This article has been published in:

Discourse, Context & Media 3 (2014) 14–26

Elsevier

https://doi.org/10.1016/j.dcm.2013.10.004

Doing sociolinguistic research on computer-mediated data: A review of four methodological issues

Brook Bolander and Miriam A. Locher

Abstract

This article focuses on four methodological issues which raise challenges for sociolinguists working with online data: (1) ethics; (2) multimodality; (3) mixed methodologies and the relationship between online and offline settings; and (4) web corpora and annotation. While there are currently numerous publications dealing with questions of ethics, data and methodology from within communication studies and social scientific research more generally, there are only a handful of publications which specifically focus on empirical linguistic research. In addition to delineating the diversity of computer-mediated data, in the course of the article we review each of these methodological issues in turn, thereby discussing key terminology and reviewing relevant literature.

Keywords: Computer-mediated communication; Methodology; Ethics; Multimodality; Web corpora

1. Introduction

This paper addresses methodological issues that represent challenges of sociolinguistic research on computer-mediated communication (CMC) and offers observations on methodological research decisions. Since the late 1990s, and particularly within the last decade, numerous publications on researching online practices have appeared from a social scientific perspective, but only a few texts dealing particularly with methodology in sociolinguistics. This paper aims at filling this gap, specifically by focusing on four methodological issues: (1) ethics; (2) multimodality; (3) mixed methodologies and moving between online and offline settings and (4) web corpora and annotation. Our decision to

focus on these four issues stems from their pervasiveness in countless discussions we have had with students and colleagues on data and methodology in connection with computer-mediated communication. Scholars wishing to use online data for sociolinguistic research will all, at one point, need to make ethical decisions; they all will also have to characterise their data with respect to modality and to reflect upon the implications of their data's mono- or multimodality for the research design, to consider the suitability of mixing methods and focusing on online and/or offline contexts in light of their research question, and to face the challenge of annotating their data when using the web as a corpus, or as a pool of data from which to create a corpus. While many of the points discussed are not restricted to sociolinguistic research only, we will address pertinent issues with sociolinguistic research in mind, where sociolinguistic is understood in the broadest sense of the term.[1]

In what follows we will make use of the existing literature from linguistics, as well as drawing on literature from media and culture studies, and sociology whenever opportune for our linguistic angle (see the list of useful and inspiring texts found in footnote 2).[2] Evidence for the heightened interest in methodological issues relating to computer-mediated data is not only provided by the upsurge in monographs, edited collections and journal articles dealing with online environments, it also becomes evident when looking at publications dealing with research methodology more generally. For example, whereas the first edition of *Qualitative Communication Research Methods* (Lindlof, 1995) did not include information on CMC, the second (2002) edition (ed. by Lindlof and Taylor, 2002) has a chapter on "Qualitative research and computer-mediated communication". And to turn to an example from within sociolinguistics, Mallinson et al. (2013) *Data Collection in Sociolinguistics: Methods and Applications*, includes a chapter on "Online data collection" (Androutsopoulos, 2013).

Despite this upsurge, there is a paucity of research addressing issues of methodology for sociolinguistic research in computer-mediated settings. In recent years, a series of edited collections on sociolinguistic research online has been published (see, for example, Thurlow and Mroczek, 2011; Herring et al., 2013a; Tannen and Trester, 2013; Tagg and Seargeant, forthcoming), yet the focus of these editions is not methodological. To the best of our knowledge, the main exceptions to this tendency are Herring's (2007) "A faceted classification scheme for computer-mediated discourse", Androutsopoulos and Beisswenger's (2008) special issue of language@internet entitled "Data and methods in computer-mediated discourse analysis", Androutsopoulos' (2013) "Online data collection", Barton and Lee's (2013, Ch. 12) chapter on "Researching language online" and Lim et al's (2013) *Innovative Methods and Technologies for Electronic Discourse Analysis*.

[1] We adopt a definition of sociolinguistics that allows for both qualitative and quantitative approaches to the study of naturally occurring data, in an endeavour to understand patterns of language practices embedded in their context (Coulmas, 2005: 10; Wardhaugh, 2002: 5).

[2] While this list on CMC methodology publications from linguistics, media and culture studies, and sociology cannot be complete, we especially recommend the following set of texts in order of appearance: Markham (1998), Jones (1999), Hine (2000), Mann and Stewart (2000), Ó'Dochartaich (2002), Best and Krueger (2004), Johns et al. (2004), Fielding et al. (2008), Hine (2005, republished in 2008), Kozinets (2009), Lazar et al. (2009), Markham et al. (2009), Hunsinger et al. (2010), Das et al. (2011), Hooley et al. (2012), Androutsopoulos (2013).

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

While strongly discourse analytic in outlook, the texts cover a range of issues including but not limited to questions of ethics, data sampling, and quantitative and qualitative approaches to studying CMC data. The need for a discussion of research methodology for sociolinguistic research online is also reflected in a number of publications currently being prepared for publication (for example, Spilioti and Georgakopoulou's (2013) *The Routledge Handbook of Language and Digital Communication*). As stated above, the aim of our paper is to add to this small body of existing research by discussing core methodological issues in connection with sociolinguistic research on computer-mediated data.

It is important to underline that the choice of methodology in empirical research (both qualitative and quantitative) co-evolves in connection with the research question. The research question and corresponding methodological steps develop in a dynamic process of decision-making. There is thus no one method which is per se better than any other (see Jucker, 2009: 1619), despite assumptions to the contrary. Rather than dismissing particular methods from the outset, "they should be evaluated carefully as to what kinds of question they can answer and what kinds of question they cannot answer" (Jucker, 2009: 1619). In this paper we do thus not wish to advocate any particular approach to sociolinguistic research on computer-mediated data. Instead, we will present key issues which cause challenges for scholars wishing to use computer-mediated data. We hope that our discussion of these issues and challenges will aid sociolinguists wishing to use CMC data in their methodological decision-making processes.

We will begin the paper by discussing the nature of computer-mediated data, and different terminologies used by scholars to refer to the settings in which they work with such data (Section 2). Subsequently, we will address ethical premises (Section 3.1), multimodality (Section 3.2), mixed methodologies and moving between online and offline settings (Section 3.3), and web corpora and annotation (Section 3.4). Finally, in Section 4 we will summarise the key arguments made in this paper and point to further research outlets.

2. The diversity of computer-mediated data, Web 2.0 and other classificatory designations

Scholars working with computer-mediated data in their study of language use online are confronted with a striking amount of diversity. Researchers nowadays recognise that there is no such thing as a monolithic variety of Internet language. Rather, language online is varied and CMC practices are changing fast. Based on Hymes' (1974) SPEAKING mnemonic, a framework guiding research conducted within an "ethnography of speaking" tradition, Herring (2007) succinctly summarises the diversity of computer-mediated data according to ten "medium factors" and eight "social factors". Her so-called "faceted classification scheme for computer-mediated discourse" can be used as a methodological tool for researchers wishing to study language use in computer-mediated environments, as each of the factors listed in her scheme has been shown to influence language use online. Yet it can also be seen as a descriptive framework which makes manifest the

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

diverse properties of computer-mediated data.

While the scheme is open and can thus be added to, the following ten medium factors are mentioned by Herring (2007):

(1) synchronicity: is the data synchronous, i.e., are exchanges performed in real time, or asynchronous, i.e., is there a time lag between the production and receipt of messages?
(2) message transmission: are messages transmitted via one-way or two-way message schemes, i.e., are they transmitted as whole entities to be read by the addressee upon completed composition by the author of the message, or read line-by-line as they are produced, respectively?
(3) persistence of transcript: how long does a written record of the interaction remain accessible on the site for other users to read?
(4) size of message buffer: are there technical restrictions on the number of characters a message has?
(5) channels of communication: via what medium are messages produced and received?
(6) anonymous messaging: does the system provide, encourage, or inhibit the production of anonymous messages?
(7) private messaging: does the system provide, encourage, or inhibit the production and reception of messages via a private channel only accessible to particular participants? ↑ 15
(8) filtering: do individuals have the technical possibility to filter out messages they do not wish to read? 16 ↓
(9) quoting: does the system provide an in-built system to quote parts of messages or entire messages without having to copy/past them or manually type them?
(10) message format: how do the messages appear on the screen; in what order?

These ten medium factors draw attention to how diverse "modes" (Murray, 1988; Herring, 2007) can be, as each mode can have any number of technological combinations characterising it (and influencing linguistic practice).

The variation becomes even more heightened when one turns to the eight social factors Herring (2007) proposes. Any number of people with different backgrounds and characteristics ("participant characteristics") can engage in various forms of interaction (e.g., one-to-one, one-to-many) at different rates of interactional intensity ("participation structure"), following alternate purposes at both the level of the group or community, as well as at the level of particular interactions ("purpose"). Moreover, the group as a whole can pursue a specific topic, as can individuals engaged in a particular interaction ("topic or theme"), and they can perform various interactional activities (e.g., debate, praise one another, give advice) when they engage in communication ("activity"); the "tone" of these activities can vary (e.g. friendly, contentious, formal, casual). The participants are also strongly guided by "norms" for social practice and language use, which may be determined top-down by individuals with (technical or social) authority, and/or emerge bottom-up through practice. Finally, the communication is realised using a specific writing system and font (in the case of text-based productions) and performed in a particular language variety ("code"). All of these medium and social factors need to be considered when collecting

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

and analysing data, as any number of them (in isolation or combination) may influence the way interlocutors use language. This diversity of language use online is a challenge for researchers (Crystal, 2011); attempts to study it prompt for methodological decisions which, in turn, are driven by research interests.

