Beyond monomodal spoken corpora

Svenja Adolphs[1] (The University of Nottingham), Dawn Knight (Newcastle University) and Ronald Carter (The University of Nottingham)

1. Introduction

The case study described in this chapter involves the incorporation of 'non-linguistic' data streams in spoken corpus analysis. Here new possibilities are outlined for how we may relate to use of language measurements of different aspects of context gathered from multiple sensors (especially, for example, of position, movement and time). Such alternative data streams are seen to be a means of generating valuable insights into discourse, by exploring the extent to which everyday language and communicative choices can be determined by different spatial, temporal social and experiential contexts and can embrace a variety of different non-linguistic sources of data.

This chapter reports on one very preliminary case study, the British Art Show study. Its incipient character means that no great claims can be made for the results drawn from this study, instead the emphasis of this paper is on the how, that is, on what kinds of tools and processes that may be needed in order to begin to undertake appropriately accented analyses of non-linguistic data sources in corpus analysis. If we do not take these steps, we may remain in a world in which we never move beyond the confines of the orthographic word and the single written text.

This chapter begins with an overview of the key challenges faced in the representation of discursive contexts in current, typically monomodal, corpora. It then proceeds to report on a case study which examines the potential for capturing and representing clearer and more accurate records of the dynamic discursive contexts that we encounter in our everyday lives. The case study focuses on video, audio and location based data collected from participants visiting 3 galleries during the British Art Show. In order to highlight the importance of conceptualising this notion of 'context', a corpus-based analysis is carried out in this chapter, examining the use of discourse from a macro level (i.e. 'beyond the text', considering the more socio-ideological and situational factors influencing language choice and use) and a micro (i.e. word-by-word, sentence and text-by-text level – with a specific focus on deictic marker use) perspective.

2. Background

The integration of the Internet with social computing, and now with mobile and ubiquitous computing, is transforming the texture of our lives in everything from games to journalism. This is driving the emergence of new forms of converged pervasive media in which the public contributes as well as consumes content 'anytime and anywhere', making mobile and ubiquitous computing ever more deeply interwoven into our daily lives. Time and space shrink from chronological measurement of time (e.g. by clocks) and space (e.g. travel by land or sea) to a world which is always 'on'; in this world, things happen or appear to happen at the same time (simultaneity), and these replace things that were previously perceived as happening in sequence (linearity). Words such as 'ubiquitous' and 'pervasive' are beginning to be increasingly collocated with 'computing'.

[1] School of English Studies, University Park, University of Nottingham, NG7 2RD *Correspondence to*: Svenja Adolphs, *e-mail*: Svenja.Adolphs@nottingham.ac.uk

The term *ubiquitous computing* (see Weiser, 1991) refers to a movement away from the workplace and the desktop PC to embed computing in the physical environment and the individual user in the many varied settings of everyday life. Ubiquitous computing is a diverse enterprise already moving beyond the research lab and 'into the wild' to explore its potential within the home, health care, environmental monitoring, education, tourism, large-group multi-player gaming, and other everyday settings and activities as well.

Ubiquitous computing embraces the use of mobile devices, including mobile and smart phones, iPads and tablets, e-books and all devices that are supported by WiFi and related forms of connectivity. It is also a term used to embrace all devices that are location-based and which enable points of connection through time and space. The term also refers to sensor-based computing wearable computing etc. (e.g. internet-enabled watches and Google Glass), and combines these with diverse interaction mechanisms including audio, video, text, and virtual reality to make possible new forms of computer-mediated relationship between people and their physical environment.

The actual and potential diversity of ubiquitous computing introduces not inconsiderable levels of complexity into the effort to understand interaction within these emerging environments. These environments are interactionally varied in nature, which is to say that people interact with one another via diverse interaction mechanisms rather than the same ones that remain forever constant. For example, one person might interact via a GPS-enabled mobile device and audio messaging, while another might respond via both an avatar in a virtual world and text messaging.

Interaction in ubiquitous computing environments may also be massively distributed, with the two parties in the above situation being located in different countries and being but two of many interacting parties operating within the environment at the same time. The nature of interaction in ubiquitous computing environments means that interaction is always to some degree *asymmetrical* and *fragmented*. It is asymmetrical in the sense that people interact via different and differentially distributed interaction mechanisms and this in turn fragments interaction. As a result of such asymmetry and fragmentation, people are obliged to reconcile the various fragments of interaction at their disposal to engage in collaborative activities.

Thus, successful interaction within these environments depends on the reconciliation of various fragments of interaction. The challenge for better understanding of what is involved is to unpack what this reconciliation turns upon and consists of as a social enterprise. However, the nature of ubiquitous computing makes this extremely difficult. Embedding computing in mobile devices, exploiting invisible sensing systems (e.g., GPS or WiFi) alongside them, connecting distributed physical and virtual environments together through them, and employing a diverse range of interaction mechanisms amongst distributed parties, raises real challenges as to how corpus linguistics and, indeed, the social sciences more generally might gather data and analyse discourse use-in-context in such complex settings.

This is especially so in the case in corpus linguistics where the main aim is to gather data for analysis which enables analysis of discourse in a variety of different contexts (discourse is defined here as language-in-use in digital contexts, observed from both a micro (i.e. word-by-word, sentence and text-by-text level) and macro (i.e. 'beyond the text', considering the more socio-ideological factors influencing language choice and use) perspective). The challenge is to ensure that rich data is captured, stored and made available interrogation but also to simultaneously ensure that the language data is aligned with the different data streams that have been collected in the kinds of 'ubiquitous' environments described above and that will have been almost inevitably

obtained as a result of the fragmented interactions. The challenge is, however, an important one if researchers are to better capture and understand the way in which language is used in these environments.

