Chapter for *Data Collection in Sociolinguistics: Methods and Applications*, eds. C. Mallinson, B. Childs & G.V. Herk (forthc. Routledge). First draft, 15-5-2012. Please do not cite or distribute.

Online Data Collection

(Jannis Androutsopoulos, University of Hamburg)

1. Introduction

In the last twenty years or so, research on computer-mediated communication (CMC) in linguistics has examined language online from a variety of aspects. Specifically sociolinguistic issues include variation and style in digital written language, processes of innovation and change, language and social identities, multilingualism and code-switching, and the relation of language, digital media, and globalization (*see further reading*). This and other research on CMC evolves in constant interaction with the socio-technological evolution of the Internet, which I find useful to divide in three broad stages: In the *pre-Web era*, until the early 1990s, CMC is largely restricted to interpersonal (dyadic or group level) exchanges carried out on applications (or modes) such as email, mailing lists, newsgroups, and Internet Relay Chat. In the *early Web era*, from the mid-1990s to mid-2000s, the emergence of the World Wide Web introduces personal homepages, web discussion forums and corporate websites, followed by blogs. In the *participatory Web era*, from the mid-2000s onwards, people draw on the infrastructure provided by blogs, social networking sites, media-sharing sites and wikis in order to both produce and consume Web content. In the course of this development, digital media evolved from socially exclusive to almost ubiquitous in the Western world, and from a small set of options for interactive written communication to a rich repertoire of multimodal and multimedia choices. The various modes of digital communication introduced in these three 'eras' accumulate in implicational ways, with each era adding on to the options offered by the previous one. These developments shape what is being viewed as typical 'Internet language', what is perceived as 'research worthy', and what counts as relevant online data.

Based on an inclusive view of sociolinguistics that encompasses variationist, interactional and discourse-oriented approaches to language in society, this chapter summarises a range of issues related to online data collection. While it is increasingly possible to draw on compiled and annotated CMC corpora (Beisswenger & Storrer 2008), this chapter focuses on issues related to the individual collection of original data. 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.

The chapter begins with a discussion of how CMC challenges methodological assumptions in sociolinguistics (section 2), and an outline of data sampling criteria in the framework of Computer-Mediated Discourse Analysis (section 3). The next two sections introduce two distinctions that impact on how we approach language online: viewing CMC as

'text' or 'place' (Section 4), and collecting data 'on screen' or through contact to users (Section 5). The following sections discuss issues related to the modes and environments being sampled (Section 6), multimodality (Section 7), social identities and participation roles (Section 8), units and sequences of online data (Section 9). We conclude with a note on research ethics (Section 10).

2. Traditions and challenges in online data collection

Language-focused CMC research faces the challenge of adapting traditions of scholarship to the technological, social and pragmatic conditions of digital communication. Familiar methods cannot be just replicated in new contexts. This is fairly well understood with respect to specific frameworks. For example, the absence of directly accessible socio-demographic information on language users and the lack of spoken-language data impose limitations to variation analysis, which call for creative solutions (Herring 2001; Androutsopoulos 2006; Paolillo 2001, Squires 2012). Likewise, the transfer of conversation-analytic categories to online data is limited due to the technological restrictions of synchronous CMC, which cancel out the familiar turn-taking system. Researchers may examine how users themselves respond to these restrictions (an empirical problem), but also have to adapt their own use of analytic categories (a methodology problem). Regardless of framework, general issues regarding the collection of online data for sociolinguistic purposes include the following:

a) The online data of interest to linguists is overwhelmingly written language data. CMC research is therefore confronted with the hitherto marginal status of written language in sociolinguistics, and at the same time contributes to raising the interest in written language data.
b) Written language online is closely related to various semiotic resources, including typography, still and moving images, and screen layout; the media-richness of contemporary digital environments increases the impact of multimodality on meaning-making (see section 7).
c) Modes of digital communication introduce new base-level units in online discourse. Categories such as 'message' or 'post' must be taken into account when collecting and analysing online data, and their relation to familiar syntactic and discourse-level units (sentence, clause, utterance, turn, adjacency pair) must be analytically examined (see sections 6, 9).
d) In CMC, social contexts can be invisible or only partially retrievable from digital exchanges themselves. Information on participants and their social relationships is often limited for both analysts and participants. New conventions of anonymous public exchange emerge, and traditional operationalization of socio-demographic may be of little use (see section 8).