We can now look back at almost three decades of (sociolinguistic) research on CMC. Just as the medium has developed and technological changes have led to new modes of CMC, much has also changed with regard to research foci. This research development is best illustrated by Androutsopoulos' (2006) "waves" approach, in which he identifies three main waves, or phases in sociolinguistic research on CMC. In a "first wave" of research "language use on the Internet [is treated] as […] distinct, homogeneous, and indecipherable to 'outsiders'" (Androutsopoulos, 2006: 420). Both the net as a whole, and individual modes or genres are conceptualised as linguistically uniform as a result of shared technological properties; these properties are seen to determine language use ("technological determinism", see also Baym, 1995; Squires, 2010; Herring et al., 2013b). In this first phase, "descriptive accounts" of online language, "the hybrid combination of written and spoken features", and "principal differences between synchronous and asynchronous modes" are prioritised (Androutsopoulos, 2006: 420). In wave 2 there is a step away from computer determinism towards an acknowledgement of "the interplay of technological, social, and contextual factors in the shaping of computer-mediated language practices" (Androutsopoulos, 2006: 421); and in wave 3 research "the role of linguistic variability in the formation of social interaction and social identities on the Internet" (Androutsopoulos, 2006: 421) is underlined. Furthermore, notions of "emergence" and "performativity" are increasingly emphasised, especially in connection with an epistemological shift towards social constructivist understandings of language use and practice.

A further central development is from what scholars call "Web 1.0" to "Web 2.0". Whereas CMC "has been the most popular and most traditional" term, first employed in the 1980s (Jucker and Dürscheid, 2012) and still widely in use (cf., for example, Herring et al., 2013a; and the title of this paper), in contemporary scholarly discourse "the buzzword 'Web 2.0' stands for a turning point that refers to the more dynamic and user-shaped development of electronic discourse" (Locher, 2014; see Jucker and Dürscheid, 2012 for a detailed discussion of terminology). This turning point occurred "in the first decade of the 21st century" with the emergence of "Web-based platforms" which "incorporate user-generated content and social interaction, often alongside or in response to structures and/or (multimedia) content provided by the sites themselves" (Herring, 2013: 4). Yus (2011: 93) similarly underlines the increase in "interactions" and "content sharing" and further points to an accompanying shift away from a "traditional 'pyramidal media communication pattern'", i.e., one "based on an authority that uni-directionally filters and delivers Internet content to the mass of users". Zappavigna (2012: 2) uses the term "social web" to describe Web 2.0, thereby drawing attention to "a shift toward the internet as an interpersonal resource rather than solely an information network", and to the Internet as a site for the performance of social relationships and not solely a virtual space catering to

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

information sharing. To this shift one can add the increase in "convergent media", i.e., the co-occurrence or co-presence of various media practices, such as "text (and voice) chat during multiplayer online games", or "text comments on photo-sharing sites" (Herring, 2013: 4–5).[3] These changes cause challenges for sociolinguists interested in studying language use online.

3. Methodological issues and challenges

We will begin this section by charting key developments with respect to ethical issues of sociolinguistic research online (Section 3.1), before turning to address multimodality (Section 3.2), mixed methodologies and the complex relationship between online and offline settings (Section 3.3), and web corpora and annotation (Section 3.4). Our discussion of these issues is not exhaustive, and the brevity of our description should by no means imply that these topics are simple. Our aim has been to outline these main issues, so as to point to some of the major challenges sociolinguists face when working with computer-mediated data. It is worth highlighting again that despite the fact that the Internet both "challenges taken-for-granted frameworks for how identities, relationships, cultures, and social structures are constructed", and how we "understand and conduct qualitative [and quantitative] inquiry", "core methodological principles do not change" (Markham, 2011: 112; see also Barton and Lee, 2013: 177). To "navigat[e] these challenges" researchers must "rel[y] on their ability […] to ask reflexive questions at critical junctures throughout the project" (Markham, 2011: 112). It is our belief that a research design and the reading of the finished report will improve when researchers feel a sense of accountability at each of these junctures; i.e. are able to explain their decisions and the path that led to these decisions both to themselves and to their readership.

3.1 Ethics

While it might be tempting for some scholars to treat the Web as an easy place to collect digital data without consideration of the
authors who originated this very data, we strongly support the point that scholars who work with data always need to ask themselves about the ethical implications of their research. This process should ideally also translate into scholarly output, so as to raise awareness in the research community about ethical considerations and also to increase transparency and understanding of the data. Herring's (1996a) introduction to the first edited collection of empirical research articles on language use in CMC states the following with regard to ethics[4]:

↑ 16
17 ↓

[3] For Herring (2013) this convergence of different media warrants the introduction of a new acronym; CMCMC which stands for "convergent media computer-mediated communication".
[4] See also Paccagnella (1997) for a further very early discussion of ethics in CMC.

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

> [B]ecause of ethical issues associated with collecting and analyzing private e-mail correspondence, most of the examples are drawn from public or semi-public group interactions […] (Herring, 1996b: 2)

In early research on language use in computer-mediated environments, the degree to which a group was public or private was a key factor steering ethical decision-making practices. While this distinction is still of paramount importance, understandings of public and private have changed. At the time Herring wrote this, public and private were understood with respect to access, i.e., as technological attributes of a particular site. This is indicated a couple of pages later in the same volume:

> The editorial policy followed in citing CMC data in this volume makes a distinction between restricted- and open-access electronic fora, the former of which are considered private, while the latter are public. (Herring, 1996b: 6)

This had implications for quoting and presenting the data in the edited collection, since pseudonyms were used for "private or semi-private sources", except where explicit permission was granted, whereas they were not used for Usenet and open-access Listservs. As the editor states, "an attempt has been made to follow common sense in respecting as much as possible the privacy of those whose messages are cited as examples, while giving credit for ideas where credit is due" (Herring, 1996b: 7). We will return to the latter point of giving credit below.

Much has changed since the mid to late 1990s with respect to research ethics, and one of the main developments concerns understandings and applications of public and private to computer-mediated environments. Key here is a shift towards conceptualising "public" and "private" in terms of both access and content, and accompanying this conceptualisation as gradable and not absolute. This move towards a more differentiated understanding is highlighted and succinctly described by Landert and Jucker (2011):

> [w]e are confronted with media texts that combine private and public aspects on various levels. They may be public in the sense that they are within the public space and can be read by a large and anonymous audience, while at the same time discussing topics which we think of as 'private' and using language which is associated with informal and private conversations. (Landert and Jucker, 2011: 1423)

In theory, a particular website, for example, can therefore be public both in terms of access and content (public; public), private in terms of both access and content (private; private), public in terms of access yet private in terms of content (public; private), or private in terms of access and public in terms of content (private; public), although this latter possibility is the least likely if we presume that a site-owner would not restrict access to his/her site if the content was not somehow private to him/her. Moreover, it is also entirely possible that parts of a site (e.g., the personal profile information on Facebook compared with wall posts) are more or less public/private in terms of both access and content than others, making it difficult to conceive of the ethical decision-making process

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

as a holistic one that can be applied to a whole site, and only once at the beginning of the research process.

In the course of the last decade, scholars have progressively tackled the complexity of ethical decision-making in CMC. Notable here is the work conducted by the Association of Internet Researchers, who have developed two extensive documents with guidelines for conducting research in computer-mediated settings, one published online in 2002 (Ess and the AOIR Ethics Working Committee, 2002) and one in 2012 (Markham, Buchanan with contributions from the AOIR Ethics Working Committee). While both documents have a core group of authors, scholars from a variety of different academic and disciplinary backgrounds contributed to the documents, a practice which has ensured a rich variety of viewpoints, a comprehensive list of core challenges researchers face, and a wide range of guiding questions, literature and case studies designed to aid researchers in ethical decision-making. In addition, there is now also a wiki by the Association of Internet Researchers, whose purpose it is to "[p]rovide a compendium of resources for ethical decision making in internet-related research", "[c]entralize guideline and updates over time" and "[b]uild a robust and open source knowledge database" (AOIR ethics wiki).

Central to both documents is a focus on guidelines, as opposed to rules and regulations. As stated by Markham et al. (2012) "[w]e advocate guidelines rather than a code of practice so that ethical research can remain flexible, be responsive to diverse contexts, and be adaptable to continually changing technologies". In campaigning for "a process approach" (Markham et al., 2012) to ethical decision-making, the authors wish to discourage a top-down, once-off application of fixed rules, since "[m]ore than one set of norms, values, principles and usual practices can be seen to legitimately apply to the issue(s) involved. […] Multiple judgements are possible, and ambiguity and uncertainty are part of the process" (Markham et al., 2012). This statement is not restricted to CMC environments, but should stand true for all research settings sociolinguists are interested in. Yet in light of the need to create guidelines which specifically tackle the particularities of online settings and the novelty of many of these settings (Ess, 2002), the authors are right to highlight the need for a dynamic process-approach to ethics in CMC.

In doing so, they put forward six main principles (Markham et al., 2012), each of which is valid for scholars doing sociolinguistic research online. The first principle maintains that "the greater the vulnerability of the community/author/participant, the greater the obligation of the researcher to protect the community/author/participant". This principle is familiar, yet with respect to CMC it can be challenging as one cannot always know the age[5] (a key factor defining vulnerability) of the individuals whose linguistic practices we are interested in studying. The second principle is equally familiar and concerns the minimisation of harm, particularly the point that scholars need to pay close attention to the context to assess the potential of causing harm, as "'harm' is defined contextually"

[5] When classic variables such as age, gender, educational or linguistic background etc. are considered important elements of the data description and research design, many CMC sources will in fact not yield reliable information on the interactants. This challenge of the anonymity of many informants has to be addressed in the study design and explained in the data description.

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

(Markham et al., 2012). The third principle concerns the issue of human subjects, and more specifically the need to follow guidelines for research on human subjects in digital settings "even if it is not immediately apparent how and where persons are involved in the research data" (Markham et al., 2012). In other words, we should be cautious not to forget that there are individuals who have authored the
contributions we are interested in analysing, even if we never encounter them as physical bodies, but only work with their practices. We will return to this point in Section 3.4 below. Related to this, we should be aware that some of the individuals whose contributions we study may wish to be acknowledged as authors, and given credit for their work (as was recognised as early as in 1996 in Herring's edited collection). This, too, will influence whether and how researchers engage in dialogue with the individuals they wish to study; as stated by Ess (2002), in cases where "subjects may be understood as authors intending for their work to be public […] then fewer obligations to protect autonomy, confidentiality, etc., will likely follow".