Describing spoken discourse is not simply a matter of collecting spoken data; it is, crucially, a matter of collecting (and accurately recording and preserving) spoken data as users of the language interact with other non-verbal data streams and then it is a matter of finding appropriate mechanisms for measuring the extent to which these other data streams determine the nature of the language that is used. Process underscores a greater understanding of the relationship between language and context.

3. Corpora and context

A key challenge faced in applied linguistics is to systematically understand how our language varies from one context to another according to changes in environment, according to different channels of communication, and according to different social contexts of human interaction. As Adolphs noted (2008: 6), 'spoken discourse is collaborative in nature and as such is more fluid and marked by emerging and changing orientations of the participants[than spoken discourse]', so it something that is particularly aligned with and affected by the context in which it is use (making this concept of context particularly relevant to studies of discourse).

Capturing, encoding and even defining context is difficult as, to a certain extent, 'the scope of interactional context is indefinite and infinite because each context is embedded into its own context that is embedded in its own context and so on', this creates a theoretical 'situation of infinite contextual regress' (Kopytko, 2003: 50). This suggests that it is somewhat impossible to fully capture the intricacies of context as, by its own definition, it is a phenomenon that is so abstract and indefinite that it does not lend itself to such definition. It is understood that current methodologies in language data analysis need to be extended to include an integrated exploration of verbal and non-verbal patterns of interaction in context.

Conventionally, contextual categories in applied linguistic research have been static in nature and focused predominantly on culturally recognised activities, such as 'business meetings', 'transactional discourse' and so on. While participants in a conversation often do make reference to such categories, 'other possible features of context that may influence linguistic choices remain largely underexplored' (Knight, 2011: 185). The affordance of new technologies have, however, recently begun to provide us with the means for capturing the subtleties of context, something that this chapter explores. From an applied linguistic perspective the analysis of computer-mediated communication (see Condon and Cech, 1996; Ko, 1996 and Herring, 2007) offer a way of gaining a better understanding of the kinds of social behaviours and relationships that are formed through the use of language in this environment and of how the increased interleaving of digital media with everyday life impact on our ability to project and manage multiple identities.

New forms of communication naturally engender new forms of language. The processes of communication in these media take place in new digital and remote environments that entail different co-constructions of interpersonal relations, different performances of the self and new adaptations and affordances in the use of language. Recent changes in the use of language have been noticed as a result of the internet and email communication (Baron, 2000) and there have been studies of chatrooms (Iwasaki and Oliver, 2003; Jepson, 2005) and on-line games (Crystal, 2004, 2011; Von Ahn, 2006; Thorne, 2008) for example.

So far research in this area has mainly concentrated on individual channels of computer-mediated communication within a single social or locational context, and has neglected the increasing use of multi-channel interaction in this area. The focus has also been on mainly *static* rather than on *dynamic* contexts. The simultaneous use of face-to-face communication and pervasive computer-mediated communication is becoming an increasingly key element of everyday discourse that is contextually dynamic and it is therefore vital to develop ways of analysing the interplay between the two modes. A corpus of interactions is needed in order to reveal significant patterns in this material.

2.3. Corpora and context

A key finding in spoken corpus analysis is that naturally occurring interaction is fragmented in nature, with participants orienting themselves to a range of transient goals throughout the course of the interaction. Linguistic descriptions of such discourse therefore have to be able to account for this *dynamic* nature of context.

Adolphs and Carter (2013) and Knight (2011) have argued that corpus evidence is needed to begin better to explicate the relationship between language and context and in doing so to provide some basis for renewed discussion about the extent to which text-external elements are invoked in our interpretations of language-in-use. However, the lack of databases and frameworks for representing such data means that they are largely under-explored.

The existence of 'system logs' are an important step in this endeavour. *System logs* refers to computational recordings of interaction from within ubiquitous computing environments; audio and text messages that people send to one another, the digital recordings of avatar movement in virtual environments, the connection and disconnection of invisible sensing systems, and the capture of locational data, to give just some examples. These digital records move beyond the current focus on capturing, synchronizing, and analyzing time-based data to focus on capturing and representing multi-dimensional data that spans both physical and digital domains, cutting across time and space. The preliminary research described in this chapter outlines some first steps for developing new means of recording and representing these kinds of multi-dimensional data with a view of discussing how, potentially, they may be effectively utilised in corpus-based studies.

The particular focus of the following case study is on the interpersonal and the interactive in communication generated in text logs that record the presence of data streams that run in parallel to language use as that language use changes from one context to another and as speakers move between locations. The case study provides an example of language operating in relation to multi-channel media, to different data streams and within dynamic contexts.

4. Case Study: The British Art Show (BAS)

4.1. Introducing the BAS

This case study involved recording the experience of three pairs of people attending the British Art Show 7, a Hayward Touring contemporary art collection. BAS showcased works from 39 British artists and artists groups across 3 art galleries across Nottingham city centre (Castle gallery (A), Nottingham Contemporary (B) and the New Art Exchange (C)), as seen in figure 1.

This show ran from 23rd October 2010 to 9th January 2011. As part of this study, researchers captured the participants' interactions when planning their routes through the city/show sites, their physical movement around the city and their uses of language in changing, locational contexts.

Figure 1: Art galleries involved in the British Art Show 7.

The theme for the show, across all coordinated sites, was 'In the Days of the Comet'. To view the complete show visitors were encouraged to visit all locations, although no specific 'recommended' order for visiting the shows was provided by the curators. However, as an incentive for visiting all three sites, visitors were given promotional fliers which not only contained information about the tour, but also spaces to collect visitor 'stamps' which allowed them to visit the 'free' galleries (New Art Exchange and Nottingham Contemporary), get their fliers stamped and then use these fliers as a free entrance ticket for the, usually fee paying, Castle gallery and grounds.

