Title: DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. [a)]

Authors: Rachel Baker and Valerie Hazan [b)]

Affiliation: Speech, Hearing and Phonetic Sciences, UCL, Chandler House, 2 Wakefield Street, London WC1E 1PF, UK.

Running head: DiapixUK

a) Portions of this work were presented in: "LUCID: A corpus of spontaneous and read clear speech in British English", a presentation at DISS-LPSS Joint Workshop, Tokyo, Japan, 24-25 September 2010

b) Electronic mail: v.hazan@ucl.ac.uk, Telephone: 0044 (0)207 679 4076

ABSTRACT

The renewed focus of attention on investigating spontaneous speech samples in speech and language research has increased the need for recordings of speech in interactive settings. The DiapixUK task is a new and extended set of picture materials based on the Diapix task by Van Engen et al. (2010) where two people are recorded while conversing to solve a 'spot the difference' task. The new task materials allow for multiple recordings of the same speaker pairs due to a larger set of picture pairs that have a number of tested features: equal difficulty across all twelve picture pairs, no learning effect of completing more than one picture task and balanced contributions from both speakers. The new materials also provide extra flexibility making them useful in a wide range of research projects; they are multi-layered electronic images that can be adapted to suit different research needs. This paper presents details of the development of the DiapixUK materials along with data taken from a large corpus of spontaneous speech that are used to demonstrate its new features. Current and potential applications of the task are also discussed.

INTRODUCTION

Speech produced during an interaction differs from read speech in lexical and grammatical content (e.g. Blauuw, 1994), in segmental phonetic characteristics and in prosodic characteristics (e.g. Lann & Van Bergem, 1993), which explains why these two speech styles are perceptually distinguishable even if they contain the same speech material. Despite these differences, many studies still use read speech to investigate research topics such as the acoustic-phonetic characteristics of different speaking styles. Undoubtedly, spontaneous interactive speech recordings are challenging to process and analyse, but the increased ability to generalise results from these data to naturally-occurring speech makes this challenge worth overcoming. Of the studies that have investigated interactive speech, many have used problem-solving tasks to elicit the speech in a controlled setting in an attempt to record conversations with similar topics and structures. This paper describes the development of a set of materials (DiapixUK), which extend the recently-developed Diapix task (Van Engen et al., 2010) – an easy way to collect spontaneous dialogs between two speakers, while having some control over the lexical content of the discourse. The task involves two people conversing to find differences between two versions of the same picture. The three picture pairs designed by Van Engen et al. (2010) are suited to American English participants and the picture content cannot be easily altered. The new DiapixUK picture materials make this methodology suitable to the needs of a wider range of researchers and are also potentially suitable for use in clinical settings. The twelve new picture pairs, which have been standardised and shown to be of equal difficulty, provide ample material for the elicitation of interactive speech over multiple sessions. They can also be adapted to suit the needs of different types of speakers due to their electronic multi-layered format, which can be easily modified. In this paper, the development and potential applications of the new task materials are presented; data from a large scale study was used to test some key features of the task set.

Existing methods for the collection of spontaneous speech corpora

Three broad approaches are generally used to record spontaneous speech in the laboratory. One approach involves recording spontaneous monologues. For example, in the Boston Directions Corpus (Nakatani, Grosz & Hirschberg, 1995) participants were asked to provide directions (of varying complexities) to a silent participant. However, this type of 'imagined' interaction may differ from real interactions in which speech is produced with clear communicative intent. In fact, Knoll, Scharrer & Costall (2009) compared infant- and foreigner-directed speech to imaginary partners produced by students and actresses and found that actresses were better able to replicate the speech styles as produced in real interactions. We therefore focus our attention in this paper on corpora that involve spontaneous interactions between pairs of participants. The simplest method to elicit such speech is to instruct two speakers to converse spontaneously about a particular topic; another is to record two speakers while they are engaged in a problem-solving task that requires verbal communication. An example of the former approach is the method used to collect speech in Spontal, a corpus of spontaneous dialog of audio, video and motion capture (Edlund et al, 2010). Pairs of participants were told that they could talk about any subject during the recording session and could cease talking whenever they wished. Although this approach results in very natural speech, for research that requires some control over the lexical content of conversations it would not be the most appropriate method. A study by Uther, Knoll & Burnham (2007) involved a more constrained 'free-conversation' session: mothers were asked to talk to a native or non-native English confederate about three toys (whose labels were the keywords of interest) that they might or might not buy for their infant. This allowed the researchers to obtain enough repetitions of the three keywords for their analysis. However, it might prove difficult to use this approach to elicit repetitions of a large range of keywords or ones that cannot be easily represented by everyday objects.

The alternative method of using a problem-solving task to collect spontaneous speech is another step away from corpora such as Spontal but it allows for more control over the speech collected, which for many speech and language research questions, is a necessity. Popular problem-solving puzzles have been used to record spontaneous speech produced by two participants. Cooke & Lu (2010) asked pairs

of participants to cooperatively complete Sudoku puzzles, which works well for obtaining many repetitions of number words. Crawford, Brown, Cooke & Green (1994) recorded two pairs of people in the same room each solving a different crossword. The resulting speech contained a wider range of lexical items than from Sudoku. However, in both approaches there is the potential problem of each person being able to solve at least part of the task without communicating with their partner because both people can see the same information. Also, there is large variation in how skilled a person is at solving these puzzles. It would be possible to use simple versions of Sudoku and crossword puzzles for recording childrens' speech. However, a number of inventive computer game scenarios have been devised that are specifically for use in child language research. In Batliner et al.'s (2004) study children interacted with a robot AIBO dog that either followed or disobeyed their commands and Bell et al. (2005) devised a fun interactive computer game where 8 – 15 year olds were asked to collaborate with cooperative and uncooperative embodied characters in the game.