e) Despite homogeneity at the level of hardware ('it's all bits and bytes'), digital language data can be strikingly heterogeneous, especially if researchers do not restrict to data from a single mode but sample across the range of digital modes, each with their respective semiotic resources, that people use in their online practices.
f) Finally, digital data is available in overwhelming amounts, making it difficult to select and focus on one specific sample or site of discourse

These are empirical conditions that CMC researchers across disciplines have to live with and adapt to in terms of their methodologies. The following sections will identify some 'best practice' solutions or alternatives that respond to these issues.

3. Data sampling in the *Computer-Mediated Discourse Analysis* framework

The first coherent framework for CMC research in linguistics was Susan Herring's 'Computer-Mediated Discourse Analysis' (Herring 2004, 2007). Herring's work includes a typology of media and social/situational factors for the classification of CMC data and an outline of six criteria for data sampling, which shall be reviewed here in some detail, elaborating on Herring's own pointers (Herring 2004: 351-354):

a) *Random sampling* means that each unit of communication from an available set of data has equal chances of being selected. A 'randomizer' tool can be used in order to select items from a numbered list of posts or messages, or items in specified intervals can be selected (e.g. every tenth message from a newsgroup). Random sampling enables representativeness and generalizability, but may result in a loss of context and coherence, for example by truncating conversations.
b) *Sampling by theme* can be used to collect data from discussion forums or other thematically organised streams of online discourse (e.g. hash-tagged tweets). The sample can consist of all messages in a particular forum thread or category, which are then compared to an equal sample from another thread in terms of e.g. language style or language choice. This method is useful within a framework that includes theme or topic as a relevant factor conditioning language variation or language choice (Androutsopoulos 2007a). However, sampling by theme has the disadvantage of excluding other co-occurring discourse activities (e.g. other topics discussed by the same users) and is therefore less useful if we are interested in language style across CMC modes and genres.
c) *Sampling by time* is required for any kind of longitudinal analysis. Researchers interested in language change online can draw samples at regular intervals across the available archives of a given newsgroup or forum. Sampling by time offers data that are rich in context, but may result in very large samples and/or truncate interactions in analytically unfortunate ways.

d) *Sampling by phenomenon* focuses on particular linguistic features or patterns of language use. For instance we could select only posts that contain emoticons or certain patterns of non-standard spelling. Such feature-based selection can be (at least partially) automatized by means of a concordance or customised script (Siebenhaar 2006). Herring's own examples are discourse-level phenomena such as joking or conflict negotiation, which must be selected manually. Sampling by phenomenon will be the method of choice for features that do not occur frequently and could therefore be absent from sample compiled on other criteria. It enables "in-depth analysis of the phenomenon" in question (Herring 2004: 351), but may result in loss of context and rule out a distributional analysis.
e) *Sampling by individual or group* can be based on socio-demographic criteria, if available or some kind of member ranking in the relevant online environment (see Section 8). It can enable analysis of selected users and user comparisons along familiar sociolinguistic lines. However, it excludes by definition exchanges to other users.
f) *Sampling by convenience*, that is, selecting "whatever data are available" (Herring 2004: 351), was popular with some early CMC research. Beyond its obvious advantage, it lacks a principle of systematic selection and may yield unsuited samples.

As this overview suggests, all alternatives have advantages and disadvantages, and the eventual choice depends on the research question and methodological practicalities. These criteria do not pre-empt the type of quantitative or qualitative analysis that will eventually be carried out. Some options (notably b, c, and e) roughly correspond to familiar 'external' or independent variables and result in data sets that will be later scanned for linguistic features of interest. Option (d) targets particular features straight away, thereby possibly ruling out a systematic control of independent variables if it is not deployed in combination with other selection criteria. In practice, however, combinations of two or more criteria are common.

