The aim of this project is to survey of the range and scope of typological variation in the intonational properties of spoken dialects of Arabic. To achieve this, directly parallel speech data will be collected for input to phonological analysis within the Autosegmental-Metrical (AM) framework (Ladd 2008). The resulting Intonational Variation in Arabic (IVAr) Corpus, together with its accompanying transcriptions, will be made available as an online resource for both academic and non-academic users.

The project adopts a sociophonetic approach (Foulkes, Scobbie & Watt 2010) to address the following questions: i) what are the intonational properties of the Arabic dialects under discussion? and, ii) what is the nature and scope of intonational variation among Arabic dialects? The answers to these questions will contribute to current theoretical debate regarding the nature and scope of cross-linguistic prosodic variation in general (Jun 2005, to appear).

Background

The experience of the ESRC-funded Intonational Variation in English (IViE) project (Grabe 2004) has shown that availability of directly parallel speech data, containing utterances of the same sentence/paragraph by a number of different speakers, is crucial in identifying in what ways (and in what contexts) the intonational patterns used by speakers of different dialects differ, whilst also highlighting shared features (Grabe & Post 2002). Grabe et al (2005) demonstrate that although speakers of different varieties of English vary greatly in their choice of which intonational contour to use in a given utterance type, they vary little in their choice of where to place main prominence in the utterance (the 'nucleus'). As they point out, one practical implication of this finding is that English language teaching can safely focus on helping learners acquire appropriate nucleus placement patterns, rather than particular contours.

Prior literature on the intonational patterns of spoken Arabic dialects is limited, but in recent years a growing number of descriptions of individual dialects have been published: for some varieties there is an AM analysis (e.g. Egyptian, Hellmuth 2006; Lebanese, Chahal 2001) but for others there is only a preliminary description (e.g. *inter alia* Syrian, Kulk et al 2003; Emirati, Blodgett et al 2007) or a description in a very different theoretical framework (e.g. Moroccan, Benkirane 1998) or no description at all (e.g. Iraqi). All of this work has been insightfully summarised in the entry on Arabic intonation in the recent *Encyclopaedia of Arabic Language and Linguistics* (Chahal 2006) but the cross-dialectal comparison therein relies on secondary analysis of published descriptions, which treat non-parallel data and are influenced by different theoretical viewpoints.

A number of experimental investigations have looked at phonetic aspects of the suprasegmental properties in various subsets of dialects, taking in rhythm (Ghazali et al 2002, 2007), local alignment of f0 peaks (Yeou et al 2007) and pitch range/register (Biadsy et al 2009). The Dynamics of Language project (Université de Lyon) investigated reliability of automatic acoustic measurements for dialectal identification purposes (Barkat et al 1999), in a corpus of semi-spontaneous speech, but the data are not in the public domain. The only study to date which has attempted to describe the basic intonational patterns of a small subset of dialects in phonological terms using broadly parallel data (Ghazali et al 2007) was very limited in both scope (read speech only) and size (two speakers per dialect), and the source data are not in the public domain.

A set of parallel data and descriptions is much needed, to facilitate meaningful comparison and as input to growing interest in Arabic dialectal variation from academic and other audiences.

Rationale

This project seeks to fill this data-gap and to do so in such a way as to set a standard for good practice in future work on Arabic intonation, leaving in place a database of spoken Arabic data as a resource for both researchers and other users. The main Arabic speech corpora currently available are intended for speech technology applications (Linguistic Data Consortium (LDC), UPenn, www.ldc.upenn.edu) and mostly comprise spontaneous speech (telephone conversations); a few include parallel utterances from speakers of different dialects, yet without the necessary metadata to accurately identify which variety is at issue (e.g. West Point Corpus). The SemArch database (University of Heidelberg, www.semarch.uni-hd.de) hosts recordings contributed by Arabic dialectologists, which are mostly monologues in rural or Bedouin-origin varieties. This project will complement these existing Arabic speech data resources by collecting parallel read and spontaneous speech in urban centres, with young monolinguals (aged 18-24 yrs).