The fourth principle listed by Markham et al. (2012) acknowledges the distinction between "authors" and "research participants", and specifically considers the possible benefits of the research relative to the rights of subjects: "researchers must balance the rights of subjects (as authors, as research participants, as people) with the social benefits of research and researchers' rights to conduct research. In different contexts the rights of subjects may outweigh the benefits of research" (Markham et al., 2012). Finally, points five and six draw attention to the need to potentially deal with ethical issues in all stages of the research (principle five) and to treat ethical decision making as "a deliberative process". This entails discussing issues with colleagues and other experts in the field and consulting resources.

The Internet constitutes a "field site" with enormous potential for sociolinguists, who have at their disposable a wide range of language use, which can be studied using different methodologies and to answer various research questions. Its appeal is enhanced by the fact that it is time-saving with regard to transcription, and, as stated by (Herring, 1996b: 5), "observers can observe without their presence being known, thus avoiding the 'Observer's Paradox' that has traditionally plagued research in the social sciences" and been key to much methodological discussion in sociolinguistics. While, as Sandler (2013: 59) points out, the "ease of recording in these environments"[6] leads to "a great temptation to collect more ethically ambiguous data", the online environment needs to be treated as both the same and different to offline ones with regard to ethics. On the one hand, we need to continue to ask the same questions when conducting empirical research as we have always done. On the other hand, we need to pay close consideration to the particular challenges raised by the settings we wish to study. Crucially, the reader of scholarly output should be made aware of the decisions that the researcher took with respect to ethical decisions (see, e.g., Barton and Lee, 2013: 274–275).

[6] Sandler (2013) is referring to virtual worlds here, yet the claim can easily be extended to the Internet in general.

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

In addition to these considerations, scholars need to acquire relevant information with respect to copyright, and their legal rights to use and disseminate certain data taken from the web. A good starting point is provided by links to 'terms of service', 'privacy policy', or copyright information typically listed at the bottom of webpages. While we do not cover these matters in this article, we encourage scholars to consult (Lipinski, 2008, 2009), who examines "the legal basis for liability on the part of researchers who 'observe' and 'collect' data from online forums such as a listserv, discussion board, blog, chat room and other sorts of web or Internet-based postings" (Lipinski, 2008: 92).

3.2 The multimodal nature of computer-mediated data

Far from being only of a graphological/textual nature, CMC data nowadays is often multimodal. This poses particular challenges for researchers with respect to capturing the dynamics of interaction. Although sociolinguists are primarily interested in linguistic signals, those among us who study communication are likely to want to take into account the complex resources that interactants draw on in the process of creating meaning (see, e.g., Kress and van Leeuwen, 2001; Jones, 2013). Scholars thus need to acknowledge the potentially multi-modal nature of their data and account for including or excluding its study in their research design.

According to Stöckl (2004a: 9), "multimodal refers to communicative artefacts and processes which combine various sign systems (modes) and whose production and reception calls upon the communicators to semantically and formally interrelate all sign repertoires present". Multimodality could thus include a mixture of language, images, sound and/or music, or of sub-types of these modes, e.g., writing (language) mixed with static or dynamic images (images) (see Image 1 in Stöckl, 2004b: 18 for a visualisation of a network of modalities, sub-modalities and characteristics in multimodal texts). A core assumption underlying multimodality "is that language is part of a multimodal ensemble" (Jewitt, 2009: 14). In other words, in multimodal CMC, language is one means through which interlocutors communicate; it may be the most central or important to the interaction at hand, but it often does not occur in isolation (cf. also Norris, 2004). Both the research question and the degree of multimodality of the computer-mediated setting one is interested in may prompt for an analysis of language as embedded within an "ensemble" of different modalities (see, for example, Scollon and Scollon, 2009 for a discussion of "retrospective" and "prospective" views on the relationship between multimodality and language).

In discourse about CMC, the term "mode" is often used to refer to blogs, wikis, social network sites, etc. (see, for example, Murray, 1988; Herring, 2002, 2007). Such modes are mono- or multi-modal, or rather more or less mono- or multi-modal, as modality is best treated as scalar and gradient. The more multi-modal a mode is the more semiotic resources (see Kress, 2010: 80) an individual has at his/her disposal, and the more likely s/he may be to draw on them when communicating via online means. In making use of these resources, an individual will be steered both by the nature of the mode/s in question,

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

as well as by the nature of the interactive context, and by his/her goals and expectations, amongst other issues.

According to Stöckl (2004a: 10), multimodality is not new. Rather "the purely mono-modal text has always been an exception while the core practice in communication has essentially been multimodal all along" (Stöckl, 2004a: 10). We agree entirely that it is dangerous to view multimodality as a new phenomenon which has arisen in connection with (computer-)mediated forms of communication. Yet we follow Ventola et al. (2004) in arguing that

> [t]he various possibilities of combining communication modes in the 'new' media, like the computer and the Internet, have forced scholars to think about the particular challenges of these modes and the way they semiotically function and combine in the modern discourse worlds. (Ventola et al., 2004: 1)

A similar claim is made by (Androutsopoulos, 2013: 237), who argues that "the media-richness of contemporary digital environments increases the impact of multimodality on meaning-making".

Moreover, looking at the development of CMC in the course of the last decades, we believe it is fair to argue that cyberspace has become more multimodal. This increased multimodality can be linked to technological advancements. For example, the virtual world Second Life, which is continuously being developed, added voice as a means for communicating among avatars in addition to chat windows and avatar gestures in 2007, several years after the virtual world was launched. Facebook nowadays combines the possibilities of interacting by means of writing status updates and leaving comments on other people's walls, as well as providing the possibilities of uploading and sharing pictures and videos, writing messages similar to email, and chatting via a chat window. In addition, as stated by Herring (2013), individuals combine different means of communication in creative ways:

↑ 18 19 ↓

> an individual may respond (asynchronously) to a YouTube video either via text comment or video; may chat (synchronously) while playing World of Warcraft via text or voice; and may send text messages (either synchronous or asynchronous, depending on whether the addressee is logged on) and/or speak (synchronously) to an interlocutor over Skype. (Herring, 2013: 16)

Moreover, "each [of these modes] in [the] multimodal ensemble is understood as realizing different communicative work" (Jewitt, 2009: 15; see also Tannen, 2013: 112 for an example of the use of different modes to communicate different messages). Which mode and combination of modes an interlocutor chooses will be intricately tied to processes of meaning-making (Jewitt, 2009: 15).

Evidently, the increase in multimodality within modes has implications for sociolinguistic methodology. The practical challenges of dealing with multimodal data are pointed out by (Herring, 2013: 19), who argues for the "need to devise parallel transcription and visualisation displays for textual and non-textual communication, which differ in a number of respects, including temporality". Since the meaning created through

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

multimodal communication “is seen as *multiplicative* rather than additive”, the “overall result is more than the sum of its parts” (Flewitt et al., 2009: 46, emphasis in original). This prompts scholars to reflect on how the communicative entity being studied can best be analysed as an entity, and not as made up of language which is analysed before any of the other accompanying displays. With respect to the interpretive process of data transcription, this means scholars need to “seek to reveal the *multimodal* basis of a text's meaning in a systematic rather than an ad hoc way” (Baldry and Thibault, 2006: 21, as quoted in Flewitt et al., 2009: 46, emphasis in original). One key starting point in this process is to place “action” in the forefront, and to be guided by the question of “[w]hat is/are the action/s that is/are being taken” (Scollon, 2001: 9, as quoted in Flewitt et al., 2009: 47; cf. also Kress and van Leeuwen, 2001). Subsequent to determining what it is that the authors of computer-mediated communicative acts are doing, one can think of sensible ways in which to record how this doing is being performed, i.e., which modes are being used to create particular meanings. This may mean working with transcription schemes which include lists of actions, lists of modes used to accomplish them, and potentially also visualisations of particular movements, gestures, etc., in addition to more standard information about time, place, number of interlocutors, etc. In their (2008) study of conversational metaphors in Hillary Clinton and Barack Obama's YouTube campaign clips for the democratic presidential nomination, Duman and Locher (2008), for example, use a transcription scheme in which they document both “linguistic” and “visual” information about particular actions; the “visual” transcription column subsumes information about the movements of the camera as well as about the candidate's body posture and physical motions. For Duman and Locher's research both “linguistic” and “visual” information appear at the same level of transcription and are discussed together. Yet there are also research questions which call for visual transcriptions, in which images or photos constitute the main information in the transcription, with text added to the individual frames (see Flewitt et al., 2009 for visualisations of different examples of transcriptions taken from a wide range of previous literature).

However, CMC researchers do not need to reinvent the wheel, since they can draw on long standing traditions from other linguistic fields and communication studies. Herring (2013: 20), for example, lists “social semiotics”, “visual content analysis” and “film studies”, as three possible methods enabling the discourse-analytic study of multimodal data. We might add multimodal conversation analysis (see, e.g. Mondada, 2007, 2012), which has long-standing experience in the transcription of video material. Goodwin (2001: 161, as quoted in Flewitt et al., 2009: 49) states that “[i]n many cases different stages of analysis and presentation will require multiple transcriptions”, as have been in use in conversation analysis for some time. In addition, important innovation in regard to managing multimodal data is provided by advancements made to “Computer Assisted Qualitative Data Analysis Software” (CAQDAS), for example, MAXQDA, NVivo, Atlas.TI, HyperRESEARCH, Ethnograph, ELAN and Kwalitan. Such software facilitates the management of one's data; one can integrate data into a project file, and analyse one's data according to customised annotation schemas. Common to many of the CAQDAS

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

currently on the market is their ability to deal with different types of data. Thus, MAXQDA, for example, allows users to work with rich text, plain text, doc/x files, pdfs, images (JPG and GIF), audio files and video files. Strikingly version 11 of MAXQDA has also introduced "emoticode", which enables the tagging of a wide range of emoticons and symbols (MAXQDA emoticode 2013). Similarly, NVivo, for example, enables users to upload and analyse rich text, and plain text files, pdfs, doc/x files, audio files, video files, spreadsheets, database tables and pictures; the newest version of NVivo advertises its ability to import both YouTube videos and comments, as well as posts from Facebook, discussions from LinkedIn and Twitter tweets. MAXQDA, NVivo and Atlas.ti are all undergoing regular upgrades, with newer versions catering explicitly to the particularities of online data.

