SAGE researchmethods

Corpus Construction: a Principle for Qualitative Data Collection

In: Qualitative Researching with Text, Image and Sound

By: Martin W. Bauer & Bas Aarts
Edited by: Martin W. Bauer & George Gaskell
Pub. Date: 2011
Access Date: April 17, 2020
Publishing Company: SAGE Publications Ltd
City: London
Print ISBN: 9780761964810
Online ISBN: 9781849209731
DOI: https://dx.doi.org/10.4135/9781849209731
Print pages: 20-37

Corpus Construction: a Principle for Qualitative Data Collection

KEYWORDS

corpus

corpus-theoretical paradox

homogeneity

population

relevance

representations (internal variety)

representative sample

sampling bias (non-coverage, response rate)

sampling frame

sampling strategy

saturation

strata and functions (external variety)

synchronicity

All empirical social research selects evidence to make a point, and needs to justify the selection that is the basis of exploration, description, demonstration, proof or disproof of a particular claim. The most elaborate guideline for selecting evidence in the social sciences is *statistical random sampling* (see Kish, 1965). The proficiency of representative sampling is uncontested. However, in many areas of textual and qualitative research, representative sampling does not apply. How do we select people for focus group research? Do we really want to represent a population in four or five focus group discussions? Unfortunately, to date this issue has been given little attention. In practice, researchers often try to fit the sampling rationale that seems misguided, like the choice of a false analogy. In this chapter we propose 'corpus construction' as an alternative principle for data collection. We use strong definitions for our basic terms: 'sampling' means statistical random sampling; 'corpus construction' means systematic selection to some alternative rationale, which will be explored below. Sampling and corpus construction are two different selection procedures. Like representative sampling, we steer the middle way between enumeration of a population and convenient selection. Unsystematic selection violates the principle of public accountability of research; however, corpus construction maintains the efficiency that is gained from selecting some material to characterize the whole. In this sense, corpus construction and representative sampling are functionally equivalent, although they are structurally different. With this use of language, we come to a positive formulation for qualitative selection, rather than defining it as a deficient form of sampling. In short, we contend that corpus construction typifies unknown attributes, while statistical random sampling describes the distribution of already known attributes in social space. Both rationales need to be distinguished carefully, to avoid confusions about, and false inferences from, qualitative research.

We develop this argument in three steps. First, we briefly review the key concepts of representative sampling, and allude to problems arising from unknowable populations. Secondly, we demonstrate corpus construction in the field in which it developed: linguistics. Thirdly, we abstract rules from this practice as guidelines for data selection in qualitative social research.

Representative sampling in social research

The act of taking stock of a population has a long history: governments have wanted to know what sort of inhabitants they govern in order to guide policy. The short history of random sampling began in the late nineteenth century in an atmosphere of conflicting opinions among researchers: some believed in complete enumeration, some in sampling, and others in single case studies. Only an unholy alliance between case-study researchers and random samplers could end the dominance of the total enumerators (O'Muircheartaigh, 1997).

Sampling secures efficiency in research by providing a rationale for studying only parts of a population without losing information - be that a population of objects, animals, human individuals, events, actions, situations, groups or organizations. How can the study of a part give a reliable picture of the whole? The key to this puzzle is *representativeness*. The sample represents the population if the distribution of some criterion is identical in the population and in the sample. Population parameters are estimated from observed sample estimates. The larger the sample, the smaller is the error margin of these estimates, while the sampling process itself may bring additional errors. In principle, there is a need to prove that the sampling criteria and the focal variables actually correlate. However, in practice, one often assumes that if the sample represents the population on a number of criteria, then it will also represent the population in those criteria in which one happens to be interested: The researcher may interview 2000 British people, carefully selected by age, sex and social class, and she will be confident in characterizing the nation's opinions on, say, genetically modified foods, with a known margin of error. This is achieved by following the rationale of sampling, which brings enormous savings in time and effort.

Sampling refers to a set of techniques for achieving representativeness. The key requirement is the sampling frame that operationalizes a population. It is a concrete list of units that are considered for selection. Every list entry represents only one member of the population, and every entry has an equal probability of being selected. A sampling frame may consist of telephone numbers, addresses and postal codes, electoral lists or lists of companies. For example, the list of students taking exams at a university is a sampling frame for the student population of that particular year. The quality of sampling frame is measured by its non-coverage. Most intensional definitions of the population are wider than its operationalization in an available list: for example, a nation's population includes its prisoners and mentally ill people, although the electoral list will exclude them. Telephone numbers create non-coverage, as some households may have no telephone while others have several. Non-coverage is the first bias of sampling.

A sampling frame is a precondition for the application of a sampling strategy. By generating 100 random numbers between 1 and 5000, and by selecting the items from the list that correspond to these 100 random numbers, a simple random sample of 100 out of 5000 is created. Take as a more elaborate example a multi-stage sample for a study into opinions on genetically modified food. The researcher may select a sample of 50 postal areas stratified by socio-economic characteristics, such as average income and urban or rural living. The assumption is that income and urban or rural living will influence opinions. In the second step, he randomly selects, in each of the 50 areas, 40 households from the postal area code, in which finally the field interviewer will talk to the one household member over 15 years of age whose birthday is nearest to the date of the interview. We have a quota sample if, at the last stage, the units are selected not randomly, but by giving the field interviewer some quota to find: the quota could be 20 women and 20 men, because we know that men and women are about equally distributed in the population.

