5 Computer-mediated Communication and Linguistic Landscapes

Jannis Androutsopoulos

Summary

This chapter discusses data collection methods in two new areas of sociolinguistic research. With *computer-mediated communication* (CMC) we cover text-based interpersonal communication via digital media, including e-mail and texting, as well as social networking sites and discussion forums. *Linguistic landscapes* (LL) cover language usage in public space, particularly on commercial and official signs. The chapter discusses procedures of data collection in these areas in terms of three *tensions*: (i) between methodological traditions and new domains of language and discourse, (ii) between focusing on language and its techno-social environment, and (iii) between text vs. participant oriented approaches to data collection. It suggests that CMC and LL extend what counts as *sociolinguistic data* and offer test-beds for the problems and challenges that arise as sociolinguistic scholarship moves on to examine language use in new environments.

Research Methods in Sociolinguistics: A Practical Guide, First Edition.
Edited by Janet Holmes and Kirk Hazen.

Introduction

Computer-mediated communication (CMC) and linguistic landscape (LL) are two recent areas of sociolinguistic research. The first covers private and public communication via digital media such as e-mail, texting, social networking sites, and discussion forums. The second deals with language use on signs and other artifacts in public space. Although CMC and LL seem to have very little in common at first sight, they both differ from traditional sites of sociolinguistic inquiry in terms of the linguistic data involved.

In particular, both CMC and LL data consist of written language in close relationship to semiotic resources such as typography, image, and layout. Moreover, their ecological conditions challenge traditional linguistic units of analysis such as clause or turn. In CMC research, categories such as "message" or "post" must be taken into account when collecting and analyzing online data, and shop windows, billboard signs, and city walls form the context for written language in the linguistic landscape. In both areas, the social contexts of language production and reception are invisible or only partially retrievable from written language data itself. Information on participants in communicative encounters is limited at first sight, and sociodemographic categories may be of little use. Finally, CMC and LL offer access to overwhelming amounts of data. From a practical angle, these are problems for data collection and analysis, which can be addressed in terms of researcher decisions and methodological procedures. From a broader perspective, CMC and LL extend what counts as sociolinguistic data and offer test-beds for the problems and challenges that arise as sociolinguistic scholarship moves on to examine language and discourse in new environments.

Another similarity between the two areas, as will be suggested in this chapter, refers to the degree of researcher engagement. Data collection in CMC and LL research can be positioned on a continuum between a "purely textual" and a more "ethnographic" approach. On the one hand, it is perfectly possible to collect data without any contact with language users. Large amounts of digital language data can be collected automatically without ever visiting the web sites they originate from, and photographs of street signs can be shot in an unobtrusive manner. Other researchers may choose to elicit data in close contact and collaboration with language users, drawing on techniques from ethnography such as observation and interviews. In both areas, procedures of data collection range from minimal or no engagement on the part of the researcher to full-fledged familiarity with relevant language users and sites of discourse.

Although both CMC and LL research can draw on existing data pools such as photography web sites or annotated CMC corpora (see Beißwenger and Storrer, 2008), this chapter focuses on the collection of original data. Issues and procedures of data collection in each area are covered separately. Each section begins with a brief outline of the research area, followed by a discussion of general strategies of data collection. As it is practically impossible to neatly separate data collection from broader issues of methodology, parts of the discussion address conceptual, methodological, and analytic conditions that may affect data collection. I outline techniques and solutions of data collection in each area, including examples from

my own research on the Internet and in the city of Hamburg, Germany. The chapter concludes with a note on research ethics.

Data Collection in Computer-mediated Communication Research

Overview

Sociolinguistic research on CMC focuses on one or more of the following interests (see Thurlow and Mroczek, 2011):

- language variation and change, especially with regard to written language;
- constraints of digital media on language use and interpersonal interaction;
- language, identity, and interpersonal relations online;
- linguistic diversity, multilingualism, and code-switching;
- language, globalization, and mobility.

This research has drawn on variationist, interactional, and discourse traditions in the field, applying both quantitative and qualitative methods. Rather than straightforwardly transferring sociolinguistic methods to CMC data, researchers need to adapt familiar methods to the conditions of digital language use. For example, technological restrictions rule out conversational processes such as turn taking in Internet data, and the absence of sociodemographic information imposes limitations to variationist analysis.

It is important to keep in mind that Internet research evolves together with the rapid sociotechnological evolution of the Internet itself. In the last 25 years, digital media developed from a small set of text-only communication modes into a rich repertoire of multimodal and multimedia choices that are almost ubiquitous in the Western world (though issues of digital divide persist). Early linguistic scholarship dealt with CMC in the pre-Web era, which was largely restricted to interpersonal exchanges carried out on language-heavy modes such as mailing lists, newsgroups, and Internet Relay Chat. Current scholarship is situated in the era of the participatory Web, where anyone can draw on the rich infrastructure provided by blogs, social networking sites, media-sharing sites, and wikis to produce and consume digital content. These developments shape what is being perceived as typical "Internet language" and what counts as relevant online data.