For the researcher, this means that contingent upon developing a sensible annotation scheme, s/he can attempt to deal with the wealth of multimodal data by embedding and subsequently annotating different types of data within the same organisational, physical space. Despite initial difficulties unfamiliar users may have with such software, the learning curves are steep. There is also a wide range of support options (FAQ sections and tutorials) provided by the individual software themselves, and even a site, "the CAQDAS Networking Project" which is devoted to providing "practical support, training and information on the use of a range of software programs designed to assist qualitative data analysis" (CAQDAS networking project, 2013).

The wish to not only work with multi-modal data but also to give readers access to it may trigger a movement towards experimenting with other forms of publication, as, for instance, done in visual anthropology, where text publications co-occur with videos and photos. This is nicely illustrated by the journal *Visual Ethnography*, an online (peer-reviewed) journal which focuses on, amongst other things, "the production and the use of images and audiovisual media in the socio-cultural practices" and "the ethnographic representation through audiovisual media and devices (film, photography, multimedia, etc.)" (Visual Ethnography homepage, 2013). Its call, rather than being a call for papers, is a call for papers, videos, photo-essays and reviews. Similar developments are also visible within linguistics, as evidenced, for example, by VARIENG (Studies in Variation, Contacts and Change in English), an open-access peer-reviewed journal published at the University of Helsinki, which "encourage[s] authors to make use of hypertext and multimedia content such as high quality images (full-colour graphs and charts, facsimiles of manuscripts and early published books, maps, etc.), audio and video streaming (dialect samples, video samples) and flash animation (interactive graphics)", as well as making it possible "to make available raw research data in the form of spreadsheet material,
wordlists, concordance tables, etc., and to include powerpoint presentations and freely available software" (VARIENG about the eSeries). Similarly, this very journal – *Discourse, Context & Media* – also encourages the publishing of multi-modal material.

While many sociolinguists may find themselves having to leave their methodological comfort zones when confronted with complex multimodal data, linguists working on CMC are actually ideally equipped to interpret what language, gestures, images, videos, etc. mean within a particular communicative act or exchange. Indeed, the challenges of

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

interpreting multimodal data are not qualitatively different from those all sociolinguists face when attempting to analyse meaning in face-to-face interaction. However, at some stage a researcher needs to make a conscious decision to what extent (if at all) the research question calls for incorporating multi-modality into his or her study design.

3.3 Mixed methodologies and moving between online and offline settings

Researchers have different interests and pursue different research questions when working with online data. While some will use (extracts of) the web as a corpus in itself and are less interested in the individual users (see Section 3.4), others cast their light on practices by particular groups of people. Depending on where a project is positioned on this continuum, sociolinguists will combine methodologies (e.g. an in-depth analysis of language on a blog with interviews of the blogger) and will make use of different types of data (e.g. observations of individuals in both offline and online contexts). When looking for the best research design, these possibilities can be freely combined and the researcher should take a stance within this spectrum and account for the decisions when preparing the results for a readership. In order to shed more light on the available possibilities, we will mention a number of approaches to mixing methodologies in this section.

Starting with data selection, we follow Androutsopoulos (2013: 240), who claims that there are "two main, and in [his and our] view complementary, sites of data collection in new media sociolinguistics": "screen-based" and "user-based". As the labels suggest, whereas the former restricts data production and collection to the screen, the latter focuses on the user and what s/he does through and with language (online, or online and offline). These possibilities are visualised in Table 1. They need to be conceived of as a continuum, ranging from no contact to contact between the researcher and the individuals producing the data s/he is interested in (relation of researcher to source of data), and from online to offline data (resulting type of data). How a researcher positions him/herself largely results from his/her research interests and foci, although as Androutsopoulos (2013: 240) claims, the rightmost column is not actually relevant for computer-mediated linguistic research. For an example, see Barton and Lee's (2013: 168) discussion of the mixed methodology employed in their Instant Messenger research, which involved a combination of online observation and contact with users, resulting in blended data.

Table 1. "Screen-based and user-based data in CMC research" (from Androutsopoulos, 2013: 241).

	Screen-based				User-based
Relation of researcher to source of data	No online observation	Systematic online observation	Online observation and contact to users	Contact to users without online observation	
Resulting type of data	Online data	Online data	Blended data	Offline data	

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

Using mixed methodologies is not particular to CMC, but reflective of a trend in linguistics in general (Angouri, 2010). One way of thinking about mixed methodologies is to use the concept of "remix", as proposed by (Markham, 2013: 64). "Remix" is not a method or a framework, but a metaphor (Markham, 2013: 66), more specifically a metaphor "that seek[s] to challenge how we envision research", and which thus prompts us to view research as "exploratory and creative, a mix of passion and curiosity" (Markham, 2013: 66). As a metaphor it offers "a powerful tool for thinking about qualitative, interpretive research practice" (Markham, 2013: 65). A linguist who uses "remix" becomes a scholar who draws on whatever tools or methods are at his/her disposal, whether they are methods traditionally associated with linguistics or not, in an attempt to answer his/her research question. Moreover, as stated by Markham (2013; see also Kincheloe, 2001 for implications of the notion of "bricolage" for research methodology),

> [t]hinking about digital culture through the lens of remix offers powerful means of resisting the focus on individuals and objects in order to get closer to the flows and connection points between various elements of the media ecology system, where meaning and assemblages and imaginaries are negotiated in relation and (inter)action. […] It allows us to embrace and grapple with complexity (rather than trying to simplify it) by focusing less on methods (as templates to either apply to experiences and organize these experiences into particular categories and structures) and more on meaning as derived from a creative process of inquiry. (Markham, 2013: 71)

This does not mean that individuals and objects are not important; nor does it imply that consideration of which methods may be useful in the course of the research is superfluous. Yet it does encourage us to adopt a processual view towards research methodology, and to treat the issue of how we do something as emergent in relation to the practices we witness interlocutors performing through language.

A look at the research literature illustrates that scholars readily mix methods when studying language use in computer-mediated environments. Many have also started to argue for the importance of such mixing, particularly with respect to "the need to go 'beyond the screen', i.e., to extend the present research focus on log file data" (Androutsopoulos and Beisswenger, 2008), where logfile data can be defined as "the stored, static records of message sequences that have been put into their particular order by a server feature and that are displayed as a message protocol on the users' screens" (Beisswenger, 2008). This can entail, for example, augmenting the study of logfiles through an analysis of the participants' kinesic behaviours while they are sitting at the computer (Beisswenger, 2008), and combining qualitative with quantitative methodologies (for the latter see, for example, Siebenhaar, 2008). This progressive move away from log-file data to inclusion of other types of data can also be tied to the tendency for early research to be more interested in formal aspects of language description than in the practices of particular individuals or groups (as outlined in Androutsopoulos' (2006) discussion of three waves of linguistics research on CMC; see Section 2).

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

Another way of going beyond logfile data is via the use of ethnographic methods, including (participant) observation and interviews. According to Androutsopoulos (2008), combinations between such methods and an analysis of logfile data "have played a somewhat peripheral role in language-focused CMC studies thus far". While this is true, a look at the literature published in the last half a decade also shows that this situation is changing (see, for example, Angouri and Tseliga, 2010; Spilioti, 2011; Lee, 2011; Barton and Lee, 2013; Bolander, in press). Increasingly research is also exploring "CMC as place", i.e., it is focusing on "digital communication as a social process and CMC environments as discursively created spaces of human interaction, which are dynamically related to offline activities" (CMC as place), whereas earlier research tended to prioritise "CMC as text", i.e., to "focus […] on the vast archive of written language provided by the internet" (Androutsopoulos, 2013: 239).[7]

A helpful framework for the exploration of "CMC as place" is provided by Androutsopoulos's (2008), "discourse-centred online ethnography" (DCOE), an approach which combines "the systematic observation of selected sites of online discourse with direct contact with its social actors", and thus one which takes into account the relationship between "digital texts and their production and reception practices". In this framework, the focus is on "everyday life on the Internet" (Androutsopoulos, 2008), which is "theoris[ed] […] as a site where culture and community are formed" (Androutsopoulos, 2008). Methodologically-speaking this means the Internet constitutes the first place where a researcher goes to begin collecting ideas and data for his/her research. During this process, scholars are encouraged to "maintain openness", "use all available technology" and "use observation insights as guidance for further sampling" (Androutsopoulos, 2008); many of these guidelines are reminiscent of what we discussed above in connection with the metaphor of "remix".

Androutsopoulos (2008) also proposes a highly recommended list of six "[p]ractice-derived guidelines for contact with internet actors" in case researchers want to employ interviews when engaging in discourse-centred online ethnography. For example, he argues for the importance of using multiple techniques where they are available, e.g., talking on the phone, engaging in a private chat and meeting face-to-face. This, too, is reminiscent of the processual approach outlined in connection with "remix" above; it is referred to by Androutsopoulos (2008) as "guerrilla ethnography", which he defines as "seizing the opportunity to use whatever methods are possible under the circumstances of each particular context". One such technique entails showing participants their own linguistic practices and discussing them together. This latter technique serves to "reconstruct […] participants' 'lay sociolinguistics', i.e., their awareness of linguistic variability and its social meanings (Niedzielski and Preston, 2000" (Androutsopoulos, 2008). By specifically asking one's informants about their own practices, the linguist can

[7] In his use of the terms "CMC as place" and "CMC as text", Androutsopoulos (2013: 239) is drawing on Milner (2011), who introduced this distinction in connection with online communication studies research.

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

gain access to participants' "own categories and distinctions", which may, or may not confirm the findings previously gained (Androutsopoulos, 2008).