Of the selected 2000 interviewees, some will not be reachable. This non-response introduces a second sampling bias. In the case of random sampling we will know how many were not approachable; but in the case of quota sampling we do not know, which makes the quota version a non-random and, for many researchers, a dubious procedure. Representative sampling reaches the best possible description of the population, despite observing only a part of it. However, it hinges on the availability of a sampling frame, a list or a combination of lists of members of the population, or knowledge of the distribution of key features in the population. Without lists or known distributions, the procedure cannot be used.

Let us consider some cases where the assumption even of a population is problematic. Some discussions of representativeness have argued for three dimensions: individuals, actions and situations (see, for example, Jahoda et al., 1951). Individuals act in situations, and, to generalize results of research to individuals acting in situations, all three dimensions need to be controlled. However, sampling focuses on individuals, not least as this is achieved most successfully. Neither for actions nor for situations are routine attempts at sampling made. Very few human actions (working, shopping, voting, playing, thinking, deciding) have been the focus of intense psychological study which has led to generalizations about human action without a basis in sampling. Equally, no attempts are made to sample situations in which people act. Why not? Neither actions nor situations seem to have a definable population. We would have to study *unknown populations*. Voting, working and shopping are important activities; however, it remains unclear how far their structure and function represent all human activity. Most social scientists regard results that are consistent across a few different situations as replicated and therefore robust. In doing so, they stick to induction for actors, but violate induction for actions and situations; sampling is applied neither for actions nor for situations (Dawes, 1997). Social science seems to rest happily with this contradictory practice.

Consider cases of *unknowable populations*. A prize of several thousand pounds was recently offered at a public lecture for anyone who came up with a sampling frame for human conversations and interactions. The speaker was confident that no-one would be able to meet the challenge. Consider the content of speech, the concatenation of words from a finite vocabulary according to a grammar. At any time the number of possible sentences is infinite, because the combinatorial space of words is an infinite resource. Speech, conversations and human interactions are open systems, with words and movements as the elements and an infinite set of possible sequences. For open systems, the population is unknowable in principle. Its elements can at best be typified, but not listed.

The rationale of representative sampling is useful for much social research, but it does not fit all research situations. There is a danger that we unduly extend the procedures of representative sampling to studies in which it may be inappropriate. We criticize certain forms of data collection as deviations from the ‘probabilistic standard’. However, even in the empire of chance, the ‘law of small numbers’ rules. Humans tend (except statisticians, of course) to overestimate the representativeness of everyday observations (Tversky and Kahnemann, 1974; Gigerenzer et al., 1989: 219ff). The moral is clear: pay more attention to sampling. However, our efforts may be misguided: the quest for representativeness may be channelling scarce resources towards strategies of selection that are inappropriate for the problem at hand.

The notion of ‘corpus’

Now we will explore what the linguist offers in constructing his corpus. Language is an open system. One cannot expect a list of all sentences from which to select randomly. The linguistics community has recently rejected a motion that language corpora be representative of language use (Johansson, 1995: 246).

The word ‘corpus’ (Latin; plural ‘corpora’) simply means ‘body’. In the historical sciences it refers to a collection of texts. It may be defined as ‘a body of complete collection of writings or the like; the whole body of literature on any subject … several works of the same nature, collected and bound together’ *(Oxford English Dictionary*, 1989); or as ‘a collection of texts, especially if complete and self-contained’ (McArthur, 1992). Examples, mainly collected during the nineteenth century, are the Corpus Doctrinae, a body of theological treatises of German ecclesiastical history; the Corpus Inscriptorum Semiticorum, a complete collection of ancient Jewish texts at the French Academy; or the Corpus Inscriptorum Graecorum of ancient Greek texts at the Berlin Academy. These collections tend to be *complete* and *thematically unified*, and to serve *research*.

Another definition of a corpus is ‘a finite collection of materials, which is determined in advance by the analyst, with (inevitable) arbitrariness, and on which he is going to work’ (Barthes, 1967: 96). Barthes, in analysing texts, images, music and other materials as signifiers of social life, extends the notion of corpus from text to any material. In his booklet on the principles of semiology he relegates considerations of selection to a few pages. Selection seems less important than analysis, but is not to be separated from it. Arbitrariness is less a matter of convenience, and more inevitable in principle. The materials ought to be homogeneous, so do not mix text and images in the same corpus. A good analysis stays within the corpus and accounts for all variation that is contained within it. In summary, while older meanings of 'text corpus' imply the complete collection of texts according to some common theme, more recent meanings stress the purposive nature of selection, and not only of texts but also of any material with symbolic functions. This selection is inevitably arbitrary to some degree: comprehensive analysis has priority over scrutiny of selection. Corpus linguistics, however, offers a more systematic discussion.

What are language corpora?

Corpora in the linguistic sense are collections of language data for the purposes of various types of language research. The term is tied to developments in computer studies of language (Johansson, 1995; Biber et al., 1998). A linguistic corpus is 'written or spoken material upon which linguistic analysis is based' *(Oxford English Dictionary*, 1989), or 'texts, utterances, or other specimens considered more or less representative of a language and usually stored as an electronic database' (McArthur, 1992). The corpora are structured along a number of parameters such as channel (spoken or written, written to be spoken, etc.), domain (art, domestic, religious, education, etc.), function (persuade, express, inform, etc.). Combinations of these subcategories may form a *hierarchical typology of registers*, as we will see. The earliest language corpora were usually of the written kind, and collected manually.

