The future of multimodal corpora

O futuro dos corpora modais

Dawn Knight*
The University of Nottingham
Nottingham / UK

ABSTRACT: This paper takes stock of the current state-of-the-art in multimodal corpus linguistics, and proposes some projections of future developments in this field. It provides a critical overview of key multimodal corpora that have been constructed over the past decade and presents a wish-list of future technological and methodological advancements that may help to increase the availability, utility and functionality of such corpora for linguistic research.

KEYWORDS: Multimodal corpus linguistics; resources; software; availability; usability.

RESUMO: Este artigo apresenta um balanço do estado da arte da linguística de corpus multimodal e propõe a projeção de desenvolvimentos futuros nessa área. Um resumo crítico dos corpora multimodais-chave que foram construídos na última década é apresentado, assim como uma lista de desenvolvimentos tecnológicos e metodológicos futuros que podem auxiliar na disponibilização e utilização, bem como na funcionalidade, de tais corpora para a pesquisa linguística.

PALAVRAS-CHAVE: Linguística de corpus multimodal; recursos; programas computacionais; disponibilidade; usabilidade.

* dawn.knight@nottingham.ac.uk

1. Introduction

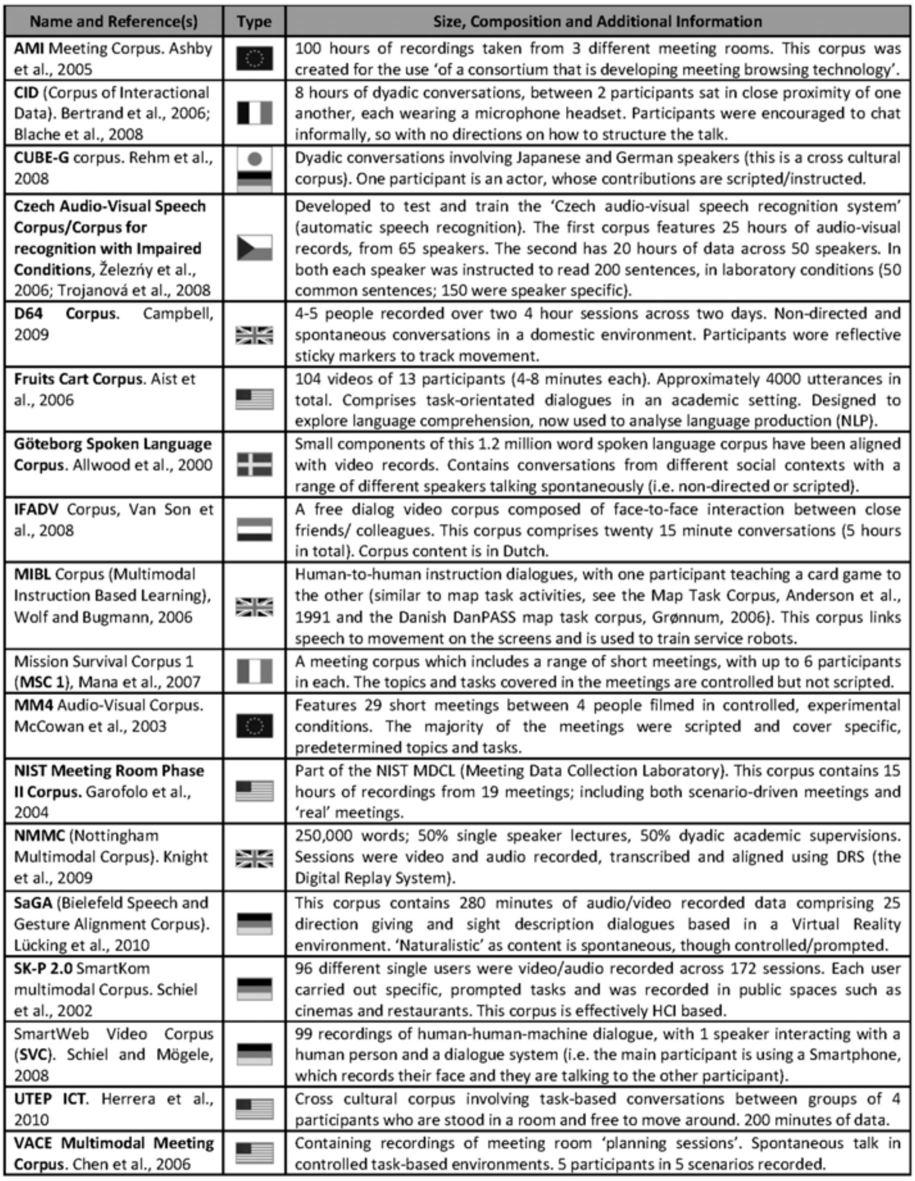
The surge in technological advancements witnessed since the latter part of the last century has provided the linguist with better tools for recording, storing and querying multiple forms of digital records. This has provided the foundations for the recent surge in interest in multimodal corpus linguistics.

A multimodal corpus, for the purpose of the current paper, is best defined as 'an annotated collection of coordinated content on communication channels including speech, gaze, hand gesture and body language, and is generally based on recorded human behaviour' (FOSTER; OBERLANDER, 2007, p. 307-308). The integration of textual, audio and video records of communicative events in multimodal corpora provides a platform for the exploration of a range of lexical, prosodic and gestural features and for investigations of the ways in which these features interact in real-life discourse.

Unlike monomodal corpora, which have a long history of use in linguistics, the construction and use of multimodal corpora is still in its relative infancy, with the majority of research associated with this field spanning back only a decade. Despite this, work using multimodal corpora has already proven invaluable for answering a variety of linguistic research questions, questions that are otherwise difficult to consider (see ALLWOOD, 2008 for further details).

The utility of corpus-based research and methods is in fact becoming popular in a range of different academic disciplines and fields of research, far beyond linguistics. For example, the processes of construction itself is of interest to computer scientists, while the tools developed can be utilised to answer questions posed by behaviourists, psychologists, social scientists and ethnographers. This means that multimodal corpora and corpus-based methods and related projects, which are often necessarily interdisciplinary and collaborative, receive ever-increasing support from academic researchers, funding councils and commercial third parties, something which is likely to be sustained well in to the future.

As a review of the current landscape, however, this paper primarily aims to provide an overview of selected multimodal corpora that have either already been built, or are currently under construction. An index of these corpora is provided in Figure 1, overleaf. The paper examines the types of data they contain, the applications of these datasets and ways in which they are limited. This is followed by a projection of ways such corpora can be further developed, improved or expanded in the future.