4 CMC as 'text' or 'place'

In a recent paper on qualitative online research, Milner argues that "the study of cultures online demands we decide whether we frame online interaction as 'place' or as 'text'." (2011: 14). Although Milner's research is in communication studies rather than linguistics, his dyad of 'place' and 'text' can be productively adapted to sociolinguistic concerns. I suggest that from the perspective of language studies, 'CMC as text' focuses on the vast archive of written language provided by the Internet. It implies a tendency towards 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) classifications and categories. By contrast, a 'CMC as place' perspective might approach digital communication as a social process and CMC environments as discursively created spaces of human interaction, which are

dynamically related to offline activities. Here online data from various modes and environments might be collected, taking into account their cross-connections in people's digital literacy practices. This approach is therefore likely to prefer ethnographic observation and blended data (see next section).

The example of twitter can be used to illustrate this distinction. Approaching 'twitter as text' may mean collecting a large set of data, possibly by means of data mining techniques, and analysing it in terms of specific linguistic variables or categories, thereby distinguishing between, say, 'private users' and 'organisations' in terms of social variable. A 'twitter as place' view could examine how particular social actors use twitter alongside other digital modes in order to report on or coordinate social action related to a particular event (say, a political rally or a natural catastrophe), thereby in effect shaping the course and meaning of that event.

One reason the text/place dyad seems useful in a sociolinguistic context is, in my view, that it echoes the familiar tension between 'system-oriented' and 'speaker-oriented' approaches, in other words the differential focus of sociolinguistic research on linguistic variation itself as opposed to speakers' language practices. The text/place dyad does not directly determine the type of (quantitative or qualitative) analysis to be carried out; rather, it defines an epistemological perspective, which in turn is likely to entail a preference for particular research questions and techniques of data collection.

5. Screen- and user-based data collection

In the second distinction, terms 'screen- based' and 'user-based' refer to the two main, and in my view complementary, sites of data collection in new media sociolinguistics. 'Screen-based' data is produced and collected online by participants, 'user-based' data is prompted by the researcher's activities and produced through their contact to CMC users. A limitation to screen-based data may seem the norm in language-focused CMC studies, but this is neither self-evident nor uncontested within the discipline, let alone from an interdisciplinary perspective. Jones (2004) argues that the notion of context in CMC should not be reduced to what is happening on screen, but requires a shift of attention to the offline social activities in which CMC is embedded. From this viewpoint, CMC is shaped by a duality of situational context with simultaneous online and offline aspects. While a limitation to digital textual data may be motivated by research questions that focus on linguistic variation rather than language practices, it is a common experience among CMC researchers that the interpretation of linguistic findings can benefit from some awareness of the social and situational contexts of the data. In my own research (Androutsopoulos 2008), I have been interested in the awareness of particular linguistic variants and choices on the part of CMC users as a complement to screen-data analysis.

Figure 1 represents the relation of screen- and user-based data on a continuum with intermediate positions, which correspond to various degrees of ethnographic engagement on

the part of the researcher. They will be briefly discussed, moving from 'left' to 'right' on the figure. (The utmost right position is not discussed in its own right, as I assume that research on CMC sociolinguistics will always encompass screen-based data.)