The most widely-used problem-solving approach is the Map Task (Brown, Anderson, Yule & Shillcock, 1983), which was developed by Anderson et al. (1991). This task involves an 'instruction giver' communicating details of a map route and of different key elements on the map to an 'instruction follower' who has no indication of the route on their map and some different map elements. The original Map Task materials consist of 16 pairs of maps. Each map contains approximately 10 key words or phrases represented with line drawings that have an accompanying label, e.g. a picture of 3 boats with the label 'moored boats'. The follower must draw a line on his or her map according to the giver's instructions. Task success in the Map Task is measured using a deviation score. This is defined as the difference in area between the route on the information follower's map (drawn during the task) and the original map described. The task has been used in a number of studies investigating, for example, word segmentation cues (White, Wiget, Rauch & Mattys, 2010), the role of visual cues in communicating information (Anderson et al., 1991). A variety of Map Task corpora exist in different languages and dialects besides the original British English Map Task corpus, e.g. Australian English (Millar, Vonwiller, Harrington, & Dermody, 1994), Japanese (Yasuo et al., 1999). Due to the conversations being, by design, fairly one-sided, the

information giver typically speaks more than the information follower. The number of word tokens produced by the information givers in all conversations in the Edinburgh Map Task corpus (without eye contact) was about 80,000 whereas the information followers only produced about 55,000 tokens (Anderson et al., 1991). In fact, Forsyth, Clarke & Lam (2008), found that longer talkspans (sequences of uninterrupted words or short phrases by the same person) by followers were negatively correlated with task success (measured by deviation score). Therefore, in order to record an equal amount of speech for each speaker in a pair, at least two Map Tasks need to be undertaken where the giver and follower roles are switched between tasks. The task provides lots of opportunities for direction-giving commands and requests but because the pictures are accompanied by labels, the opportunity and need for detailed description of the pictures is limited.

A recent alternative to the Map Task, which has many of its advantages but is arguably even more flexible is the Diapix task (Van Engen et al., 2010). Diapix involves pairs of participants engaged in a 'spot the difference' picture task. Each participant is presented with a different version of the same cartoon-style picture and the two participants have to collaborate to find ten differences between the two pictures without seeing each other's picture. As with the Map Task, the differences in the Diapix task can be designed to encourage the production of specific words or phrases so that segmental aspects of speech can be analysed, e.g. vowel space. The recordings can also be used to investigate more global characteristics of speech such as speech rate, fundamental frequency etc. Success or communicative efficiency in the Diapix task can be assessed using transaction time, which is the time taken to complete the task (Van Engen et al., 2010). One important difference between the Diapix task and the Map Task lies in the role of the two participants. The Diapix task can be completed with no defined role given to each participant, i.e. both participants are encouraged to work collaboratively and to equally contribute to the conversation. This means that a roughly equal amount of speech should be produced from each participant and the types of conversations that are elicited are closer to natural communication than the 'instruction giving/receiving' feature of the Map Task. The dynamics of the Diapix conversations can also be manipulated so that one person talks more than the other, i.e. giver/follower structure, by instructing one participant to 'take the lead' in the conversation. Another

advantage of Diapix is that the richness of the picture content necessitates the participants to use a wide range of linguistic structures in order to identify the differences (Van Engen et al., 2010). Although the conversations can involve instructions, they often involve an exchange of information without one leader emerging. This is, in part, due to the variety of differences that occur in the pictures, i.e. the type of difference and the level of complexity, and also due to the nature of the pictures, which contain a lot of information, e.g. objects and people that can be described in a variety of ways.

We believe that there is potential in the Diapix task to be useful for a wide range of studies as an alternative to existing tasks. The Diapix materials have been used in the Wildcat Corpus (Van Engen et al., 2010) an extensive corpus examining interactions across different combinations of native and nonnative speakers; these materials are highly appropriate for further studies that would require the recording of a single dialog for any given speaker dyad. However, the existing materials are limited by the fact that they include a small set of pictures that were not standardised for difficulty and were hand-drawn so they cannot be modified for specific needs of other research studies. In order to increase the flexibility of the Diapix task for our own research aims and also for other future projects (both research and clinical), a new and extended set of Diapix materials, the DiapixUK materials, has been developed. A number of key objectives were addressed in the development of the new materials, i.e. increasing the number of Diapix pictures to twelve pairs of equal difficulty and establishing the lack of a significant learning effect of completing more than one picture pair. Equal difficulty and the lack of a learning effect would make it feasible for the same people to do multiple Diapix tasks (either in the same session or across sessions). This is imperative for studies involving the same speakers in different test conditions or for longitudinal studies, as it has to be established that any effect obtained across repetitions of the task are not due to differences in task complexity. Other objectives were to maintain the balanced nature of the task, i.e. equal contributions from both participants, to design the pictures so that a set of keywords could be elicited in one or both speakers, and to construct the materials in such a way that individual elements within the pictures could be easily changed.

In the following sections we first describe the process of developing the new set of Diapix materials. Next, we present an investigation of the planned key features of the materials using a large corpus of spontaneous speech. Finally other applications of the new Diapix materials are discussed.