Name and Reference(s)	Type	Size, Composition and Additional Information
AMI Meeting Corpus. Ashby et al., 2005	🇪🇺	100 hours of recordings taken from 3 different meeting rooms. This corpus was created for the use 'of a consortium that is developing meeting browsing technology'.
CID (Corpus of Interactional Data). Bertrand et al., 2006; Blache et al., 2008	🇫🇷	8 hours of dyadic conversations, between 2 participants sat in close proximity of one another, each wearing a microphone headset. Participants were encouraged to chat informally, so with no directions on how to structure the talk.
CUBE-G corpus. Rehm et al., 2008	🇯🇵 🇩🇪	Dyadic conversations involving Japanese and German speakers (this is a cross cultural corpus). One participant is an actor, whose contributions are scripted/instructed.
Czech Audio-Visual Speech Corpus/Corpus for recognition with Impaired Conditions, Železný et al., 2006; Trojanová et al., 2008	🇨🇿	Developed to test and train the 'Czech audio-visual speech recognition system' (automatic speech recognition). The first corpus features 25 hours of audio-visual records, from 65 speakers. The second has 20 hours of data across 50 speakers. In both each speaker was instructed to read 200 sentences, in laboratory conditions (50 common sentences; 150 were speaker specific).
D64 Corpus. Campbell, 2009	🇬🇧	4-5 people recorded over two 4 hour sessions across two days. Non-directed and spontaneous conversations in a domestic environment. Participants wore reflective sticky markers to track movement.
Fruits Cart Corpus. Aist et al., 2006	🇺🇸	104 videos of 13 participants (4-8 minutes each). Approximately 4000 utterances in total. Comprises task-orientated dialogues in an academic setting. Designed to explore language comprehension, now used to analyse language production (NLP).
Göteborg Spoken Language Corpus. Allwood et al., 2000	🇸🇪	Small components of this 1.2 million word spoken language corpus have been aligned with video records. Contains conversations from different social contexts with a range of different speakers talking spontaneously (i.e. non-directed or scripted).
IFADV Corpus, Van Son et al., 2008	🇳🇱	A free dialog video corpus composed of face-to-face interaction between close friends/ colleagues. This corpus comprises twenty 15 minute conversations (5 hours in total). Corpus content is in Dutch.
MIBL Corpus (Multimodal Instruction Based Learning), Wolf and Bugmann, 2006	🇬🇧	Human-to-human instruction dialogues, with one participant teaching a card game to the other (similar to map task activities, see the Map Task Corpus, Anderson et al., 1991 and the Danish DanPASS map task corpus, Grønnum, 2006). This corpus links speech to movement on the screens and is used to train service robots.
Mission Survival Corpus 1 (**MSC 1**), Mana et al., 2007	🇮🇹	A meeting corpus which includes a range of short meetings, with up to 6 participants in each. The topics and tasks covered in the meetings are controlled but not scripted.
MM4 Audio-Visual Corpus. McCowan et al., 2003	🇪🇺	Features 29 short meetings between 4 people filmed in controlled, experimental conditions. The majority of the meetings were scripted and cover specific, predetermined topics and tasks.
NIST Meeting Room Phase II Corpus. Garofolo et al., 2004	🇺🇸	Part of the NIST MDCL (Meeting Data Collection Laboratory). This corpus contains 15 hours of recordings from 19 meetings; including both scenario-driven meetings and 'real' meetings.
NMMC (Nottingham Multimodal Corpus). Knight et al., 2009	🇬🇧	250,000 words; 50% single speaker lectures, 50% dyadic academic supervisions. Sessions were video and audio recorded, transcribed and aligned using DRS (the Digital Replay System).
SaGA (Bielefeld Speech and Gesture Alignment Corpus). Lücking et al., 2010	🇩🇪	This corpus contains 280 minutes of audio/video recorded data comprising 25 direction giving and sight description dialogues based in a Virtual Reality environment. 'Naturalistic' as content is spontaneous, though controlled/prompted.
SK-P 2.0 SmartKom multimodal Corpus. Schiel et al., 2002	🇩🇪	96 different single users were video/audio recorded across 172 sessions. Each user carried out specific, prompted tasks and was recorded in public spaces such as cinemas and restaurants. This corpus is effectively HCI based.
SmartWeb Video Corpus (**SVC**). Schiel and Mögele, 2008	🇩🇪	99 recordings of human-human-machine dialogue, with 1 speaker interacting with a human person and a dialogue system (i.e. the main participant is using a Smartphone, which records their face and they are talking to the other participant).
UTEP ICT. Herrera et al., 2010	🇺🇸	Cross cultural corpus involving task-based conversations between groups of 4 participants who are stood in a room and free to move around. 200 minutes of data.
VACE Multimodal Meeting Corpus. Chen et al., 2006	🇺🇸	Containing recordings of meeting room 'planning sessions'. Spontaneous talk in controlled task-based environments. 5 participants in 5 scenarios recorded.

FIGURE 1: An index of multimodal corpora

2. Multimodal Corpora: analysing discourse 'beyond the text'

2.1. Current multimodal corpora

There are two broad 'types' of researchers who are interested in multimodal corpus linguistics, as identified by Gu (2006). Firstly, there are those who are interested in undertaking 'multimodal and multimedia studies of discourse', addressing more social science based issues, with a concern on 'human beings' (GU, 2006, p. 132).

Secondly, there are those interested in the construction of multimodal corpora as an explorative exercise, tackling specific technological challenges of assembling and (re)using these datasets, and evaluating how this is best achieved; that is, which software and hardware tools to use etc. Many of these researchers are more interested in 'how to improve human-computer interaction' (GU, 2006, p. 132, also see KNIGHT *et al.*, 2009 for further discussion and associated examples).

Similar to current monomodal corpora, the contents of multimodal corpora, the ways in which they are recorded, their size, and so on, are highly dependent on the aims and objectives that they are intended to fulfil; the specific research questions that want to be explored or the specific technological or methodological questions that require answering by those developing and/or using the corpus. Given this, there are a variety of different forms of multimodal corpora and related research projects, all with, to some degree, bespoke characteristics regarding:

- **Design and infrastructure**: Concerning what the data in the corpus looks like; what sorts of recordings are included and the basic design methodology used to collect, compile, annotate and represent this data.
- **Size and scope**: Amount of data (in terms of hours and/or word count) and the variation in the types included (in terms of the range of speakers or different contexts included and so on).
- **Naturalness**: How 'natural' or 'real' (authentic) the data is perceived to be; whether it is scripted and/or structured or more spontaneous.
- **Availability and (re)usability**: Access rights to data, whether corpora are published and can be utilised and analysed by other researchers.

Each of these will be discussed at length in the subsequent sections of this paper.

2.2. Design and infrastructure

While research using audio recordings of conversation has had a long history in corpus-based linguistics, the use of digital video records as 'data' is still fairly innovative. The specific strategies and conventions used to compile (record), annotate and represent/replay video records for a multimodal corpus therefore generally differ from one to the next (for further discussions on each of these processes, see KNIGHT *et al.*, 2009).

No formally agreed, standardised approach exists for recording data for multimodal corpora and although each current corpus, as seen in figure 1, tends to utilise a range of highly specialised equipment in a fixed, predefined, thus *replicable* set-up, the exact nature of this setting is not necessarily consistent from one to the next. Specific forms of equipment, where they are located and even the file formats that they record in are subject to variation.

Further to this, as discussed extensively in Knight *et al.* (2009), various different schemes exist to mark up, code and annotate multimodal data, and as yet no standard approach is used across all multimodal corpora (although the International Standards for Language Engineering, ISLE project acknowledges the need for such, DYBKJÆR; OLE BERNSEN, 2004, p. 1). As Baldry and Thibault note (2006, p. 148):

> In spite of the important advances made in the past 30 or so years in the development of linguistic corpora and related techniques of analysis, a central and unexamined theoretical problem remains, namely that the methods adapted for collecting and coding texts isolate the linguistic semiotic from the other semiotic modalities with which language interacts.... [In] other words, linguistic corpora as so far conceived remain intra-semiotic in orientation.... [By] contrast multimodal corpora are, by definition, inter-semiotic in their analytical procedures and theoretical orientations.