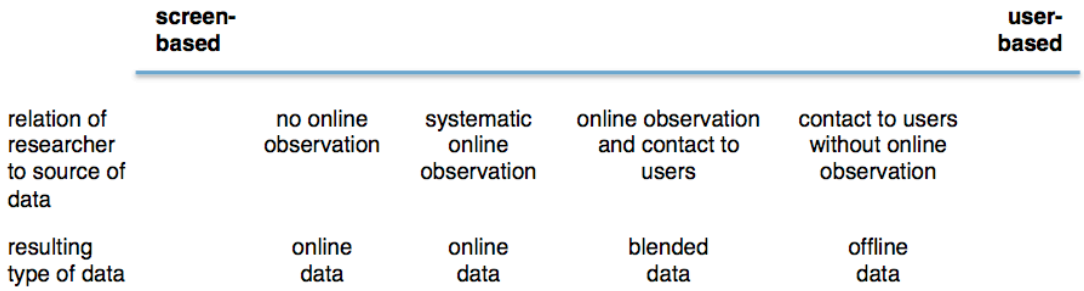

Figure 1: Screen-based and user based data in CMC research

Collecting screen data depends on both the options provided by various modes and environments (see Section 6) and the technological sophistication brought along by researchers. Synchronous applications such as IRC and IM come with the convenience of logfiles. Forum pages can be manually downloaded and then have to be cleaned up from html code in preparation for concordance or other software treatment. Content from social networking sites can be saved in PDF files, or relevant portions can simply be copy-pasted. Besides these more or less simple techniques, large portions of screen data can be *mined* by means of web crawlers, application program interfaces (APIs), customised scripts or other resources. Digital data can also be *delivered* to researchers by users themselves, for example students, members of the general public who donate data, or acquaintances by one member of a research team (Dürscheid & Stark 2011, Schmidt & Androutsopoulos 2004, Tagliamonte & Denis 2008, Tsiplakou 2009). This option specifically concerns private digital data exchanged on one-to-one applications and comes with the bonus of available socio-demographic information. Depending on research question, the selection of screen data may proceed on any of the six sampling criteria (or combinations thereof) reviewed above.

Strategies of online data collection differ not just in terms of technology, but regarding the degree of researcher engagement with the relevant site(s) of online communication. The researcher's position on a cline between no or minimal observation to fully-fledged familiarity with the online research site is in principle independent of the technique of screen data collection. Data mining of course rules out a simultaneous online observation; what is relevant, however, is whether any prior engagement with the original sites of this data has taken place, by which a selection of data to be mined has been determined. In the extreme opposite case (which I am tempted to label the 'take the data and run' approach), a researcher may harvest large amounts of digital data without ever visiting the sites where

they originate. However, a complete lack of familiarity with the original site of the data may limit the available contextual information, resulting to a preference for standardised ('etic') user categorisations and perhaps a replacement of socio-demographic categories by modes (Section 6).

Online observation refers to the process of 'virtually being there', with or without active participation, and watching the digital communication you will eventually analyse as it unfolds in a website or a network of connections across sites. Online observation is implicitly part of much linguistic CMC research, but often not explicitly acknowledged. I distinguish three aspects of online observation: 're-visit', 'roam around', 'explore resources for participation'. 'Re-visit' stands for paying regular, iterative visits to the selected site, tracing both routine activities and changes. 'Roam around' suggests exploring the virtual ground, browsing around sites, sections, threads or profiles. Whether to lurk or actively participate is open to debate in the literature (Milner 2011, Garcia et al. 2008). What is important, in my view, is that researchers do not end up analysing their own data or data that incurred as a direct outcome of their own contributions. 'Explore resources of participation' stands for trying out all resources afforded by an online environment of choice, such as search facilities, user lists, statistics, tags and tag-related hit lists. Across these activities, online observation involves a systematization of vernacular digital literacy practice, and the collection of screen data is complemented by the digital equivalent of ethnographic fieldnotes (which may involve tools like Zotero or Evernote).

Observation, the bottom line of any 'virtual fieldwork', comes in degrees. I suggest that even limited online observation offers a (limited) degree of ethnographic grounding, which can be further expanded and refined, and whose benefit can only be assessed within a particular project. In the absence of direct contact to users, the ethnographic information thus gained will of course be limited to what can be elicited in, or inferred from, the online environment. But especially when it comes to public (and semi-public) web spaces where participants' mutual background knowledge is incomplete and fragmented anyway, systematic observation can offer considerable insights that can subsequently be used to interpret surface data, to identify new objects of analysis or to articulate new research questions (Androutsopoulos 2008). Such insights may concern intertextual references or running gags, common and rare discussion topics, the usual pace or rhythm of discursive activities, categories of participation (e.g. core and peripheral members), the distribution of particular features across members, the trajectory or career of particular threads, and so on. With a bit of luck, researchers may even witness trends in a community's online talk as they emerge (see Kytölä & Androutsopoulos 2012). As in any ethnographic endeavour, systematic observation allows researchers to acquire some of the 'tacit knowledge' underlying the semiotic practices of regular members.

'Blended data' refers to any combination of screen-data to data collection through direct contact to selected users. I focus here on cyclical procedures of blended data

collection, assuming that user-based data will come to complement and interpretively frame the analysis of screen-based data. User-based data is of course not 'online data' in the narrow sense of the term. Depending on question and contact, its collection may even take the researcher far off the computer to the offline environments where the social activities that participants 'entextualize', i.e. document and turn into digital text, can be observed (Jones 2009).