Once constructed, corpora can be used as databases for linguistic research. When the first corpora were constructed, data retrieval also had to take place by hand. So, for example, a researcher who was interested in working on verbs of perception in English (verbs like 'see', 'hear', etc.), would have to go through the corpus manually in order to find these verbs. Later corpora were computerized: the first was the Brown Corpus, constructed in the 1960s at Brown University in Providence, Rhode Island. Nowadays all corpora are computerized and allow automatic searches.

The early computerized corpora emerged at an interesting point in the history of linguistics, namely the beginning of the Chomskyan era. Chomsky's book *Syntactic Structures* (1957) is the seminal publication in this period. Chomsky claimed that all humans are endowed with an innate language capacity, which he labelled 'universal grammar'. Ever since the beginnings of Chomskyan linguistics there has been an emphasis on how linguists can go about constructing abstract representations of each and every speaker's knowledge of language. Because the theory is all about abstract representations, this field of linguistics is characterized by a move away from empiricism, and by a reliance on the internal knowledge of language that we have as native speakers. Chomsky made a distinction between what he called *competence*, which is the innate knowledge that speakers have of language, and *performance*, the way that they make use of their innate knowledge. More recently he has introduced the term ‘I-language’ (internalized language) and ‘E-language’ (externalized language). Chomsky's theory is a competence theory (a theory of I-language) and not a performance theory (a theory of E-language). In the Chomskyan model, any particular language constitutes an epiphenomenon, with the term ‘language’ now reserved to mean I-language.

Early followers of Chomsky very much railed against empirically oriented linguistics. Nelson Francis, the compiler of the Brown Corpus, was asked at a conference by Robert Lees, one of Chomsky's followers, what he was working on. Francis replied that he was compiling a corpus of American written and spoken English. This was met with exasperation by Lees who held that this was a complete and utter waste of time. Lees's view, and that of many Chomskyans at the time, was that one only needs to reflect for a moment in order to come up with one's own examples of particular linguistic phenomena in English. Chomskyan linguists have always insisted that the only interesting data for language study are introspective data, that is, data that are made up on the basis of one's native-speaker knowledge of a language. The aversion to real data has persisted to this day. Chomsky, when he was recently asked by one of us what he thought of modern corpus linguistics, simply replied ‘it doesn't exist’. The collection of data in a corpus is regarded by Chomsky as being on a par with butterfly collecting.

Corpus linguists in turn claim that corpora can be useful for linguists who are not native speakers, and may contain examples that are hard to think of, because they are rare. They feel that linguistics should concern itself with real language data, that is, performance data, and not with made-up, artificial competence data. Of course, the question of what kind of data to use was not the only bone of contention. Corpus linguists are, on the whole, inductivists, while Chomskyans are deductivists. The dispute, then, is also a methodological one.

What is corpus linguistics, and how can corpora be used in linguistic research?

The field of linguistics is vast, and includes subdisciplines such as psycho-linguistics, neurolinguistics, forensic linguistics, socio-linguistics, formal or theoretical linguistics, semantics and so on. People now also speak of ‘corpus linguistics’. One might wonder whether corpus linguistics is to be regarded on a par with the other branches of linguistics. Strictly speaking, corpus linguistics is not really a branch of linguistics in itself: it is a linguistic methodology that might be used in all branches of linguistics. So, for example, a syntactician might turn to a corpus to study particular grammatical structures, while a socio-linguist might want to study telephone conversations in a corpus to see whether people speak differently on the phone from the way they do in face-to-face situations. In fact, with this use in mind, some corpora contain not merely one but various categories of phone conversations: for example, conversations between people of the same social status, and between people of different social status. Another use that socio-linguists have made of corpora is to study the differences between the ways in which men and women speak (see, for example, Tannen, 1992a; 1992b; Coates, 1996). Linguists or socio-linguists who are interested in the phenomenon of ‘handwritten notices’, one of the categories found in the original Survey of English Usage Corpus at University College London, might have been amused by the following notice found on the door of a public lavatory at Euston station in London: ‘Toilets out of order, please use platform 6.’

How do researchers go about using a corpus? And what do they search for? Obviously, this depends on their research goals. In any case, a computer program is needed that can make intelligent searches. The simplest kind of search is for a particular lexical item, say the word ‘the’. Things become more complex if a search is made, say, for all the nouns in a corpus. To do this, the corpus needs to be parsed. In the early days, parsing was done by hand; now it is done automatically. The first stage of parsing is called tagging. During this process each and every word is given a word class label, such as noun, verb, adjective, etc. This can be done automatically by a computer program. The results are around 90 per cent accurate, and need to be corrected manually. The second stage of parsing involves analysing the corpus into grammatical constructions. For example, in a sentence such as ‘The dog bit the postman’, the program must analyse ‘the dog’ as the subject of the sentence, and ‘the postman’ as the direct object. Again, the automatic parsing has to be corrected manually. Once parsing is complete, queries can be formulated. For this a search program is needed. For example, the search program can be instructed to find all the direct objects following the verb ‘see’. Researchers at University College London have developed a tagger, a parser and also a search program. The search program is called the ICE Corpus Utility Program, or ICECUP for short.

An example of a corpus: the International Corpus of English

As an example of a corpus, Figure 2.1 shows the text categories in the International Corpus of English (ICE), developed in the English Department at University College London. ICE is international in the sense that identically constructed corpora have been set up, or are in the process of being set up, in some 20 English-speaking countries, among them the USA, Canada, Australia, New Zealand, Kenya and Nigeria. This corpus was designed to contain both spoken and written material, and both the spoken and written categories are further subdivided. The ICE-GB corpus of British English is now complete, and is available on CD-ROM; the other national corpora are still under construction. (See www.ucl.ac.uk/english-usage).

