White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/81602/

Version: Published Version

Proceedings Paper:
Gargett, Andrew, AlGethami, Ghazi and Hellmuth, Sam (2014) DiVE-Arabic: Gulf Arabic Dialogue in a Virtual Environment. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), pp. 4434-4439.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

DiVE-Arabic: Gulf Arabic Dialogue in a Virtual Environment

Andrew Gargett[1], Ghazi AlGethami[2], Sam Hellmuth[3]
School of Computer Science[1], Department of English[2], Department of Language & Linguistic Science[3]
University of Birmingham[1], Taif University[2], University of York[3]
A.D.Gargett@cs.bham.ac.uk, Ghazi.Algethami@gmail.com, Sam.Hellmuth@york.ac.uk

Abstract

Documentation of communicative behaviour across languages seems at a crossroads. While methods for collecting data on spoken or written communication, backed up by computational techniques, are evolving, the actual data being collected remain largely the same. Inspired by the efforts of some innovative researchers who are directly tackling the various obstacles to investigating language in the field (e.g. see various papers collected in Enfield & Stivers 2007), we report here about ongoing work to solve the general problem of collecting *in situ* data for situated linguistic interaction. The initial stages of this project have involved employing a portable format designed to increase range and flexibility of doing such collections in the field. Our motivation is to combine this with a parallel data set for a typologically distinct language, in order to contribute a parallel corpus of situated language use.

Keywords: Arabic, corpus linguistics, field linguistics resources

1. Background

This paper reports on progress in building a new corpus, and on development of an associated field-based methodology for collecting situated linguistic data. We combine qualitative with quantitative methods of describing instruction-giving and -following dialogues, with the long-term aim of collecting all of this into a searchable database of both the language and actions of the dialogue participants.

We are motivated in this project by the following observations:

(1) Established corpora of instruction-giving dialogues (e.g. TRAINS [1], Map Task [2]) typically require interlocutors to interact in a relatively stable setting, in a situation that is fairly static and that requires very little monitoring of the surrounding environment.

(2) Communication as characterised by such data collections is relatively less situated, interactive in varying degrees, and typically yields a limited range of data (e.g. typically verbal code and vocal channel, but sometimes also including paralinguistic data, as well as non-vocal gestures, gaze, proxemics, etc).

(3) Such collections rarely contain much detail about the actions taking place in and around language, which has led to a paucity of information about how people use situational features, including environment and actions, during such forms of dialogue.

(4) Extending models of interaction to incorporate such information is likely to provide qualitatively distinct accounts of what is going on in dialogue.

In addition to methodological limitations in the work on dialogue, there are few existing corpus collections for Arabic dialogue, and those that do exist are either unsituated and lack prosodic annotation (e.g. CallHome phone conversations, from the Linguistic Data Consortium[3]) or prosodically annotated but collected in a static environment, and without an accompanying record of actions performed (e.g. map task data, in the IVAr database under construction [4]). To our knowledge therefore, the DiVE-Gulf Arabic corpus is the first to collect together actions with intonational data.

Further, collection of situated data collection has, until recently, remained a lab-based activity, somewhat removed from actual field locations. Certainly, a more portable method for carrying out such data collections, which nevertheless replicates laboratory conditions as far as possible, is sorely needed. Finally, to our knowledge there are no existing parallel Arabic-English corpora combining data on actions and language, captured during a fully situated task. Section 2 describes our approach to filling such gaps in current work. Section 3 outlines some of our ongoing work on the data we have collected, and also provides a concrete example of how we intend to use the finished database for Arabic, by way of presenting previous work we have carried out for situated dialogue in English. Section 4 concludes the paper.

2. Method

This paper reports ongoing work toward a novel corpus of human-to-human real-time spoken instruction giving in 3D environments for Gulf Arabic. Our corpus has specific features designed into it for the purpose of filling gaps we have identified in previous work (see section 1 for details):

(1) The corpus employs technology for making detailed recordings of **both** the actions **and** speech of Arabic-speaking interlocutors in a 3D environment

(2) Working toward a parallel corpus of Arabic and English, we have replicated the approach taken for building the SCARE corpus (Stoia et al. 2008), [5] an English corpus of instruction giving in a virtual world. Achieving this combined database is then a matter of properly combining the DiVE and SCARE corpora.