Two distinctions that affect how we approach language online are whether we view CMC as "text" or "place," and whether data are collected "on screen" or by contact with participants. *Screen data* are both produced and collected online, whereas *user-based data* are produced through direct contact with Internet users – for example, by means of interviews or focus groups. CMC research might seem obviously limited to screen data at first sight, but researchers are increasingly interested in the social activities in which CMC is embedded and in people's own awareness and evaluations of their language practices. Research questions that focus on

linguistic variation rather than language practices may justify a restriction to screen data; nonetheless, it is common experience among CMC researchers that the analysis of digital language can benefit considerably from insights into social and situational contexts of the data at hand. Screen and user-based data are therefore best seen as complementary sites of data collection in new media sociolinguistics.

The second distinction comes from qualitative online research in communication studies. Milner argues that "the study of cultures online demands we decide whether we frame online interaction as 'place' or as 'text'" (2011: 14). A "CMC as text" view approaches the Internet as a vast archive of written language, whereas from a "CMC as place" perspective, digital communication is a social process that unfolds in discursively created spaces of human interaction, which are dynamically related to offline activities. From a sociolinguistic angle, this binary echoes the familiar tension between *system-oriented* approaches that focus on linguistic variation and *speaker-oriented* ones that focus on interactional language practices (see Hazen, this volume). The example of Twitter can be used to illustrate how this distinction fits language analysis. Approaching "Twitter as text" would imply collecting a large set of data and focusing on specific linguistic features or categories, taking social variables such as, say, "private user" as opposed to "organization" into account. By contrast, a "Twitter as place" approach would examine how particular social actors use this medium in order to engage in social activities in the context of a particular event (say, a political rally), thereby shaping the course and social meaning of that event.

The "text vs. place" distinction identifies an epistemological perspective, which in turn is likely to entail a preference for particular research questions, techniques of data collection, and types of (quantitative or qualitative) analysis. A "CMC as text" approach may imply a tendency toward screen-based data, a view of digital modes as containers of written language, and a preference for *etic* (researcher-oriented) rather than *emic* (participant-oriented) categories. A "CMC as place" approach is more likely to prefer ethnographic observation and blended data collection, in which online language data from various modes and environments is collected, taking into account the digital literacy practices in which they originate.

Online observation

Online observation refers to the process of "virtually being there," watching the digital communication you will eventually analyze as it unfolds on a web site or in a network of connections across sites. Though often not explicitly acknowledged in research publications, observation is the bottom line of any "virtual fieldwork" and the ground pillar of most linguistic CMC research. In my own experience, systematic online observation is particularly useful in public digital spaces, such as discussion forums or virtual worlds, where participants' shared background knowledge is incomplete and fragmented anyway. Here, systematic observation is the key to gaining initial insights into participants' language practices, such as their common discussion topics, usual pace of discursive activities, categories of participation (e.g., core and peripheral members), distribution of particular linguistic features among members, and so on. As in any ethnographic fieldwork, systematic observation allows researchers to acquire some of the tacit knowledge that underlies the semiotic practices of regular members. This knowledge can be used to interpret patterns of

usage, to identify new objects of analysis, or to articulate new research questions (Androutsopoulos, 2008; Garcia *et al.*, 2009).

Three practices of online observation can be distinguished: *revisit*, *roam around*, and *try out*. The first suggestion is to make regular and iterative visits to the target site of data selection, documenting routine activities as well as changes. A second suggestion is to explore the virtual ground, browsing around web sites and their sections, threads, or profiles. Whether to lurk or actively participate is open to debate (see Markham, 2008; Milner, 2011). What is important, in my view, is that researchers do not end up analyzing their own data or data that occurred as a direct outcome of their own contributions. A third suggestion is to explore all available resources of participation by trying out all options afforded by an online environment, such as search facilities, user lists, statistics, tags, and tag-related hit lists. For ethnographic field notes and collections of digital resources (e.g., web links and screenshots), software tools like Zotero or Evernote can be used.

Screen-data collection

Screen data refers to digital written language produced by people online. The practicality of collecting screen data depends on the options provided by various modes and environments as much as on the technological sophistication brought along by researchers. There are some common simple solutions:

- Applications for synchronous chatting and messaging often save conversations automatically in a time-stamped logfile.
- Web forum pages can simply be copy-and-pasted or downloaded in HTML format, but then have to be 'cleaned up' from HTML code in order to be fed into a concordancer or other software program for further treatment.
- Content from social networking sites can be saved in HTML format, as a PDF file, or as a screenshot, the latter being the least preferred option because it doesn't allow exporting the language data.

At a more sophisticated and technical level, large portions of screen data can be mined by means of web crawlers, application program interfaces (APIs), customized scripts, or other resources (see Hundt, Nesselhauf, and Biewer, 2007). Digital data can also be delivered to researchers by users themselves (e.g., students or members of the general public who donate private digital data together with relevant sociodemographic information).

Depending on the research question, the selection of screen data may proceed on various sampling criteria. Susan Herring's framework for computer-mediated discourse analysis distinguishes six criteria for data sampling (Herring, 2004: 351–354):

1 *Random sampling*, by which each unit from a set of data has equal chances of being selected, enables representativeness and generalizability. Researchers can select items at specified intervals (e.g., every tenth message from a newsgroup) or use a "randomizer" tool to select items from a numbered list of posts. Random sampling may result in loss of context and coherence, for example by truncating conversations.