On this basis it was expected that visitors would possibly visit the free sites first before visiting the castle, although this was not always the case as visitors may have had other passes for the castle, had chosen to pay instead, or just have decided not to visit all the sites. Therefore the order the sites were visited in, the time period over which they were visited and the total number of sites visited by individuals was potentially highly variable; however, this design was a deliberate strategy to create a dataset which was as naturally derived as possible. Each participant had at least some intention of visiting the show independently, but did not necessarily plan to go to all three sites. It was the intention of the research team to ensure that visitors had the opportunity to pick and choose sites as required. Out of the three pairs only one did not visit all three sites but this was owing to the fact the final site was located some distance away from the city centre and both participants were beginning to feel tired after having already spent four hours recording. As part of this study, the following data was captured:

- Verbal interactions throughout the planning phases (i.e. how pairs collaborated to discuss the routes they would take, how they would move between sites etc.).
- Language use in changing geographical contexts, from the starting point (a coffee shop); through the city to the galleries; in the galleries; on the tram etc.

- Variations in language when alone, both with their partners and with external members of the research team.

The subsequent analyses of the data were intentionally corpus-driven, with no specific research questions or hypotheses devised in advance of the data collection phase (i.e. conducting general analyses of corpus data and utilizing the results as a means of formulating more specific lines of enquiry/more targeted research questions). The reason for opting for this approach and not a corpus-based one (where the linguist aims to utilize corpus data to answer a particular question/line of enquiry) is that this is a very experimental project, one which is likely to throw up a range of different practical and methodological challenges, rather than providing conclusive answers/responses. A key challenge faced, for example, once data was collected was how to align the different modes of information to enable us to make 'sense' of the data. Following the alignment, basic word frequency lists were created to determine whether any interesting patterns of word usage emerged across the different speakers and locations over time. From this, a more structured, corpus-based approach to analysis can carried out.

4.2. Participants, devices and the 'Field Work Tracker'

The following pairs of participants were recorded in this study:

No	Date	Participant information						Order Visited		
		Ref.	Gender	Age	Occupation	Nationality	Relationship	A	B	C
BAS. 1	14/12/10	<$M3>	Male	50s	Artist	British	Partners	2nd	3rd	1st
		<$F2>	Female	40s	Unknown	Canadian				
BAS. 2	7/1/11	<$M4>	Male	20s	Student	British	Friends/ Colleagues	3rd	2nd	1st
		<$M5>	Male	20s	Student	British				
BAS. 3	8/1/11	<$F3>	Female	50s	Secretary	British	Mother - Daughter	1st	2nd	N/ A
		<$F4>	Female	20s	Unknown	British				

Table 1: Participants recorded for the British Art Show study.

On average each pair took at least 3 hours to visit the show which amounted to over 10 hours of audio data being collected in total (audio was recorded continuously on Sony 4GB SX Series Linear PCM digital voice recorders throughout the study). As seen in table 1, the order in which the galleries were visited differed from one pair to the next and indeed the final pair failed to visit all three of them. In spite of this, a large amount of data was still collected for the final pair, as discussed below.

Each pair was given an iPhone on which to run the Field Work Tracker, an application that continuously records the phone's GPS position in a time-stamped log. Users can take photographs or movies, record audio, and make textual notes. Each of these media items appears in the log with a timestamp and location. A screenshot of this application can be seen in figure 2:

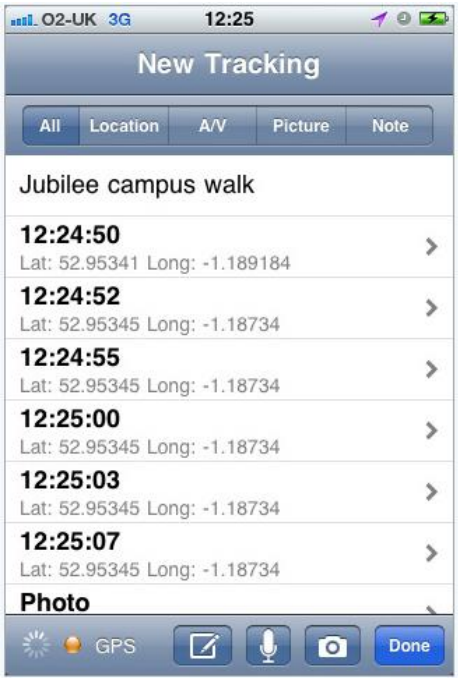

Figure 2: The Field Work Tracker application.

The Field Work tracker is a bespoke mobile application which creates detailed location-based logs. It was developed to support the capture for qualitative analysis of fieldwork data, providing a cheap and simple multi-function recorder which allows for the automated synchronisation of data (see Knight et al., 2010). The Fieldwork Tracker is compatible with the iPhone and the iPod Touch, thus allowing data capture with a single user device.

The Field Work tracker was designed to be specifically compatible with the Digital Replay System (DRS - see Greenhalgh et al., 2007 and French et al., 2006). DRS is a freeware tool which was built to support the annotation and analysis of multimodal linguistic *corpora*, and/or the requirements of corpus-based querying and analysis. It allows users to construct some form of time-stamped transcripts, align these with video, audio and other forms of digital records, and to encode features of interest within and across each stream of data, within individual coding tracks. These coding tracks are tied, by time, to the video and transcript.

Captured logs can be uploaded either by email or through a dropbox.com account straight into DRS and as part of this process, the device time is linked with the computer, which in turn is linked with apple.time (this is Apple's definition of chronological time on which all of their devices, applications and systems are synchronised with). This means that all the data collected from multiple devices is systematically synchronised.

As far as possible the Field Work Tracker is set to run continuously in the background without any discernible impact on the use of the device by participants, although at certain intervals throughout the recording process researchers were required to intervene as on occasion the application would stall or fail. Participants were encouraged to use the Field Work Tracker to take photos, record notes and audio recording as desired, although they were also provided with a video camera (and dictaphone) between each pair so that this additional data could also be sourced.