The importance of data triangulation is also underscored by Angouri and Tseliga (2010) in their study of disagreement and impoliteness in two online fora. The authors worked with a set of 200 postings, containing explicit disagreement, and constituting a representative sample with respect to the fora's users, as well as with interview data. The interviews, held via Skype, and reminiscent of those of Androutsopoulos (2008), also entailed having core users "'talk through' the threads and provide background context on the relationships of the people involved" (Angouri and Tseliga, 2010: 65–66). Another method of gaining insight into user perceptions is provided by Bolander (in press) in her study of language and power in a set of eight diary blogs. Combining an analysis of logfiles (48 posts and 841 comments) with written questionnaires (conducted via an online tool), Bolander (in press) asked bloggers to provide an example from their own blog which they felt best illustrated the phenomenon under analysis; e.g., following a question about agreements, the bloggers were asked to link to a post/series of comments in their blog illustrating their viewpoint. Again, this enabled her to gain more in-depth insight into the phenomenon being studied, and to either receive ratification for tendencies she had discovered through discourse analysis, or an alternate image which then led to explorations of possible reasons for such differences.

Such mixing of methodologies, notably an analysis of logfiles with interviews, evidently also means that some of the research takes place in offline contexts, as, for example, with the majority of the semi-structured interviews conducted by Androutsopoulos (2008) in his study of hip-hop in German-based websites. It can also involve the analyst in participant observation of offline as well as online practices. This is demonstrated, for example, by Spilioti's (2011) ethnographic study of SMS interactions between individuals belonging to three young peer-groups in Athens. In order to analyse perceptions towards the appropriateness of closings in text messages, as well as the linguistic characteristics of different types of closings, she worked with SMS texts, as well as participant observation and interviews. The latter took place offline. Spilioti (2011: 71) informs us that she wishes to "move beyond the examination of SMS as log data (i.e., a corpus of randomly collected individual texts) and to probe more into their analysis as contributions to sequences embedded into the participants web of face-to-face and mediated interactions". The methodological treatment of these text messages as embedded in both face-to-face and mediated interactions is intrinsically tied to her aim, namely to "explore the participants' perceptions of politeness and norms of appropriateness in their use of closings" (Spilioti, 2011: 71). Moreover, it is through her mixing of these particular methodologies that she is able to show that the presence/absence and type of closing formulae are intrinsically tied to both "the position of the text in the SMS sequence" and "to the participants' relational concerns" (Spilioti, 2011: 80).

Spilioti (2011: 70) refers to her own methodology as a "blended ethnography", i.e. "a blend of online and offline ethnography, with offline activities receiving equal or even more attention than online ones" (Androutsopoulos, 2008). For Androutsopoulos (2008)

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

this type of ethnography “focuses on the Internet in everyday life, asking how new communications technologies are integrated into the life and culture of a community”; it thus differs from his discourse-centred online ethnography which prioritises “everyday life *on* the Internet, theorising the Internet as a site where culture and community are formed” (Androutsopoulos 2008, emphasis in original). Whereas the former researches linguistic practices in both offline and online spaces, the latter focuses on online spaces; researchers may also conduct research offline, for example, in the form of interviews, but may also mix different methodologies within an online space (e.g., discourse analysis of blog comments with online interviews of bloggers). Again, neither
is per se better than the other; the value of mixed methodologies, and the choice of research setting and decisions about mixed settings should emerge in connection with the design of the study. Borders between offline and online may also not always be relevant: for example, Hine's (2000) virtual ethnography, as stated by Androutsopoulos (2008) “takes as its point of departure an offline event (a court case) and follows the online activities related to that event […]”.

↑ 21 22 ↓

Whereas in this section we predominantly discussed a shift from qualitative discourse-analytic and ethnographic research predominantly working with screen-based data to combinations of screen-based and user-based research, in the next section we will turn to focus on mainly quantitative research.

3.4 Web corpora and annotation

In contrast to the examples discussed in the previous section, the individual user and his or her practices are generally not at the heart of corpus linguistic approaches. Instead, corpora are compiled in order to get at general patterns of naturally occurring linguistic data across speakers and genres, using a quantitative perspective.

If we take the term corpus to mean “a collection of texts or parts of texts upon which some general linguistic analysis can be conducted” (Meyer, 2002: xi, as quoted in Meyer, 2008: 1), there are “two types of corpora that meet this definition: pre-electronic and electronic corpora” (Meyer, 2008: 1). In the case of web corpora, we are referring to the second type of corpus. The history of electronic corpora goes back to the 1960s, and particularly to the 1970s and 1980s, where “we find an explosion in the quantity and variety of texts prepared for analysis by computer”, and “used by a fast increasing number of researchers and for a wide range of purposes” (Johannson, 2008: 33). Much of the theory and methodology underlying corpus linguistics can be applied to the study of web corpora, although web corpora also pose particular challenges and require adaptations to the existing procedure; some of these will be discussed in this section.

While a web corpus is not suitable to every linguistic research endeavour, “there are cases in which the data needed to answer or explore a question cannot be found in a standard corpus”, notably “because the phenomenon under consideration is rare (sparse data), belongs to a genre or register not represented in the corpus, or stems from a time that the corpus data do not cover (for example, it is too new)" (Lüdeling et al., 2007: 7; cf.

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

also Hundt et al., 2007; Fletcher, 2007). In addition, language use online "may be a major source of influence for ongoing language change"; in order to empirically assess this relationship between online language use and language change more generally, more information on the nature of online language use is needed (Hundt et al., 2007: 2). A further important reason for using the web as/for corpus (we will return to the distinction below) concerns the speed of language change. As stated by Hundt et al. (2007: 2), "[i]t takes a long time and considerable financial resources to compile standard reference corpora which, ironically, are quickly out of date when it comes to recent or ongoing change".

While this should by no means imply that the use of web corpora is a simple endeavour, the web provides an amazing wealth of already-compiled data, which can be utilised to answer a whole range of linguistic research questions. Moreover, much of this data has not yet been accessible for large-scale corpus based studies, for example, data on many outer and expanding circle varieties of English (Hundt et al., 2007: 1). Indeed, as Fletcher (2007: 27) states, one of the "powerful reasons to supplement existing corpora or create new ones with online materials" is "linguistic diversity", or more specifically the fact that "languages and language varieties for which no corpora have been compiled are accessible online". In all such instances, "the web seems a good and convenient source of data" (Lüdeling et al., 2007: 7).[8]

The research literature highlights two main approaches, commonly referred to under the headings "web for corpus" and "web as corpus" (see De Schryver, 2002). Following De Shryver (2002, as paraphrased in Fletcher, 2007: 28) the "web for corpus" approach treats the web "as a source of machine-readable texts for corpus compilation", rendering web data static by taking it offline and preserving it. In contrast, the "web as corpus" approach refers to the direct consultation of the web, a process which retains the dynamic nature of the data (Fletcher, 2012). Fletcher (2007: 28; see also Fletcher, 2012) uses the term "web corpus" in an inclusive way to encompass both notions. For both types of web corpus studies, scholars can seek and retrieve information via either "hunting", or "searching directly for specific information", "grazing" or "using ready-made data sets composed and maintained by an information provider" and "browsing" or "coming across useful information by chance" (Fletcher, 2007: 28).

Both "web for corpus" and "web as corpus" approaches obviously have advantages and limitations. Yet the list of methodological limitations in the case of the latter runs longer than those listed for the former. As stated by Hundt et al. (2007: 2–3) "the main problems with this [web as corpus] approach are that we still know very little about the size of this 'corpus', the text types it contains, the quality of the material included or the amount of repetitive 'junk' that it 'samples'". Our lack of an overview of what the ever-growing and ever-changing web consists of makes it difficult to have the requisite knowledge about the characteristics of the broader data pool from which a corpus is sampled, and thus to know

[8] The use of web corpora should not preclude the continued employment of non-web corpora, nor discourage the improvement of existing corpora (see Leech 2007 for treatment of this topic).

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

how what we receive (as a result of our use of crawlers or search-engines) relates to what could potentially have been sampled. A further problem relates specifically to the use of crawlers to search the web for data. Not only can crawlers not "access all web pages because some pages are 'invisible'", "more worrying still – the commercial crawlers have an inbuilt local bias", meaning that the information sampled depends strongly on where the user is accessing the web from; similarly, since crawlers 'learn', information about past activities of the user are stored, and also feed into subsequent searches, skewing results further (Hundt et al., 2007: 3). In addition, as pointed out by Lüdeling et al. (2007: 14), "all search engines perform some sort of normalisation", which may not but can be problematic when it comes to counts of frequency; typically the searches do not take account of capitalisation, they "automatically recognise variants ('white-space' finds *white space*, *white-space* and *whitespace*)" and they "implement stemming for certain languages (as in *lawyer fees* vs. *lawyer's fees* vs. *lawyers' fees* […])". Related to this is the issue of "duplication", which means that the search engine may return multiple duplicate results, giving rise to a need to, for example, perform manual checks of results, a "time-consuming" endeavour which "is hampered by artificial limits that Google imposes on the number of search results returned" (Lüdeling et al., 2007: 14). Finally, the fact that the web is constantly changing prevents linguists from being able to reproduce results (Hundt et al., 2007: 3), and thus also compare and contrast their findings.

Having said this, the direct use of search engines as a means to search the entire web for a particular linguistic phenomenon is the most frequent approach currently adopted by scholars (Lüdeling et al., 2007: 8), and while not without disadvantages, "can also be used fruitfully as a place where we may quickly find back-up for previously more or less anecdotal evidence" (Hundt et al., 2007: 3); moreover, the study of certain subject matters, for example, neologisms (Hundt et al., 2007: 3), can rewardingly be pursued via the use of search engines. There have also been attempts to "'improve on Google by making web data more suited for linguistic work" (Lüdeling et al., 2007: 16), notably via the use of pre-processing systems, i.e., those that "pre-process queries before they are sent to search engines and post-process the results to make them more linguist-friendly", for example, through the use of systems which support substring queries (like WebCorp) or "systems that try to dispense with search engines completely, by building and indexing their own web corpora" (Lüdeling et al., 2007: 16). For further discussion of the web as corpus see also the ACL SIGWAC homepage (2013) – "The Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus"; and Gatto's (2011) paper on "The 'body' and the 'web': The web as corpus ten years on."