2 *Sampling by theme* is useful for selecting data from discussion forums or other thematically organized streams of online discourse (e.g., hashtagged tweets). Thematic samples from two or more sources can be compared in terms of language style or language choice. This criterion excludes other co-occurring discourse activities (e.g., other topics discussed by the same users) and is therefore less useful for the study of language style across various modes and genres.
3 *Sampling by time* is necessary for any kind of longitudinal analysis. A common procedure is to select samples at regular intervals from the archives of a newsgroup or forum. It offers data that are rich in context but may result in large samples and truncated interactions.
4 *Sampling by phenomenon* focuses on particular features or patterns of language use. Features such as emoticons or non-standard spellings can often be automatically selected by means of a concordance or customized script. Discourse-level phenomena such as joking or conflict negotiation (Herring's 2004 examples) involve qualitative analysis and so must be identified manually. This is the method of choice for features that are rare in a sample. It enables in-depth analysis of the selected phenomenon, but it may rule out a systematic control of independent variables and result in loss of context.
5 *Sampling by individual or group* can draw on sociodemographic information, if available, or explore member categories in the relevant online environment, such as forum member rankings. It enables focused analysis of selected users and user networks, but it may exclude the study of interactional exchanges.
6 *Sampling by convenience* means selecting "whatever data are available" (Herring, 2004: 351). This was popular with some early CMC research, but it obviously lacks a principle of systematic selection and may yield unsuited samples.

These criteria do not preempt the type of analysis to be carried out. Some (notably 2, 3, and 5) roughly correspond to familiar independent variables and yield data sets that will be later scanned for linguistic features of interest. In practice, combinations of two or more criteria are common.

Research with participants

Depending on the research question and the type of data, contacting Internet users can be either an initial or a later step of the research process. In research on private or semi-public data such as e-mails, text messages, or social networking sites, contacting people and obtaining their permission to use the data is a precondition to further analysis. In research on publicly accessible language data (e.g., unrestricted web forums or blogs) where such permission is not legally required, contact with participants can be initiated at a later point after a period of online observation, in which the researcher can identify core members or users who "stand out" in some way in their online community. In the next step, screen data can be collected and preliminary analyses can be carried out, preparing the ground for contacting selected participants. Such contact will obviously follow criteria of feasibility and pay due attention to how relations of power and/or solidarity between researcher and participants are negotiated (Androutsopoulos, 2008).

Box 5.1 Discussing samples of online writing with participants

In research on the language practices of German hip-hop artists and fans on the Internet (Androutsopoulos, 2007, 2008), a productive technique for eliciting participants' awareness of language style online was to have them discuss excerpts from hip-hop web sites or discussion forums. Asking them to identify what they saw as "typical features of hip-hop writing" helped me understand the categories and distinctions that mattered to them in tailoring their language style. This approach can sometimes confirm the analyst's interpretations but can also offer new, unexpected insights. A case in point are stylized "hip-hop English" features, such as the spelling variant <z> for the noun plural marker <s>, which was very popular among German hip-hop fans at that time. Discussions with my informants revealed that their knowledge about this feature was variable, focused on aesthetics and social values rather than linguistic aspects, and overall more localized than I initially assumed. For example, a 15-year-old girl who used spellings such as *friendz* on her home page said that <z> "is what Wu Tang use," thereby alluding to a rap group, whereas a 19-year-old boy explained, "this is how my buddies write." Rather than linking <z> to the "global hip-hop nation," as I was expecting them to, these youngsters foregrounded quite specific sources of inspiration and digital literacy practices in their local community.

Research with CMC users can draw on interviews, group discussions, or questionnaires. Interviews in particular can be semi-structured or narrative, and conducted face to face or via Skype or e-mail. A useful prompt in order to elicit participants' awareness of and attitudes to language use online are samples of online content that are already analyzed by the researcher (see the example in Box 5.1). Participant observation of user activities can focus on their online practices at home or in Internet cafés, but it can also take the researcher far from the computer to people's offline activities, which they later entextualize online.

In linguistic CMC research, user-based data are typically not the single source of available data but a complement to screen data that is collected before or after contacting participants. Collecting blended data – that is, combined sets of online and offline data – is typically a cyclical process, oscillating between screen/online and users/offline contexts. An interview or other form of user contact follows up on screen-data analysis and can help to deepen and contextualize the analyst's interpretation of those data. In turn, insights gained in the interview can also trigger further screen-data collection.

In my own research I have experimented with various sequences of screen and participant data. In early research on multi-party Internet Relay Chat (IRC), a period of familiarization involving observation of and some active participation in the channel of choice was followed by contact with selected individuals by means of the one-to-one ("whisper") mode afforded by chat software; disclosing my researcher identity, I could then discuss language issues with these individual chatters or ask them to fill in a short questionnaire. In research on private home pages and discussion forums, the strategy was to observe these sites first, then contact and interview their producers or webmasters, then return to and refine screen-data analysis. In research on social networking sites, the first step is an initial contact with likely participants, by which permission to access their social network profiles is sought. This is then followed by a period of observing profile activities, in which digital language samples are collected and preliminary analyses carried out. This is followed by individual interviews or group discussions.