DEVELOPMENT OF DIAPIXUK TASK MATERIALS

Like the original Diapix task (Van Engen et al., 2010), each DiapixUK task consists of two versions of the same cartoon picture that are different in small ways; two people are each given a different version and have to cooperate to find all of the differences without seeing each other's picture. The number of differences between pictures was increased from 10 to 12 so that three differences could be positioned in each quadrant of the picture. Increasing the number of differences by two also allows for more speech to be recorded without lengthening the task to the extent that participants become bored. Twelve picture-pairs were created, which belong to one of three themes: beach (B), farm (F), and street (S) scenes, with four pairs per theme. See Figure 1 for examples of one picture pair per scene. (See supplementary material for all pictures.) The large picture set means that the task can be completed multiple times by the same person. The multiple pictures per theme increase the chance of similar vocabulary being produced in a set of task sessions, e.g. beach scene 1, farm scene 1 and street scene 1 could be completed in an initial session, and beach, farm, and street scenes no. 2 could be completed in a subsequent session etc. The scenes were hand-drawn in a cartoon style and were moderately humorous to maintain interest in the task.

Differences involve either a change in an object or action across the two pictures (e.g. a red ball of wool in picture A vs. a blue ball in picture B; girl holding a beach ball in picture A vs. girl sitting on a beach ball in picture B) or a missing item in one of the pictures (e.g. a sign on a shop in picture A vs. no sign in picture B). The differences involving missing items were evenly distributed over pictures A and B to encourage equal involvement from both participants during the task. Where possible, differences were represented pictorially but due to constraints outlined below, some of them involve text (e.g. signs on posters or shop fronts). To allow for greater flexibility of using these materials in future projects, the pictures were scanned to create digital line drawings, which were coloured in

using Adobe Photoshop. Each item in the pictures, e.g. an object or person, was assigned to a separate layer in Photoshop so that objects can be added, removed or modified according to particular research needs.

Many aspects of speech during an interaction can be examined using the Diapix task, e.g. global acoustic-phonetic characteristics such as fundamental frequency range and mean, vowel space and speech rate, or discourse functions such as uses of backchannels, hedges etc. To also allow for segmental speech analyses, the original Diapix task had differences that were based around keywords containing a selection of vowels. The new pictures retain this characteristic except that the keywords begin with one of four consonants (/p,b,s,ʃ). Each of the 36 keywords is a monosyllabic CV(C) word that belongs to a (near) minimal word pair (e.g. *pear*/*bear*; *sign*/*shine*). The /p-b/ VOT contrast and the /s-ʃ/ contrast were chosen because they have previously been used to examine speaker variability in speech production (e.g. Newman, Clouse & Burnham, 2001; Allen, Miller & DeSteno, 2003; Theodore, Miller & DeSteno, 2009). The 36 keywords were divided into three sets and each set of twelve was used for a different picture theme, i.e. beach, farm or street, so that completion of three tasks (a beach, a street and a farm scene) is likely to result in the production of all keywords. As far as possible, words were assigned to appropriate themes, e.g. 'sheep' in the farm theme, 'shell' in the beach theme. A training picture pair of a park scene was also created to familiarise participants with the task procedure. It has twelve differences (three in each quarter) but the differences are not related to the keyword set.

FIGURE 1: DiapixUK task materials. Top: Beach scene 3; Middle: Farm scene 3; Bottom: Street scene 3. (Version A on left; version B on right)

TESTING THE KEY FEATURES OF THE DIAPIXUK TASK USING THE LUCID CORPUS

Following minor revisions made to the pictures to address shortcomings identified in a pilot study, the DiapixUK materials were used in the LUCID corpus (London UCL Clear Speech in Interaction Database: Baker & Hazan, 2010), which contains recordings of 40 native Southern British English speakers as they carry out the Diapix task in a range of different communicative situations. The dataset that will be examined here is a set of recordings of pairs of friends doing a series of 3 Diapix tasks in succession in good listening conditions. Planned key features of the DiapixUK materials were assessed using these recordings and the five features are outlined in Table 1. Van Engen et al. (2010) had found that the original Diapix materials did lead to balanced speech between speakers, but none of the other features had previously been assessed.

Table 1: Planned key features of the DiapixUK task materials and related advantages.

Key feature	**Advantage of feature**
A. Each task lasts for a minimum of approximately five minutes	Provides sufficient speech material for acoustic-phonetic and other linguistic analysis
B. Balanced speech between both speakers within a pair	Speech data can be collected for both speakers during one task
C. No learning effect of completing more than one picture in a session	Allows large amounts of speech data to be collected using multiple tasks
D. Equal difficulty across picture pairs	Allows any combination of picture pairs to be used in a study
E. Reliable production of keywords within each task by both speakers	Allows detailed phonetic analysis to be conducted on aspects of the keywords

Participants

Participants comprised forty native speakers of Southern British English (SBE, mean = 22.6 years, *SD* = 2.75, age range: 19 – 29, 50% women). All were students or staff from the University of London. They volunteered with a friend of the same gender so there were 20 'friend' pairs (10 M, 10 F). Each potential participant was recorded reading 'accent-revealing' sentences before being accepted onto the study, to ensure that s/he was from the appropriate accent group. All participants were screened for normal hearing thresholds (20dB HL or better for the range 250 – 8000Hz), reported no history of speech or language disorders and all but two had no specific experience communicating with people

with speech and language difficulties. Participants were naive as to the purpose of the recordings but were debriefed afterwards and paid for their participation.