(3) All recordings are made using portable devices,

[1] http://www.cs.rochester.edu/research/speech/trains.html

[2] http://groups.inf.ed.ac.uk/maptask/

[3] www.ldc.upenn.edu

[4] www.york.ac.uk/res/ivar

[5] http://slate.cse.ohio-state.edu/quake-corpora/scare/

including laptops, head-mounted microphones and digital audio recorders.

The 3D worlds used to record this corpus were the same as those used for the SCARE corpus, the latter being a collection of instruction giving dialogues in a 2-level virtual world, each level having between 7 and 9 rooms, these rooms having buttons for opening cabinets that contained objects to be retrieved (see Figure (1) below for a screenshot). The corpus employs the QuakeII gaming software.

The SCARE corpus consists of 15 sessions, with interlocutors taking roles of either instruction giver (IG) or instruction follower (IF). They had to complete a series of 5 simple tasks (retrieving or manipulating objects), with the IG verbally guiding the IF through the world, but only the IG having access to a map of the world, and a list of the tasks to be completed. The 19 male and 11 female participants had an average age of 30, and identified as native speakers of North American English. Sessions ranged from 10 minutes in length to over half an hour.

Data collection for the DiVE-Arabic corpus was carried out alongside other data gathering for Intonational Variation in Arabic (IVAr, www.york.ac.uk/res/ivar), a project itself inspired by earlier work on English (www.phon.ox.ac.uk/IViE/), and served the purpose of collecting situated dialogue in Gulf Arabic (one of the dialects included in IVAr). In Al Ain (UAE) and Buraimi (Oman), we recorded dialogue between participants trying to solve instruction giving tasks within a virtual world. For this corpus collection, we replicated the SCARE task and number of participants recorded, while also approximating as closely as possible the population demographics. For the Gulf Arabic corpus, sessions ranged from just under 10 minutes to just over half an hour. The 17 female and 9 male participants were university students aged in their early- to late-twenties, who interacted in same gender pairs. The DiVE-Arabic corpus is then commensurate in terms of extra-linguistic factors with the SCARE corpus (except that it involves no gender mixing of interlocutors).

A key motivation for replication was to make possible a parallel English-Arabic corpus of situated dialogue when the SCARE and DiVE corpora are combined, making this a highly unique contribution to the area of cross-linguistic dialogue studies. Our plan is to construct the parallel Arabic-English corpus, by collecting together signals and annotations from both the DiVE and SCARE corpora as a stand-off database using the Nite NXT toolkit (Carletta et al.). The Nite NXT approach is particularly useful for us due to its rich structuring of data, including a data set model for structuring a corpus in terms of (i) observations, (ii) agents, (iii) the interaction, as well as, (iv) annotations of the signal. We divide annotations of the signal into segmental and supra-segmental components, each being stored separately in line with the stand-off approach. In particular, observations can be multi-layered, either directly aligned to the timing level, or else symbolically linked to other levels (e.g. annotations of dialogue acts can be linked to actual utterances, which in turn can be directly aligned with the timing of the original audio and video signal). Aside from allowing us to adequately model the rich information from the dialogue data, this also allowed access to a very useful library of Java classes bundled with the Toolkit (e.g. for searching NXT-formatted corpus files). It should be stressed that augmenting the SCARE corpus like this in no way impedes the compatibility of the two corpora. Indeed, we have also analysed the original audio signal from SCARE, splitting this into two channels, whereby we are able to incorporate tracks for both instruction-giver and follower, into the resulting database we have built for the combined corpus (after checking and cleaning this data).