Extensive deliberation also exists about what aspects should actually be marked up and how; so which specific non-verbal behaviours (patterns of gesticulation) or prosodic features should be annotated and so on. This problem is also true for the software used in order to undertake the processes of coding, annotation, synchronisation and representation (for a more in depth discussion on each of these processes please refer to KNIGHT, 2011).

While an increasing number of multimodal projects, particularly those linked to the multimodal corpora workshop series,[1] are using the software tool Anvil[2] (KIPP, 2001; KIPP *et al.*, 2007), others favour ELAN[3], DRS[4] (FRENCH *et al.*, 2006; GREENHALGH *et al.*, 2007) or EXMARaLDA[5]. Given this, standardised procedures for carrying out these processes would thus be welcomed and are perhaps a priority for the future of research in this field.

2.3. Size, scope and range

Figure 1 indicates that few multimodal corpora extend beyond a few thousand words in size. While the AMI corpus (see ASHBY *et al.*, 2005) comprises an impressive 100 hours of video, the majority of this data exists solely as video records. In other words many videos have yet to be transcribed, so the actual *size* of this corpus as a functional multimodal (i.e. text and video based) tool is not especially large. Other multimodal corpora contain only a few hours of video and/or a limited number of words.

This issue of size is especially noteworthy because current monomodal corpora pride themselves on the fact that they extend into multi-million word datasets, such as the British National Corpus (BNC), the Bank of English (BoE) and the Cambridge International Corpus (CIC). The BNC contains 100 million words of British English (90% written, 10% spoken); the BoE stands at over 650 million words (75% written, 25% spoken) and the CIC corpus has recently hit the 1 billion word mark.

Obviously, the advantage of using text-based discourse in the compilation of corpora is that large quantities of data are readily available,

[1] Details of the multimodal corpora workshop series on multimodal corpora, tools and resources can be found at: <http://www.multimodal-corpora.org>.

[2] ANVIL is a frame accurate multimodal annotation and visualisation tool, available for free from: <http://www.dfki.de/~kipp/anvil/>.

[3] ELAN is a 'professional tool for the creation of complex annotations on video and audio resources' which is available to download for free at: <http://www.lat-mpi.eu/>.

[4] DRS, The Digital Replay System, is a multimodal corpus construction and replay tool which is available to download for free at: <http://sourceforge.net/projects/thedrs/>.

[5] Exmeralda, Extensible Markup Language for Discourse Annotation, 'is a system of concepts, data formats and tools for the computer assisted transcription and annotation of spoken language, and for the construction and analysis of spoken language corpora' which is available to download for free at: <http://www.exmaralda.org/en_index.html>.

already machine-readable and/or relatively easy to get hold of, so the process of assembling such databases is relatively straightforward. The process of compiling spoken components or indeed purely spoken corpora is renowned as being a more lengthy process. This is because spoken data needs to be recorded before it is transcribed, annotated and coded before it is integrated into the corpus. As, it is estimated, the process of transcription alone takes a trained researcher up to ten hours to tackle one hour of audio, compiling spoken corpora is often a long and arduous process. For this reason spoken corpora tend to be of a smaller size, such as the five million word CANCODE[6] corpus.

Adding further 'multimodal' levels and perspectives to corpora compounds this problem as recording, aligning and transcribing (if at all) different streams of data is naturally more time consuming and technically difficult than when dealing with a single stream. Furthermore, if specific gestures are to be annotated, the processes of defining, marking up and coding these add further complexity to the construction of these datasets as, it is generally considered, 'the most labour-intensive part for acquiring a multimodal corpus is the annotation of the data, in particular for the visual modality' (FANELLI *et al.*, 2010, p. 70). However, over time we have witnessed an increase in the availability of technical resources for not only recording but also processing, aligning and archiving multimodal corpora, so it is likely that these limitations will become less inhibiting in the future.

Further to size, current multimodal corpora are somewhat limited in terms of *scope*. The majority of the corpora seen in figure 1 tend to be domain specific, mono-lingual (aside from CUBE-G) and/or of a specialist nature, so built of one form of data recorded in a given discourse context. AMI, the MM4 Audio-Visual Corpus, MSC1, the VACE Multimodal Meeting Corpus and the NIST Meeting Room Phase II Corpus all feature records of interaction from a professional meeting room. In these meeting-based corpora, the primary motivation behind the associated research (and corpus construction) is to enable the development and integration of technologies for displaying and researching

[6] CANCODE stands for Cambridge and Nottingham Corpus of Discourse in English. This corpus has been built as part of a collaborative project between The University of Nottingham and Cambridge University Press with whom sole copyright resides. CANCODE comprises five million words of (mainly casual) conversation recorded in different contexts across the British Isles.

meeting room activity specifically. In some of these corpora, the content is scripted or pre-planned to a certain extent and/or the conditions in which the recordings take place are controlled and experimental, with participants being told specifically where to sit and so on.

So, despite the commendable size of AMI, the utility of this corpus for general corpus linguistic research is perhaps limited. As with specialised monomodal corpora such the MICASE corpus[7] of academic discourse and the Wolverhampton Business English Corpus,[8] the contextual and compositional specificity of the data included means it is not necessarily appropriate for addressing research questions that focus on the more interpersonal aspects of communication (for example), beyond this formal, professional contextual domain. This is because the meeting room environment is generally understood as not being particularly conducive to the frequent occurrence of more informal, interpersonal language and/or behaviours. The specialised nature of these corpora potentially affects the spontaneity of the content included (a facet discussed in more detail below), as the constrained nature of the discourse context influences the content and structure of the discourse.

A similar criticism is valid for the NMMC which includes only lecture and supervision data (i.e. academic), and can also be extended to the map or task-based corpora, which prompt highly structured and sometimes scripted content (examples include CUBE-G, the Czech Audio-Visual Speech Corpus, Fruits Carts Corpus, MIBL, SaGA and UTEP ICT).

Further to this, the NMMC was initially designed to allow the application of a 2D digital tracker onto the derived images (see Knight et al., 2006 for further details), as a means of defining patterns of gesticulation. Therefore, recordings are all close up, focusing mainly on the head and torso of participants in order to produce high *quality* images to support the use of the tracking software. Thus while patterns of hand, arm and head movements can be analysed in this data, other bodily actions and spatial positions (i.e. proxemics), for example, cannot. Therefore researchers interested in

[7] MICASE, the Michigan Corpus of Academic English, is a 1.7 million word corpus of transcribed interactions recorded at the University of Michigan. For more information, see: <http://lw.lsa.umich.edu/eli/micase/index.htm>.

[8] The Wolverhampton Business English Corpus is comprises 10 million words of written English from the business domain. These texts were collected between 1999 and 2000. For more information, see: <http://www.elda.org/catalogue/en/text/W0028.html>.

researching a range of different behaviours would perhaps find the NMMC dataset limited (see BRÔNE *et al.*, 2010 for further discussion). This is true for other examples of corpora using more laboratory based and/or situated, static, recording methodologies, as detailed in Figure 1.

If the NMMC utilised recordings of participants at a greater distance away, thus capturing more aspects of the bodily movement, it is unlikely that the tracking system, which required a face-on and in-focus image, could be utilised. This would make the data recorded unfit for its original intended purpose. Overall, it is difficult to maintain a balance between the quality of corpus data and its potential usability, a balance which is somewhat constrained by the limitations of recordings equipment used to collect it. This makes the criticisms of the balance between the relative quality and reusability of multimodal corpus data particularly difficult to resolve/overcome.