Some user contacts offer access to data in the first place, others are initiated and established after an initial period of online observation and screen-data collection. In the first case contacts may be decided in advance, as part of the overall research design; in the latter case their selection will depend on previous observation and selection, e.g. by focusing on core members or users who 'stand out' in some way. Depending on research question and the researchers' familiarity with social-scientific methods, user-based data can be elicited in direct (face to face) or mediated contact by means of various instruments, including interviews, group discussions, questionnaires or by observation of people's literacy practices in front of their computers. Interviews (narrative or semi-structured) can be also carried out on Skype, phone or email. Each choice has implications in terms of further methods of data handling, including recording and transcription.

Cyclical procedures of blended data collection can begin with observation, followed by screen data collection and preliminary analysis, then establishing contact to selected participants. In the contact situation, samples of online content can serve as a prompt in order to elicit participants' awareness of and attitudes to language use online. The cycle can be extended, or repeated, by additional data collection, perhaps following new hints to language features or patterns identified in the interview. User contacts can thus be the last or an intermediate step between two layers of screen data analysis. My own experience includes various patterns of sequencing screen and user-based data. One pattern is to observe private homepages or discussion forums, then contact and interview their producers/webmasters, then return to and refine screen data analysis. Similarly, research on social networking sites may involve an initial contact (off or online) to likely participants, gaining permission to access their profiles, then observing profile activities and collecting samples; carrying out preliminary analyses; then conducting individual or group interviews. In research on multi-party Internet Relay Chat, a period of familiarization involving observation of and some active participation in the channel of choice was followed by contact to 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 this case, screen and user-based portions of CMC data were collected in parallel and simultaneous, but separate online activities.

6. Modes and environments

Broadly defined as applications that offer a standardised user interface and a set of options for digital interaction, modes are key components of CMC for both users and researchers. Modes are traditionally classified on the parameters of synchronicity (synchronous/asynchronous) and publicness (one-to-one, one-to-many or many-to-many), thereby distinguishing Instant Messaging (synchronous, 1:1) from Internet Relay Chat (synchronous, many:many) from email (asynchronous, 1:1 or 1:many) and so on (Herring 2001).

Modes usually serve as invariant parameters for digital data selection, and a lot of data reported in the literature is specified for or even restricted to particular modes, e.g. IRC, Instant Messaging or email. Analysis of mode-specific online data 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 external (independent) variables, based on the assumption of more or less stable relations between modes and patterns of online language use. In particular, the hypothesis that synchronous modes of CMC resemble spontaneous spoken language more than asynchronous ones has been tested for variation of standard/vernacular and spoken/written features as well as for the occurrence of conversational code-switching (Androutsopoulos 2007b, Paolillo 2011). Such *inter-mode* analysis compares data from two or more CMC modes (e.g., messaging vs. email or chatting vs. newsgroups) while controlling other social and situational factors. By contrast, an *intra-mode* design compares data from the same mode across varying social and/or situational conditions, e.g. informal online chat to moderated chat sessions with politicians. Provided the primacy of mode effects on language over social and/or situational factors is not being assumed by default, modes offer an invaluable handle for data collection and exploration.

The usefulness of modes as building blocks of online data collection is weakened by the growing importance of participatory web environments, where old modes are integrated and new genres cannot be distinguished on synchronicity and publicness alone. Such environments include online portals that host edited content and user discussion forums; social network sites with user profiles, walls and groups; and content-sharing platforms for photos and videos. Due to their sheer size and diversity of contributors, genres and interactive applications, web environments create new problems of comparability. To put it simply, comparing *YouTube* to *Vimeo* 'as such' makes little sense for linguistic perspective. Developing a meaningful comparison relies here on systematic online observation by which to identify relevant types of content, genres or users prior to the actual data collection. Examples include a comparison of three asynchronous genres on a hip-hop portals for colloquial markers in spelling (Androutsopoulos 2007b), the analysis of status updates as a prominent small genre on facebook walls (Bolander & Locher 2011; Lee 2011), and the selection of YouTube videos and comments based on user tags (Pihlaja 2011).

