovery of its role in subtle but regular and systematic changes in perception and/or production; these, in turn, have consequences for social uses of language and for language variation and change (e.g., Docherty and Foulkes 1999; Hay, Warren, and Drager 2006). This has necessitated the use of both stratified sampling, to recruit speakers from the varieties of interest, and snowball sampling, to recruit further speakers from the same networks (Milroy and Gordon 2003: 32; Scobbie and Stuart-Smith 2012: 611) Medium-sized annotated corpora (forty to sixty speakers) that have been designed for phonetic analysis using stratified sampling techniques include the Buckeye Corpus (Pitt et al. 2005), the Nationwide Speech Project (Clopper and Pisoni 2006), and the Kiel Corpus of Speech (Kohler 2001).

In terms of technological advances, more portable professional recording and improved sound transmission technology now allow researchers to combine higher-quality field recordings with more naturalistic conditions and larger participant pools, even for physiological investigations which were traditionally confined to the lab due to heavy articulatory equipment (Gick, Bird, and Wilson 2005; Scobbie, Wrench, and van der Linden 2008; Scobbie, Stuart-Smith, and
Lawson 2009). New automated measurement techniques, now available through popular software programs such as Praat (Boersma 2001) and EMU (Harrington 2010), have also made analyses of much larger datasets feasible. Further technological development and guidance are still needed to enable researchers to process these datasets and normalize across speakers (Foulkes, Scobbie, and Watt 2010: 733).

At the other end of the spectrum, advances in sociophonetic theory have also highlighted the need to look at the individual and their identity in order to understand their phonetic behavior (Johnstone and Bean 1997; Wassink and Dyer 2004; Docherty 2007) and the subtle phonetic patterns they employ to signal group affiliation (e.g., Docherty and Foulkes 1999; Stuart-Smith 1999, 2007; Local 2003; McDougall 2004); this justifies the use of smaller numbers of participants in order to carry out more detailed analyses on each individual. Sampling in sociophonetics has also been concerned with the characteristics of the field-worker/interviewer, as these have been shown to influence the perception and production behavior of participants, and can actually be used as part of the experimental design/research question (e.g., Hawkins and Smith 2001; Clopper and Pisoni 2004; Hay, Warren, and Drager 2006; Delvaux and Socquet 2007; Foulkes, Scobbie, and Watt 2010).

### 3.5 Child language research

Sampling in child language research is inextricably linked to the age and stage of development of the child, the type of data that are possible to elicit, and the level of inference one is prepared to make about the verbal and non-verbal behavior that can be gathered.

Research with infants requires the use of indirect perceptual methodologies such as the High Amplitude Sucking, head-turn, and preferential looking techniques (e.g., Fernald and Kuhl 1987; Eilers, Wilson, and Moore 1977) to draw inferences about the children’s “raw” perceptual abilities and their journey toward tuning in to their input languages (see Menn and Ratner 2000 for an overview of data collection methods in child language research). Sample sizes are normally between ten and twenty in this age group, but attrition is very high due to having to discard non-cooperating participants, so the number of infants initially recruited for the research can be much higher.

At the onset of speech (nine to twelve months), individual differences are key, and the rate at which children show comprehension and/or production of early words can vary widely, which makes studies of language development in the second year most suited to case studies, in order to thoroughly document the child’s transition to language (e.g., Leopold 1939; Brown 1973). What also necessitates the use of small numbers is the relative difficulty in collecting, transcribing, and analyzing child speech compared with adult speech (Khattab and Roberts 2010). Young children’s speech is often unintelligible and difficult to transcribe, and, for children who are making the transition from babbling to early...
words, a thorough analysis of word identification (e.g., Vihman and McCune 1994) is required to establish whether or not a given utterance is a real word. Children’s smaller vocal tracts and mobile nature also offer challenges for acoustic analysis (Khattab and Roberts 2010).

The most common sampling methods for preschool children consist of cross-sectional designs using relatively large samples of children (a hundred or more), and capturing their production and/or perception abilities in order to derive norms for the development of various aspects of the grammar. Many such studies are carried out by speech and language therapists interested in early diagnosis of speech and language impairment (see Menn and Ratner 2000 for a summary). Since language development in this age range is very fast, researchers are careful to categorize children into narrow age bands, normally using six-month intervals. What tends to fall by the wayside in these studies, however, is a thorough discussion of individual differences or a detailed understanding of each child’s overall (socio)linguistic abilities (Docherty and Khattab 2008).

At the other end of the spectrum are small group and single case studies of child language. What these studies lack in numbers, they more than make up for in meticulous analysis and attention to detail, particularly in longitudinal design (see Chapter 22). Understanding the process of language acquisition requires close attention to the input the child receives and the role of social context (home, day care center, neighborhood, etc.), and therefore also requires documentation and analysis of both adult (caregiver) and child language.

Finding children to participate in research can be more challenging than finding adults, due to the need to find willing parents first, followed by a strict ethical approval process in many countries. For recruiting young children, playgroups and early day care centers are often the first point of contact for the researcher, while schools and community centers can be targeted for older children. Demographic information about the institutions can provide the researcher with clues about various social characteristics of the target families, and recruiting from different institutions is a common way of tapping into different social categories. This sampling strategy can yield a representative sample for the subgroups that the researcher is interested in. However, a school choice based on the perceived social class, religious, or language background of its pupils according to official statistics can yield a population that is actually far from homogeneous in that respect, and the researcher needs to supplement this information with detailed language background questionnaires or interviews with the parents. Moreover, while categorization of the child’s ethnicity, religion, social class, and so on, is normally derived from the caregiver’s background, we need studies that capture children’s individual expression of these social characteristics, in much the same way that children have been found to express gender that is above and beyond what is expected based on their biological development, based on socialization processes (Sachs, Lieberman, and Erickson 1973; Eckert 1997). This is possible with older children using ethnographic methods of data collection, but is much more challenging with younger children.
4 Conclusions

Any researcher in the field of linguistics has “specific social, [cognitive] or linguistic questions in mind when they start their research, and in order to ensure that their research adequately addresses those questions, they stratify their sample somewhat” (Meyerhoff and Schleef 2010: 7). Given that research in different linguistic subfields tends to ask very different questions, it is no surprise to find that the sampling strategies that have traditionally dominated these subdisciplines are vastly dissimilar too. More recently, however, linguistics as a discipline has witnessed a readiness to leave behind methodological orthodoxy, leading to a relaxation of the concept of statistical representativeness and a convergence of sampling methods across subdisciplines. In this chapter we have aimed not only to describe the most prevalent types of sampling tools used in linguistic research, but also to root these strategies in the questions and concerns they originated from and help to resolve.

References