Recording set-up and procedure

Participants were seated in different rooms and communicated via Beyerdynamic DT297PV headsets equipped with cardioid microphones. Speech was recorded at a sampling rate of 44,100 Hz using an E-MU 0404 USB audio interface and Adobe Audition. The speech of each participant was saved on a separate audio channel to facilitate transcription and acoustic analysis. Each pair began the session with the training task and then did three Diapix tasks in succession: 1 Beach scene, 1 Farm scene and 1 Street scene. The pictures were selected from the whole set of DiapixUK task materials and were chosen so that scenes (B, F, S) and number (pictures 1, 2, 3, 4) were counterbalanced across pairs.

During the pilot study, participants used different strategies to talk about the pictures, e.g. going from left to right, top to bottom or going clockwise from the top left. To increase the chance of the conversations being more comparable across speaker pairs in terms of the order in which the content was discussed, participants were told to start each task in the top left corner of the picture and work in a clockwise manner around the scene. Both participants were encouraged to contribute to finding the differences. The experimenter monitored the recording from outside both of the recording rooms and stopped each recording either once the twelve differences were found, or after at least 15 minutes had lapsed and participants could not locate the final differences.

Sound file post processing

There were 120 single channel files (3 conversations for 20 participant pairs, with 2 single channel files per conversation). For each file, the speech was orthographically transcribed using freeware transcription software (Wavescroller) to a set of transcription guidelines based on those used by Van Engen et al. (2010). The transcripts were automatically word-aligned to the sound files using NUAligner software, which created a Praat TextGrid. Both transcription and alignment software were developed by Northwestern University. The word-level alignment was hand-checked in

approximately two-thirds of the file set. All audio files were normalized to a mean amplitude of 15dB (with soft limiting) in Adobe Audition.

A: Does the task provide enough speech material for acoustic/linguistic analysis?

The mean length of recording for each conversation is 7.7 minutes, with the mean duration of actual speech (i.e., linguistic material only, excluding silences and pauses) per participant per picture being 2.6 minutes. So across the 3 pictures, there is approximately 8 minutes of speech per participant. The mean number of words produced per participant per conversation, calculated over all pictures and participants, is 613 (see supplementary material for details of individual participant means). Using the amount of task materials per speaker from Smiljanic and Bradlow (2005, 2008) as a guide, the data suggest that the DiapixUK materials provide sufficient speech material for acoustic and linguistic analysis.

B: *Within a conversation is the speech contribution balanced?*

The balanced nature of the conversations in the original Diapix task materials is one of the features that distinguished it from the Map Task so this feature was tested in the new materials using the number of words produced by each speaker. In the subset of the Edinburgh Map Task corpus (Anderson et al., 1991) that is the most comparable to the recording setup in the LUCID corpus (where there was no eye contact), the percentage of words produced by all instruction givers was 68% and the percentage produced by followers was 32%. This contrasts with the more balanced contributions by the speakers in the current corpus where the percentage of total words produced by 'A' speakers is 51% and the percentage produced by 'B' speakers is 49%. There is also a positive relationship between the number of words produced between speaker A and speaker B within conversations (Pearson's $r = 0.64, p < 0.01$: Figure 2), which is also evident for male and female speakers separately (male: $r = 0.61$, $p < 0.01$; female: $r = 0.71$, $p < 0.01$).

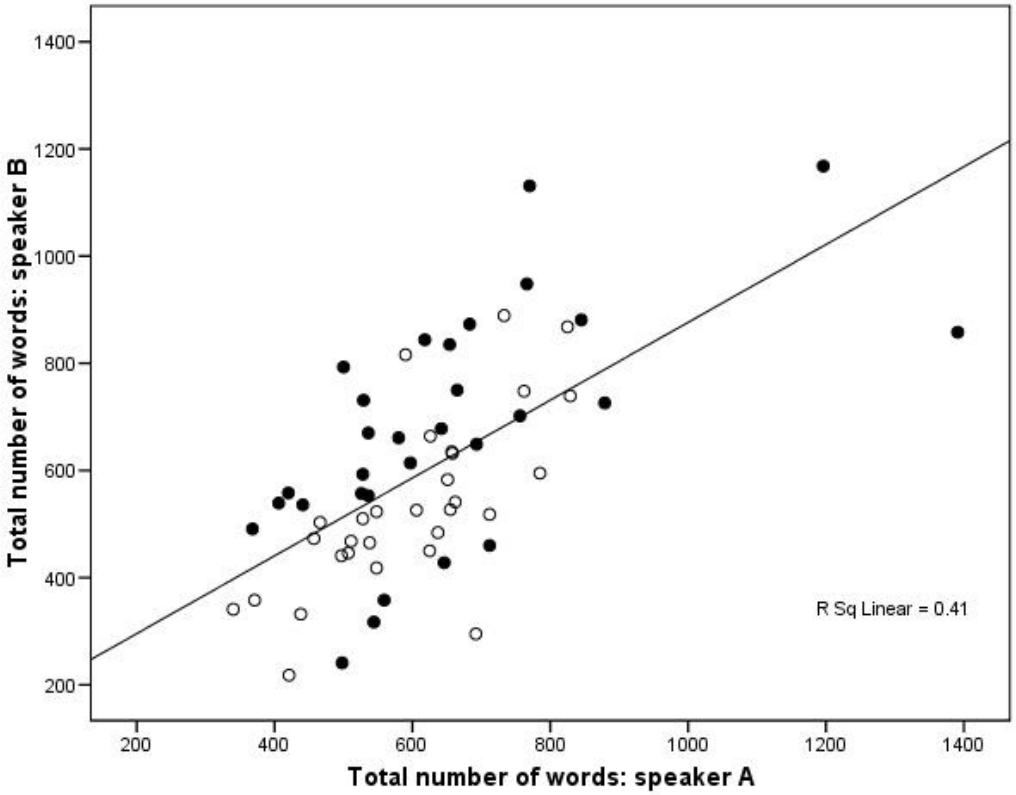

Figure 2: Scatterplot of total number of words for speaker A versus total number of words for speaker B for each of the 60 conversations. Male speaker pairs are represented by black dots and female speaker pairs by white.