As pointed out above, DiVE has been deliberately constructed to mirror as far as possible the SCARE corpus, in order to provide a parallel Arabic-English corpus of situated communication. It is therefore useful to more explicitly compare various features of the corpora:

	SCARE	**DiVE**
Gender	19 male, 11 female	17 female, 9 male
Avge age	30 yrs	25 yrs
Education	tertiary	tertiary
Quantity	3hrs, 41mins	4hrs, 40mins

It is worth pointing out that DiVE has only same gender interlocutors, although as with SCARE, participants were recruited in pairs. Note also that 3 of the 15 DiVE sessions were recorded by participants who took part in more than one session; that is, we have a subset of data from participants who acted as Instruction-Giver (IG) having previously acted as Instruction-Follower (IF), and vice versa. These and other aspects of data collection have been carefully documented.

We recorded audio signals of each interlocutor, video of what each participant sees during their interaction, as well as detailed information about instruction follower movements in the virtual world (the computer continuously records orientations and positions in the virtual world).

The corpus is a combination of signals and annotations of these signals. Details for recordings of the signals are as follows:

(1) Spoken Arabic, one channel per speaker (using a Marantz PMD661 Solid State digital audio recorder, and two Shure SM10 unidirectional head-mounted microphones). Annotations are in progress and are aligned with the signal.

(2) Actions of the instruction follower in a virtual world (the same world as used for the SCARE corpus). Instruction giving sessions took place through the Jake2[6] platform (a freely available Java version of QuakeII), and these sessions were recorded on a PC laptop running Windows 7. Information about actions is automatically acquired from the computer log files, and incorporated in the corpus.

(3) Video signal of the monitor output for the virtual world, showing the location and actions of the instruction follower as they move through the world,

[6] http://jake2.cvs.sourceforge.net/

and both participants can view this while interacting. There are no immediate plans for comprehensive annotation of the video data, although this data will also be incorporated in the corpus.

Annotation work is ongoing (completion 2015), but we can report on the progress of this for each of the three levels of the corpus as follows:

(1) Audio:
   a. Turns-at-talk are labelled automatically in a Praat [7] TextGrid using a script which identifies periods of low intensity in the signal; after manual correction of interval labels, each turn is annotated in broad phonetic transcription using an ASCII character set by a native speaker of Arabic (annotation at this level completed in 2014).
   b. Turn-level transcriptions are force-aligned to the signal using the ProsodyLab Aligner (Gorman et al. 2011) to yield word-by-word labelling within each turn in a Praat TextGrid; manual prosodic annotation in a ToBI annotation system developed for Gulf Arabic (annotation at this level is expected to be completed in 2015).
   c. The Textgrid files from steps a. and b. are then parsed into XML files, and included in the corpus database.

(2) Video: for qualitative analysis of actions, the audio and video signal are synchronized within ELAN [8], along with turn-level annotations of the audio imported via a Praat TextGrid; this permits searches for e.g. lexical items or fillers, and manual annotation in ELAN of accompanying actions in the video, as illustrated in the case studies in section 3 below.

(3) Actions: for quantitative analysis of actions, an existing parser created for analysis of the dm2 files from the SCARE corpus has been adapted for analysis of the DiVE dm2 files, in order to parse these files into XML and incorporate within the database.

All annotations will be incorporated into a single, comprehensive XML database, and by aligning each annotation level with the signal, we link all these levels in a multi-dimensional way. As shown in Gargett (2012), this way of presenting the information enables a flexible set-up for carrying out analyses across modalities (we come back to this in section 3.2 below). The database will be hosted permanently under the umbrella of the IVAr project[9], and will be made available in 2015 (see website for updates).

In summary, all of this information allows us to capture exactly what the interlocutors said to each other while interacting, exactly what they could see at the time, and exactly what they did and when.

[7] www.praat.org

[8] http://tla.mpi.nl/tools/tla-tools/elan/

[9] www.york.ac.uk/res/ivar

3. Results

The richness in data resulting from situated data collection comes from interlocutors engaging linguistically to successfully complete assigned tasks, while negotiating the environment of the virtual world to solve various problems they encounter along the way. It should be stressed that in this setup, instruction givers know things which instruction followers don't (e.g. only instruction givers have a map giving locations of the things that need to be found), but only instruction followers can move, forcing interlocutors to work together. As an example of the kind of data we collected, in Figure (1), the task for the instruction follower (m1) is to move the picture on the wall, and only one of the nearby buttons performs this action; the correct button is marked on the instruction-giver's map.