The different corpora in the ICE project are being constructed in order to allow researchers to study particular aspects of the English language in different varieties of English. Identical construction of the different national corpora is being implemented so as to allow meaningful statistical comparisons between the varieties of English. To give an example, somebody who is interested in comparing the use of verbs of perception in Australian English with their use in British English would be able to use the ICE-GB and ICE-Australia corpora for her research.

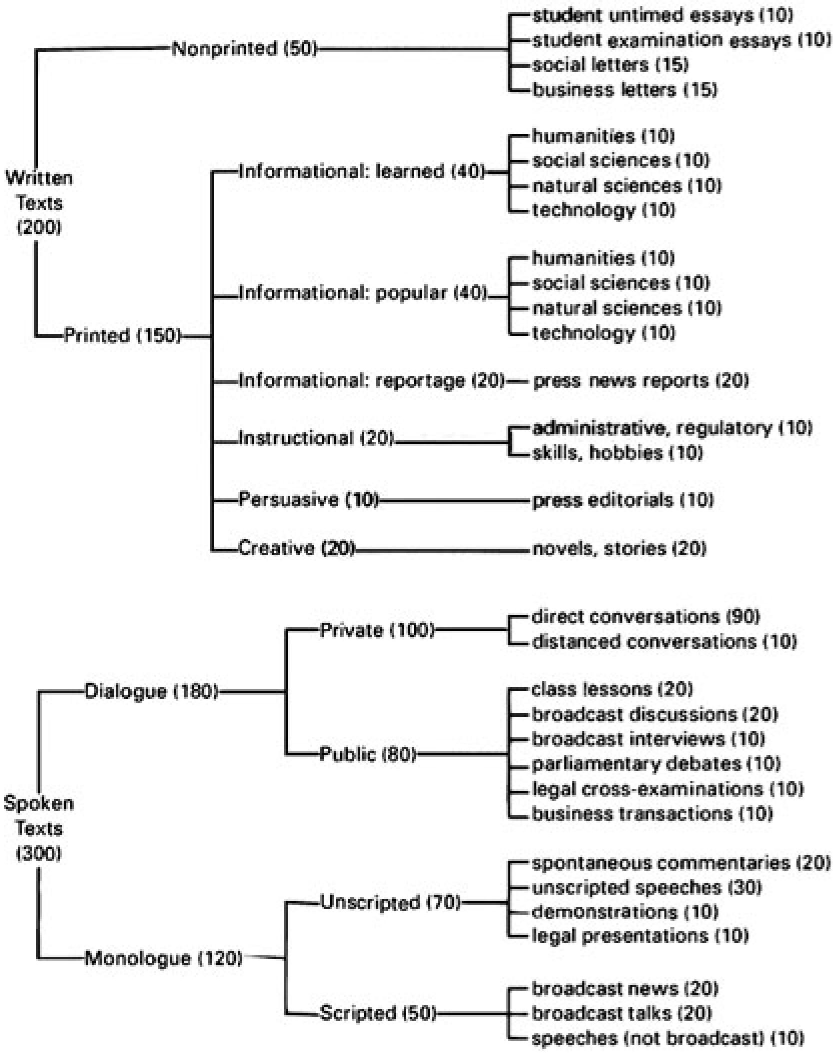
How are language corpora constructed?

One would have thought that corpus linguists would have concerned themselves from the early days with the issue of how to construct corpora, and with related issues such as statistical representativeness. Surprisingly, this is not so. Fairly fundamental papers dealing with this issue were published as recently as the early 1990s

(Atkins et al., 1992; Biber, 1993). The rationale for corpus construction developed autonomously from solving practical problems. Statistical sampling had little influence on the development of corpus linguistics; indeed, the merits of a statistical rationale for language corpora are contested. A motion that 'language corpora should be based on statistical representation' was voted down at a meeting of linguists in Oxford some years ago. Standard approaches to statistical sampling are hardly applicable to building a language corpus (Atkins et al., 1992: 4).

Issues of corpus construction that are discussed are as follows. Which categories of speech and writing should be included? How big should the samples be for each category of writing or speech, in terms of number of words? How large should the corpus be in terms of number of words? It is commonly accepted that corpus size is a less relevant consideration, while representativeness deserves more attention.

Figure 2.1 An outline of the ICE map



Corpus linguists recognize two important dimensions of representativeness in corpus design (Biber, 1993: 243). First, a corpus should include 'the range of linguistic distributions in a language' (1993: 243) - for example, a comprehensive range of grammatical constructions. What exactly constitutes 'the range of linguistic distributions' is something that is hard to determine *a priori*, but one could say that this locution refers to the sum total of the empirically established and diachronically accumulated knowledge that working grammarians have of a particular language - in other words, the material that most linguists would agree should be covered by a wide-ranging grammar of English, such as Quirk et al. (1985). This internal variation of language is called type or dialect variation.

Secondly, a corpus should include a sufficient range of text in the target population, where target population is taken to mean a bounded, that is rigidly defined, collection of textual material from different contexts. These variations are also called registers, genres or functions, and differ according to situational and thematic variables. This necessitates careful thought. The choice of target population depends on one's research aims: a linguist who is interested in language development will construct a corpus in a different way from someone who wants to study, say, dialectal variation (see Aston and Burnard, 1998: 23). The classification of registers or functions of speech that may bear internal variation is a matter of linguistic intimation and intuition: the question arises of how one decides whether or not the target population is sufficiently diverse. Atkins et al. (1992: 7) note that the range of text types to choose from is open-ended as well as culturally specific. For example, one could imagine that someone constructing a corpus representative of a society in which religion plays a pivotal part may want to include sermons, while in other corpora this category may be of much less interest. In the end the decisions as to which text types to include and which to exclude in a corpus are arbitrary.