Inter- and intra-mode designs

In CMC research, modes of digital communication such as instant messaging, Internet Relay Chat, and e-mail often serve as invariant parameters for digital data selection. Much data reported in the literature is restricted to particular modes, for example IRC, Instant Messaging, or e-mail. Analysis of CMC data by mode ties in with the practice of dividing "Internet language" to mode-specific components, which are then discussed in separate textbook chapters, and so on. In sociolinguistic practice, modes have also played the role of independent variables, based on the assumption of more or less stable relations between modes and patterns of online language use. In such an *inter-mode* analysis, data from two or more CMC modes (e.g., messaging vs. e-mail or chatting vs. newsgroups) are compared in terms of one or more sociolinguistic variables while controlling for other social and situational factors. Here, the data collection design is primarily defined by user networks and subdivided by mode, as in the following examples:

- In research on CMC by university students, their instant messaging conversations (synchronous, among students) are compared to e-mails (asynchronous, addressed to lecturers; Lee, 2007).
- In research on Punjabi-background users, their language use on IRC (synchronous) is compared to a newsgroup (asynchronous; Paolillo, 2011).
- In research on German hip-hop on the Internet, various genres on a big hip-hop web site are compared: for example, amateur artist home pages (asynchronous, unidirectional) and forum discussions in the same online community (asynchronous, interactive; Androutsopoulos, 2007).

By contrast, an *intra-mode* design compares data from the same CMC mode and varying social and/or situational conditions, as in the following examples:

- A corpus of e-mails among university students can be compared to a corpus of e-mails exchanged between students and lecturers.
- A data set with informal (non-moderated) public chat sessions can be compared to a data set of institutional (moderated) chat sessions, for example with a politician.
- A study of spelling variation in instant messaging compares data sets that vary by interlocutors' gender (female–female, male–male, and female–male conversations; Squires, 2012).

Provided the primacy of mode effects on language over social and/or situational factors is not assumed by default, modes offer an invaluable handle for CMC data collection and exploration. However, their usefulness is weakened by the growing importance of participatory web environments, such as social networking sites and

content-sharing platforms, which integrate old modes and give rise to new genres which cannot be distinguished on the criteria of synchronicity and publicness alone. Due to their sheer size and diversity of participants, genres, and interactive applications, participatory online environments create new problems of comparability. Developing meaningful comparisons depends here on systematic online observation, by which relevant types of content, genres, or users within a web environment can be identified prior to screen-data collection.

Social identity variables

CMC complicates the process of social identity ascription for both researchers and participants. Digital communication, especially of the public type, is often carried out anonymously and among interlocutors who lack cues for mutual social categorization. This is a problem for any sociolinguistic analysis that depends on clear-cut sociodemographic information on gender, social class, and so on. It can be addressed or circumvented in a number of ways. First, researchers can contact relevant users and collect relevant sociodemographic information post hoc, though this is not always practically feasible, especially in public domain CMC. Second, researchers can work with the social identity cues offered by users themselves. Depending on mode and genre, these include propositional information and indexical cues such as screen names and associated "virtual identity" signs such as avatars and member signatures. The theoretical and analytical challenge here is how to handle the tension between online and offline identities and whether to conceive of users as "behaving like" or rather "performing" a particular social identity. Alternatively, researchers can abandon external sociodemographic categories altogether and turn to online-specific categories such as types and degrees of membership (regulars vs. newbies, admins vs. normal users) to which sociolinguistic variation is then correlated. Another alternative could be to focus on the discourse practices by which participants ascribe and negotiate social identities to selves and others, which however usually implies an interpretive approach and rules out a quantitative analysis of language variation.

In sum, the main challenges of data collection in new media sociolinguistics are the shift to written language data and the lack of information about language users. I argued that a degree of ethnographic engagement can help researchers gain contextual knowledge that might help with making data collection decisions, as well as with developing research questions and interpreting findings. In the next section we will see how research on linguistic landscapes follows a similar trajectory from an exclusive focus on public written language to increasing ethnographic engagement with the community.

Data Collection in Linguistic Landscapes Research

Overview

Linguistic landscapes (LL) is a recent area of sociolinguistics and interdisciplinary scholarship that focuses on how language constructs public space. Its main empirical object is language use on street signs. According to one oft-cited definition, "The language of

public road signs, advertising billboards, street names, place names, commercial shop signs, and public signs on government buildings combines to form the linguistic landscape of a given territory, region, or urban agglomeration" (Landry and Bourhis, 1997: 25). Building on earlier work on minority languages and multilingual urban environments, LL has now become the dominant paradigm in the study of visible language in urban settings (for state-of-the-art publications see Shohamy and Gorter, 2009; Jaworski and Thurlow, 2010; Shohamy, Ben-Rafael, and Barni, 2010; Shohamy, 2012).