The only corpora featured in figure 1 (above) that are exempt from this criticism of 'scope' are D64, components of the Goteborg Spoken Language Corpus, IFADV, SK-P and the SmartWeb Video Corpus. These corpora are either mobile based, so are not fixed to specific geographical or social contexts (SK-P and the SmartWeb Video Corpus) or include data which is seen to be 'spontaneous' and 'naturalistic'; featuring speakers who are static but who are discussing a range of self-selected topics (elements of the Goteborg Spoken Language Corpus and IFADV) and are perhaps, as is the case of D64, recorded in relaxed and familiar domestic settings.

2.4. Naturalness

Support for using corpora in linguistic research was traditionally founded on the notion that while 'introspective data is artificial…..corpora are natural, unmonitored sources of data' (McENERY; WILSON, 1996, p. 8, also see McCARTHY, 2001, p. 125 and MEYER, 2002, p. 5). Corpora therefore provide records of discourse as it is used in real-life contexts, that is, language as it is performed; rather than relying on more rationalistic, intuitive accounts (as previously advocated by CHOMSKY, 1965).

Constructing and utilising authentic, *naturalistic* language records is also a real aim for those working with multimodal data; an aim which has proven to be difficult to fully achieve. By definition alone, this notion of naturalness is abstract and interpretive. As an idealised concept, it is best described as that language which is used in our daily lives; unconstrained and fluid, changeable from one context to the next.

Following this definition, and given the matters discussed in section 2.4, the proposed naturalness of the data contained in those corpora listed in figure 1 can be brought under scrutiny. As the recording set-ups used are generally fixed, laboratory based and/or feature specialist environments with participants; they are thus far from 'unconstrained' and 'context-free'. Oertel et al. suggest that current set-ups effectively exist on a cline, a 'spectrum', as seen in Figure 2 (2010, p. 28).

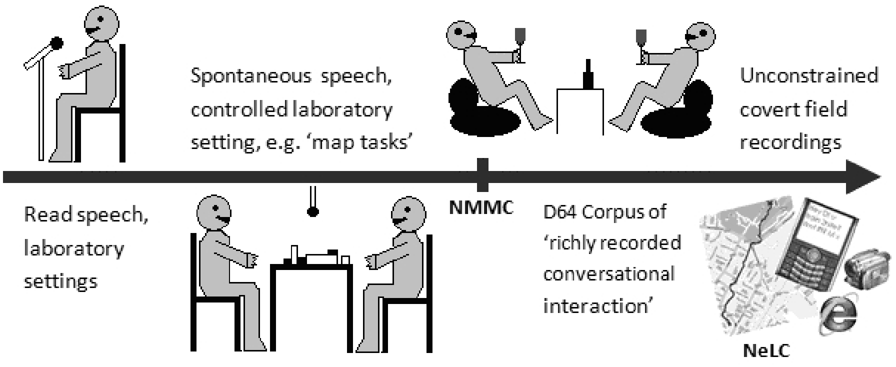

Spontaneous speech, controlled laboratory setting, e.g. 'map tasks'
Unconstrained covert field recordings
Read speech, laboratory settings
NMMC
D64 Corpus of 'richly recorded conversational interaction'
NeLC

FIGURE 2: 'Spectrum of observation scenarios ranging from highly controlled to truly ethological' (based on OERTEL *et al.*, 2010, p. 28)

At the extreme left of the spectrum exists the highly conditioned and scripted forms of corpora such as CUBE-G and Czech Audio-Visual Speech corpus. This progresses to dyadic, but situated, records of speakers in controlled scenarios (such as the Fruit Carts Corpus, MIBL Corpus and SaGA Corpus) through to more spontaneous forms of 'richly recorded' datasets taken from more informal contexts, such as domestic settings (the D64 corpus for example). At the right side of the spectrum we see unconstrained covert field recordings.

To develop corpora which are as *naturalistic* as possible then, it is suggested that the form of recording set-up positioned to the far right of this figure would be most effective. This would thus include data recorded in dynamic environments; on the move and in a variety of different contexts, away from the standardised, fixed and situated setting. While no corpus of this nature has been fully developed as yet, plans to do so are currently underway at the University of Nottingham (see section 3.1 for more details).

Not only does the recording context, that is the physical setting, potentially compromise this notion of naturalness in corpus development, but so too does the equipment used in this context. Audio and video recorders can impact on the data due to the 'observer's paradox' (LABOV, 1972), whereby participants may (sub)consciously adjust their behaviours because they are aware that they are being filmed. Given that video cameras, in particular, are quite obtrusive, and technically it is not ethical to 'hide' these or other recording devices (without the participant's consent), it is difficult to minimise the potential effect this will have on how *naturalistic* behaviours are.

In addition to cameras and microphones, in order to track gestures the D64 corpus, for example, also required participants to wear reflective sticky markers during the recording phase. Again these markers are somewhat invasive and detrimental to the perceived naturalness of the recorded data as they are 'not only time-consuming and often uncomfortable' to wear but 'can also significantly change the pattern of motion' (FANELLI *et al.*, 2010, p. 70, also see FISHER *et al.*, 2003). However, as a means for capturing bodily movements and sequences of gestures accurately, the use of these markers is unavoidable, as they provide the best method for accurately capturing patterns of discrete body movements. So, as a means of fitting the future research needs of this particular corpus, the use of these devices cannot be legitimately criticised (although in terms of multimodal corpora as 'generic' tools, the reverse is the case).

Fanelli et al. suggest the utility of 3D capture techniques for gesture tracking as an alternative, more unobtrusive alternative to sticky markers. This is something that is still under development by a range of different researchers (i.e. a proven accurate version of such a utility has yet to be released).

Arguably the most naturalistic of the those multimodal corpora listed in figure 1 are the CID, UTEP ICT, SVC and the D64 corpus (despite its' use of sticky markers). The CID contains recordings of German interaction between dyads of people sitting next to each other. The participants are encouraged to discuss any topic or issue they wish, in a bid to provide accounts of conversational data which is as true to 'real-life' as possible. However, again, the conditions in which these recordings took place are to a certain extent experimental, with participants sitting in a closed laboratory and wearing headset microphones.

Participants in the UTEP ICT corpus were also required to wear microphones, although these were wireless and pin-on. For this corpus, cameras are placed around the room as unobtrusively as possible, with participants standing in the middle of the room, able to move freely around the room as desired. Although the content is described as spontaneous, a key limitation of this corpus is that discussions are task based and specifically 'designed to elicit a range of dialog behaviours' (HERRERA *et al.*, 2010, p. 50).

The SVC adopts a recording approach which is even less context-specific and more 'mobile'. It uses portable Smartphone devices to record a range of different public spaces, both inside and outside, with varying light conditions and acoustic properties (SCHIEL; MÖGELE, 2008, p. 2). However, the Smartphone devices are only used to record single participants in these corpora, despite the fact the SVC is based on dyadic conversations. This limits the potential for exploring patterns in dyadic or group behaviour in the data. Furthermore the quality of these recordings is not particularly good and only specific sequences of behaviour, facial expressions and head movements are captured at a high resolution. So for potential reuse in studies which look at other forms of gesticulation, proxemics or other features, this dataset is limited. Though, in truth, this is perhaps more a limitation of the equipment specifications than the recording design methodology. An additional, more general limitation of these corpora is that they are both task-orientated, so although discourse is occurring in real-life contexts, the prescribed nature of the tasks again affects the spontaneity and perceived *naturalness* of the data.