Participants were recruited by word-of-mouth and eventually 6 participants from a range of different backgrounds and ages were recruited. Participants were instructed to meet the researchers in the city centre in order that they could be appropriately

prepared. As far as possible the same instructions were given across the groups. Participants were informed that researchers were interested in collecting language in location, so the use of language across the different sites they would be moving between. It was emphasised that their reactions to given works in the show were not the main purpose (although some attention may be given to this during the analysis of the data). This reassurance was provided in an effort to make them feel as at ease as possible with the recordings as well as to relieve any anxieties that they were being 'tested' on their reactions to the show.

3.3. Questionnaires

Some basic art-related questions were asked at the start of the session and a more detailed discussion was held post hoc. Participants were then shown how to operate the equipment and were subsequently given the chance to ask any questions of the research team. Relevant consent forms were signed by each participant and full permission to use of their data (including biographical information) was provided in advance of the collection period.

The researchers adopted the role of passive bystanders throughout the recording process, following the participants as they moved from venue to venue and waiting in the coffee bar or entrance as the participants moved around within the galleries. At times the researchers took short video clips and photos, but were mainly available to receive any queries throughout the recording sessions and/or to check that the software was still functioning adequately.

3.4. Transcription Conventions

Problems associated with poor battery life, losses of GPS signal and other uncontrollable factors meant that complete accounts of experiences were collected for only one person in each pair, while some of the data for the other person was partial and incomplete. Parts of the journeys were not recorded and some of the audio records were inaudible. For this reason we decided, in the first instance, to transcribe, synchronise and align only data recorded from those individuals in a pair whom had assembled the best record of their experience. The 'best record' was defined as in terms of the largest number of photos and video recordings taken, the most complete GPS logs recorded and the longest and most detailed audio accounts recorded (these were $M4, $F2 and $F3). It was decided that recordings from the other participants ($M5, $M3 and $F4) could be used to supplement this core dataset during the analysis phase.

The audio files were recorded in Transana[2] using the same transcription conventions as used for the CANCODE[3] corpus (see Adolphs, 2008: 137-8 for full details of these conventions). A summary of these are seen in Table 2:

Actions / nonverbal utterances	<$E> smokes cigarette <\$E> <$E> pause <\$E> <$E> Sighs <\$E> <$E> sings <\$E>
Environmental factors	<$E> Background noise (in city) <\$E> <$E> wind interference <\$E>

[2] Transana is qualitative analysis software for video and audio data, developed by the University of Wisconsin-Madison Centre for Education Research. See: www.transana.org/

[3] CANCODE stands for Cambridge and Nottingham Corpus of Discourse in English, a 5 million word corpus of spoken English taken from different contexts across the British Isles. CANCODE was built in collaboration by The University of Nottingham and Cambridge University Press (with whom sole copyright resides).

	<$E> Background noise <\$E>
Guess	<$G?> wunce <$G?>
Inaudible content	<$G?>
Incomplete word	wa= wa= wanting
Interrupted sentence	<$M3> They must do ehm promotion and programme+ <$F1> Yeah. <$F2> +out of there.
Laughter	<$E> Laughs <\$E>
Pause	<$E> pause <\$E> ... = short pause (for breath)
Restarts	<$=> it's the same <\$=> it's the same
Single repeated words	<$F> They they must do this
Speaker Codes	<$M1> = Male researcher 1 <$M2> = Male researcher 2 <$M3> = Main male participant for recordings BAS.1 <$M4> = Male participant in BAS.2 <$M5> = Male participant in BAS.2 <$M> = Male cafe workers/ bus conductors (not a central part of recordings) <$F1> = Female researcher <$F2> = Main female participant in BAS.1 <$F3> = Female participant in BAS.3 <$F4> = Female participant in BAS.3 <$F5> = Female curator in BAS.3 <$F> = Female cafe workers/ bus conductors (not a central focus of recordings)

Table 2: Some transcription conventions used in the BAS data.

3.5. Processing the data

The raw data taken from the Field Work Tracker and transcript can be uploaded into DRS, as seen in Figure 3. Additional datasets can also easily be added to the record and then hand synchronised using DRS's comprehensive synchronisation tools. In Figure 3 we see the mapped route of participant <$F3>, through the centre of Nottingham during the British Art Show. Each individual GPS point is shown as a square on the map. Points where photos/videos/annotations were taken along this route are flagged on the map and can be selected, zoomed in or examined in more detail, as seen with the photos in this figure.