We begin by reviewing theoretical and empirical developments in LL scholarship that have had an impact on data collection strategies. Early LL research focused on minority languages and coined a distinction between "communicative" and "symbolic" uses of minority languages in the linguistic landscape (Ben-Rafael *et al.*, 2006), by which the use of a minority language either indexes the spatial presence of its ethnolinguistic community (communicative use) or is intended as a symbol of that community (symbolic use). Later research suggested that the relation between linguistic landscape and minority communities is more complex, depending among other things on strategic entrepreneurial decisions and power relationships among majority and minority groups. The empirical scope of contemporary LL research encompasses all linguistic resources in the landscape, notably including globalized uses of English.

Moreover, LL research is going beyond its early exclusive concern on linguistic signs to include their visual, material, and spatial properties. Questions about how signs are designed, how they coexist in urban space, and how various semiotic resources contextualize language choices are now part of the LL research agenda. This shift foregrounds issues of materiality (i.e., how the material a sign is made of indexes types of institutional authority) and granularity (i.e., how design encodes different viewing distances, which correspond to different types of recipients) (Auer, 2010). Finally, the relation between textual data and ethnographic research is changing too. Early LL research was restricted to photographic documentation and content analysis of street signs drawing on the coding categories discussed below (see the section "Coding categories"). Contacts with the people who design the linguistic landscape and encounter it in their daily lives were limited. However, ethnographic research revealed that shop owners are not always aware of the semiotic choices of their own shop signs (Malinowski, 2009). Involving participants is now increasingly seen as necessary in order to understand the relation between semiotic choices on signs and their social context (Shohamy, Ben-Rafael, and Barni, 2010).

Data collection phases in LL research

LL research resembles CMC research in its "tension" between textual data and participant-driven research, but it differs in that all data collection depends on physical fieldwork. In his study of the LL in Tokyo, Backhaus (2007) followed a sequence of three steps in fieldwork focused on textual data: determine the survey area, the items to be surveyed, and the coding categories. We discuss these below. In ethnographic fieldwork focusing on participants rather than the signs themselves, Garvin (2010) suggests the following data collection stages:

- selection of sites;
- photographic documentation;

- selection of and contact with participants;
- conducting individual "walking tour" interviews on the selected site;
- transcription and analysis of interviews and field notes;
- follow-up meetings with participants in order to validate findings and offer opportunities to continue the dialogue with the researcher.

While it is possible to do LL fieldwork on a single site, such as a public monument, LL data collection is typically carried out in a vast urban environment that cannot be surveyed exhaustively. LL research therefore begins by determining a survey area together with the institutional domains and types of sign to be covered. The survey area is often a district or neighborhood, specified down to a set of street blocks or a trajectory in urban space, which can be determined by a set of orientation markers such as subway stations. Comparative designs are common, by which similarities and differences in LL patterns within a city or across different cities are explored.

Linguistic landscape fieldwork in Hamburg

Figures 5.1 and 5.2 are data collected during LL fieldwork in Hamburg, Germany. The fieldwork design included a comparison of shopping streets in various districts: trendy inner-city neighborhoods, working-class immigrant areas, and affluent suburbs. A research hypothesis was that the frequency of various languages, notably German, English, and various migrant languages, would vary across the LL of these areas. Figures 5.1 and 5.2 were taken in St Georg, an inner-city multi-ethnic neighborhood. Both types of shops shown here – that is, an ethnic supermarket and a "cheap phone calls" shop – are common in this area (see Scarvaglieri *et al.*, in press).

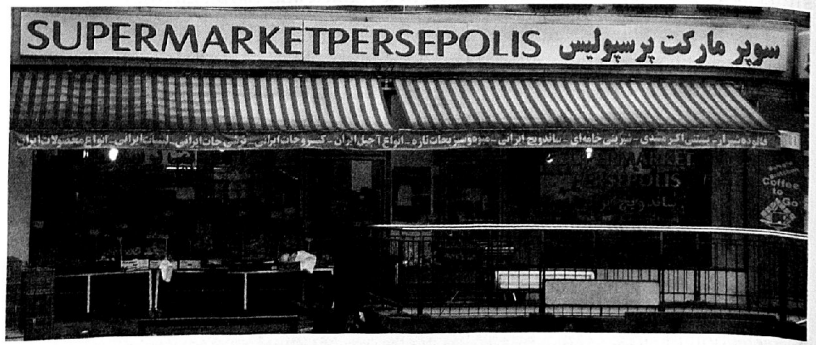

Figure 5.1 Front of "Persepolis Supermarket" in St Georg, Hamburg.

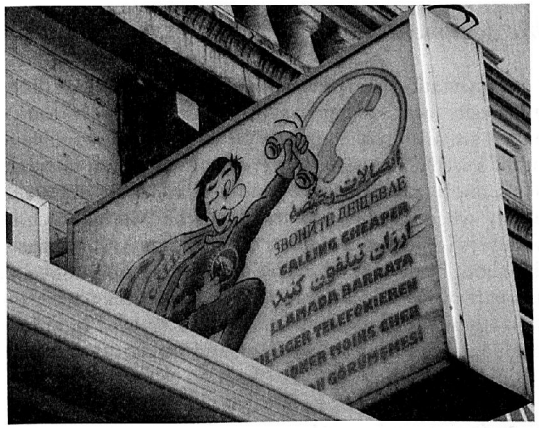

Figure 5.2 Multilingual call shop sign in St Georg, Hamburg.