In the following section, 3.1, we provide case studies of the type of insights that can be gained from qualitative analysis of data of this type, showing how the record of actions (here, from the video signal) enriches our understanding of how language and actions interact in dialogue contexts. In section 3.2 we illustrate how quantitative analysis of the data (from parsing of the dm2 actions data, in a database set up) has been used to analyse data from the SCARE corpus, as a model for future parallel work on the DiVE-Arabic corpus.

3.1 Case studies

We present two qualitative case studies: the first illustrates how interlocutors' monitoring of actions interacts with actual dialogue produced; the second examines how the prosodic design of requests differs systematically, depending on whether the request triggers a verbal response (i.e. a verbal answer to a question) or whether the request triggers an action response (only).

3.1.1 Case Study 1

Figure 1 illustrates how situated action can disambiguate prosody. Here the Instruction Follower (IF-M1) is trying to complete an assigned task of moving the picture on the wall; only one button performs this action and the correct button is marked on the Instruction Giver (IG-M2)'s map. Multi-level recording provides a record of participant actions as well as utterances, and the links between them, and unlocks vital information. It would be impossible to interpret the interactional value of the sequences produced by IG-M2 in lines 4 and 5 in Figure 1, based on audio alone, but with the information provided by the video/movements record, we can make inferences about the interaction involved. In a conventional approach presenting textual modality alone, the two major silences, which precede turns 4 and 5, respectively, could only really be interpreted via the verbal code. In both cases, we have a lengthy within-speaker pause and no apparent response from the interlocutor. This is a potential sign of trouble, but is this so in this case?

Our approach provides added situational information, revealing an additional dimension of such silences, which can disambiguate. During a linguistic silence, the IF may

move around, push buttons, turn and change orientation, or indeed stay absolutely still and do nothing; that is, linguistic silences can now be interpreted via an additional dimension, in terms of the non-linguistic actions. For example, the IF saying nothing but moving is likely to be interpreted quite differently by an IG monitoring the IF's behaviour, compared to when the IF speaks without moving, or says and does nothing.

In the present example, the IF is active during both silences: during the first silence IF changes orientation several times with gaze towards different buttons; during the second silence there is backwards movement away from the button and a change of orientation towards the picture (to see if the task has been completed).

3.1.2 Case Study 2

The second case study involves the way in which dialogue context cues can yield independent evidence to support analysis of prosodic phenomena. Figure 2 illustrates this patterning of context and prosody.

Prosodic analysis of the acoustic signal reveals that the speakers systematically realise the last accented syllable of utterances at distinct levels of pitch, finishing at a pitch level which is either high (H), mid (M) or low (L) in their pitch range. These categories illustrate the type of labels that will be used during prosodic annotation of the corpus (i.e. these are high/mid/low 'boundary tones'). The f0 trace in Figure 3 illustrates the clear difference in final pitch, M vs H, for speaker IGM2.

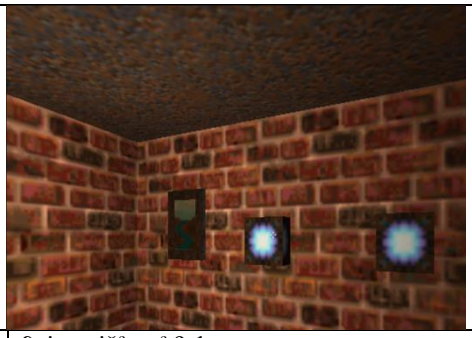
Inspection of the accompanying verbal text suggests that there is a correlation between the choice of final pitch level used in a contribution to the dialogue and its function or role: responses are low (L), requests are mid (M) and questions are high (H). In our corpus we are able to provide an additional layer of independent evidence to support this classification.

Inspection of the video data shows that contributions with H and M final pitch are treated differently by the interlocutor: H-final contributions (putative questions) always receive a verbal response (generally unaccompanied by an action response), but M-final contributions (putative requests) always precede actions (and generally no verbal response). One explanation for this kind of systematic difference is that interlocutors are using prosody to distinguish requests for information (i.e. questions), from requests for action. This difference is revealed only because we can recover both linguistic and non-linguistic information from our corpus.