C: *Is there a learning effect of doing more than one picture task?*

The pilot study indicated that there is no significant learning effect of doing more than one DiapixUK task (after having completed a training task). To assess the possible existence of a learning effect in the main corpus, a measure of task difficulty was used to compare performance on each picture pair in each position. Following Van Engen et al. (2010), transaction time was used. In this case, it was defined as the time taken to find eight differences. Transaction time does not significantly differ between tasks 1, 2 and 3 within a recording session for the whole dataset [$F(2,72) = 2.2$, $p = 0.115$] (see Figure 3) or for male and female speakers separately [male: $F(2,38) = 0.46$, $p = 0.64$; female: $F(2,34) = 3.1$, $p = 0.06$]. The lack of a significant learning effect, i.e. the fact that participants do not get significantly better at the task after solving one or two pictures, means that several picture tasks

can be carried out by a same pair of speakers during a session. As stated above, this is imperative for studies that require a large amount of speech data per participant.

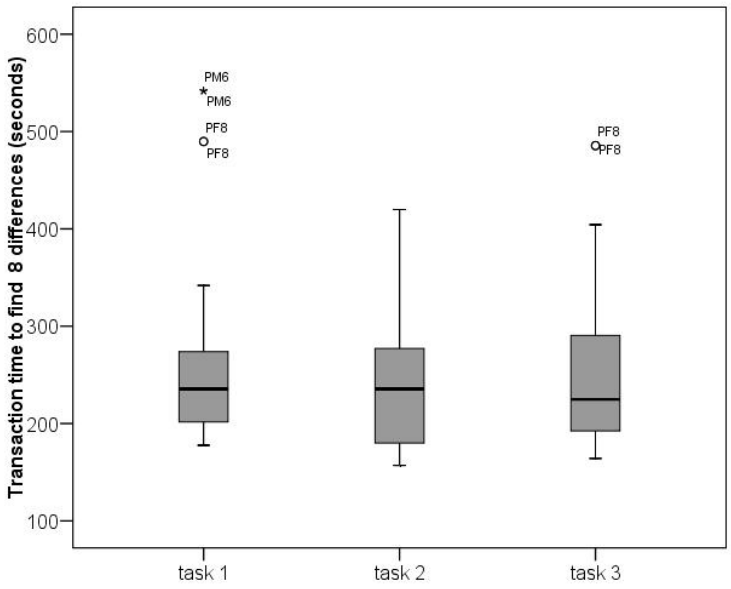

Figure 3: Transaction time taken to find 8 differences in the DiapixUK tasks in positions 1, 2 and 3 for all participant pairs. Black bars show median values; grey boxes show the interquartile range and lines show the range excluding any outliers, which are indicated by an asterisk (male speakers) or a circle (female speakers).

D: *Is there an equal level of difficulty across pictures?*

After the pilot study, three pictures (i.e. Beach, 4, Farm 4, Street 4) were simplified because they were shown to take more time to solve than the other pictures. Transaction time was compared across the revised picture pairs to test if the modifications had made these pictures more similar to the other pictures. Figure 4 shows that although these pictures have a larger inter-quartile range than the other pictures none of the pictures take significantly more time to solve than others: picture number: [$F(3,15) = 1.2$, $p = 0.37$]; picture scene: [$F(2,30) = 2.0$, $p = 0.15$]; number by scene interaction: [$F(6,30) = 1.6$, $p = 1.9$]).

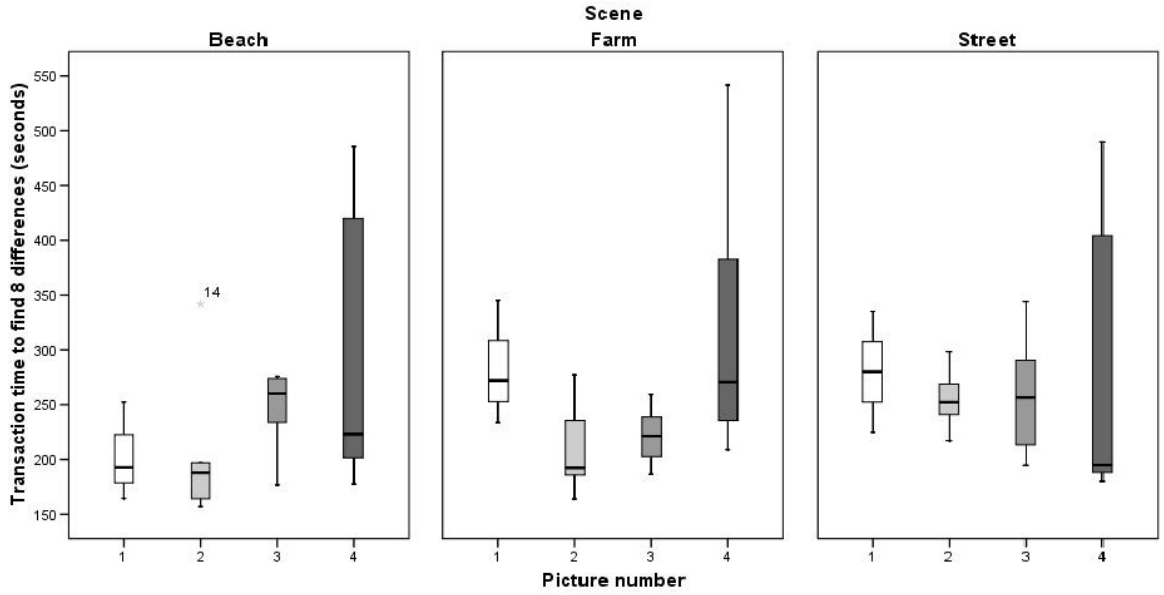

Figure 4: Transaction time (seconds) split by picture scene (B,F,S) and picture number (1,2,3,4). Black horizontal bars show median values; boxes show the interquartile range; lines show the range excluding any outliers, which are indicated by an asterisk and speaker pair number.