↑
22
23
↓

In instances where linguists create their own corpora with selected data from the web, the web is not used as a corpus, but *for* corpus creation. This process can take place in either a random or controlled manner (Lüdeling et al., 2007: 8). In the case of the former procedure, the corpus is automatically constructed via the downloading of pages from the web, for example, through Google's query option (Lüdeling et al., 2007: 8; see Blachman, 2013 for information on how to enter and write queries using Google Guide). The latter entails the creation of a corpus via either "manual or semi-automatic selection of pages

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

downloaded from the web", depending on one's research objectives (Lüdeling et al., 2007: 8). As the authors state, "[t]his procedure is not so different in principle from building a corpus such as the BNC or Brown Corpus, and has the same advantages and disadvantages […]" (Lüdeling et al., 2007: 8).

An alternative way of conceptualising this distinction between automatic or manual and semi-automatic is provided by (Beisswenger and Storrer, 2008: 294), who differentiate between "corpora for general use", on the one hand, and "project-related corpora", on the other. As the labels imply, the selection procedure underlying the former is not determined by specific aims or research desiderata but "established rather as a data pool for the investigation of diverse potential research questions"; whereas the choice of data to feed into the latter results directly from the research questions, and is thus "compiled as an empirical basis for questions in a particular project" (Beisswenger and Storrer, 2008: 294). See Mair (2013) for an example of project-related corpora compiled from three online forums in West Africa and the Caribbean, with the aim of researching World Englishes online.

Both types of corpora may or may not subsequently be annotated, as shown in Table 2. A corpus designed to be project-related can either be comprised of raw or annotated data, as can a corpus designed for general use. Beisswenger and Storrer (2008: 294) maintain that project-related corpora containing raw data constitute the most widely used corpora in contemporary CMC research, both because of the novelty of CMC research as a whole, and because of the fact that there is a lack of "large balanced corpora" containing CMC genres. What characterises Type 1 and 3 corpora in Beisswenger and Storrer's, (2008: 295) view is that their size is "manageable", they are generally only used by the person who compiles them and not by other scholars, who do not have access. In addition, their documentation is often scant.

Table 2. Types of CMC corpora (from Beisswenger and Storrer, 2008: 294).

The corpus originally designed to be	Data elicited for purposes of analysis	
	No	Yes
Project-related	1 Corpora of raw data	3 Annotated corpora
For general use	2 Corpora of raw data	4 Annotated corpora

Type 2 corpora, on the other hand, provide "scholars [with] a data pool for the empirical study of diverse research questions" (Beisswenger and Storrer, 2008: 295). In this sense, they differ from Type 1 corpora. While the authors do not provide further information on the characteristics of such corpora, the list of links next to the examples the authors present suggest that many do not need to be accessed via the person/s who originally compiled the corpora. For example, at the Apache SpamAssasin Project (2013) site, individuals can see roughly 6000 emails from this project. By virtue of this difference, one can also expect that these Type 2 corpora are better documented than those belonging to Type 1.

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

Types 3 and 4 corpora are those which have been annotated, either with more specific (Type 3) or more general (Type 4) purposes in mind. In the case of annotated corpora (Types 3 and 4), "the data are subjected to a coding process, which facilitates both the work with the corpus and the access to and analysis of the data" (Beisswenger and Storrer, 2008: 295; see also Lüdeling et al., 2007). This annotation process is contingent upon prior "develop[ment] [of] appropriate description categories and document grammars, which grasp the linguistic particularities of CMC genres, and modify existent tools for the linguistic preprocessing of speech data […]" (Beisswenger and Storrer, 2008: 301). This is particularly necessary in light of the fact that the particularities of CMC, for example, non-standard spellings and abbreviations, make it impossible to work with "[t]ools developed for the automatic annotation of linguistic data (sentencizers, POS taggers, lemmatizers, chunk parsers)" (Beisswenger and Storrer, 2008: 302). As this makes manifest, careful prior thinking about which labels might be the most appropriate for annotation is needed, particularly if one wishes to produce annotation standards which apply not to just one mode, e.g., blogs, or one type of communication, e.g., synchronous versus asynchronous forms of communication, but to CMC as a whole.

This call for deliberation strongly stems from the need to develop adequate terminology to describe CMC data, since it is debatable whether "it is appropriate to describe conversation structures in synchronous CMC by uncritically using categories such as 'turn', 'turn taking' and 'sequentiality'" (Beisswenger and Storrer, 2008: 301). In instances where particular labels are rejected, appropriate alternatives must be found, for example, by carefully reflecting on the existence of functional overlaps between behaviours linguists know and regularly document in offline conversational and written settings and those pervasive to online contexts. As Beisswenger and Storrer (2008: 301) argue, the fact that "simultaneous backchannel feedback is not possible" in synchronous CMC "does not inevitably mean that in synchronous CMC there are absolutely no functional equivalents to the backchannel behaviour in face-to-face conversations".

The aim to develop appropriate annotation standards for CMC has prompted an upsurge of research dealing with coding. It is even partly responsible for the establishment of the German-based scientific network "Empirikom", a group of 15 researchers from twelve German universities and research institutions, and funded by the German Research Foundation (DFG). In their introductory statement their president Beisswenger (2011) underlines both the need for such annotation standards, and the Network's aim to fill this gap:

> Due to the digital source format of linguistic data on the internet, datasets of internet-based communication are initially simple to collect; however, up until now empirical research on internet-based communication has been lacking well-established formats, standards and description categories for representing and capturing the linguistic and interactional phenomena in new genres […]. Moreover, the existing
> procedures for the automatic processing of linguistic data—procedures that have been developed for standard written texts—need to be adapted to the linguistic characteristics of internet-based writing. (Beisswenger, 2011, emphasis removed)

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

While their focus is on “German internet-based communication”, the challenges the Network outlines and deals with are not language specific, and are thus of relevance to all linguists wishing to use the web as/for corpus. The Network has produced a range of publications on these issues (see Empirikom, 2013).

Linking back to points raised in Section 3.1, we would like to highlight particular challenges that ethics and copyright considerations play when collecting and annotating quantitative computer-mediated data for linguistic analysis. Just as in the case of the classical corpora like the BNC, where copyright issues and approval for recording had to be granted, the same concerns should be raised for large dataset compilations. In the most recent AoIR document published (see above), the authors discuss the challenges the Internet poses to “the fundamental ethics question of personhood” (Markham et al., 2012). This question appears to be easier to answer for some computer-mediated settings than for others. As the authors point out, in settings where the researcher obtains data directly from individuals, for example because s/he interviews them, “we are likely to naturally define the research scenario as one that involves a person” (Markham et al., 2012). Yet in contexts where this collection of data occurs via less direct means, “there may be a tendency to define the research scenario as one that does not involve any persons” (Markham et al., 2012). If one thinks back to the discussion above, it is reasonable to argue that the use of search engines for the automatic compilation of large amounts of data, and even the downloading of all the information on a particular webpage or series of webpages can take our attention away from the person who produced the language we subsequently annotate for our research. This may be exacerbated by the relative lack of demographic information we often have with respect to persons interacting online. In many computer-mediated settings a person may only appear in the form of text+nickname. Yet as Markham et al. (2012) underline, “there is considerable evidence that even ‘anonymised’ datasets that contain enough personal information can result in individuals being identifiable”. This raises the question of whether “the connection between one's online data and his or her physical person enable psychological, economic, or physical, harm” (Markham et al., 2012). As avoidance of harm is one of the cornerstones of ethical reflection, this factor raises clear challenges for researchers wishing to use the web as/for corpus, thereby adding an additional challenge to those mentioned in this section. Having said that, as stated at the beginning of this section, the Internet constitutes a striking amalgamation of different accents, dialects and varieties, the study of which can add to and enrich existing corpus-based research.

4. Concluding remarks

In this article we set out to discuss four methodological issues which are important to performing sociolinguistic research on computer-mediated communication. Our overall focus in this paper was triggered by the fact that there is a paucity of papers dealing with methodology in CMC research from a linguistic angle; those that do exist tend to focus on a particular subject matter, or to approach data and methodology from a particular

(typically discourse-analytic or corpus linguistic) angle. Our choice of the four issues we covered stems from discussions with students and colleagues, and our belief that these matters challenge scholars wishing to study the social facets of language use in online contexts. Throughout the article we stressed that issues of data and methodology cannot be evaluated in a vacuum, as they strongly depend on the research question a linguist wishes to answer. Acknowledging that all research projects go through stages and develop continuously, we want to stress once more that scholars need to keep track of their methodological decisions and be able to explain them to themselves and their readers. This sense of accountability will both improve research design and, ultimately, readability of research reports. So as to embed our discussion of these four issues within a broader context, we also dealt with the diversity of CMC, the history of linguistic research on CMC, and the development from Web 1.0 to Web 2.0, or the "social web".

Many open questions remain, and our wish to highlight a range of challenges relating to the four issues we addressed means we could not deal with all of the issues in the detail they deserve. Thus, we did not, for example, address in depth the repercussions of the relative anonymity of many online environments for the elicitation of core information on social variables like age, gender and social class background. Similarly, more discussion on how to quote from one's data in publications is needed in light of research ethics. Without quotes, linguists cannot exemplify their results and provide support for their arguments, yet quotes can easily be traced via google searches, rendering the practice of anonymisation a pro forma act.

The steady upsurge in publications on methodological and ethical issues in social scientific research is encouraging. The fact that there is an increase in publications on language use in CMC in journals not solely dedicated to CMC (for example in the *Journal of Sociolinguistics* and the *Journal of Pragmatics*) suggests that the time is ripe for increased collaboration and an exchange of ideas between linguists with shared research interests whether they explore them in offline or online settings, or a combination of settings.

Acknowledgements

We would like to thank Greg Myers and the two anonymous reviewers for their helpful and encouraging feedback. In addition, our gratitude goes to our students and to the participants of the summer school on Researching Computer-mediated Communication in Linguistics, in Ascona 2012, who inspired this work.

References

ACL SIGWAC Homepage, ACL SIGWAC Homepage. Available from: http://www.sigwac.org.uk/ (accessed 10.10.13).

Atlas.ti. Available from: http://www.atlasti.com/index.html (accessed 18.11.13).