Figure 1

Example illustrating Case Study 1 (including picture of items referred to).

Broad phonetic transcription in IPA.



1.	IGM2:	ʔajwa iðˤɣatˤ ʕala:: *yes push on…*
2	IFM1:	ʔajji waħdˤɑ (0.4) [il-ʔu:la] *which one? the-first?*
3	IGM2:	[az-] (0.3) izzur- izzur iθθa:ni *the- the-button the-button the-second*
	⇒	(1.3)
4	IGM2:	ʔajwa ha:ða laʔ (1.4) *yes this no*
	⇒	(1.3)
5	IGM2:	ʃu:f ajwa sˤɑħ *look yes correct*
6	IFM1:	nze:n Ok

Figure 2

Example illustrating Case Study 2 (including picture of items referred to).

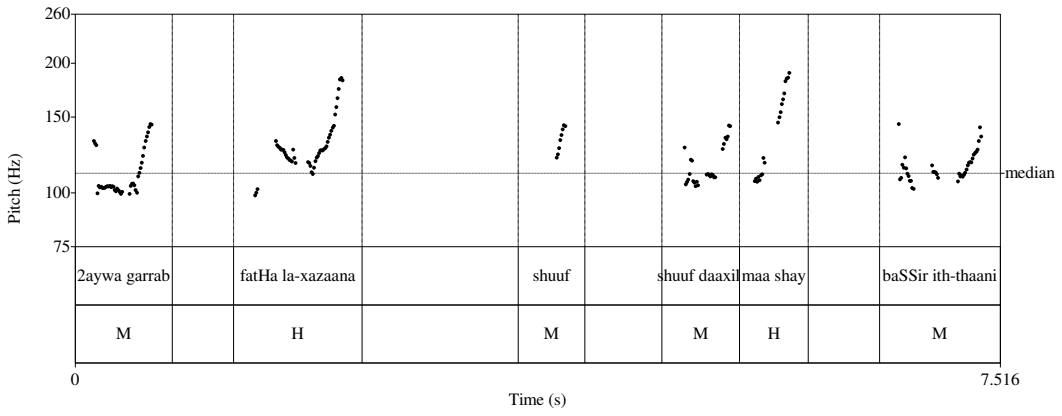
Broad phonetic transcription in IPA.

1	IFM1:	ʔaðˤɣatˤ ʕalay-h haːða (0.7) *push on-it this?*	H
2	IGM2:	ʔaywa garrab (0.8) *yes try*	M
3	IGM2:	fatħa l-xazaːna (0.7) *opened the-cupboard?*	H
4	IFM1:	ʔaywa (0.6) *Yes*	L
5	IGM2:	ʃuːf (0.8) *look*	M
6	IGM2:	ʃuːf daːxil *look inside*	M
7	IGM2:	maː ʃay (0.6) *NEG thing?*	H
8	IFM1:	maː fiː-ʃ *NEG there-NEG*	L
9	IGM2:	basˤˤir iθ-θaːni (2.3) *look the-second*	M
10	IGM2:	maːshi ʔaywa ʃiːl haːða *ok yes carry this*	L

Figure 3

Pitch trace for turns 2-9 of case study 2, for speaker IGM2 only.

Axes show this speaker's min/max/median pitch range measured across whole conversation.
Transcription illustrates ASCII character set used in annotation of turns-at-talk in corpus.



3.2 Comparison with SCARE

3.2.1 Feedback patterns

Gargett (2012) considered a range of backchannels produced by interlocutors from the SCARE corpus, as these overlapped with pauses in actions by the instruction follower, and pauses in speech by both instruction giver and follower. For this purpose, an NXT-style database for the SCARE corpus was built, and its design forms the basis for that of the DiVE database currently underway.