Large, general-purpose corpora differ in the taxonomy of texts they include, and this variety reflects their different objectives. The Brown Corpus defined the target population for written material as all English texts printed and published in the USA in 1961. It included 15 text genres with subgenres. An example of a text genre would be ‘learned science’, and a subgenre of this could be ‘natural sciences’. Another example of a text genre could be ‘newspaper language’, with ‘sports commentary’ as a possible subgenre. Samples were chosen from a list of all the items in the Brown University Library and Providence Athenaeum that were published in 1961. The Survey of English Usage Corpus at University College London, which dates from roughly the same period as the Brown Corpus, had educated spoken and written adult English as its target population (see Figure 2.1).

Also, as far as corpora that are aimed to represent a particular language as a whole are concerned, it should be clear that for linguistic research a corpus constructed proportionally, that is following a random sampling rationale on the basis of all language use, would not be suitable. Such a corpus would consist predominantly of spoken language, because an estimated 90 per cent of all language produced is conversation (Biber, 1993: 247). Rather, linguists require a range of samples of language use that are sufficiently diverse, and contain the full range of grammatical structures. So, in addition to samples of conversation, there should be samples of material that is not produced in great quantities, such as highly technical scientific language (see Figure 2.1). Linguistic corpus construction is highly overselective of certain functions of speech and genres of text, because of their significance in revealing specific type variety. Linguists consider the rare event, while representative sampling would suggest ignoring it.

The corpus-theoretical paradox

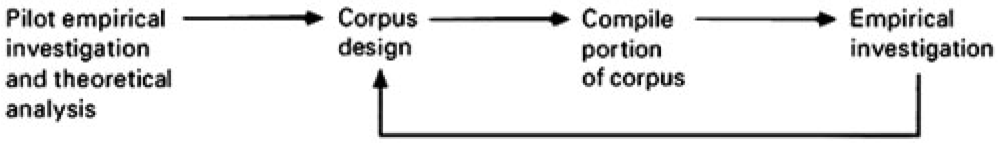
In corpus design, the genres and functions of text and speech are set up on what seem to be intuitive grounds. Josef Schmied, a German corpus linguist, has called this a ‘corpus-theoretical paradox’:

> On the one hand, a corpus is more representative of language use in a community if its subdivisions reflect all the variables that determine language variation in that community; on the other hand, we need results from a representative corpus in order to determine these variables empirically. (1996: 192)

In order to remedy such problems, corpus design is regarded by Biber as a cyclical process (see Figure 2.2), as one cannot determine *a priori* what a representative corpus looks like. In other words, the construction of successive corpora with a particular focus should lead to something like an industry standard for the ‘balanced corpus’. For Atkins et al. (1992), a balanced corpus is finely tuned on the basis of multiple user feedback, so that a manageable small-scale model of the linguistic material emerges. Balancing means that successive corrections are implemented to compensate the biases that are being identified. A cyclical process will give due recognition to two rules of corpus construction. Biber observes that external variation precedes awareness of internal variation, and so corpus construction has to start from different contexts (rule 1). According to Atkins et al., the aim is to maximize internal dialect variety through the extension of the functions, registers or genres that are being considered (rule 2). A corpus is balanced when additional efforts add little dialect variance. The problem is to determine those external variations that add significantly to internal type variety.

A future standard of corpus construction may include documentation of cyclical improvements, working towards a standard taxonomy of texts and speech situations, and conventions to mark the selected token texts and speech examples with standard codes. Transparency will not change the inevitable arbitrariness in the selection, but it brings it to light so that we might avoid false claims and suggest further improvements (Atkins et al., 1992).

Figure 2.2 Corpus design as a cyclical process (Biber, 1993: 256)


Pilot empirical investigation and theoretical analysis
Corpus design
Compile portion of corpus
Empirical investigation

Corpora in the social sciences

The question now arises of what we might learn from linguists in thinking about how to select data for qualitative research. ‘Corpus’ is not a technical term that is widely used in the methodology of the social sciences. With qualitative research gaining critical mass, the selection of interviews, textual and other materials requires a more systematic treatment comparable to that of survey research.

One may distinguish general-purpose corpora from topical corpora. A *general-purpose corpus* is designed with a broad range of research questions in mind, and serves as a resource in the widest sense. Most large-scale linguistic corpora constitute such projects. Judged by the effort involved, these corpora are resources comparable to the 10-year census or the annual labour force survey conducted in many countries.

Archival collections constitute general-purpose corpora for research. We may think of the many national libraries that contain complete collections of newspapers and magazines published in that country, as hard copy and/ or on microfiche. The British Newspaper Library in London stores every daily and weekly newspaper printed in the British Isles since the early nineteenth century. In recent years, online services have emerged that provide complete collections of newsprint on a daily basis, such as FT-Profile or Reuters, or with a regular update on CD-ROM directly from the newspaper publishers. Many of these sources are nearly complete and are listed, and so lend themselves to representative, even strict random, sampling. Classical content analysis profits from these developments.