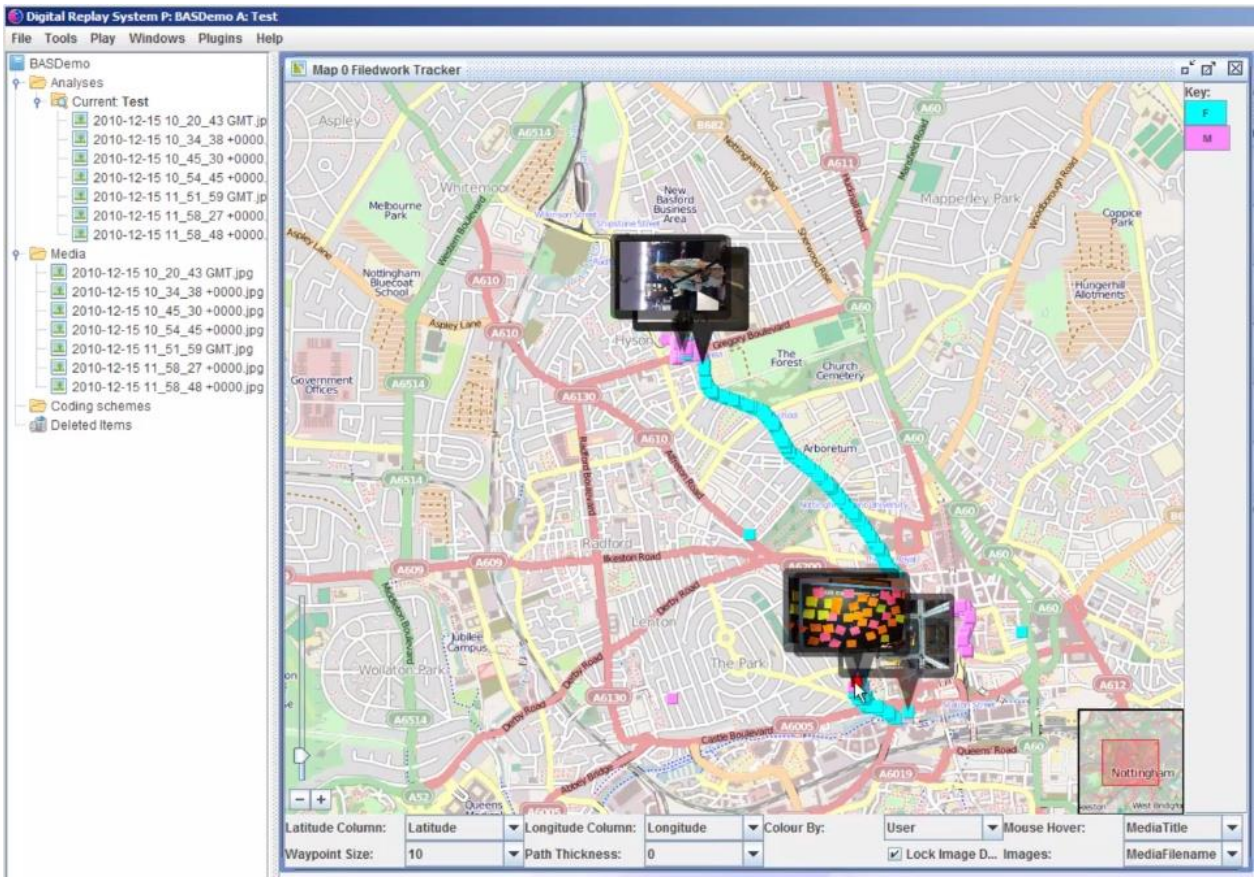

Figure 3: Uploading the Field Work Tracker logs into DRS

3.6. DRS, the corpus and location

DRS enables users to query corpora from a micro level, that is, according to a specific word, phrase, tag or code, to a more global level, that is, according to a particular type of media used when recording or according to a particular physical location. DRS also allows users to map routes and graphically represent frequencies and/or the incidence of specific words and behaviours. The key utilities of DRS are summarised below:

- Tools for searching data **and** metadata in a principled and specific way
- GPS based mapping tools
- Transcription tools
- Graphing tools for mapping the incidence of words or events, for example, over time and location (space) and for comparing sub-corpora and domain specific characteristics
- Concordancing tools

4. **Analysis**

4.1. Approach to analysis

As a starting point for analysing the data we decided to explore patterns of word use outside galleries (i.e. when visitors were walking and travelling to locations), as a point of contrast to when they were inside the galleries (i.e. examining the art and/or sitting in the coffee shop talking about the art). This contrastive approach enabled an initial purchase on potentially different patterns in the data and allowed a systematic test of the corpus query tools.

Such a strategy allowed the exploration of both changes in patterns of word use across geographical locations and in particular patterns of word use when in the defined space of the gallery compared with language used 'on the move'. The filtering tools within DRS allow researchers to do this in a relatively simple manner by highlighting locations on the mapped GPS outputs (as can be seen in Figure 4), and categorising them as either 'inside' or 'outside'. This process was carried out for each of the three individual datasets. Table 3 illustrates how this process is carried out in DRS.

Figure 4: Filtering data by location.

These maps are fully interactive as researchers are able to select a specific part of the route (indicated by the boldface mapped out route seen here), or even a single node, in order to view the accompanying video and/or transcript and to investigate, for example, specific patterns of language use in given contexts. This allows for a micro-level analysis of the corpora, allowing users to search the corpora according to a specific word, phrase, tag or code.

In this example, the area highlighted would be a point of specific focus. After selecting this area, the user can go back to any data tables, graphs or concordance outputs that have been generated in association with this media (and those synchronised with it) and DRS will indicate which elements on the table, graphs or concordance outputs were enacted/spoken/recorded within this given point of reference.

Using this facility, it was possible to highlight and select on the map where each of the galleries were, along with the cafe that was used as the meeting point for two of the studies. For the purposes of this case study the selected locations are designated 'inside' and all non-highlighted locations are designated 'outside'. The division into 'inside' and 'outside' is of course, to some degree, crude and arbitrary but makes for demarcation within the data set that can subsequently be modified and developed. Also included in the marking up of the data were the journeys to and between the galleries which sometimes involved using public transportation (e.g. tram). By segmenting these features on the map, DRS also automatically segmented the transcripts (around the

time-stamped points corresponding to the GPS locations). The amount of data included in each of these segmented transcripts is seen in Table 3:

	Start	Gallery 1	Gallery 2	Gallery 3	TOTAL
BAS.1		New Art Exchange (gallery/cafe)	Castle (cafe then gallery)	Contemporary (gallery then cafe)	**12759**
		New Art Exchange to Castle	***Castle to contemporary***		***3694***
BAS.2	Coffee shop at start	New Art Exchange (gallery)	Contemporary (gallery then cafe)	Castle (gallery then cafe)	**16587**
	Journey to NAE	***NAE to contemporary***	***Contemporary to castle***		***5940***
BAS.3	Coffee shop at start	Castle (gallery)	Contemporary (gallery then cafe)		**18536**
		Castle to contemporary	***Outside contemporary***		***1872***
					59392

Table 3: Word counts for the 'inside' and 'outside'.

Table 3 also shows that the word count for the 'outside' parts of the study was far less than for the 'inside' sub-corpus. This was only to be expected as the main task given to the participants was to view the art show. The data from BAS.3 is also missing as during the journey to the castle the participants decided to turn off the dictaphones. The same pair did not visit the New Art Exchange Gallery, so no was data recorded for the 'inside' part of their experience.