The database will have a two-tier structure: a set of *main survey* data will be collected in one city in each of five regionally-defined dialectal areas (Versteegh 2001: Egyptian, Levantine, Peninsular, Mesopotamian and Maghrebi) with an additional in-depth *cluster survey* in two regions. The cluster dataset will comprise the same materials as the main survey collected in two more local varieties in two regions, with 6 young (18-24) and 6 older (50+) speakers, yielding an apparent time picture of intonational change, as well as experimental data for fine-grained phonetic analyses.

The decision to collect the main survey dataset with young speakers in urban centres is in part practical, since it will greatly facilitate recruitment of suitable participants (via urban English language schools) with the necessary ease and speed to fit within the life-cycle of a three-year project. More importantly, the choice represents a robust response to Bassiouney's (2009:116) call for studies on the language of young adult urban speakers (cf. Miller 2004). There has been little prior investigation of the individual dialects of Arabic-speaking youth, who represent 20% of the population of the Middle East and North Africa region (by the UN definition of 'youth' as 15-24 years, Assaad & Roudi-Fahimi 2007) and no cross-dialectal studies have targeted this key population group. Finally, this data collection strategy also allows us to replicate the successful data collection methodology of the IViE project (collected with youth in same-sex friendship pairs) to yield a sister corpus to IViE. In the cluster survey we will have apparent time data, in the form of directly parallel data from both younger and older speakers, to document any intonational change in progress; for the data collected for individual dialects in the main study it will nonetheless be possible to make indirect comparison with existing datasets (e.g. LDC, SemArch).

The choice of *main survey* data collection locations will be finalised during Phase I of the project. The PI has personal contacts with researchers and/or English language schools in urban centres in Egypt, Morocco, Tunisia, Jordan, Syria, Lebanon, Saudi Arabia, the Emirates, Qatar and Yemen, with whom contacts will be made in the first instance. A probable *cluster survey* data collection scenario will be a cluster of Egyptian varieties (Cairo in main survey + Alexandria & Upper Egypt) and a cluster of Levantine varieties (Damascus in main survey + Amman & Beirut). However, the two cluster regions will be chosen (during Phase I) to best exploit the complementary expertise and interests of the eventual RA and PI, as well as to maximise the sociolinguistic value of the results (e.g. to be able to compare the degree of dialectal homogeneity in two countries in which the variety of the capital city acts as a national standard, Miller 2004).

In Phase I (months 1-6) dialect-specific parallel stimuli will be devised and tested with native Arabic speakers recruited locally in York (alongside consultation with potential non-academic end-users, a literature review and recruitment of local fieldworkers to finalise the choice of regions). Although the project is modelled closely on IViE, there are issues specific to data collection in Arabic which require attention.

Firstly, diglossia between written formal and regional spoken varieties means that textual prompts risk eliciting speech in formal Modern Standard Arabic (MSA), and this will be mitigated by using local lexis and orthographic norms (Siemund et al 2002). Nonetheless, variation in lexis, syntax and in both segmental and metrical phonology across Arabic dialects means that creation of utterances which will be directly parallel in different dialects is not a simple task. It is not a solution to employ MSA terms and structures since this risks eliciting a non-local register of speech from participants. We will therefore invest time during Phase I to identify lexical items and structures which are sufficiently parallel across dialects and which also meet other desired criteria such as containing mostly sonorant speech sounds (to facilitate generation of continuous pitch traces) and a similar relative frequency of lexical items. These factors will feed into design of read speech sentences, a 'folk tale' narrative text and individual landmarks in a Map Task layout.

Secondly, the degree of dialectal diversity varies in different urban settings (Bassiouney 2009:112ff.), with greater homogeneity observed in, say, Cairo than in other cities where a speaker may speak a local regional 'standard' but only as a second dialect, with another regional (often rural) variant spoken in the home. It is necessary to take steps in data collection to encourage speakers to use their home variety and to include a means of determining what that is. In a pilot study on Yemeni Arabic, funded by the University of York Research Priming Fund, an Arabic version of the Sense Relation Network (SRN) tool (Llamas 2007), in which a pair of speakers collaboratively complete and discuss a spider diagram of related vocabulary items, was created for use in data collection in the capital city San'aa. This technique successfully elicited speech in the colloquial register, whilst also providing tokens of a number of key local shibboleth vocabulary items for each speaker. The SRN tool will be used in the present project alongside more familiar IViE-style stimuli: read speech sentences, a narrative read and re-told from memory (cf. Hellmuth 2007) and a Map Task.