E: *How frequently are the keywords produced when doing the task?*

Table 2 shows the median number of tokens of keywords produced for each picture pair per speaker. For all the beach scenes, the median number of tokens was between 0 and 2.5, for the farm scenes the median was between 0 and 4 and for the street scenes the median was between 0 and 8.5. The median range for the street scenes is large due to large numbers of tokens being produced of the keywords 'sign' and 'shop'.

Table 2: Median number of tokens per keyword produced for each picture pair per speaker (with max number in parentheses).

keywords	***bin***	***beach***	***ball***	***Paul***	***peach***	***push***	***sea***	***sock***	***seat***	***shack***	***shell***	***shore***
Beach 1	1(1)	1.5(3)	1.5(2)	1.5(3)	1(2)	1(3)	2(3)	1(3)	1.5(2)	2(4)	1(1)	0(1)
Beach 2	1(3)	1(2)	1(7)	1(1)	1(2)	1(3)	2(4)	3(7)	1.5(3)	1(5)	1(2)	0(1)
Beach 3	0.5(4)	1(2)	1(6)	1(1)	2(4)	2(4)	0(3)	1(2)	2(5)	1(3)	1(4)	0(2)
Beach 4	1(4)	2.5(3)	1.5(4)	1(5)	2(4)	0.5(3)	1(3)	2(3)	0.5(2)	2(3)	1(3)	0(2)
keywords	***bee***	***bush***	***buy***	***pear***	***pin***	***peas***	***Sue***	***saw***	***sack***	***sheep***	***sheet***	***shoot***
Farm 1	1(2)	1(4)	1(2)	1(3)	0(2)	3(5)	1(2)	0.5(1)	2(4)	1.5(3)	0(4)	0(1)
Farm 2	2(4)	0.5(2)	1(2)	2(5)	1(2)	1(3)	0(1)	1(3)	1(3)	1(3)	1.5(4)	1(3)
Farm 3	2(4)	1(3)	1(1)	0(4)	1(3)	1(1)	1(3)	0(3)	0(2)	1(3)	0(2)	2(5)
Farm 4	4(11)	2(4)	0.5(1)	2(3)	1(6)	1.5(6)	1(1)	0.5(2)	2(3)	2(6)	1(3)	1.5(2)
keywords	***bear***	***bet***	***bill***	***pill***	***pie***	***pet***	***suit***	***sell***	***sign***	***shine***	***shop***	***shoe***
Street 1	1(2)	2(4)	0(0)	0.5(1)	1(2)	1.5(4)	1(3)	1(2)	8(11)	1(2)	4(11)	3(4)
Street 2	1(2)	3(7)	1(2)	1(3)	1(3)	0.5(2)	0(1)	1(3)	4(7)	1.5(6)	5(10)	2(5)
Street 3	1(2)	0.5(2)	0(0)	0.5(2)	2(6)	1(3)	1(3)	0(2)	5(12)	2(3)	4(9)	2(4)
Street 4	1(2)	1(3)	0(0)	1(2)	2(4)	1(5)	1.5(6)	1(2)	8.5(17)	1(4)	4(10)	3(6)

The data show that it is difficult to reliably elicit multiple repetitions of each keyword using this type of task, largely because nouns are often replaced with pronouns after first mention. However, as there were 9 different keywords starting with each of the four phonemes under investigation (/p/-/b/ and /s/-/ʃ/), a number of repetitions of these phonemes are likely to be obtained even if a few of the keywords were not produced. It is also worth remembering that analyses of global acoustic-phonetic features of speech, e.g. fundamental frequency measures, speech rate, vowel space, long-term average spectrum can be carried out on the entirety of the speech recorded in the Diapix session and are a useful addition to segmental analysis.

In summary, four out of the five planned key features of the DiapixUK materials have been confirmed. Each conversation provides enough speech material for acoustic-phonetic and linguistic analysis, the contribution of each speaker is roughly balanced, there is no significant learning effect of completing more than one task in a session and the 12 pictures are of approximate equal difficulty. Moreover, there is no difference between male and female speakers in terms of balance of speech and task learning. The feature that is the least robust is the elicitation of multiple tokens of each keyword. However, even with a small number of repetitions, phonetic investigation of sounds that occur in each of the minimal word pair sets (/b/ and /p/, /s/ and /ʃ/) is still achievable, as is more global acoustic-phonetic analysis of longer stretches of speech.