A *topical corpus* is designed for a narrowly defined research purpose; it may become a general research resource for secondary analysis. Most textor interview-based social science research is of this kind. An example of a topical corpus is Ulm Textbank (Mergenthaler and Kaechele, 1988). The collection includes verbatim transcripts of over 8000 sessions of psychotherapy, from over 1000 patients and around 70 therapists, from Germany, Austria, Sweden, Switzerland and the USA. It was conceived as a tool for psychotherapy research, for studying the dynamics of interaction and experience. While the largest part of the material is psychoanalytically oriented, not all of the recordings are. Psychotherapy is a particular form of human interaction that happens worldwide, and in this corpus representativeness is not the principle of data selection: such a rationale would have to consider urban world centres of psychotherapeutic activities such as New York, Zurich, Vienna and Buenos Aires as locations of sampling. By contrast, the criteria that guided the selection are therapeutic orientation (register 1), diagnosis of the patient (register 2), the success of the treatment (register 3) and a minimum length of 300 to 500 hours (register 4). The selection aims at balancing different registers to enable comparative research. It does not aim to be representative, either by the real-life distribution of success and failure, or across the 600 different therapy schools, but aims to have sufficient examples across 34 text types relating to therapeutic interactions. The focus of analysis is the verbal activity, the expressions of various forms of emotionality during the course of therapy. The purpose is to relate particular initial diagnosis and subsequent patterns of verbal dynamics to the outcomes of therapy. The corpus is designed to maximize internal variety of verbal dynamics during sessions across the external registers of the therapist's orientation, diagnosis, therapy outcome and length of treatment (Mergenthaler, 1996).

How to construct a corpus in the social sciences

Linguists and qualitative researchers face the corpus-theoretical paradox. They set out to study the varieties in the themes, opinions, attitudes, stereotypes, worldviews, behaviours and practices of social life. However, as these varieties are as yet unknown, and therefore also their distribution, the researchers cannot sample according to a representativeness rationale. But paradoxes often resolve when we consider time. Linguists suggest a stepwise procedure:

(a) to select preliminarily

(b) to analyse this variety

(c) to extend the corpus of data until no additional variety can be detected.

In other words, they conceive the corpus as a system that grows. This is the first lesson for qualitative selection:

Rule 1 Proceed stepwise: select; analyse; select again.

Relevance, homogeneity, synchronicity

Barthes's (1967: 95ff) suggestions for corpus design may be helpful for qualitative selection: relevance, homogeneity, synchronicity. First, materials must be theoretically relevant, and should be collected from *one point of view* only. Materials in a corpus have only one thematic focus, only one particular theme. For example, a study of science and technology news requires a corpus of news items that refer to science and technology and that excludes all other news items. It is a different problem from determining the proportion of science news among all news: this would require a representative sample of all news. Trivial as this criterion may be, it serves as a reminder to be focused and selective.

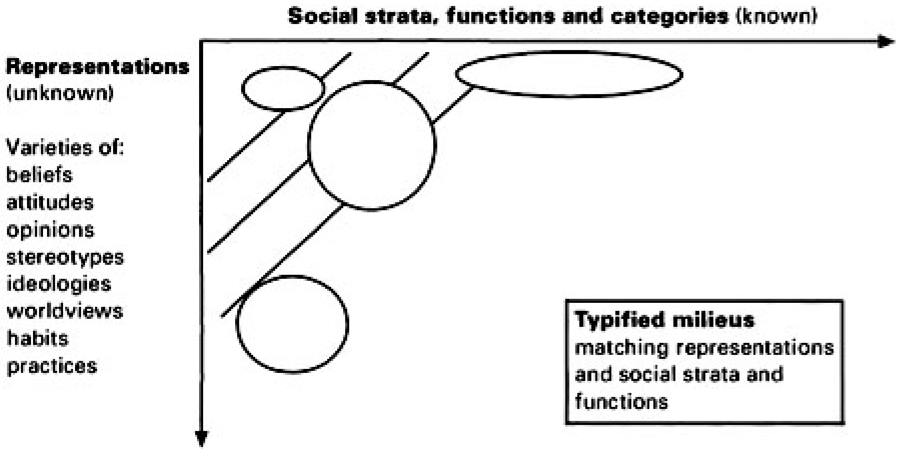
Secondly, materials in a corpus must be *as homogeneous as possible*. This refers to the material substance of data. Textual materials should not be mixed with images, nor should communications media be mixed; transcripts of individual interviews should not be grouped with transcripts of focus group interviews. Images, texts and individual and focus group interviews may address parts of the same research project; however, they must be separated into different corpora for comparison.

Thirdly, a corpus is a cross-section of history. Most materials have a natural cycle of stability and change. The materials to be studied should be chosen from within one natural cycle: they must be *synchronous*. The normal cycle of change will define the time interval within which a corpus of relevant and homogeneous materials should be selected. For example, family patterns are likely to be stable over one or two generations; clothing fashions change on a yearly cycle; newspaper and television editorial policy may have a cycle of a few years; opinion has a short cycle of days or weeks. For corpus construction, several materials within one cycle are preferable to one type of material over several cycles of change. Changes across cycles are studied by comparing two corpora, not within one single corpus.