7. Multimodality

Multimodality can be understood in at least three different ways in the context of CMC. First, it can refer to user activities during the production of and interaction with online content. In research that includes photographs or video-recordings of users in front of their screens (see papers in Androutsopoulos & Beisswenger 2008), methods of multimodal analysis of embodied interaction can be used to examine the relation between users' face expression and posture to the online content they type in or read. In a second sense that relates to the concept of mode in the previous section, multimodality refers to the simultaneous use of more than one applications in people's digital literacy practice. Screen movies (recorded by means of special software) can be used to document how users multitask on various applications, and what this means in terms of e.g. style-shifting. This technique is not (yet) widespread in sociolinguistics, but could offer an interesting addition to blended data. In a third sense, multimodality refers to the coexistence of resources from more than one semiotic mode in digital content itself. The evolution of CMC brought about increasingly complex forms of multimodal communication, and while language-heavy modes such as email predominate in early language-focused research, the contemporary integration of written language with pther semiotic resources (spoken language, audio, static and moving image, video, colour, pictograms, typography, …) presents a methodological challenge. Researchers interested in self-presentation online have long been alert to how users draw on all semiotic resources at their disposal in order to construct their identities on homepages and blogs. In contemporary web environments, an increasing amount of written or spoken language comes embedded in visual or audiovisual texts (think of lolcat images and *Youtube* videos), and written-language exchanges are often prompted by multimodal texts, as can be observed on *Flickr* or social networking sites (Lee and Barton 2011). Even when the research question is concerned with the language part, taking into account multimodal prompts may help interpret patterns of variation or style choice. In the absence of widely accepted standards for multimodal online data collection, page-long screenshots and automated video/comment download are viable techniques, though ethics considerations may restrict the types of content that can be downloaded.

8. Social identities and participation frameworks

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 information for mutual social categorization. This is a serious problem for any sociolinguistic analysis that depends on clear-cut socio-demographic information (such as male/female, middle/working class, and so on) as a guideline to data collection itself. It can be addressed or circumvented in a number of ways. First, researchers can contact relevant users and collect relevant socio-demographic information post-hoc, though this is not always practically feasible, especially in public CMC.

Second, researchers can take data offered by users themselves as a basis for speaker categorization. Depending on mode and genre, these may range from fairly straightforward information to a range of indexical cues in screen names and associated virtual identity signs such as avatars, member profiles or signatures. One challenge here is how to handle the tension between online and offline identities, and whether to conceive of users "behaving like" or rather "performing" a particular social identity; however, this issue goes beyond data collection. Alternatively, researchers can abandon external socio-demographic factors and turn to environment-specific categories such as regulars/novices or admins/normal users, to which sociolinguistic variation is then correlated (Paolillo 2001). A further alternative is to focus on the discourse processes by which participants ascribe and negotiate social identities to selves and others, thereby drawing on interpretive methods of data collection and analysis.

This discussion suggests that the more we depart from 'offline' socio-demographic variables as a basis for the sociolinguistic analysis of CMC data, the more we need to reconstruct participation roles in various digital modes and environments, thereby going beyond a medium-specific replication of the simple 'sender : receiver' (i.e. writer : reader) model. This has been an issue for analysis rather than data collection so far. In his study of participation roles in French newsgroups, Marcoccia (2004) distinguishes between 'host' and 'casual sender' based on various diagnostic criteria. Hosts send and answer more messages than other senders; are often on friendly terms with each other; manage (e.g. initiate, regulate) online interactions, and often play the role of experts. On the reception side, Marcoccia distinguishes between the (explicitly) addressed recipient, the favoured recipient (which he takes to coincide with the host), and the 'eavesdropper', i.e. the non-addressed, but ratified recipient that is commonly referred to as 'lurker'. Data collection in public or semi-public online environments can anticipate these (or adequately modified) participation roles, especially in terms of their relation to institutional conditions of communication online and/or theoretical frameworks of choice.

9. Units, sequences, intervals

Sociolinguistic studies of CMC data usually focus on micro-linguistic and interactional units, and data collection is therefore geared towards collecting material that contains these units. However, familiar units of linguistic analysis are embedded and reframed in larger-scale units of digital mediation which are defined by CMC applications or environments. These include the categories of messages (units in one-to-one exchanges) and post (units of contribution to public, multi-party exchanges), which are in turn embedded in larger, multi-authored structures such as threads or lists of comments. Messages and posts are indispensable units of data collection, but their relation to familiar linguistic or conversation-analytic categories like sentence, utterance or turn is neither trivial nor straightforward. For example, a

conversational turn can be divided into several online posts, and one post can accommodate more than one turns depending on its composition by the poster.