In Phase II (months 7-16) data collection will be carried out in the main survey locations. Six male and six female participants will be recruited in each location via mid-price English language schools whose students are likely to be of similar socioeconomic status, among those at pre-intermediate level in English or lower. As in IViE, recordings will be carried out in same-sex friendship pairs. It is essential that the researcher making the recording is a speaker of the desired variety, to avoid linguistic interference and to minimise the effects of the 'observer's paradox' (Labov 1972), but it is also crucial that materials are collected in parallel fashion across dialects and with due care paid to the need to obtain fully informed consent from participants.

We will implement a model used successfully by Dr Ghada Khattab (Newcastle), in her ESRC First Grant funded data collection in Lebanon, whereby the PI or RA travel to a local centre to train and oversee a local fieldworker; the PI/RA can monitor the quality of recordings and also carry out initial analysis in the field (e.g. of SRN data, to identify shibboleth vocabulary items).

This model permits data collection to take place in more than one site at a time, shortening the data collection period and maximising PI/RA time for the lengthy process of data transcription. Initial preparation of data for archiving will begin during this period, with editing of recordings for input to the eventual online database and preparation of basic orthographic transcriptions (Roman script transliteration, Arabic script and English translation).

In Phase III (months 17-26) the key work of prosodic transcription of the main survey data will be carried out by the PI and RA, in a structured sequence of inter-transcriber comparison designed to motivate a phonological description of each individual variety.

Our prosodic analysis will be carried out within the widely accepted AM framework, the current standard for cross-linguistic comparison of intonational patterns (Gussenhoven 2004, Jun 2005, Ladd 2008; cf. Nolan 2009). An important contribution of the IViE project was to propose a labelling tier (Grabe 2001) for fine-grained transcription of the type of variation that is most relevant for English dialects (in the local shape of f0 contours). Work on a similar system for different varieties of French showed the need for a further tier, to capture a type of variation relevant for French (in the global f0 contour, across longer stretches), arguably not relevant in English (Post & Delais-Roussarie 2006). Pilot comparative work suggests that both of these tiers (local and global) are relevant for Arabic but that manual annotation of both in a full corpus is time-consuming and unnecessary (Hellmuth & Chahal 2009); instead an interim first transcription should resemble a 'broad' transcription but include an "alternatives tier" (Brugos et al 2008) to identify which broad category labels are prone to ambiguity and in which contexts. The subset of the data thus identified can then usefully be submitted to a 'narrow' transcription, using IViE-style local and global f0 tiers, to establish the particular properties of a given 'broad' label in a particular variety. In this way it will be possible, in a reasonable time-scale, to establish the inventory of tonal shapes observed in each variety in a transparent and reproducible manner.

Transcription will start with the most controlled portions of the data (read speech sentences and narratives), then move on to portions of more spontaneous speech from the retold narratives, Map Task and SRNs. The result of the transcription process will be a phonological description within the AM framework of the intonational patterns of the five main survey varieties. Formal descriptions will be disseminated through conference presentations (such as *Arabic Linguistics Society, ICPhS, Manchester Phonology Meeting, Tone and Intonation in Europe, Speech Prosody*) and by means of submissions to relevant publications, to include general descriptions of the intonational patterns of one or more individual dialects (to journals such as *BSOAS* or *Zeitschrift für Arabische Linguistik*) and studies of the prosodic properties of one or more individual dialects in the context of wider prosodic typology (to journals such as *Journal of Phonetics*, *Journal of the International Phonetic Association*, *Journal of Linguistics*, *Language and Speech*).

Publication of phonological descriptions of individual dialects will facilitate future research on topics as diverse as the intonational resources employed in talk-in-interaction in Arabic (unlocking the potential of the LDC corpora) and the degree of transfer of L1 properties in the L2 intonation patterns of Arabic learners of English (cf. Hellmuth to appear).

These emerging descriptions will also be used within the project to inform the design of fine-grained experimental investigations to be carried out alongside collection of the main parallel

corpus stimuli in cluster survey locations. Cluster survey fieldwork data collection will be completed by the end of Phase III.