Specifically, acknowledgements like "ok" or "yeah", tacit agreements like "mhm", and fuller expressions of agreement like "yep" or "alright", as well as interjections "um" and "uh". The following questions were used to focus the studies:

(1) In the context of pause in actions, what IG backchannels are most likely?
(2) In the context of pause in actions, what IF backchannels are most likely?
(3) In the context of pause in actions and words, what is the likely backchannel to be used, and who is most likely to use it?
(4) In the context of pause in actions and words, how does the choice between backchannels "um" vs. "uh" affect word pause duration?

The first question aimed to evaluate overlap of backchannels of instruction givers with the activity of instruction followers, while the second question targeted overlap of backchannels of instruction followers with their own activity. Results here suggested that use of backchannels is not independent of the role of speaker. The purpose of question 3 was to evaluate overlap of backchannels and cessation of both actions and language in instruction givers and followers, with results suggesting that independence of use of tokens from speaker role could not be refuted. Finally, question 4 aimed to evaluate overlap of interjections (specifically, "um" and "uh") with cessation of both actions and language. Results suggested that while speech pauses are projected by pauses in speech, this tends to disappears in the context of actions pauses.

In summary, Gargett (2012) showed that, in English, on average, an instruction giver is more likely to produce a further verbal response after a linguistic silence if there is no accompanying action, than if there is accompanying action. Full elaboration of the DiVE-Arabic corpus will allow us to determine whether the same pattern is in fact also observed in Arabic, or whether there are cross-linguistic differences in such matters.

3.3 Summary

Our corpus makes several novel contributions:

a. It captures information about how the prosodic realization of talk relates to other linguistic levels, in situated dialogue in Arabic, for the first time.
b. It enables qualitative *and* quantitative exploration of how the linguistic design of interlocutors' productions link to the actions being undertaken.
c. Unlike comparable approaches (e.g. GIVE-2[10]), we are able to explore in detail the interaction between all levels of communicative behaviour, both spoken language and actions carried out while interlocutors are interacting, enabling richer possibilities for investigating how such behaviour is grounded in the surrounding environment during communication.

4. Conclusion

Our project has the potential to allow researchers to recover both linguistic and non-linguistic information about the situatedness of dialogue, which has previously been unavailable. In addition, the methods employed are explicitly chosen to yield a cross-linguistic corpus which is genuinely parallel, so that results for one language can be usefully compared to those in the other.

One limitation of the current setup, which we are now addressing, is that the task itself, while well-suited to collecting data sufficient for modelling the range of phenomena we are interested in, may not result in amounts of data sufficient for all possible dialogue projects. We are currently revising the task and setup of the worlds, in order to tackle this issue, with the aim of trialling it on a wider range of languages (e.g. Mandarin and Tamil, as well as English and Arabic). A further limitation is that there may be an issue to do with mobility in very remote research locations, and we have plans to develop a version of this approach that may be deployed on mobile technology (e.g. tablets, smart phones).

Acknowledgments

Thanks to Dr Rana Almbark (University of York) for help in preparing data collection materials. We also wish to thank the Department of Linguistics, University of the UAE, Al Ain, for hosting S.H. during her fieldwork visit.

5. References

Carletta, J., S. Evert, J. Kilgour, C. Nicol, D. Reidsma, J. Robertson & H. Voormann, "Documentation for the NITE XML Toolkit"[11].

Enfield, N.J. & T. Stivers. (2007). Person Reference in Interaction: Linguistic, Cultural and Social Perspectives. Cambridge: Cambridge University Press.

Gargett, A. (2012). "Feedback and activity in dialogue: signals or symptoms?" Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog, September 7 -8, 2012, Oregon.

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-Aligner: A tool for forced alignment of laboratory speech. Proceedings of Acoustics Week in Canada, Quebec City.

Stoia, Laura, Darla M. Shockley, Donna K. Byron, & Eric Fosler-Lussier. (2008). "SCARE: A Situated Corpus with Annotated Referring Expressions." Proceedings of the 6th International Conference on Language Resources and Evaluation.

[10] http://www.give-challenge.org/research/page.php?id=give-2-corpus

[11] http://http://groups.inf.ed.ac.uk/nxt/documentation.shtml