Finally, the D64 corpus is an English based corpus which has been recorded in arguably the most naturalistic setting; that is a domestic living room (see CAMPBELL, 2009), aiming to record language in a truly social situation, so 'as close to an ethological observation of conversational behavior as technological demands permit' (OERTEL *et al.*, 2010, p. 27). Conversations were recorded over long periods of time, the topics of which were not scripted or prompted. As with the UTEP ICT, participants were able to move around the room as they so wished, although they notably did remain seated for the majority of the time. Interestingly 'to add liveliness to the conversation, several bottles of wine were consumed during the final two hours of recording' (OERTEL *et al.*, 2010, p. 27). While the raw data for this corpus is now available, the edited version, complete with transcriptions, codes, tags and so on has yet to be released.

2.5. Availability and (re)usability

As Brône et al. note even now 'truly multimodal corpora including visual as well as auditory data are notoriously scarce' (2010, p. 157), as few have been published and/or are publicly available and no ready-to-use large corpus of this nature is currently commercially *available*.

This is due to a variety of factors, but is most strongly linked to 'privacy and copyright restrictions' (van SON *et al*., 2008, p. 1). Corpus project sponsors or associated funding bodies enforce restrictions on the distribution of materials, and prescriptions related to privacy and anonymity in multimodal datasets reinforce such constraints. Although, notably, plans to publish/release data contained within the D64 (CAMPBELL, 2009) and NOMCO corpora (an 'in-progress' cooperative corpus development project between Sweden, Denmark and Finland focusing on human-human interaction, see BOHOLM; ALLWOOD, 2010) have been confirmed for the near future, these have yet to come to fruition.

2.6. Section overview

In brief, shortcomings of current multimodal corpora and related research approaches and methodologies can be summarised as follows:

- **Design:** Multimodal corpora tend to include synchronised video, audio and textual records designed and constructed primarily to meet a specific research need and/or to answer particular questions.
- **Infrastructure**: Strategies and conventions used to record, mark-up, code, annotate and interrogate multimodal corpora vary dramatically from one corpus to the next. Standardised procedures for each of these processes have yet to be developed and/or agreed.
- **Size**: They are all fairly limited in size, compared to their monomodal equivalents. Multi-million word multimodal corpora do not exist as yet.
- **Scope**: The majority of these corpora tend to be domain specific, mono-lingual and/or are of a specialist nature (i.e. recorded in one discourse context). In some of these, the content is also pre-planned or scripted, and the conditions under which they are recorded are experimental and controlled.

- **Naturalness**: The controlled recording conditions, settings and obtrusive equipment used may compromise the extent to which the data contained within the majority of multimodal corpora is spontaneous and 'naturalistic'.
- **Availability and (re)usability**: No widely available, large scale corpus has been published to date.

The next section outlines ways in which these may be overcome in the future of research in this field.

3. Future developments for multimodal corpora

3.1. Making multimodal corpora 'bigger' and 'better'

While section 2 focused on outlining some limitations related to current multimodal corpus linguistics, the following section seeks to propose some solutions which may help to change the landscape of this area of research for the future.

Firstly, perhaps the obvious solution to criticisms related to the size, scope and availability of multimodal corpora is to strive for the development of bigger, more diverse datasets. Paradoxically, 'what is meant by large corpora is however quite a relative notion' in conventional linguistic research (BLACHE *et al.*, 2008, p. 110). 'In some linguistic fields such as syntax, for instance, corpora of several million words are used, whereas in prosody where most of the annotations are made manually, a few hours of speech are considered as a large corpus' (BLACHE *et al.*, 2008, p. 110). So the appropriate size of a corpus, whether it be mono or multimodal, can only really be determined in the light of what it is to be used for. This means it is perhaps ill informed to qualify size as a strength or shortcoming of those corpora in figure 1 (as addressed in section 2.3) given that, as with the monomodal counterparts, the data in multimodal corpora tends to be research specific, specialist and/or domain specific.

Further to this, 'since language text is a population without limits, and a corpus is necessarily finite at any one point; a corpus, no matter how big, is not guaranteed to exemplify all the patterns of the language in roughly their normal proportions' (SINCLAIR, 2008, p. 30). Corpora are necessarily 'partial', as it is impossible to include *everything* in a corpus as the methodological and practical processes of recording and documenting natural language are selective; ergo 'incomplete' (THOMPSON, 2005, also see OCHS, 1979; KENDON, 1982, p. 478-479 and CAMERON, 2001, p. 71). This is true irrespective of whether a corpus is specialist or more general in nature.

Yet, in an ideal scenario, current multimodal corpora would be larger and more extensive in order to allow them to be more representative of a wider range of language samples/types, to enable the linguist to make better informed observations of language-in-use from a multitude of different perspectives. Further to this, multimodal corpora should accommodate a range of other forms of media, beyond the standard of video, audio and textual data and associated metadata. This projected strand of corpus research and compilation thus works on the understanding that 'communication is not only a linguistic process, but also a multimodal exchange of meaningful information' (BOYD; HEER, 2006). Communication in the digital age is performed via a multitude of multimedia platforms with real-life, everyday discourse witnessing an ever increasing use of digital devices in a variety of different contexts. It is thus vital that we attempt to embrace this evolution in the next phase of multimodal corpus development.

As already noted, early efforts to capture the fluidity and complexity of context (see GOODWIN, 2000, 2007) in real-life discourse have been made by researchers who developed the SVC corpus. The DReSS II project,[9] based at the University of Nottingham, builds on this further. The project is focusing on assembling a corpus of everyday (inter)actions from various different resources, incorporating not only text-based data, such as SMS messages, interaction in virtual environments (for example instant messaging logs and entries on personal notice boards), but also audio and video records from face-to-face conversation, as well GPS logs and a range of other media types. This project is still in progress.

The compilation of such heterogeneous data may enable us to extrapolate further information about communication across a range of different speakers, mediums and environments. In theory, this could assist in the questioning of the extent to which language choices are determined by different spatial, temporal and social contexts in communication.

In reality, there are obviously a whole host of ethical, practical and methodological problems that need to be faced when constructing such

[9] For more information, results and publications from DReSS, please refer to the main project website: http://web.mac.com/andy.crabtree/NCeSS_Digital_Records_Node/Welcome.html

corpora. The realisation of these aims and the successful development of heterogeneous multi-context corpora is heavily reliant on: technological advancements; on the constant refinement of systems that will enable the capture and structuring of natural language-in-use; as well as software that will promote the interrogation of different multimodal datasets. Constraints attributed to questions of scalability are also obviously inherent to the practical implementation of this 'next-step', since, as already identified, the processes of recording, transcribing, time-stamping and coding data remain very time-consuming despite the availability of software for this (for detailed discussions and specific examples of these, see KNIGHT *et al.*, 2009).

Such problems may deter linguists from attempting to create multimodal corpora of this nature because, to date, simple solutions to these problems have yet to emerge. This includes matters of what and how behaviours are quantified, queried and represented to the linguist, and how patterns are statistically assessed and/or analysed.