Decisions on the institutional domains to be surveyed involve the established distinction between "top-down" signs (those issued by public authorities) and "bottom-up" signs (those produced by commercial businesses). To these, the domain of transgressive signs, notably graffiti, is sometimes added. Regarding commercial signs, decisions with an aim to narrow down the sample can be linked to research assumptions about the degree of linguistic innovation, semiotic creativity, or ethnocultural stereotyping that can be expected in certain business sectors. In LL research on minority languages, we often find examples from gastronomy (restaurants, food shops), so-called "ethnic" shops, and telecommunications (cheap call shops, Internet cafés), which, for partially different reasons, are likely to draw on multilingualism and ethnocultural stereotypes on their signage.

Determining the unit of analysis for data collection involves a complex set of procedural decisions, including questions like the following:

- Is the unit of analysis the individual sign, the shop window, or a specific chunk of space on the street?
- What aspects of the materiality (physical shape) of signs shall be taken into account in analysis?
- Are multilayered shop windows documented in their entirety or do we just focus on the main signboard?
- Are mobile signs (e.g., those placed on the street for the day) to be included?

Decisions of this sort are closely related to the research questions and, at the same time, impact directly on the photographic documentation to be carried out. An example for comprehensive coverage is Backhaus (2007), who documented "anything from the small, handwritten sticker attached to a lamp-post to huge commercial billboards outside a department store" (2007: 66), including stickers at entrance doors and lettered foot mats.

Photographic documentation

Photographic documentation lies at the heart of LL data collection. Besides some basic hardware requirements such as a digital camera with sufficient resolution, an adequate documentation will strive for sequential completeness under favorable contextual conditions. Within the selected area to be surveyed, it is important to document complete sequences of signs, one by one. Adequate conditions for doing this are sometimes hard to meet, especially when research is carried out on a busy commercial street. In Hamburg, researchers went to these streets on early Sunday mornings so as to obtain the best possible shots, being as unobtrusive as possible. Archiving and displaying the collected data is part of the documentation process. Besides photo storage software, Google Maps or other web-based map services can be used in order to display the photos at their topographic location (see Barni and Bagna, 2009).

Coding categories

Being aware of the coding categories that will be applied to the collected data is useful in anticipating certain details of the photographic documentation. The three examples below illustrate the range of coding criteria that are employed in the research literature:

- Cenoz and Gorter (2006) focus their coding on linguistic aspects of signs. Main categories include: type of sign; branch; number of languages on the sign; and the distinction between top-down vs. bottom-up signs. Multilingual signs are additionally coded for the following variables: first language on the sign; amount of information per language; semantic relation between the two languages on the sign; and fonts used on the sign.
- Backhaus (2007) categorizes his items for the following criteria: monolingual vs. multilingual; languages on the sign; top-down vs. bottom-up; geographic distribution; and semantic relations between language elements on a sign.
- Barni and Bagna (2009) used five main criteria to enter their items into a database: mono- vs. multilingual signs; textual genre (e.g., advertisement, warning sign); location; domain (e.g., educational or work-related); and place (e.g., catering places, including kiosks and bars).

Collecting language policy documents

When LL is studied from the angle of language policy, access to policy documents is an important additional dimension of data collection. Relevant policy documents can relate to any institutional decision by which language use on public signs is regulated. Examples are legislation acts or public authority manuals that regulate top-down signs at an airport or a city's subway system. Some countries or regions also control by law the languages that may be used on commercial signs. Language policy documentation can also be an important resource for historical research on the linguistic landscape (Backhaus, 2007; Pavlenko, 2010).

Involving participants in LL research

LL research that involves participants draws especially on interviews, but telephone questionnaires and field notes of fieldwork observations are also used. Participant numbers are usually small, and the overall approach is qualitative. An example of how various methods can be combined is Malinowski's (2009) research in California, featuring interviews with local business owners, participant observation, photograph and media analysis, and interpretive walking/driving tours.

Participant research can focus on either producers or recipients, or both. Research with the people who commission and/or design signs can examine their motivations for the choice of particular languages and other semiotic resources, their own interpretations of shop signs, and the impact of factors such as business sector, district, or target customers. Interviews with shop owners can feature questions such as: Who makes these signs? Who decides on their language choice, naming patterns, design, material, and so forth? What is the division of labor between commissioners and designers? In designing the interviews, researchers can draw on their analysis of relevant photographic data, and participants can be asked to share their views on the analytic findings.

Research with local residents and/or passersby uses a range of techniques. In a study of the LL of San Sebastian, Basque Country, Aiestaran, Cenoz, and Gorter (2010) did short interviews with randomly selected local people. Their questions covered the respondents' backgrounds and their views on the city's linguistic landscape, including their observed frequency of the relevant languages (Basque and Spanish) and their preferences on the language that they thought ought to be used in public space. Other researchers use so-called "walking tour" interviews, where interviews are conducted while walking (or driving) through the selected area. In research on LL in Memphis, Tennessee, Garvin (2010) did walking tours with a small sample of local residents, thereby eliciting their "self-reported emotional understandings and visual perceptions" (2010: 258) of the LL around them. Questions included: "How do you feel when you see languages other than English?" and "Do you go into stores that advertise in languages other than English?" as well as "What do you think these languages say about the people in this area?" Here, too, it makes sense to have photographed the signs on these routes prior to the walking interview itself.