Each of the three pairs decided to take different routes and this, along with the variability in word count and data collected, means that it is difficult to test the reliability of analyses of the 'inside' vs 'outside' contrast given that the two segments are highly variable, small and not wholly comparable. In spite of such inevitable problems with first data runs and facility testing together with the behavioural unpredictability of human subjects it was nonetheless felt to be beneficial to carry out some basic corpus-based comparisons of specific words in-use across this data to provide a starting point for discussions of language use in context and its variation according to different locational features.

4.2. Corpus comparisons: 'inside' and 'outside'

DRS is equipped with simultaneous concordancers that allow users to compare separate, filtered categories against each other based on linguistic features. That is, it is possible to generate frequency searches of segmented parts of the transcripts of conversation that were spoken outside the galleries, in comparison to that spoken inside. Here the data from speakers can be captured and compared as they move dynamically across time and space and across task and location. Upon an initial inspection of the BAS data, the use of deictic markers was shown to be particularly frequent in the speech, as is common in frequency lists. The word deixis is derived from the ancient Greek word for pointing; it is a key component of 'orientational' language and in

marking changing reference to location. For this purpose it was decided that the focus should be on deictic markers.

'Inside' and 'outside' here also represent two main activities on the part of participants ---- inside the gallery conversation is more goal- and task-oriented; outside the gallery conversation is looser and more casual with reference occurring in less predictable ways. In both cases participants occupy different spaces and relations to time and different uses of language are therefore enacted, most particularly, it might reasonably be hypothesised, in relation to deixis and to the relative frequency of different deictic markers for purposes of reference.

Terms which were compared are the following deictic markers:

- *that*
- *this*
- *you*
- *it*
- *them*
- *he*
- *I*
- *the*
- *and*
- *a*
- *here*
- *there*
- *she*
- *they*
- *him*
- *her*
- *their*
- *we*

Deictic markers such as the personal pronouns *you, it, them, he, she, him, her, they, their, I, we,* determiners *the, and, a,* adverbs *here* and *there* and demonstrative adverbs, *this* and *that,* are forms of linguistic reference. These are used to refer to speakers, incidents and objects in discourse, according to their specific spatial and temporal locations. In this study we might expect to see an increase in the use of '*this, that*' and equivalent references increasing when we look at the people in the museums while 'outside' it is reasonable to expect a greater incidence of personal pronouns as the individuals are involved in more interpersonal exchanges, while discussing, for example, which routes to take between the galleries.

The raw frequencies of each of these terms, within the segmented versions of the 'outside' and 'inside' sub-corpora are seen in Table 4. The relative frequencies of these terms are also tabulated here, these denote the number of times the specific search term (i.e. 'word') is used at a 'per word' rate in the entire sub-corpus. Naturally, the raw frequencies between the sub-corpora differ dramatically due to the differences in the word count for each of these. However by comparing the relative frequencies of these terms we see that the terms *that, this, her* and *you* are more significantly more frequent in the 'inside' segregated corpus while *I* and *we* are more frequent (relatively) in the 'outside' corpus. The rate of difference at which these terms occur from one sub-corpus to the next is >+3 in L.L. score for each, with a rate of +8.55 for *that,* +5.26 for *this,* +6.64 for *her,* +6.48 for *you,* +4.20 for *I* and +5.10 for *we* (note that since the table is comparing the first dataset ('inside') to the second ('outside'), LL scores that are marked as + denote that there is a statistically more frequent use of the specific search term in the 'inside' data, while those marked with a – denote a statically more frequent use of the search term in the 'outside' data.

	Inside (I)			**Outside (O)**		**LL Scores (I Vs O)**
	Raw Freq	**Rel. Freq**		**Raw Freq**	**Rel. Freq**	
that	1070	2.23		207	1.80	+ 8.55
this	403	0.84		73	0.63	+ 5.26
you	1219	2.55		246	2.14	+ 6.48
it	1393	2.91		366	3.18	- 2.27
them	101	0.21		21	0.18	+ 0.38

he	169	0.35		50	0.43	- 1.61
I	1666	3.48		444	3.86	- 3.69
the	1719	3.59		396	3.44	+ 0.58
and	794	1.66		194	1.69	- 0.04
a	875	1.83		214	1.86	- 0.05
here	93	0.19		15	0.13	+ 2.26
there	375	0.78		96	0.83	- 0.30
she	50	0.10		13	0.11	- 0.06
they	310	0.65		91	0.79	- 2.72
him	23	0.05		6	0.05	- 0.03
her	28	0.06		1	0.01	+ 6.64
their	19	0.04		5	0.04	- 0.03
we	411	0.86		125	1.09	- 5.10

Table 4: Raw and relative frequencies of deictic markers in the BAS corpora.

The corpus utilities provided by the DRS tool are limited compared to other standard corpus analysis toolkits such as Wordsmith Tools (Scott, 1999) and WMatrix (Rayson, 2003). Thus, as an extension to these comparisons, patterns in word frequency were also examined more widely across the dataset. Table 5 charts the keywords (i.e. words which occur at a statistically significantly more frequent rate in one corpus than the one with which it is compared) that emerge when comparing the 'inside' Vs 'outside' sub-corpora and the 'outside' Vs 'inside' sub-corpora:

		Inside		**Outside**		**LL Score**		**Outside**		**Inside**		**LL Score**
		Freq	**Rel. Freq**	**Freq**	**Rel. Freq**			**Freq**	**Rel. Freq**	**Freq**	**Rel. Freq**	
1	art	148	0.31	10	0.09	+ 22.06	chocolate	17	0.15	4	0.01	+ 37.04
2	installation	63	0.13	1	0.01	+ 20.14	cheese	6	0.05	0	0	+ 19.68
3	of	766	1.59	123	1.06	+ 19.07	ducks	6	0.05	0	0	+ 19.68
4	video	81	0.17	3	0.03	+ 18.88	wind	6	0.05	0	0	+ 19.68
5	wall	30	0.06	0	0	+ 12.93	dark	11	0.09	5	0.01	+ 18.37
6	artist	27	0.06	0	0	+ 11.64	cigarette	5	0.04	0	0	+ 16.4
7	British	27	0.06	0	0	+ 11.64	harrowing	5	0.04	0	0	+ 16.4
8	looks_like	51	0.11	2	0.02	+ 11.52	walking	12	0.1	8	0.02	+ 15.9
9	question	26	0.05	0	0	+ 11.21	town	10	0.09	5	0.01	+ 15.78
10	mmm	70	0.15	5	0.04	+ 9.84	vegan	9	0.08	4	0.01	+ 15.2
11	show	44	0.09	2	0.02	+ 9.08	love	11	0.09	7	0.01	+ 15.05
12	work	44	0.09	2	0.02	+ 9.08	my	48	0.41	100	0.21	+ 14.08
13	different	60	0.12	4	0.03	+ 9.07	oh	106	0.91	285	0.59	+ 13.66
14	painting	21	0.04	0	0	+ 9.05	Byron	4	0.03	0	0	+ 13.12
15	paintings	21	0.04	0	0	+ 9.05	Dogville	4	0.03	0	0	+ 13.12
16	strange	21	0.04	0	0	+ 9.05	Nancy	4	0.03	0	0	+ 13.12
17	that	1070	2.23	207	1.80	+ 8.55	clock	4	0.03	0	0	+ 13.12
18	like	535	1.11	95	0.82	+ 7.98	dream	4	0.03	0	0	+ 13.12
19	basically	18	0.04	0	0	+ 7.76	evil	4	0.03	0	0	+ 13.12
20	paper	18	0.04	0	0	+ 7.76	extreme	4	0.03	0	0	+ 13.12

Table 5: The most common words used in the 'inside' Vs 'outside' sub-corpus and the 'outside' Vs 'inside' sub-corpus.

BAS	**BNC**	

		Freq	Rel. Freq	Freq	Rel. Freq	LL SCORE	
1	its	651	1.09	228	0.02	+	2741.08
2	art	158	0.26	25	0	+	759.97
3	yeah	1175	1.96	9494	0.97	+	437.94
4	like	630	1.05	3743	0.38	+	437.28
5	mmm	75	0.13	3	0	+	403.54
6	okay	311	0.52	1147	0.12	+	401.61
7	hmm	77	0.13	11	0	+	375.04
8	installation	64	0.11	4	0	+	335.8
9	gallery	58	0.10	0	0	+	331.47
10	contemporary	57	0.10	2	0	+	308.52
11	video	84	0.14	68	0.01	+	279.07
12	castle	53	0.09	11	0	+	245.46
13	yep	80	0.13	132	0.01	+	191.8
14	tram	35	0.06	1	0	+	191
15	quite	188	0.31	928	0.09	+	172.09
16	liked	49	0.08	40	0	+	162.29
17	film	51	0.05	56	0.01	+	149.98
18	looks_like	53	0.09	71	0.01	+	142.01
19	kind_of	41	0.09	39	0	+	128.07
20	interesting	56	0.07	104	0.01	+	125.15

Table 6: The most common words used in the BAS corpus compared to a spoken component of the BNC

5. Discussion

Too much should not be read into limited data sets (circa 60,000 words) involving a limited number of participants but the corpus analysis does suggest interesting points for further investigation. There are differences between inside and outside locations in terms of deictic marking with the use of singular personal pronouns outside indicating more personal comment and reference. The 'task' of viewing 'inside' is more goal-directed and leads to more referential 'pointing', to more shared observations (inflected in plural personal pronouns) and to the drawing of analogies through the word *like*.

my favourite as well <$M5> +I quite	like	that . Erm <$F1> Its quite small tho
ooks like if you keep looking at it	like	its got a pile to it <$F4> It does d
iece of art <$F3> Yeah <$F5> People	like	it from the mechanical side of it+ <
it you can get .... I still don't	like	it but yeah <$F5> No I 'll tell you
er yeah <$F5> Er the harpsicord is	like	a 3D version of that . Which I ca n'
r paintings on the stairs <$F3> I I	like	my paintings to look like <E> pause
e recognisable things <$F5> You 'll	like	it in here then <$F3> Something like
when we come out . That is lovely I	like	that . <$F5> But its missing arms an
hat is absolutley beautiful <$M1> I	like	that as well . That 's extraordinary
read them yeah <$M1> Then you can	like	er <$F4> Ah ha .... I see what it is

Figure 5: Concordance output of *like* in the 'inside' sub-corpus.

The concordance output in figure 5 shows the use of like in the 'inside' sub-corpus. Here we can see there that the use of the use of quotative *like* – to introduce direct speech – is only use in 1 of the 10 cases. This relative infrequency is something that is witnessed in this sub-corpus as whole. The relative frequency of the word *like* here may

be both a verb and a preposition linked to evaluative comparison and analogy (note the presence of *like* at point 4, *liked* at point 16 and *looks like* at point 18 in Table 6). *I like*, *I quite like* and *I really like* are among the most frequent clusters in this dataset, providing summaries of the individuals' perceptions of the art they are looking at. In comparison, the use of like in the outside sub-corpus witnesses a greater number of the use of the quotative form, rather than being evaluative.

In any case, there appears to be a greater concentration on evaluation in the 'inside' corpus linked to evaluative adjectives regarding what items *look like* such as *interesting* , *strange* and *different*, as well as the presence of specific reference to painting(s), types of media (*wall*, *video*, *installation*, *art*). Lexical variation is more marked in the 'outside' corpus (*cheese*, *chocolate*, *ducks*) with the inside corpus understandably evidencing a narrower range built around the art installations. And the greater concentration in the 'inside' corpus of backchannels of what might be reasonably taken to be support and agreement would suggest a more collaborative conversational interaction on the part of participants.