Saturation

A procedure to overcome the corpus-theoretical paradox is depicted in Figure 2.3. The social space is unfolded in two dimensions: strata or functions, and representations. The horizontal dimension comprises the *social strata, functions* and *categories* that are known and are almost part of common sense: sex, age, occupational activity, urban/rural, income level, religion and so on. These are the variables by which social researchers usually segment the population; they are external to the actual phenomenon in question. Qualitative researchers' main interest is in typifying the varieties of representations of people in their life world. The ways in which people relate to objects in their life world, their subject-object relation, is observed with concepts such as opinions, attitudes, feelings, accounts, stereotypes, beliefs, identities, ideologies, discourse, worldviews, habits and practices. This is the second or vertical dimension of our scheme. This variety is unknown, and worth investigating. Representations are particular subject-object relations tied to a social milieu. The qualitative researcher wants to understand different social milieus in social space by typifying social strata and functions, or combinations thereof, together with particular representations. Social milieus occupy a social space and may have a project of common interest and investment that underpins their particular representations. External and internal variety, strata and representations may correlate, but need not. There are old social milieus and new ones emerging in a dynamic society. It requires sociological imagination and historical intimation to recognize new social milieus and to identify the traditional milieus that make a difference for the representation of a new issue in society (Bauer and Gaskell, 1999).

Figure 2.3 The two dimensions of social space: strata and representations


Social strata, functions and categories (known)
Representations (unknown)
Varieties of:
beliefs
attitudes
opinions
stereotypes
ideologies
worldviews
habits
practices
Typified milieus
matching representations and social strata and functions

To select interviewees or documents for qualitative research, we choose people and sources according to external criteria: social strata, functions and categories. For example we may invite interviewees for a focus group study on moral issues of human cloning by sex, age and education. However, the focus of research is not the difference between the sexes or the age groups, but the variety of moral issues and their argumentative structure. In other words, qualitative research tends to maximize the variety of the unknown phenomenon, in this case the moral issues of cloning. This is different from sample survey research: there, the opinions and attitudes are *a priori* framed in the questions and compared across known strata of people. For example, research will report the differences in opinions according to levels of education, sex or age. Following these considerations we formulate three further rules:

Rule 2 In qualitative research, strata and function variety precedes variety of representations.

> *Rule 3* Characterizing variety of representations has priority over anchoring them in existing categories of people.
>
> *Rule 4* Maximize the variety of representations by extending the range of strata/functions considered.

An implication of these rules may be that certain strata are deliberately overselected, so that a particular group that offers complex views may be given an overproportional attention in the research. If, for example, in focus group discussions on human cloning, women show much greater concern and diversity of views, one would not hesitate to explore different strata and functions among women only - for example, with or without children, religious background, etc. One would ignore the fact that the corpus contained more women-talk than men-talk. However, to avoid false conclusions, any judgement about the distribution of opinions should be avoided. Only representative sampling of opinions allows us to describe conclusively the distribution of opinions. In this sense corpus construction helps typifying unknown representations, while by contrast representative sampling describes the distribution of already known representations in society. Both rationales need to be distinguished with care in order to avoid confusion and false conclusions.

In order to overcome the initial paradox of corpus construction, research starts with the external strata and functions (rule 2). In focus group research, one may consider age groups or educational strata following an initial hunch about what would make a difference for representations of an issue. However, researchers are well advised not to rely on their hunches alone when they segment social space. They need to maintain an open mind for further strata and functional distinctions that may not be obvious in the first instance. They may start with sex, age and education, but may have to consider ethnicity, religion and urban/rural divisions in order to identify and maximize variety in people's representations of an issue. Here the law of diminishing returns may apply: adding further strata may make only a small difference with regard to additional representations. When this occurs, the corpus is saturated. Rule 1 stipulates that selection for qualitative research is a cyclical process, and a cyclical process requires a stopping criterion, otherwise a research project has no end. *Saturation* is the stopping criterion: one searches for different representations only until the inclusion of new strata no longer adds anything new. We assume that representational variety is limited in time and social space. To detect additional variety adds disproportionately to the costs of research; thus the researcher decides to stop studying additional strata. The dangers of this criterion are local maxima: it may be that talking to another regular in the local pub does not add any new facet of opinion; however, going to a very different neighbourhood or going out of town might well do so. Researchers live in a life world; and they have to ask themselves whether the variety they have detected covers their locale or wider ground.

Size of corpus

Little can be said about the size of corpora for qualitative research. One needs to consider the effort involved in data collection and analysis, the number of representations one would like to characterize, and some minimal and maximal requirements, for example in automatic text analysis, as criteria for the size of a corpus.

Most limitations come from the effort that is required to run a large number of focus groups or in-depth interviews, or to collect documents. The time available to conduct these interviews, and to analyse them, will be the first constraint on corpus size. Qualitative research involving large amounts of material has been rightly identified as an 'attractive nuisance' (Miles, 1979). Researchers easily collect more interesting material than they can effectively handle within the time of a project. This leads to the usual complaint that the project ends without the materials having been analysed in any depth. This results in the creation of 'data dungeons': materials collected but never really analysed. A considered assessment of the time required for selection and analysis procedures will increase the realism of many researchers.

The more representations that the researcher expects on a particular issue, the more different strata and functions of people or materials need to be explored, and the larger the corpus. The researcher will have to decide to study one or many representations in detail. Equally, if automatic textual analysis is considered, including the application of statistical procedures, this may require a minimum number of words in a corpus to reach reliable results. For example, ALCESTE (see Kronberger and Wagner, Chapter 17 in this volume) will require a text with a minimum of 10,000 words. Such procedures may also put an upper limit on the corpus size, beyond which the procedures either do not work or take a very long time to run.