A Note on Research Ethics

Both CMC and LL research face ethical issues related to the tension between privacy and publicness. Respecting and protecting the privacy of informants is a basic legal and ethical requirement in social-scientific fieldwork, and our research must observe legal requirements of "privacy." At the same time, our considerations should not marginalize informants' own understandings of the boundaries between privacy and publicness.

There is no general consensus on how to protect individual privacy in CMC research, and the relevant ethics guidelines for researchers and students vary considerably by country and institution. It should be common sense among CMC researchers that

protecting the anonymity of our informants entails avoiding disclosure of their offline identities and the publishing of any clues that may lead to their identification. Various CMC modes and user groups pose different conditions for achieving this aim. Maintaining anonymity for private online data is easier than for public and semi-public data. Asking participants for permission to use private data is the rule, but it is not always feasible for data collected from or available on public sites of CMC. Moreover, the researcher's (technical) definition of what constitutes publicness may not be shared by participants themselves, resulting in diverging interpretations of what data can be treated as "public domain." Some scholars treat publicly posted screen names (e.g., on YouTube) as publishable. However, these can be easily traced back to other publicly available utterances posted under the same screen name. Even when screen names are anonymized, verbatim quotations from publicly accessible material may also lead back to original posts via web search. A complete anonymization of public CMC data may even be technically impossible. On the other hand, we have to consider that not all online communicators may wish to stay anonymous in academic publications; famous bloggers could be a case in point. This should not be understood as an excuse not to anonymize but, rather, it should act as a reminder that participant and researcher views do not forcibly coincide. (Readers are also referred to the ethics guidelines of the Association of Internet Researchers; latest review draft at http://aoirethics.ijire.net.)

Linguistic landscape is part of the public space, and its basic documentation technique – photographing shop signs on the street – should be legally unproblematic in most parts of the world. There are, however, limitations to this. Photographing certain kinds of top-down signage, such as military sites, is strictly forbidden in many countries. Likewise, photographing individuals without their permission may be against the law. Photographing on the streets, especially by zooming in on shop windows, can be felt as offensive by shop owners, particularly when the researcher is clearly not part of the local community. However, asking each and every shop owner for permission could be unrealistic under certain fieldwork conditions. Doing the documentation at an unobtrusive time of day is a practical solution to this. Overall, issues of ethics in LL research seem to depend on local legislation as much as on sensitivity to local concerns and habits in the community to be surveyed.

Project Ideas

In addition to the findings presented in this chapter, consider doing a small project to explore the following questions:

1 *Collect a small data sample* to compare an individual's CMC writing style in one synchronous and one asynchronous mode. Identify the linguistic variables that best reflect inter-mode differences in your sample, taking into account inter-mode differences in addressee and topic.
2 *Collect tweets* that comment on a specific media event, such as a television show or a sporting event, as it happens. You will need to know the particular hashtag (#) for that event and could use a collecting service such as TwapperKeeper (now to be found at HootSuite: http://hootsuite.com). Examine your data in terms of what stances they express to that event and how they reflect different phases of the event as it unfolds.

3 *Document and compare* the linguistic landscape of two main streets in different neighborhoods of your city or town. To keep this feasible, you may want to limit your documentation to a small number of street blocks and the main sign of each shop. Work out the linguistic repertoire and language ranking for each street, taking variation in branches into account, and draw on sociodemographic data, if available, to interpret your findings.

Further Reading and Resources

Herring, S.C. 2004. Computer-mediated discourse analysis: an approach to researching online communities. In *Designing for Virtual Communities in the Service of Learning*, ed. S.A. Barab, R. Kling, and J.H. Gray, 338–376. Cambridge and New York: Cambridge University Press.

Shohamy, E. 2012. Linguistic landscapes and multilingualism. In *The Routledge Handbook of Multilingualism*, ed. M. Martin-Jones, A. Blackledge, and A. Creese, 538–551. London: Routledge.

Shohamy, E. and Gorter, D. (eds) 2009. *Linguistic Landscape: Expanding the Scenery*. London: Routledge.

Thurlow, C. and Mroczek, K. (eds) 2011. *Digital Discourse: Language in the New Media*. Oxford: Oxford University Press.

References

Computer-mediated communication

Androutsopoulos, J. 2007. Style online: doing hip-hop on the German-speaking Web. In *Style and Social Identities*, ed. P. Auer, 279–317. Berlin: De Gruyter.

Androutsopoulos, J. 2008. Potentials and limitations of discourse-centered online ethnography. *Language@Internet*, 5. www.languageatinternet.org/articles/2008/ (last accessed March 27, 2013).

Beißwenger, M. and Storrer, A. 2008. Corpora of computer-mediated communication. In *Corpus Linguistics*, Vol. 1, ed. A. Lüdeling and M. Kytö, 292–309. Berlin: De Gruyter.