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

AOIR ethics wiki. Available from: http://ethics.aoir.org/index.php?title=Main_Page (accessed 04.05.13).

Androutsopoulos, Jannis, 2006. Introduction. Special issue on Sociolinguistics and computer-mediated communication. J. Socioling., 10 (4) (2006), pp. 419-438.

Androutsopoulos, Jannis, 2008. Potentials and limitations of discourse-centered online ethnography. Language@internet 5. Available from: http://www.languageatinternet.org/articles/2008/1610 (accessed 06.07.12).

Androutsopoulos, Jannis, 2013. Online data collection. In: Mallinson, Christine, Childs, Becky, Van Herk, Gerard (Eds.), Data Collection in Sociolinguistics: Methods and Applications, Routledge, London (2013), pp. 236-249.

Androutsopoulos, Jannis, Michael, Beisswenger (Eds.), 2008. Special issue on data and methods in computer-mediated discourse analysis. Language@internet 5. Available from: http://www.languageatinternet.org/articles/2008/1609 (accessed 10.05.11).

Angouri, Jo, 2010. Quantitative, qualitative, or both? Combining methods in linguistic research. In: Litosseliti, Lia (Ed.), Research Methods in Linguistics, Continuum, London (2010), pp. 29-45.

Angouri, Jo, Tseliga, Theodora, 2010. You have no idea what you are talking about! From e-disagreement to e-impoliteness in two online fora. J. Politeness Res., 6 (1) (2010), pp. 57-82.

Apache SpamAssasin Project. Available from: http://spamassassin.apache.org/publiccorpus/ (accessed 01.06.13).

Barton, David, Lee, Carmen, 2013. Language Online. Investigating Digital Texts and Practices. Routledge, London.

Baldry, Anthony, Thibault, Paul J., 2006. Multimodal Transcription and Text Analysis. Equinox, London.

Baym, Nancy, 1995. The emergence of community in computer-mediated communication. In: Jones, Steven G. (Ed.), Cybersociety: Computer-Mediated Communication and Community. Sage, Thousand Oaks, pp. 138-163.

Beisswenger, Michael, 2008. Situated chat analysis as a window to the user's perspective: aspects of temporal and sequential organization. Language@internet 5. Available from: http://www.languageatinternet.org/articles/2008/1532 (accessed 03.07.12).

Beisswenger, Michael, 2011. Scientific network: empirical research on Internet-based communication. Available from: http://www.empirikom.net/bin/view/Main/WebHomeEn (accessed 02.02.13).

Beisswenger, Michael, Storrer Angelika, 2008. Corpora of computer-mediated communication. In: Lüdeling, Anke, Kytö, Merja (Eds.), Corpus Linguistics HSK, vol. 29.1, Walter de Gruyter, Berlin, pp. 292-309.

Best, Samuel J., Krueger, Brian S., 2004. Internet Data Collection. Quantitative Applications in the Social Sciences, vol. 141, Sage, London.

Blachman, Nancy. Google Guide Making Searching Even Easier. Part I: Query Input. Available from: http://www.googleguide.com/category/query-input/ (accessed 01.06.13).

Bolander, Brook. Language and Power in Blogs: Interaction, Disagreements and Agreements. Pragmatics & Beyond New Series, vol. 237. John Benjamins, Amsterdam, in press.

CAQDAS Networking Project. Available from: http://www.surrey.ac.uk/sociology/research/researchcentres/caqdas/ (accessed 04.03.13).

Coulmas, Florian, 2005. Sociolinguistics: The Study of Speakers' Choices. Cambridge University Press, Cambridge.

Crystal, David, 2011. Internet Linguistics. Routledge, London.

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

Das, Marcel, P.,Ester, Kaczmirek, Lars, 2011. Social and Behavioral Research and the Internet: Advances in Applied Methods and Research and Strategies Routledge, London.

De Schryver, D.-M., 2002. Web for/as corpus: a perspective for the African languages. Nordic J. Afr. Stud., 11 (2), pp. 266-282.

Duman, Steve, Locher, Miriam A., 2008. So let's talk. Let's chat. Let's start a dialog: an analysis of the conversation metaphor employed in Clinton's and Obama'sYouTube campaign clips. Multilingua, 27 (3), pp. 193-230.

Ethnograph. http://www.qualisresearch.com/ (accessed 18.11.13).

Ess, Charles, the AOIR Ethics Working Committee, 2002. Ethical decision-making and Internet research. Recommendations from the aoir ethics working committee. Available from: www.aoir.org/reports/ethics.pdf (accessed 07.09.10).

ELAN. Available from: http://tla.mpi.nl/tools/tla-tools/elan/ (accessed 19.06.13).

Empirikom. Available from: http://www.empirikom.net/bin/view/Publikationen/WebHome (accessed 17.10.13).

Fletcher, William H., 2007. Concordancing the web: promise and problems, tools and techniques. In: Hundt, Marianne, Nesselhauf, Nadja, Biewer, Carolin (Eds.), Corpus Linguistics and the Web. Rodopi, Amsterdam, pp. 25-46.

Fletcher, William H., 2012. Corpus analysis of the World Wide Web. In: Chapelle, Carol A. (Ed.), The Encyclopaedia of Applied Linguistics, Blackwell, Oxford. (Available from: http://webascorpus.org/Corpus_Analysis_of_the_World_Wide_Web.pdf (accessed 05.07.13))

Flewitt, Rosie S., Hampel, Regine, Hauck, Mirjam, Lancaster, Lesley, 2009. What are multimodal data and transcription? In: Jewitt, Carey (Ed.), The Routledge Handbook of Multimodal Analysis. Routledge, London, pp. 40-53.

Fielding, Nigel, Lee, Raymond M., Blank, Grant (Eds.), 2008. The SAGE Handbook of Online Research Methods. Sage, London.

Gatto, Maristella, 2011. The 'body' and the 'web': the web as corpus ten years on. ICAME J., 35, pp. 35-58. (Available from: http://icame.uib.no/ij35/Maristella_Gatto.pdf (accessed 17.10.13))

Goodwin, Charles, 2001. Practices of seeing visual analysis. An ethnomethodological approach. In: Van Leeuwen, Theo, Jewitt, Carey (Eds.), Handbook of Visual Analysis. Sage, London, pp. 157-182.

Herring, Susan C., 2002. Computer-mediated communication on the Internet. Ann. Rev. Inf. Sci. Technol., 36, pp. 109-168.

Herring, Susan C., (Ed.), 1996a. Benjamins, Amsterdam.

Herring, Susan C., 1996b. Introduction. In: Herring, Susan C. (Ed.), Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives. Benjamins, Amsterdam, pp. 1-13.

Herring, Susan C., 2007. A faceted classification scheme for computer-mediated discourse. Language@internet 4. Available from: http://www.languageatinternet.org/articles/2007/761/ (accessed 10.07.10).

Herring, Susan C., 2013. Discourse in Web 2.0: familiar, reconfigured, and emergent. In: Tannen, Deborah, Trester, Anna M. (Eds.), Georgetown University Round Table on Languages and Linguistics 2011: Discourse 2.0: Language and New Media. Georgetown University Press, Washington, DC, pp. 1-25.

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

Herring, Susan C., Stein, Dieter, Virtanen, Tuija (Eds.), 2013a. Handbooks of Pragmatics, vol. 9. Mouton De Gruyter, Berlin and New York.
Herring, Susan C., Stein, Dieter, Virtanen, Tuija, 2013b. Introduction to the pragmatics of computer-mediated communication. In: Herring, Susan, C., Stein, Dieter, Virtanen, Tuija (Eds.), Pragmatics of Computer-mediated Communication. Handbooks of Pragmatics, vol. 9, Mouton de Gruyter, Berlin and New York.
Hine, Christine, 2000. Virtual Ethnography. Sage, London.
Hine, Christine, 2000. Virtual Methods. Issues in Social Research on the Internet. Bergs, London.
Hooley, Tristram, Marriott, John, Wellens, Jane, 2012. What is Online Research: Using the Internet for Social Science Research. Bloomsbury Academic, London.
Hundt, Marianne, Nesselhauf, Nadja, Biewer, Carolin (Eds.), 2007. Corpus Linguistics and the Web. Rodopi, Amsterdam.
Hunsinger, Jeremy, Klastrup, Lisbeth, Allen, Matthew (Eds.), 2010. International Handbook of Internet Research, Springer, London.
Hymes, Dell, 1974. Foundations in Sociolinguistics: An Ethnographic Approach. University of Pennsylvania Press, Philadelphia.
HyperRESEARCH. Available from: http://www.researchware.com/products.html (accessed 18.11.13).
Jewitt, Carey, 2009. What is multimodality? In: Jewitt, Carey (Ed.), The Routledge Handbook of Multimodal Analysis. Routledge, London, pp. 14-27.
Jones, Rodney, 2013. Multimodal discourse analysis. In: Chapelle, Carol E. (Ed.), The Encyclopedia of Applied Linguistics. Wiley-Blackwell, Oxford, UK. http://dx.doi.10.1002/9781405198431.wbeal0813.
Jones, Steve, 1999. Doing Internet Research: Critical Issues and Methods for Examining the Net. Sage, London.
Johns, Mark D., Chen, Shing-Ling, Jon Hall, G., 2004. Online Social Research. Methods, Issues and Ethics. Digital Formations, vol. 7. Lang, New York.
Johannson, Stig, 2008. Some aspects of the development of corpus linguistics in the 1970s and 1980s. In: Lüdeling, Anke, Kytö, Merja (Eds.), Corpus Linguistics HSK, vol. 29.1. Walter de Gruyter, Berlin, pp. 33-53.
Jucker, Andreas, 2009. Speech act research between armchair, field and laboratory: the case of compliments. J. Pragmat., 41.8, pp. 1611-1635.
Jucker, Andreas H., Dürscheid, Christa, 2012. The Linguistics of Keyboard-to-screen Communication. A New Terminological Framework. Linguistik Online 56. Available from: http://www.linguistik-online.org/56_12/juckerDuerscheid.html (accessed 14.03.13).
Kincheloe, Joe L., 2001. Describing the bricolage: conceptualizing a new rigor in qualitative research. Qual. Inq., 7.6, pp. 679-692.
Kress, Gunther, 2010. Multimodality: A Social Semiotic Approach to Contemporary Communication. Routledge, London.
Kress, Gunther, van Leeuwen, Theo, 2001. Multimodal Discourse: The Modes and Media of Contemporary Communication. Arnold, London.
Kozinets, Robert, 2009. Netnography: Doing Ethnographic Research Online. Sage, London.
Kwalitan. Available from: http://www.kwalitan.nl/engels/ (accessed 18.11.13).
Landert, Daniela, Jucker, Andreas H., 2011. Private and public in mass media communication. From letters to the editor to online commentaries. J. Pragmat., 43, pp. 1422-1434.