Towards basic standards of corpus construction and reportage

As in corpus linguistics, one may renounce any hopes for a fully representative general-purpose corpus on a topic. A multitude of topical corpora may emerge from a flourishing practice of qualitative research. The problem arises of how to make these materials comparable and accessible for secondary analysis. A way forward in this direction is the development of guidelines for corpus construction and reporting. Survey research has developed elaborate standards of reporting representative sampling procedures, and analogous standards may be necessary for qualitative research. These may include:

- a description of the substance of materials involved: text, images, sounds, etc.
- a characterization of the research topic, e.g. moral issues of human cloning
- a report on the modalities of the stepwise extension of the open corpus
- the social strata, functions and categories that were used at entry
- the social strata, functions and categories that were added later
- evidence for saturation
- the timing of the cycles of data collection
- the place of data collection.

Indeed, the ESRC Data Archive at Essex University (Heaton, 1998; or ESRC at http://www.essex.ac.uk/qualidat/) is already undertaking to build an archive for qualitative research for which standards of reportage are required, and which protects the privacy of the informants - an issue that is very sensitive in qualitative research.

STEPS IN CONSTRUCTING A CORPUS

1 Decide on the topic area, and consider the four rules of corpus construction:

> Rule 1 Proceed stepwise: select; analyse; select again.
>
> Rule 2 In qualitative research, strata and function variety precedes variety of representations.
>
> Rule 3 Characterizing variety of representations has priority over anchoring them in existing categories of people.
>
> Rule 4 Maximize the variety of representations by extending the range of strata/ functions considered.

2 Consider a two-dimensional social space: strata and functions; and representations of the topic. List as many social strata and functions as possible.

3 Explore representations of the topic, with one or two strata or functions to start with.

4 Decide on whether these strata are likely to exhaust the variety of representations, or whether additional strata or social functions need to be explored.

5 Extend the corpus accordingly. Check whether you have achieved a saturated variety. Which strata are left unconsidered?

6 Conduct the final analysis and revise the social space in the light of the findings, and report your findings; or follow a cyclical procedure by returning to step 4.

Martin W.Bauer

BasAarts

References

Aston, G. and **Burnard, L.(1998)** The BNC Handbook: Exploring the British National Corpus with Sara.Edinburgh: Edinburgh University Press.

Atkins, S. Clear, J. Ostler, N. *'Corpus design criteria'* 7(1992)1–16.

Barthes, R.(1967) Elements of Semiology.New York: Hill and Wang, The Noonday Press [translation from

French original, 1964].

Bauer, M.W. Gaskell, G. *'Towards a paradigm for research on social representations'* 29(2)(1999)163–86.

Biber, D. *'Representativeness in corpus design'* 8(4)(1993)243–57.

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

Chomsky, N.(1957) Syntactic Structures.The Hague: Mouton.

Coates, J.(1996) Women Talk.Oxford: Blackwell.

Dawes, R.(1997) 'Qualitative consistency masquerading as quantitative fit'. Paper presented at the 10th International Congress of Logic, Methodology and Philosophy of Science, Florence, Italy, August 1995.

Gigerenzer, G., **Swijtink, S.**, **Porter, T.**, **Daston, L.**, **Beatty, J.** and **Krueger, L.(1989)** The Empire of Chance: How Probability Changed Science and Everyday Life.Cambridge: Cambridge University Press.

Heaton, J.(1998) 'Secondary analysis of qualitative data', Social Research Update, issue 22 (autumn), Sociology at Surrey, University of Surrey.

Jahoda, M., **Deutsch, M.** and **Cook, S.W.(1951)** Research Methods in Social Relations, Vols 1 and 2.New York: Dryden.

Johansson, S. *'ICAME - Quo vadis? Reflections on the use of computer corpora in linguistics'* 28(1995)243–52.

Kish, L.(1965) Survey Sampling.New York: Wiley.

McArthur, T(1992) The Oxford Companion to the English Language.Oxford: Oxford University Press.

Mergenthaler, E.(1996) 'Computer-assisted content analysis', in Text Analysis and Computers.Mannheim: ZUMA Nachrichten Spezial. pp. 3–32.

Mergenthaler, E. and **Kaechele, H.(1988)** 'The Ulm Textbank management system: a tool for psychotherapy research', in Edited by: **H. Dahl**, **H. Kaechele** and **H. Thomae** (eds), Psychoanalytic Process Research Strategies.Berlin: Springer.

Miles, M.B. *'Qualitative data as an attractive nuisance: the problem of analysis'* 24(1979)590–601.

O'Muircheartaigh, C.(1997) 'Measurement error in surveys: a historical perspective', in Edited by: **L. Lynberg**, **RR Biemer**, **M. Collins**, **E. deLeeuw**, **C. Dippo**, **N. Scharz** and **D. Trewin** (eds), Survey Measurement and Process Quality.New York: Wiley. pp. 1–25.

Quirk, R, **Greenbaum, S.**, **Leech, G.** and **Svartvik, J.(1985)** A Comprehensive Grammar of the English Language.London: Longman.

Schmied, J.(1996) ‘Second-language corpora’, in Edited by: **S. Greenbaum** (ed.), Comparing English Worldwide: the International Corpus of English.Oxford: Clarendon. pp.18296.

Tannen, D.(1992a) You Just Don't Understand: Men and Women in Conversation.London: Virago.

Tannen, D.(1992b) That's Not What I Meant! How Conversational Style Makes or Breaks your Relationships with Others.London: Virago.

Tversky, A. Kahnemann, D. *‘Judgement under uncertainty: heuristics and biases’* 185(1974)1124–31.

http://dx.doi.org/10.4135/9781849209731.n2