Acknowledging messages/posts as an additional level in the organisation of online data is, in turn, indispensable for working with sequences, i.e. temporally arranged chains of posts/messages that are exchanged in a particular interactional configuration. A sequence is either collected 'as such' in a public CMC environment (e.g. a facebook wall conversation, or forum threads) or reconstructed ('zipped together') from data exchanged between separate digital interlocutors. Any research question that takes its cues from pragmatics and interactional sociolinguistic is more or less dependent on collecting sequences rather than isolated messages/posts. As a consequence, the interactional processes usually examined in sequential analysis (e.g. adjacency pairs) are reframed within a sequence of posts/messages. When researching code-switching online, for example, post-internal and post-external code-switching (i.e. within or across posts) form an additional level of analysis that does not coincide with either turns or sentences (Androutsopoulos 2007a). This reframing has also an impact on intervals, i.e. the time distance between individual contributions in the flow of a dyadic or multi-party exchange. Much has been written on intervals from the viewpoint of constraints determined by technology, resulting in transmission gaps or leading to an order of posts that disrupts expectations of sequential coherence. But relatively little is known about the active management and interpretation of intervals by participants themselves (Jones 2005, Schmidt & Androutsopoulos 2004). In practice, the time-stamps contained in the online data or noted down by researchers or participants are a useful resource for reconstructing intervals, which can be analysed as indexes to participants' footings in text-based interaction.

10. A note on research ethics

Respecting and protecting the privacy of informants is a basic legal and ethical requirement in social-scientific fieldwork. There is no general consensus on just how to achieve this in CMC research, and ethnics guidelines for researchers and students vary by country and institution. It is common sense among CMC researchers that we need to protect the anonymity of our informants by not directly disclosing their offline identities and avoiding to publish any cues that may lead to their identification. Various modes, environments 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 and publish is the rule regarding private data, but 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 to diverging interpretations on 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 as a reminder that participant and researcher views do not forcibly coincide. Our ethic decisions must ultimately observe legal requirements of 'privacy', but our considerations should not neglect informants' views on the shifting boundaries of privacy and publicness. (Readers are also referred to the ethics guidelines of the Association of Internet Researchers, latest review draft at: http://aoirethics.ijire.net.)

References

Androutsopoulos, Jannis (ed.) (2006) *Sociolinguistics and computer-mediated communication*. Theme Issue, *Journal of Sociolinguistics, 10/4*.

Androutsopoulos, Jannis (2007a) Language choice and code-switching in German-based diasporic web forums. In: Danet, Brenda, and Susan C. Herring (eds.) *The Multilingual Internet*, 340-361. Oxford: Oxford University Press.

Androutsopoulos, Jannis (2007b) Style online: Doing hip-hop on the German-speaking Web. In: Peter Auer (ed.), *Style and Social Identities*, 279-317. Berlin, NY: de Gruyter.

Androutsopoulos, Jannis (2008) Potentials and limitations of discourse-centered online ethnography. *Language@Internet*, 5. http://www.languageatinternet.org/articles/2008/

Androutsopoulos, Jannis / Michael Beißwenger (eds.) (2008) *Data and Methods in Computer-Mediated Discourse Analysis*. Special Issue, *Language@Internet* 5 (2008). http://www.languageatinternet.org/articles/2008

Beißwenger, Michael & Angelika Storrer 2008 Corpora of Computer-Mediated Communication. In Lüdeling, Anke / Kytö, Merja eds. *Corpus linguistics* ,Vol. 1 292-309. Berlin, New York: Mouton de Gruyter

Bolander, Brook / Miriam A. Locher (2010) Constructing identity on Facebook: Report on a pilot study. *SPELL Swiss Papers in English Language and Literature* 24, 165-185.

Dürscheid, Christa & Elisabeth Stark 2011. SMS4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland. In: Thurlow & Mroczek (eds.), 299–320.