3.2. Software and hardware requirements

Given that NELC (Nottingham eLanguage Corpus, developed as part of the DReSS II project) is to include multiple forms of varying media types, there are lots of issues to be addressed regarding the optimum ways in which these are recorded, processed, stored and accessed/interrogated by the linguist. The methods employed at each of these stages naturally differ from each media type because they are stored in a variety of file formats, and are typically visualised and represented in different ways. Therefore better devices for recording multiple forms of data, in synchronicity and at a high quality, need to be developed. This will help to enhance the speed at which corpora are composed, giving researchers the chance to extend the size of their corpora at speed.

While cameras and Dictaphones and other recording hardware of an ever increasingly higher specification are constantly being developed, the mobility and functionality of these still recommend that the situated forms of laboratory type recordings will yield the best results. Numerous cameras can be positioned in various locations around the room in order to capture participants from multiple perspectives, from close up and head on (which would support the use of tracking software on resultant images when analysing the data) to birds eye views or more panoramic shots. Similarly eye or movement tracking equipment (such as the sticky markers discussed earlier) can be worn, as required, by participants, in static environments.

More mobile toolkits, as called for here, are becoming increasingly available, although they are still somewhat primitive as the quality of recordings, or the length allowed for recordings, for example, is limited (for an example of such a toolkit under development, see the DReSS II website for more information, also see CRABTREE; RODDEN, 2009).

It would also prove beneficial to look to develop more enhanced tools for the automatic transcription of data. While such tools are currently in existence (such as Speechware[10]), it is widely acknowledged that these are far from accurate, especially when recording spontaneous dyadic or group conversation. Given this, these tools are rarely used in monomodal or multimodal corpus construction.

Thirdly, semi-automated processes of annotating data would also ease the speed at which multimodal corpora are developed and analysed. This may take the form of those digital tracking devices discussed above, designed to allow users to automatically define and subsequently encode specific features of interest in video data (according to specific parameters set by the analyst), to allow for larger scale explorations of language and gesture-in-use to be undertaken with ease (see KNIGHT *et al.*, 2006; BRÔNE *et al.*, 2010 and JONGEJAN, 2010). Although in practice, since such technologies are still 'developing', these tracking techniques are far from perfect, so at present they remain a speculative *potential* rather than *functional* part of the multimodal Corpus Linguistic approach.

Finally, software to support the representation and meaningful interrogation of these datasets needs to be developed as again no standard procedures exist for this in current multimodal corpus methodology. Knight *et al.* identify the following features as being essential to interrogate heterogeneous corpus toolkits, although utilities are likely to need to extend beyond these (2010, p. 17):

- The ability to search data **and** metadata in a principled and specific way, within and/or across the three global domains of data:
    - Devices/ data type(s)
    - Time and/or 'location
    - Participants' given contributions

[10] Speechware is an automatic transcription and speech recognition tool. For more information, visit the following website: <www.speechware.be/en/company.php>.

- Tools that allow for the frequency profiling of events/ elements within and across domains (providing raw counts, basic statistical analysis tools, and methods of graphing such).
- Variability in the provisions for transcription and the ability for, for example, representing simultaneous speech and speaker overlaps.
- Graphing tools for mapping the incidence of words or events, for example, over time and for comparing sub-corpora and domain specific characteristics.

These will seek to build on, combine and extend the functionalities of common monomodal corpus analytical tools, such as those provided by WordSmith Tools (SCOTT, 1999), Sketch Engine (KILGARRIFF *et al.*, 2004) and WMatrix (RAYSON, 2003), as well other forms of social science and qualitative data research software (as mentioned in section 3.1 above). Ideally, such tools should also be free/open source since, to date, much of the field has been monopolised by pay-for-prescription tools and datasets as monies are perhaps necessary to fund the development, maintenance and sustainability of corpus infrastructure (as although funding is often available, commercialisation is often a by-product of this). This somewhat inhibits the accessibility of tools to certain users. Open source software and uiltiities will, in comparison, enhance accessibility for all and will promote the cross fertilization of corpus based methods into other linguistic fields and beyond.

Thankfully, a range of sophisticated corpus tools are being developed in this research 'space', aiming to support some or all of the utilities listed above, within an open-source corpus workbench, including ELAN, DRS, Exmeralda and Anvil. While these tools mainly support corpus construction, maintenance and analysis without providing any corpus 'data' of their own, they set a great example of the potential for the availability of corpus tools for the future.

4. Summary

Multimodal corpora are an important resource for studying and analysing the principles of human communication' (FANELLI *et al.*, 2010). Multimodal datasets function to provide a more lifelike representation of the individual and social identity of participants, allowing for an examination of prosodic, gestural and proxemic features of the talk in a specific time and place. They thus reinstate partial elements of the reality of discourse, giving each speaker and each conversational episode a specific distinguishable identity. It is only when these extra-linguistic and/or paralinguistic elements are represented in records of interaction that a greater understanding of discourse can be generated, following linguistic analyses.

This paper has outlined various strengths shortcomings of current (early) multimodal linguistic corpora. It has focused on outlining characteristics of the basic design and infrastructure of (early) multimodal corpora; their size and scope; the quality and authenticity/naturalness of data contained in them and their availability and (re)usability. The paper has offered some reflections on the strengths of current multimodal corpora alongside some recommendations and a projective 'wish-list' for key areas of development that are likely to be addressed in the future of this area.

The successful implementation of these prospective advancements is heavily reliant on institutional, national and international collaborative interdisciplinary and multidisciplinary research strategies and funding. This is because 'modern research is increasingly complex and demands an ever widening range of skills…..often, no single individual will possess all the knowledge, skills and techniques required' (for discussion on the advantages of cross and multi-disciplinary research see NEWELL, 1984; KATZ; MARTIN, 1997 and GOLDE; GALLAGHER, 1999, p. 281). It is difficult to gauge whether all or any of these projections will ever be fully met, or how the multimodal landscape will look in the next decade or so, although it can be asserted with a fair amount of confidence, that interest in these corpora and associated methodologies will attract an ever increasingly amount of interest as time goes on and our digital worlds continue to expand.

Acknowledgments

The research on which this paper is based is funded by the UK Economic and Social Research Council (ESRC), grant numbers RES-149-25-0035 & RES-149-25-1067.

References

AIST, G.; ALLEN, J.; CAMPANA, E.; GALESCU, L.; GÓMEZ GALLO, C.; STONESS, S.; SWIFT, M.; TANENHAUS, M. Software architectures for incremental understanding of human speech. In: Interspeech 2006. *Proceedings*… Pittsburgh PA, USA: Interspeech, 2006.

ALLWOOD, J. Multimodal corpora. In: LÜDELING, A.; KYTÖ, M. (Ed.). Corpus Linguistics: An International Handbook. *HSK - Handbücher zur Sprach und Kommunikationswissenschaft*, v. 29, n. 1-2, p. 207-225, 2008.

ALLWOOD, J.; BJÖRNBERG, M.; GRÖNQVIST, L.; AHLSEN, E.; OTTESJÖ, C. The Spoken Language Corpus at the Department of Linguistics, Göteborg University. *Forum: Qualitative Social Research*, v. 1, n. 3, 2000. Available at: < http://www.qualitative-research.net/index.php/fqs/article/view/1026>. Retrieved: 12 Jul. 2010.

ANDERSON, A.; BADER, M.; BARD, E.; BOYLE, E.; DOHERTY, G. M.; GARROD, S.; ISARD, S.; KOWTKO, J.; McALLISTER, J.; MILLER, J.; SOTILLO, C.; THOMPSON, H. S; WEINERT, R. The HCRC Map Task Corpus. *Language and Speech*, v. 34, p. 351-366, 1991.