Before too many claims might be made for such insights, however, the transcribed data would need to be mined more qualitatively using broadly discourse analytical and conversation analytical insights and, where relevant, set alongside ancilliary data such as the *pre* and, especially, *post hoc* questionnaires where participants' attitudes and responses to the tasks and exhibition content can be explored more ethnographically.

The data assembled is also suggestive in other ways and could form the basis for further exploration based on more extensive data sets. For example, are there points in such comparative data sets where transition occurs from one location to another? Is there more 'orientational' language 'outside' and more evaluative and analogical language 'inside'? Are there connections between mode of transport (e.g. walking v. by tram) which evidence different deictic reference to location? Does conversation differ between the movement of the tram and the stopping the tram in a station; or the walk between one art installation and another? In other words capturing context dynamically is a process that involves numerous gradations. Our understanding is enhanced by comparisons between one context and another but finding ways of accurately capturing multiple contexts is a more appropriate way of measuring the truly dynamic nature of contexts and movements between contexts. What can be claimed here is that the adaptation of DRS to mobile and hand-held devices does facilitate such a research focus.

6. Limitations and future directions

One limitation of the case study in this chapter is perhaps the question of how 'real' or natural it is. How far can the tracking of individuals be said to embrace typicality in the use of language or in the capture of forms of language that can be said to evidence a connection between place, space, experience and language choices? And to what extent is the subjection of participants to markedly rare physical and affective experiences or enforced moves between 'inside' and 'outside' locations likely to produce results that are of limited utility and generalisability? On the other hand, are there not opportunities here for more precise contextually-related description of language and for much enhanced understanding of key forms such as deixis. The rather loose terms for deixis such as 'orientational' language can begin to be revaluated and reaccented within evidence-based frameworks which allow for much more dynamic accounts, leading, for example, to an enhanced understanding of speakers' orientation when they are engaged in more than one channel of communication at the same time.

It has not been our purpose in this chapter to offer definitive correlations between language use and non-linguistic factors. 'Results' are not therefore the point. The aim is to suggest methods, processes and starting points for further development and further analysis. The analysis of this data provides a good, albeit crude, starting point for outlining an approach to the analysis of word use and linguistic patterning across different forms of media and in terms of time, space and place. It provides an example of future lines of enquiry for a corpus linguistics that aims to move beyond text and language as conventionally conceptualised and to embrace the many other data streams that intersect with language use. This chapter represents no more than a beginning but, it is argued here, it is a significant beginning with numerous possibilities for further development and extension.

7. References

Adolphs, S. (2008). *Corpus and Context: Investigating Pragmatic Functions in Spoken Discourse.* London: John Benjamins Publishing Company.

Adolphs, S. and Carter, R. (2013). *Spoken Corpus Linguistics: From Monomodal to Multimodal.* London: Routledge.

Baron, N. (2000). *Alphabet to Email: How Written English Evolved and Where it's Heading*. London: Routledge.

Condon, S.L. and Cech, C.G. (1996). Profiling turns in interaction. In Proceedings of the thirty-fourth *Annual Conference of the Hawaii International Conference on System Sciences.* Los Alamitos, California: IEEE Computer Society Press.

Crystal, D. (2004). *A glossary of Netspeak and Textspeak*. Edinburgh: Edinburgh University Press.

Crystal, D. (2011). Back to the Future. *The Linguist*. January 2011: 10-13

French, A., Greenhalgh, C., Crabtree, A., Wright, W., Brundell, B., Hampshire, A. and Rodden, T. (2006), Software Replay Tools for Time-based Social Science Data. *Proceedings of the 2nd Annual International e-Social Science Conference.* Available at: www.cs.nott.ac.uk/~axc/work/eSS-1-06.pdf

Greenhalgh, C., French, A., Tennant, P., Humble, J. and Crabtree, A. (2007), From replay tool to digital replay system. *Proceedings of the 3rd International Conference on e-Social Science ESRC/ NSF* [online]. Available at: http://ess.si.umich.edu/papers/paper161.pdf.

Herring, S.C. (2007). A faceted classification scheme for computer-mediated discourse. *Language@Internet 4*(1): 1-37.

Iwasaki, J., and Oliver, R. (2003). Chatline interaction and negative feedback. *Australian Review of Applied Linguistics 17:* 60-73.

Jepson, K. (2005). Conversations and negotiated interaction in text and voice chat rooms. *Language Learning and Technology 9*(3): 79-98.

Knight, D. (2011). *Multimodality and Active Listenership: A Corpus Approach*. London: Bloomsbury.

Knight, D., Tennent, P., Adolphs, S. and Carter, R. (2010), Developing ubiquitous corpora using the digital replay system (DRS). *Proceedings of the LREC 2010 (Language Resources Evaluation Conference) Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, May 2010, Giessen, Germany, pp. 16–21.

Ko, K. (1996). Structural characteristics of computer-mediated language: a comparative analysis of InterChange discourse. *Electronic Journal of Communication 6*(3). Available at: http://www.editlib.org/p/83178/

Rayson, P. (2003), *Matrix: A Statistical Method and Software Tool for Linguistic Analysis Through Corpus Comparison*. Unpublished PhD thesis. Lancaster University.

Scott, M. (1999), *Wordsmith Tools* [Computer program]. Oxford: Oxford University Press.

Thorne, S.L. (2008). Transcultural communication in open Internet environments and massively multiplayer online games. In Magnan, S.S. (Ed.), *Mediating Discourse online.* London: John Benjamins Publishing Company.

Von Ahn, L. (2006). Game with a purpose. *Computer 39*(6): 92-94.

Weiser, M. (1991). The Computer for the Twenty-First Century. *Scientific American*, September 1991, pp.94-110.