Garcia, Angela Cora et al. (2009) Ethnographic Approaches to the Internet and Computer-Mediated Communication. *Journal of Contemporary Ethnography* 38(1), 52-84.

Herring, Susan C. (2001) Computer-mediated discourse. In: Schiffrin, Deborah et al. (eds.) *The Handbook of Discourse Analysis,* 612-634. Malden: Blackwell.

Herring, Susan C. (2004) Computer-Mediated Discourse Analysis: An Approach to Researching Online Communities. In: Barab, Sasha A. et al. (eds.) *Designing for Virtual Communities in the Service of Learning*, 338-376. Cambridge/New York: Cambridge University Press.

Herring, Susan C. (2007) A faceted classification scheme for computer-mediated discourse. *Language@Internet*, 4. http://www.languageatinternet.org/articles/2007/

Jones, Rodney (2004) The problem of context in computer-mediated communication. In: LeVine, P., and Scollon, R. (eds.) (2004) *Discourse and Technology: Multimodal Discourse Analysis*, 20-33. Washington, DC: Georgetown University Press

Jones, Rodney (2005) 'You show me yours, I'll show you mine': The negotiation of shifts from textual to visual modes in computer mediated interaction among gay men. *Visual Communication* 4 (1): 69-92.

Jones , Rodney 2009 Dancing, skating and sex: Action and text in the digital age. *Journal of Applied Linguistics*, Vol 6, No 3 (2009), 283-302.

Kytölä, Samu and Jannis Androutsopoulos (2012) Ethnographic Perspectives on Multilingual Computer-Mediated Discourse. In Marilyn Martin-Jones, Marilyn & Sheena Gardner (eds) *Multilingualism, Discourse, and Ethnography*, 179-196. London: Routledge.

Lee, Carmen (2011) Texts and Practices of Micro-blogging: Status Updates on Facebook. In C. Thurlow and K. Mroczek. (eds). *Digital Discourse: Language in New Media*. Oxford: Oxford University Press (in print).

Lee, Carmen & David Barton (2011) Constructing glocal identities through multilingual writing practices on flickr.com. *International Multilingualism Research Journal* 5(1): 39-59.

Locher, Miriam A. (ed.) (2010) Politeness and impoliteness in computer-mediated communication. Theme Issue, *Journal of Politeness Research* 6:1.

Marcoccia, Michel (2004) On-line polylogue: Conversation structure and participation framework in Internet newsgroups. *Journal of Pragmatics* 36: 115-145.

Markham *,* Annette (2008). *The* methods, politics, and ethics of representation in online ethnography. In *Denzin*, Norman K. (ed.) *Collective and interpreting qualitative materials*, 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. http://gjss.org/index.php?/

Paolillo, J. C. (2001). Language variation on Internet Relay Chat: A social network approach. *Journal of Sociolinguistics*, *5*(2), 180-213.

Paolillo, J. C. (2011). "Conversational" codeswitching on Usenet and Internet Relay Chat. Language@Internet, 8, article 3 (2011). http://www.languageatinternet.org/articles/2011/

Pihlaja, Stephen (2011) Cops, popes, and garbage collectors: Metaphor and antagonism in an atheist/Christian YouTube video thread. Language@Internet, 8, article 1 (2011) http://www.languageatinternet.org/articles/2011/Pihlaja

Schmidt, Gurly & Androutsopoulos, Jannis (2004) löbbe döch. Beziehungskommunikation mit SMS. *Gesprächsforschung*, 5. (www.gespraechsforschung-ozs.de)

Siebenhaar, Beat (2006) Code choice and code-switching in Swiss-German Internet Relay Chat rooms. *Journal of Sociolinguistics* 10(4): 481-509.

Squires, Lauren (2012) Whos punctuating what? Sociolinguistic variation in instant messaging. In Jaffe, Alexandra et al. (eds.) *Orthography as social action: Scripts, spelling, identity and power*, 289-323. Berlin, New York: de Gruyter.

Tagliamonte, Sali A., and Derek Denis (2008) Linguistic ruin? LOL! Instant messaging and teen language. *American Speech* 83(1): 3-34.

Tsiplakou Stavroula (2009) Doing bilingualism: Language alternation as performative construction of online identities. *Pragmatics* 19(3): 361-391.