ASHBY, S.; BOURBAN, S.; CARLETTA, J.; FLYNN, M.; GUILLEMOT, M.; HAIN, T.; KADLEC, J.; KARAISKOS, V.; KRAAIJ, W.; KRONENTHAL, M.; LATHOUD, G.; LINCOLN, M.; LISOWSKA, A.; MCCOWAN, I.; POST, W.; REIDSMA, D.; WELLNER, P. The AMI Meeting Corpus. In: Measure Behaviour 2005. *Proceedings*…Wageningen, NL: Measuring Behavior, 2005.

BALDRY, A.; THIBAULT, P.J. *Multimodal Transcription and Text Analysis*: A multimedia toolkit and course book. London: Equinox, 2006.

BERTRAND, R.; BLACHE, P.; ESPESSER, R.; FERRE, G.; MEUNIER, C.; PRIEGO-VALVERDE, B.; RAUZY, S. Le CID: Corpus of Interactional Data -protocoles, conventions, annotations. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix en Provence* (TIPA) v. 25, p. 25-55, 2006.

BLACHE, P.; BERTRAND, R.; FERRÉ, G. Creating and exploiting multimodal annotated corpora. In: LREC 2008. *Proceedings*…Marrakech, Morocco: Sixth International Conference on Language Resources and Evaluation (LREC), 2008. p. 110-115. Available at: < http://www.lrec-conf.org/proceedings/lrec2008/>. Retrieved: July 12, 2010.

BOHOLM, M.; ALLWOOD, J. Repeated head movements, their function and relation to speech. In: LREC 2010. *Proceedings*…Mediterranean Conference Centre, Malta: LREC Workshop on Multimodal Corpora, 2010.

BOYD, D.; HEER, J. Profiles as conversation: Networked identity performance on Friendster. In: HICSS 2006. *Proceedings*... Hawaii: Hawaii International Conference of System Sciences (HICSS-39), 2006.

BRÔNE, G., OBEN, B.; FEYAERTS, K. InSight Interaction- A multimodal and multifocal dialogue corpus. In: LREC 2010. *Proceedings*... Mediterranean Conference Centre, Malta: LREC Workshop on Multimodal Corpora, 2010.

CAMERON, D. *Working with spoken discourse*. London: Sage, 2001.

CAMPBELL, N. Tools and Resources for Visualising Conversational-Speech Interaction. In: KIPP, M.; MARTIN, J.-C.; PAGGIO, P.; HEYLEN, D. (Ed.). *Multimodal Corpora*: From Models of Natural Interaction to Systems and Applications. Springer: Heidelberg, 2009.

CHEN, L.; TRAVIS-ROSE, R.; PARRILL, F.; HAN, X.; TU, J.; HUANG, Z.; HARPER, M.; QUEK, F.; MCNEILL, D.; TUTTLE, R.; HUANG, T. VACE *Multimodal Meeting Corpus*. Lecture Notes in Computer Science, v. 3869, p. 40-51, 2006.

CHOMSKY, N. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press, 1965.

CRABTREE, A.; RODDEN, T. Understanding interaction in hybrid ubiquitous computing environments. In: ACM 2009. *Proceedings*... Cambridge, ACM: 8th International Conference on Mobile and Ubiquitous Multimedia. Available at: <http://portal.acm.org/toc.cfm?id=1658550&type=proceeding&coll= GUIDE&dl =GUIDE&CFID=96741701&CFTOKEN= 20154123>. Retrieved: July 12, 2010.

DYBKJÆR, L.; OLE BERNSEN, N. Recommendations for natural interactivity and multimodal annotation schemes. In: LREC 2004. *Proceedings*...Lisbon: Language Resources and Evaluation Conference (LREC) Workshop on Multimodal Corpora, 2004.

FANELLI, G.; GALL, J.; ROMSDORFER, H.; WEISE, T.; VAN GOOL, L. 3D Vision Technology for Capturing Multimodal Corpora: Chances and Challenges. In: LREC 2010. *Proceedings*...Mediterranean Conference Centre, Malta: LREC Workshop on Multimodal Corpora, 2010.

FISHER, D.; WILLIAMS, M.; ANDRIACCHI, T. The therapeutic potential for changing patterns of locomotion: An application to the acl deficient knee. In: ASME 2003. *Proceedings*... Miami, Florida: ASME Bioengineering Conference, 2003.

FOSTER, M.E.; OBERLANDER, J. Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Language Resources and Evaluation*, v. 41, n. 3/4, p. 305–323, 2007.

FRENCH, A.; GREENHALGH, C.; CRABTREE, A.; WRIGHT, W.; BRUNDELL, B.; HAMPSHIRE, A.; RODDEN, T. Software Replay Tools for Time-based Social Science Data. In: ICeSS 2006. *Proceedings*… Manchester, UK: 2nd annual international e-Social Science Conference, 2006. Available at: <http://www.ncess.ac.uk/events/conference/2006/papers/>. Retrieved: July 12, 2010.

GARFOLO, J.; LAPRUN, C.; MICHEL, M.; STANFORD, V.; TABASSI, E. The NIST Meeting Room Pilot Corpus. In: LREC 2004. *Proceedings*…Lisbon, Portugal: 4th Language Resources and Evaluation Conference (LREC), 2004.

GOLDE, C.M; GALLAGHER, H.A. The challenges of conducting interdisciplinary research in traditional Doctoral programs. *Ecosystems*, v. 2, p. 281-285, 1999.

GOODWIN, C. Action and embodiment within situated human Interaction. *Journal of Pragmatics*, v. 32, n. 10, p. 1489-522, 2000.

GOODWIN, C. Participation, stance and affect in the organisation of activities. *Discourse and Society*, v. 18, n. 1, p. 53-73, 2007.

GREENHALGH, C.; FRENCH, A.; TENNANT, P.; HUMBLE, J.; CRABTREE, A. From ReplayTool to Digital Replay System. In: ICeSS 2007. *Proceedings*… Ann Arbor, Michigan, USA: 3rd International Conference on e-Social Science, 2007. Available at: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.100.755>. Retrieved: July 12, 2010.

GRØNNUM, N. DanPASS - a Danish phonetically annotated spontaneous speech corpus. In: LREC 2006. *Proceedings*…Genoa, Italy: 5th LREC conference, 2006.

GU, Y. Multimodal text analysis: A corpus linguistic approach to situated discourse. *Text and Talk*, v. 26, n. 2, p. 127-167, 2006.

HERRERA, D.; NOVICK, D.; JAN, D.; TRAUM, D. The UTEP-ICT Cross-Cultural Multiparty Multimodal Dialog Corpus. In: LREC 2010. *Proceedings*… Mediterranean Conference Centre, Malta: LREC Workshop on Multimodal Corpora, 2010.

JONGEJAN, B. Automatic face tracking in Anvil. In: LREC 2010. *Proceedings*… Mediterranean Conference Centre, Malta: LREC Workshop on Multimodal Corpora, 2010.

KATZ, J.S.; MARTIN, B.R. What is research collaboration? *Research Policy*, v. 26, p. 1-18, 1997.

KENDON, A. The organisation of behaviour in face-to-face interaction: observations on the development of a methodology. In: SCHERER, K.R.; EKMAN, P. (Ed.). *Handbook of Methods in Nonverbal Behaviour Research*. Cambridge: Cambridge University Press, 1982.