OTHER APPLICATIONS OF THE DIAPIXUK TASK

The DiapixUK materials are freely available to the research and clinical community and can be accessed by requesting a free user account on the Northwestern Online Speech/Corpora Archive and Analysis Resource (OSCAAR, Kendall (2010): http://oscaar.ling.northwestern.edu/). The pictures can be used in their current form or they can be modified to suit a particular requirement using basic functions in Adobe Photoshop. The advantages of using the picture materials in their current form have been highlighted above. That is, the materials are of equal difficulty and normative data for a variety of transactional and acoustic-phonetic measures are available for a set of 40 native British English speaking participants. This normative data could be compared with new recordings of spontaneous speech from Diapix tasks recorded with other populations of speakers such as children, L2 and bilingual speakers or speakers with hearing or language impairments. However, the flexibility of the new DiapixUK materials means that they can be modified to suit a particular research need, e.g. for a language other than English, for much younger or much older age groups. This could be particularly useful in clinical settings if the intended task participants have low cognitive abilities as the pictures could be simplified to an appropriate level. However, it is advisable to be aware of the potential impact of any changes on the key characteristics of the pictures if trying to retain the features of the existing task materials.

In terms of language and cultural background, in their current form, the materials are well-suited to a British audience, and they would be suitable for use with speakers of American English after making some small adjustments (e.g., the shape of garbage cans, which appear in the beach scenes, is different in the UK and USA). The materials have been used successfully in a study assessing the speech of L2 English speakers whose native language is Finnish (Granlund, 2010). Text in the pictures can also be changed if the task is to be done in a language other than English, and Granlund (2010) successfully adapted the pictures for her speakers when they did the task in Finnish. The task materials can also be used for different age groups. In the data collected so far, adults aged between 18 and 39 years were

recorded doing the task but the materials have also been found to be suitable in pilot recordings with 10 year olds.

Various linguistic phenomena that are particularly prevalent in conversation can be investigated using the Diapix materials. For example, a subset of pictures was adapted for a small-scale study that assessed phonetic convergence of vowels across two different British English accents (Evans, Hazan, Baker & Cyrus, 2010). Some of the objects were changed in the pictures to elicit words containing the target vowels. Conversations collected using the Diapix task could also be useful for discourse or conversation analysis of conversations in constrained settings. For example, another study investigated gender effects in discourse markers such as back-channels (Chan, 2010).

A potential clinical application of the DiapixUK task materials would be to use them in therapy sessions or as part of an assessment battery. Conversational tasks have previously been used in studies assessing the referential abilities of hearing-impaired children. Ibertsson et al. (2009) used a referential task to assess requests for clarifications in conversations between children with cochlear implants and those with no hearing impairment, who took turns to be 'information giver' and 'receiver'. Reuterskiöld-Wagner, Nettelbladt & Sahlen (2001) used the same task to assess the referential abilities of children with specific language impairment. A variant on this referential task involves listening to a puppet describe some pictures and selecting the correct picture from an array (Arnold, Palmer & Lloyd, 1999).

The Diapix task differs from these referential tasks in that the relationship between participants can be varied, i.e. both people can be asked to contribute to finding the differences or one person can be given the task of describing the picture to a partner. The Diapix task would be particularly useful in therapy sessions where the same task needs to be repeated on successive sessions. The existing set of DiapixUK materials could be used or simplified versions could be created if necessary. For example, objects could be removed from the pictures to simplify the scenes and to make them suitable for use with children. It is quite common for a speech and language therapy client to be asked to complete the

same picture task multiple times over different therapy sessions. The availability of multiple pictures of equal difficulty that are likely to elicit similar vocabulary would relieve boredom and would potentially engage the client more than repeating exactly the same task. In terms of assessment, where there is an interest in a global measure of communication efficiency before and after therapy, a measure of communication ease, e.g. transaction time, could be used. A small-scale study has already used a subset of the picture pairs to investigate relative communication difficulty in different cochlear implant simulations. These are just a few of the ways in which the new task materials could be used by other researchers and clinicians. One of the aims in developing these materials was to make them as flexible as possible for future work so that they become a useful resource.

CONCLUSION

In this paper, the DiapixUK task, which is a new set of task materials for the problem-solving Diapix task by Van Engen et al. (2010) have been presented. The development of a 'standard' set of picture pairs will likely be of use to other researchers and clinicians who need a large set of task materials of equal difficulty. Furthermore, the ability for the pictures to be modified to suit different research topics and participant groups extends their usefulness. However, it is important to note that there is a trade-off between flexibility and standardisation and the former compromises the latter, meaning that some of the DiapixUK features might not hold if modifications are made. An analysis of the spontaneous speech collected as part of the LUCID corpus has shown that the objectives that we set out for the DiapixUK materials have mostly been achieved, although the keywords were not all produced by all of the participants in the interactions. Both the LUCID corpus and the DiapixUK materials are available as part of the OSCAAR archive facility that is hosted by Northwestern University (Kendall, 2010: http://oscaar.ling.northwestern.edu/).

ACKNOWLEDGEMENTS

Thanks to our collaborator Ann Bradlow and colleagues at Northwestern University, designers of the original Diapix task, for their contributions to the development of the DiapixUK materials and for giving us access to some of the software and other facilities used for our corpus analysis. Thanks also

to Tyler Kendall for his help in making the LUCID corpus and DiapixUK materials available on the OSCAAR website. This project is funded by the UK Economics and Social Research Council (RES-062-23-0681).

REFERENCES

Allen, J.S., Miller, J.L., & DeSteno, D. (2003). Individual talker differences in voice onset time. *Journal of the Acoustical Society of America*, **113**, 544-552.

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. (1991). The HCRC Map Task Corpus. *Language and Speech*, **34**, 351-366.

Arnold, P., Palmer, C. & Lloyd, J., (1999). Hearing-impaired children's listening skills in a referential communication task: an exploratory study. *Deafness and Education International*, **1**, 47–55.