Garcia, A.C., Standlee A.I., Bechkoff, J., and Yan, C. 2009. Ethnographic approaches to the internet and computer-mediated communication. *Journal of Contemporary Ethnography* 38(1): 52–84.

Herring, S.C. 2004. Computer-mediated discourse analysis: an approach to researching online communities. In *Designing for Virtual Communities in the Service of Learning*, ed. S.A. Barab, R. Kling, and J.H. Gray, 338–376. Cambridge and New York: Cambridge University Press.

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

Lee, C.K.M. 2007. Linguistic features of email and ICQ instant messaging in Hong Kong. In *The Multilingual Internet*, ed. B. Danet and S.C. Herring, 184–208. New York and Oxford: Oxford University Press.

Markham, A.M. 2008. The methods, politics, and ethics of representation in online ethnography. In *Collecting and Interpreting Qualitative Materials*, ed. N.K. Denzin, 247–284. Los Angeles: SAGE.

Milner, R.M. 2011. The study of cultures online: some methodological and ethical tensions. *Graduate Journal of Social Science* 8(3): 14–35.

Paolillo, J.C. 2011. "Conversational" codeswitching on Usenet and Internet Relay Chat. *Language@Internet*, 8. www.languageatinternet.org/articles/2011/ (last accessed March 27, 2013).

Squires, L. 2012. Whos punctuating what? Sociolinguistic variation in instant messaging. In *Orthography as Social Action: Scripts, Spelling, Identity and Power*, ed. A. Jaffe, J. Androutsopoulos, M. Sebba, and S. Johnson, 289–323. Berlin: De Gruyter.

Thurlow, C. and Mroczek, K. (eds) 2011. *Digital Discourse: Language in the New Media*. Oxford: Oxford University Press.

Linguistic landscapes

Aiestaran, J., Cenoz, J., and Gorter, D. 2010. Multilingual cityscapes: preferences of inhabitants. In *Linguistic Landscape in the City*, ed. E. Shohamy, E. Ben-Rafael, and M. Barni, 221–236. Clevedon, UK: Multilingual Matters.

Auer, P. 2010. Sprachliche Landschaften. Die Strukturierung des öffentlichen Raums durch die geschriebene Sprache. In *Sprache intermedial: Stimme und Schrift, Bild und Ton*, ed. A. Deppermann and A. Linke, 271–300. Berlin: De Gruyter.

Backhaus, P. 2007. *Linguistic Landscapes: A Comparative Study of Urban Multilingualism in Tokyo*. Clevedon, UK: Multilingual Matters.

Barni, M. and Bagna, C. 2009. A mapping technique and the linguistic landscape. In *Linguistic Landscape: Expanding the Scenery*, ed. E. Shohamy and D. Gorter, 126–140. New York: Routledge.

Ben-Rafael, E., Shohamy, E., Amara, M.H., and Trumper-Hecht, N. 2006. Linguistic landscape as symbolic construction of the public space: the case of Israel. *International Journal of Multilingualism* 3(1): 7–30.

Cenoz, J. and Gorter, D. 2006. Linguistic landscape and minority languages. *International Journal of Multilingualism* 3(1): 67–80.

Garvin, R. 2010. Postmodern walking tour. In *Linguistic Landscape in the City*, ed. E. Shohamy, E. Ben-Rafael, and M. Barni, 254–276. Clevedon, UK: Multilingual Matters.

Jaworski, A. and Thurlow, C. (eds) 2010. *Semiotic Landscapes: Language, Space, Image*. London: Continuum.

Landry, R. and Bourhis, R.Y. 1997. Linguistic landscape and ethnolinguistic vitality: an empirical study. *Journal of Language and Social Psychology* 16: 23–49.

Malinowski, D. 2009. Authorship in the linguistic landscape: a multimodal, performative view. In *Linguistic Landscape: Expanding the Scenery*, ed. E. Shohamy and D. Gorter, 107–125. New York: Routledge.

Pavlenko, A. 2010. Linguistic landscape of Kyiv, Ukraine: a diachronic study. In *Linguistic Landscape in the City*, ed. E. Shohamy, E. Ben-Rafael, and M. Barni, 133–154. Clevedon, UK: Multilingual Matters.

Scarvaglieri, C., Redder, A., Pappenhagen, R., and Brehmer, B. In press. Capturing diversity: linguistic land- and soundscaping. In *Linguistic Super-diversity in Urban Areas – Research Approaches*, ed. I. Gogolin and J. Duarte. Amsterdam and Philadelphia: Benjamins.

Shohamy, E. 2012. Linguistic landscapes and multilingualism. In *The Routledge Handbook of Multilingualism*, ed. M. Martin-Jones, A. Blackledge, and A. Creese, 538–551. London: Routledge.

Shohamy, E., Ben-Rafael, E., and Barni, M. (eds) 2010. *Linguistic Landscape in the City*. Clevedon, UK: Multilingual Matters.

Shohamy, E. and Gorter, D. (eds) 2009. *Linguistic Landscape: Expanding the Scenery*. London: Routledge.

Part II Methods of Analysis