KILGARRIFF, A.; RYCHLÝ, P.; SMR•, P.; TUGWELL, D. The sketch engine. In: EU-RALEX 2004. *Proceedings*... International Congress, Lorient, France: In Proceedings of EU-RALEX, 2004.

KIPP, M. Anvil – A generic annotation tool for multimodal dialogue. In: INTERSPEECH 2001. *Proceedings*... Aalborg, Denmark: 7th European Conference on Speech Communication and Technology 2nd INTERSPEECH Event, 2001.

KIPP, M.; NEFF, M.; ALBRECHT, I. An annotation scheme for conversational gestures: how to economically capture timing and form. *Language Resources and Evaluation*, v. 41, n. 3/4, p. 325-339, 2007.

KNIGHT, D. *Multimodality and active listenership*: A corpus approach. London, UK: Continuum Books, 2011.

KNIGHT, D.; BAYOUMI, S.; MILLS, S.; CRABTREE, A.; ADOLPHS, S.; PRIDMORE, T.; CARTER, R. Beyond the Text: Construction and Analysis of Multimodal Linguistic Corpora. In: ICeSS 2006. *Proceedings*... Manchester, UK: 2nd International Conference on e-Social Science, 2006. Available at: <http://www.ncess.ac.uk/events/conference/2006/papers/>. Retrieved: July 12, 2010.

KNIGHT, D.; EVANS, D.; CARTER, R.; ADOLPHS, S. Redrafting corpus development methodologies: Blueprints for 3rd generation "multimodal, multimedia" corpora. *Corpora*, v. 4, n. 1, p. 1-32, 2009.

KNIGHT, D.; TENNENT, P.; ADOLPHS, S.; CARTER, R. Developing heterogeneous corpora using the Digital Replay System (DRS). In: LREC 2010. *Proceedings*... Mediterranean Conference Centre, Malta: LREC Workshop on Multimodal Corpora, 2010.

LABOV, W. *Sociolinguistic Patterns*. Philadelphia, PA: University of Pennsylvania Press, 1972.

LÜCKING, A.; BERGMAN, K.; HAHN, F.; KOPP, S; RIESER, H. The Bielefeld Speech and Gesture Alignment Corpus (SaGA). In: LREC 2010. *Proceedings*... Mediterranean Conference Centre, Malta: LREC Workshop on Multimodal Corpora, 2010.

MANA, N.; LEPRI, B.; CHIPPENDALE, P.; CAPPELLETTI, A.; PIANESI, F.; SVAIZER, P.; ZANCANARO, M. Multimodal Corpus of Multi-Party Meetings for Automatic Social Behavior Analysis and Personality Traits Detection. In: ICMI 2007. *Proceedings*... Nagoya, Japan: Workshop on Tagging, Mining and Retrieval of Human-Related Activity Information, ICMI'07.

McCARTHY, M.J. *Issues in Applied Linguistics*. Cambridge: Cambridge University Press, 2001.

MCCOWAN, S.; BENGIO, D.; GATICA-PEREZ, G.; LATHOUD, F.; MONAY, D.; MOORE, P.; WELLNER; BOURLAND, H. Modelling Human Interaction in Meetings. In: IEEE ICASSP 2003. *Proceedings...* Hong Kong: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2003.

McENERY, T.; WILSON, A. *Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1996.

MEYER, C.F. *English corpus linguistics*: An introduction. Cambridge: Cambridge University Press, 2002.

NEWELL, W.H. Interdisciplinary curriculum development in the 1970's: the paracollege at St. Olaf and the Western College Program at Miami University. In: JONES, R.M.; SMITH, B.L (Ed.). *Against the current*: reform and experimentation in higher education. Cambridge: Schenkman, 1984.

OCHS, E. Transcription as theory. In: OCHS, E.; SCHIEFFELIN, B.B. (Ed.). *Developmental Pragmatics*. New York: Academic Press, 1979.

OERTEL, C.; CUMMINS, F.; CAMPBELL, N.; EDLUND, J.; WAGNER, P. D64: A Corpus of Richly Recorded Conversational Interaction. In: LREC 2010. *Proceedings...* Mediterranean Conference Centre, Malta: LREC Workshop on Multimodal Corpora, 2010.

RAYSON, P. Matrix: *A statistical method and software tool for linguistic analysis through corpus comparison*. (Doctoral thesis) – Department of Linguistics and English Language/Lancaster University, Lancaster, 2003.

REHM, M.; NAKANO, Y.; HUANG, H-H.; LIPI, A-A.; YAMAOKA, Y.; GRÜNEBERG, F. Creating a standardized corpus of multimodal interactions for enculturating conversational interfaces. In: IUI ECI 2008. *Proceedings...* Gran Canaria: IUI-Workshop on Enculturating Interfaces (ECI), 2008.

SCHIEL, F.; MÖGELE, H. Talking and Looking: the SmartWeb Multimodal Interaction Corpus. In: LREC 2008. *Proceedings...* Sixth International Conference on Language Resources and Evaluation (LREC), 2008. Available at: <http://www.lrec-conf.org/proceedings/lrec2008/>. Retrieved: July 12, 2010.

SCHIEL, F.; STEININGER, S.; TÜRK, U. The SmartKom Multimodal Corpus at BAS. In: LREC 2002. *Proceedinngs...* Las Palmas, Gran Canaria, Spain: 3rd Language Resources and Evaluation Conference (LREC), 2002.

SCOTT, M. *Wordsmith Tools* [Computer program]. Oxford: Oxford University Press, 1999.

SINCLAIR, J. Borrowed ideas. In: GERBIG, A.; MASON, O. (Ed.). *Language, people, numbers* - Corpus Linguistics and society. Amsterdam: Rodopi BV, 2008.

THOMPSON, P. Spoken Language Corpora. In: WYNNE, M. (Ed.). *Developing Linguistic Corpora*: a Guide to Good Practice. Oxford: Oxbow Books, 2005.

TROJANOVÁ, J.; HRÚZ, M.; CAMPR, P.; ŽELEZNÝ, M. Design and Recording of Czech Audio-Visual Database with Impaired Conditions for Continuous Speech Recognition. In: LREC 2008. *Proceedings*... Marrakech, Morocco: Sixth International Conference on Language Resources and Evaluation (LREC) 2008. Available at: < http://www.lrec-conf.org/proceedings/lrec2008/>. Retrieved: July 12, 2010.

VAN SON, R. J. J. H.; WESSELING, W.; SANDERS, E.; VAN DER HEUVEL, H. The IFADV corpus: A free dialog video corpus In: LREC 2008. *Proceedings*... Marrakech, Morocco: Sixth International Conference on Language Resources and Evaluation (LREC), 2008. Available at: <http://www.lrec-conf.org/proceedings/lrec2008/>. Retrieved: July 12, 2010.

WOLF, J.C.; BUGMANN, G. Linking Speech and Gesture in Multimodal Instruction Systems. In: IEEE RO-MAN 2006. *Proceedings*... Plymouth, UK: 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN06), 2006.

ŽELEZNY, M.; KRNOUL, Z.; CÍSAR, P.; MATOUŠEK, J. Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis. *Signal Processing*, v. 83, n. 12, p. 3657-3673, 2006.

Recebido em 13/07/2010. Aprovado em 08/05/2011.