Baker, R. & Hazan, V. (2010). LUCID: A corpus of spontaneous and read clear speech in British English. *Proceedings of the DiSS-LPSS Joint Workshop 2010*, Tokyo, Japan.

Batliner, A., Hacker, C., Steidl, S., Noth, E., D'Arcy, S., Russell, M. & Wong, M. (2004). "You stupid tin box" - Children interacting with the AIBO robot: a cross-linguistic emotional speech corpus. In *Proceedings of the 4th International Conference of Language Resources and Evaluation*, Lisbon, Portugal.

Bell, L., Boye, J., Gustafson, J., Heldner, M., Lindstrom, A., & Wiren, M. (2005) The Swedish NICE Corpus: Spoken dialogues between children and embodied characters in a computer game scenario. In *Proceedings of Interspeech*, (pp. 2765-2768). Lisbon, Portugal.

Blauuw, E. (1994). The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. *Speech Communication*, **14**, 359-375.

Brown, G., Anderson, A., Yule, G. & Shillcock, R. (1983). *Teaching Talk*. Cambridge, UK: Cambridge University Press.

Chan, L. (2010). *The effect of gender, conversational role, difficult communicative situations and their interactions on the production of backchannels, hedges and tag questions*. Unpublished MSc thesis, UCL, London.

Cooke, M. & Lu, Y. (2010). Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. *Journal of the Acoustical Society of America*, **128** (4), 2059-2069

Crawford, M. D., Brown, G. J., Cooke, M. P. & Green, P. D. (1994). The design, collection and annotation of a multi-agent, multi-sensor speech corpus. *Proceedings of the Institute of Acoustics*, **16** (5), 183-189.

Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Stronbergsson, S. & House, D. (2010). Spontal: A Swedish spontaneous dialogue corpus of audio, video and motion capture. *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta

Evans, B., Hazan, V., Baker, R., & Cyrus, T. (2010). Investigating the effects of regional accent background on phonetic alignment in spontaneous speech. *Experimental Approaches to Perception and Production of Language Variation (ExAPP2010)*, Groningen, Netherlands.

Forsyth, R. S., Clarke, D. D., and Lam, P. (2008) Timelines, talk and transcription: A chronometric approach to simultaneous speech. *International Journal of Corpus Linguistics*, **13** (2), 225-250.

Granlund, S., (2010). *An acoustic-phonetic comparison of late bilinguals' Finnish and English clear speech using spontaneous and read speech*. Unpublished MSc thesis, UCL, London

Ibertsson, T., Hansson, K., Maki-Torkko, E., Willstedt-Svensson, U., & Sahlen, B. (2009). Deaf teenagers with cochlear implants in conversation with hearing peers, *International Journal of Language and Communication Disorders*, **44**, 319-337.

Kendall, T. (2010). Developing Web Interfaces to Spoken Language Data Collections. In *Proceedings of the Chicago Colloqium on Digital Humanities and Computer Science*, **1** (2).

Knoll, M., Scharrer, L. & Costall, A. (2009). Are actresses better simulators than female students? The effects of simulation on prosodic modifications of infant- and foreigner-directed speech. *Speech Communication*, **51**, 296-305.

Lann, G.P.M. & Van Bergem,D.R. (1993). The contribution of pitch contour, phoneme durations and spectral features to the character of spontaneous and read aloud speech. *Proceedings of Eurospeech*, (pp. 569-572). Berlin, Germany.

Millar, J., Vonwiller, J. Harrington, J., & Dermody, P. (1994). The Australian national database of spoken language. In *Proceedings of the ICASSP-94* (pp. 97–100).

Nakatani, C., Grosz, B., & Hirschberg, J. (1995). Discourse structure in spoken language: Studies on speech corpora, *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, Stanford, USA.

Newman, R.S., Clouse, S.A. & Burnham, D. (2001). The perceptual consequences of within-talker variability in fricative production. *Journal of the Acoustical Society of America*, **109** (3). 1181-1196.

Reuterskiöld-Wagner, C., Nettelbladt, U. & Sahlén, B. (2001). Giving the crucial information: Performance on a referential communication task in Swedish children with language impairment. *International Journal of Language and Communication Disorders*, **36** (4), 433-445.

Smiljanić, R., & Bradlow, A.R. (2005). Production and perception of clear speech in Croatian and English, Journal of the Acoustical Society of America, **118**, 1677–88.

Smiljanić, R., & Bradlow, A. R. (2008a). Temporal organization of English clear and plain speech. *Journal of the Acoustical Society of America*, **124** (5), 3171-3182.

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *Journal of the Acoustical Society of America*, **125**, 3974-3982.

Uther, M., Knoll, M.A., & Burnham, D. (2007). Do you speak E-N-G-L-I-S-H? Similarities and differences in speech to foreigners and infants. *Speech Communication*, **49**, 1-7.

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M. & Bradlow, A. R. (2010). The Wildcat corpus of native- and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech*, **53** (4), 510-540.

White, L., Wiget, L., Rauch, O., & Mattys, S.L. (2010). Segmentation cues in spontaneous and read speech. In *Proceedings of the Fifth Conference on Speech Prosody*, Chicago, USA.

Yasuo, H., Yukiko, N., Hanae, K., Masato, I., Hiroyuki, S., Michio, O., Makiko, N., Syun, T., & Akira, I. (1999). The design and statistical characterisation of the Japanese Map Task Dialogue Corpus, *Journal of Japanese Society for Artificial Intelligence*, **14** (2), 261-272.