In Phase IV (months 27-36), the phonological descriptions of the five main varieties, and the directly comparable transcriptions on which they are based, will be exploited within the project to begin to establish the scope and range of typological variation among the varieties under study. The cluster survey data will undergo basic and prosodic transcription for analysis of local micro- and sociolinguistic-variation in the context of wider cross-dialectal prosodic variation in Arabic. Finally, the main survey dataset and its accompanying transcriptions will be transferred to a database server to be made available via a searchable web interface by the end of the project.

There is good reason to suspect that Arabic dialects stand at different ends of a number of theoretically relevant prosodic continua. For example, Egyptian Arabic is known to display very rich accent distribution, having an accent on almost every content word, whereas Lebanese Arabic does not (Chahal & Hellmuth to appear), but it is not known where other varieties stand with respect to this parameter. Similarly, Jun (2005) points out that there is an important typological distinction between intonation languages which use pitch movements to mark the 'head' stressed syllable of each word (e.g. English) vs. the edge of a (roughly word-sized) Accentual Phrase (e.g. Korean). Descriptions in the literature suggest that some Maghrebi dialects may be edge-marking rather than head-marking (e.g. Boudlal 2001), which would set them apart from all other varieties of Arabic thus far described.

We anticipate that inter-dialectal variation observed among Arabic dialects will enable us to contribute significantly to the current prosodic typology debate, by revealing which parameters of prosodic variation are purely phonological and which are intrinsically related to syntactic or other structural phenomena; this is particularly true of any prosodic micro-variation observed among local clusters of dialects (which are likely to be somewhat more homogenous with regards to morphosyntactic properties).

The cluster survey apparent time datasets will also permit us to document the degree of intonational homogeneity in the two cluster regions under study, for comparison to existing studies of the degree of homogeneity based on morphosyntactic, lexical and segmental phonological properties (Miller 2004).

The results of these comparative studies, at inter-regional, intra-regional and intra-dialectal levels, will be disseminated by means of conference presentations (again, to events such as *Arabic Linguistics Society, Manchester Phonology Meeting, Tone and Intonation in Europe, Speech Prosody*) and in scholarly articles for submission to journals (such as *Journal of Phonetics, Journal of the International Phonetic Association, Journal of Linguistics, Language and Speech, Language Variation and Change* or *Linguistic Typology*). We anticipate publication of one or more pairwise comparisons of dialects that illustrate particular typological contrasts, as well as one or more general overview articles surveying the range and scope of variation across Arabic dialects, as evidenced by the IVAr corpus sample.

The searchable web interface will be created by an external sub-contractor (we have approached the contractor who did the work on the IViE corpus) and the project website will be maintained subsequently by the PI. The Department of Language and Linguistic Science at the University of

York already hosts a number of large language corpora and the IVAr Corpus will thus benefit from infrastructure provisions already in place.

The main survey speech recordings and their accompanying basic and prosodic transcriptions will be made available to the general public on the website by the close of the project. As a contingency against unexpected delays to the project, we will prioritise publication of prosodic transcriptions for a subset of speakers across all dialects (cf. the IViE corpus), rather than for all speakers for a subset of dialects. Cluster survey speech recordings of the parallel corpus data and accompanying transcriptions will similarly be made available on the website by the close of the project. The full dataset will also be made available within three months of the end of the project for archiving with ESDS.

Project Timetable

	months	
Phase I	1-6	- design and testing of cross-dialectally parallel recording materials - planning of data collection and data management - consultation with ESDS and web designer - consultation with potential non-academic end-users of the database - literature review
Phase II	7-16	- data collection in the five main survey locations - orthographic transcription of main survey data: Roman script transliteration, Arabic script, English translation - prosodic transcription of pilot subset of main survey data (to finalise transcription protocols)
Phase III	17-26	- prosodic transcription of main survey data - phonological analysis of individual dialects - writing up of intonational descriptions of individual dialects for publication - planning of data collection in cluster survey locations - data collection in two regional cluster survey locations - preparation of non-academic outputs on individual dialects
Phase IV	27-36	- writing up of inter-regional comparative analyses for publication (from main survey) - orthographic & prosodic transcription of cluster survey data - writing up of analyses of inter-dialectal micro-variation and/or sociolinguistic variation for publication (from cluster surveys) - preparation of non-academic outputs comparing across dialects

Sam Hellmuth
University of York
27.1.11