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

Lazar, Jonathan, Feng, Jinjuan Heidi, Hochheiser, Harry, 2009. Research Methods in Human–Computer Interaction. Wiley, Chichester.

Lee, Carmen, 2011. Micro-blogging and status updates on facebook: texts and practices. In: Thurlow, Crispin, Mroczek, Kristine (Eds.), Digital Discourse. Language in the New Media, Oxford University Press, Oxford, pp. 110-128.

Leech, Geoffrey, 2007. New resources, or just better old ones? The Holy Grail of representativeness. In: Hundt, Marianne, Nesselhauf, Nadja, Biewer, Carolin (Eds.), Corpus Linguistics and the Web. Rodopi, Amsterdam, pp. 133-150.

Lim, Hwee, Ling, Sudweeks, Fay (Eds.), 2013. Innovative Methods and Technologies for Electronic Discourse Analysis. IGI Global, Hershey.

Lindlof, Thomas R., 1995. Qualitative Communication Research Methods (First ed.). Sage, London.

Lindlof, Thomas R.,Taylor, Bryan C., 2002. Qualitative Communication Research Methods (Second ed.). Sage, London.

Lipinski, Tomas A., 2008. Emerging legal issues in the collection and dissemination of Internet-sourced research data: Part I, basic tort law issues and negligence. Int. J. Internet Res. Ethics, 1 (1), pp. 92-114.

Lipinski, Tomas A., 2009. Emerging legal issues in the collection and dissemination of Internet-sourced research data: Part II, tort law issues involving defamation. Int. J. Internet Res. Ethics, 2 (1), pp. 57-72.

Locher, Miriam A., 2014. Electronic discourse. In: Schneider, Klaus Peter, Barron, Anne (Eds.), The Pragmatics of Discourse. Mouton, Berlin.

Lüdeling, Anke, Evert, Stefan, Baroni, Marco, 2007. Using web data for linguistic purposes. In: Hundt, Marianne, Nesselhauf, Nadja, Biewer, Carolin (Eds.), Corpus Linguistics and the Web. Rodopi, Amsterdam, pp. 7-24.

Mair, Christian, 2013. Corpus approaches to the New English Web: post-colonial diasporic forums in West Africa and the Caribbean. (Available from: http://journals.covenantuniversity.edu.ng/jls/published/Mair2013.pdf (accessed 19.10.13)). Covenant J. Lang. Stud., 1.1, pp. 17-31.

Mann, Chris, Stewart, Fiona, 2000. Internet Communication and Qualitative Research: A Handbook for Researching Online. Sage, London.

Mallinson, Christine, Childs, Becky, Van Herk, Gerard (Eds.), 2013. Data Collection in Sociolinguistics: Methods and Applications. Routledge, London.

Markham, Annette, 1998. Life Online. Researching Real Experience in Virtual Experience. AltaMira, California.

Markham, Annette, 2011. Internet research. In: Silverman, David (Ed.), Qualitative Research: Theory, Method, and Practices (Third ed.). Sage, London.

Markham, Annette, 2013. Remix cultures, remix methods: reframing qualitative enquiry for social media contexts. In: Denzin, Norman, K., Giardina, Michael D. (Eds.), Global Dimensions of Qualitative Enquiry. Left Coast Press, California, pp. 63-81.

Markham, Annette, Baym, Nancy (Eds.), 2009. Internet Enquiry: Conversations about Method, Sage, London.

Markham, Annette, Buchanan, Elizabeth, with contributions from the AOIR Ethics Working Committee, 2012. Ethical Decision-making and Internet Research 2.0: Recommendations from

the Aoir Ethics Working Committee. Available from: www.aoir.org/reports/ethics2.pdf (accessed 05.04.13).
MAXQDA emoticode. Available from: http://www.maxqda.com/products/maxqda11/emoticode (accessed 15.04.13).
MAXQDA. Available from: http://www.maxqda.com/ (accessed 18.11.13).
Meyer, Charles F., 2002. English Corpus Linguistics: An Introduction. Cambridge University Press, Cambridge.
Meyer, Charles F., 2008. Origins and history of corpus linguistics—corpus linguistics vis-à-vis other disciplines. In: Lüdeling, Anke, Kytö, Merja (Eds.), Corpus Linguistics HSK, vol. 29.1, pp. 1-14.
Milner, Ryan M., 2011. The study of cultures online: Some methodological and ethical tensions. Grad. J. Soc. Sci., 8.3, pp. 14-35.
Mondada, Lorenza, 2007. Multimodal resources for turn-taking: pointing and the emergence of possible next speakers. Discourse Stud., 9, pp. 194-225.
Mondada, Lorenza, 2012. Coordinating action and talk-in-interaction in and out of video games. In: Ayaß Ruth, Gerhardt, Cornelia (Ed.), The Appropriation of Media in Everyday Life. John Benjamins, Amsterdam, pp. 231-270.
Murray, Denise, 1988. The context of oral and written language: a framework for mode and medium switching. Lang. Soc., 17, pp. 351-373.
Niedzielski, Nancy A., Preston, Dennis R., 2000. Folk Linguistics. Mouton de Gruyter, Berlin and New York.
Norris, Sigrid, 2004. Analyzing Multimodal Interaction: A Methodological Framework. Routledge, London.
NVivo. Available from: http://www.qsrinternational.com/products_nvivo.aspx (accessed 18.11.13).
ÓDochartaich, Niall, 2002. Internet Research Methods: A Practical Guide for Students and Researchers in the Social Sciences. Sage, London.
Paccagnella, Luciano, 1997. Getting the seats of your pants dirty: strategies for ethnographic research on virtual communities. J. Comput.-mediated Commun., 3.1. Available from: http://jcmc.indiana.edu/vol3/issue1/paccagnella.html (accessed 10.02.13).
Sandler, Randall, 2013. Vignette 3d. Real ethical issues in virtual world research. In: Mallinson, Christine, Childs, Becky, Van Herk, Gerard (Eds.), Data Collection in Sociolinguistics: Methods and Applications. Routledge, London, pp. 58-62.
Scollon, Ron, 2001. Mediated Discourse: The Nexus of Practice. Routledge, London.
Scollon, Ron, Scollon, Suzie Wong, 2009. Multimodality and language: a retrospective and prospective view. In: Jewitt, Carey (Ed.), The Routledge Handbook of Multimodal Analysis. Routledge, London, pp. 170-180.
Siebenhaar, Beat, 2008. Quantitative Approaches to Linguistic Variation in IRC: Implications for Qualitative Research. Language@internet 5. Available from: http://www.languageatinternet.org/articles/2008/1615 (accessed 05.06.12).
Spilioti, Tereza, 2011. Beyond genre: closings and relational work in text-messaging. In: Thurlow, Crispin, Mroczek, Kristine (Eds.), Digital Discourse: Language in the New Media. Oxford University Press, Oxford, pp. 67-85.
Spilioti, Tereza, Georgakopoulou, Alexandra. (Eds.) 2013. Routledge, London, in preparation.
Squires, Lauren, 2010. Enregistering internet language. Lang. Soc., 39.4, pp. 457-492.

Published in: https://doi.org/10.1016/j.dcm.2013.10.004

Stöckl, Hartmut, 2004a. In between modes: Language and image in printed media. In: Ventola, Eija, Cassily, Charles, Kaltenbacher, Martin (Eds.), Perspectives on Multimodality. Benjamins, Amsterdam, pp. 9-30.
Stöckl, Hartmut, 2004b. Typographie: Gewand und Körper des Textes—Linguistische Überlegungen zu typographischer Gestaltung. Z. Angew. Linguist., 41, pp. 5-48.
Tannen, Deborah, 2013. The medium is the metamessage. Conversational style in new media interaction. In: Tannen, Deborah, Trester, Anna Marie (Eds.), Discourse 2.0. Language and New Media. Georgetown University Round Table of Language and Linguistics, Georgetown University Press, Georgetown, pp. 99-118.
Tannen, Deborah, Trester, Anna Marie, 2013. Discourse 2.0. Language and New Media. Georgetown University Round Table of Language and Linguistics. Georgetown University Press, Georgetown.
Tagg, Caroline, Seargeant, Philip, 2013. The Language of Social Media: Communication and Community on the Internet. Palgrave Macmillan, New York, in press.
Thurlow, Crispin, Mroczek, Kristine (Eds.), 2011. Digital Discourse: Language in the New Media. Oxford University Press, Oxford.
VARIENG about the eSeries. Available from: http://www.helsinki.fi/varieng/journal/about.html (accessed 05.03.13).
Visual Ethnography homepage. Available from: http://www.vejournal.org/?journal=vejournal (accessed 05.03.13).
Ventola, Eija, Charles, Cassily, Kaltenbacher, Martin, 2004. Introduction. In: Charles, Cassily, Ventola, Eija, Kaltenbacher, Martin (Eds.), Perspectives on Multimodality. Benjamins, Amsterdam, pp. 1-8.
Wardhaugh, Ronald, 2002. An Introduction to Sociolinguistics, 4th ed. Blackwell, Oxford.
Yus, Francisco, 2011. Cyberpragmatics: Internet-Mediated Communication in Context. Benjamins, Amsterdam.
Zappavigna, Michele, 2012. Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web. Continuum, London.

Published in: https://doi.org/10.1016/j.dcm.2013.10.004