OrienTel – Arabic speech resources for the IT market

Rainer Siemund[1], Barbara Heuft[1], Khalid Choukri[2], Ossama Emam[3], Emmanuel Maragoudakis[4], Herbert Tropf[5], Oren Gedge[6], Sherrie Shammass[6], Asuncion Moreno[7], Albino Nogueiras Rodriguez[7], Imed Zitouni[8], Dorota Iskra[9]

[1] Philips Speech Processing, [2] ELDA, [3] IBM, [4] University of Patras, [5] Siemens, [6] Natural Speech Communication, [7] Universitat Politècnica de Catalunya, [8] Lucent Technologies, [9] SPEX

c/o Rainer Siemund, Philips Speech Processing, Kackertstr. 10, D-52072 Aachen, Germany
rainer.siemund@philips.com
http://www.orientel.org

Abstract

A survey of the language resources market clearly shows that the Arabic language is still a stepchild of international R&D efforts in the field of speech recognition. *OrienTel* for the first time makes an effort to create speech data on a large scale. It does so by profiting from the experience of previous *SpeechDat* projects and from the European Commission's policy to embrace non-EU Mediterranean and surrounding countries. The participants of *OrienTel* will collect Standard and Colloquial varieties of Arabic in Saudi Arabia, the UAE, Egypt, Israel + Palestine, Tunisia and Morocco, supplemented by other languages of the region. Help in creating an Arabic network of speech experts is appreciated.

1. Introduction and goal of the paper

Like all other members of the *SpeechDat* family of data collections, *OrienTel* is driven by an international industrial and academic consortium.[1] This time the co-ordinator is Philips Speech Processing, the other participants being ELDA, IBM, Knowledge, the University of Patras, Siemens, NSC, Universitat Politècnica de Catalunya, and Lucent Technologies. *OrienTel* resembles previous *SpeechDat*-like undertakings[2] insofar as the recordings are supposed to serve the broadest possible application areas, ranging from simple command and control services to unified messaging, information retrieval, customer care, banking, WAP and service portals. It is different from previous projects, however, as it takes *SpeechDat* to a variety of non-European languages such as Arabic and Hebrew, which require far-reaching adaptations to database design and annotation standards. The aim of the present paper is to introduce the broad setup of *OrienTel*, give an account of the present status of database design, and to call for a joint effort in producing language resources for the Mediterranean and the Middle East. In addition to the rather general description of the project presented in the LREC2002 Proceedings, the present paper provides some Arabic-specific problems.

2. Why *OrienTel* now?

The *OrienTel* region is a region of extremes. While Israel ranks as one of the World Bank's 26 so-called 'developed' economies, its next door neighbour, the Palestine Authorities, are having to cope with a GDP per capita figure of just US$ 1,634.[3] The contrasts are even more extreme among the Gulf states. The United Arab Emirates possess the highest GDP per capita of the region - which at US$ 16,800 is on one level with most Western European countries - yet just a few hundred kilometres away Yemen reports an average GDP figure of only US$ 304. From a commercial point of view, therefore, not all countries in the *OrienTel* region are currently of equal interest to the present consortium. In anticipation of high growth rates for future mobile communication services, however, the project boldly aims at covering the whole area between Morocco in the West and Kuwait in the East, from Turkey in the North to Yemen in the South of the *OrienTel* region. Through the development of dialect adaptation techniques it will be possible in the near future, we hope, to adapt acoustic models of one language variety of Arabic to a related one. For the time being, a preselection process based on linguistic and commercial judgements as well as on considerations of sheer manpower has picked 9 out of potential 19 countries in which *OrienTel* will become active.

Despite the diversity between individual markets certain general trends are apparent. Mirroring what has happened in the rest of the developed world, cellular telephony has grown rapidly, particularly in markets where the fixed line infrastructure is inadequate. A case in point is Egypt. With 62 million inhabitants Egypt is the second most populous country in the *OrienTel* region: its fixed line teledensity of 10.97% places it at place 13 in the regional league table of lines per 100 inhabitants, yet it comes third in terms of number of mobile subscribers. While the Egyptian government has attempted to improve the availability of fixed line telephony by setting ambitious targets for state-owned Telecom Egypt, the private sector has been allowed a relatively free rein in the mobile sector. The result has been an explosion in the

[1] Thanks go to the European Commission, who are funding *OrienTel* as an R&D project under the 5th Framework Programme (Contract IST-2000-28373).

[2] Infos on *SpeechDat* and related projects can be gathered from http://www.speechdat.org . Publications focusing on specific members of the *SpeechDat* family are, for example, Höge/Tropf 1996 (*SpeechDat M*), Höge et al. 1999 (*SpeechDat II*), Pollak et al. 2000 (*SpeechDat East*), Moreno et al. 2000a (*SpeechDat Car*), and Moreno et al. 2000b (*SALA*).

[3] All figures of the present section were taken from CIT Publications (2000).

number of cellular subscribers. During 1999 the number of subscriptions rocketed from 187,000 to 890,000. According to Egypt's two mobile operators, France Télécom-backed MobilNil and Vodafone AirTouch's Misrfone, the market doubled in size again during 2000 to 1.8 million. Turkey is in a similar situation. With a fixed line network of 17.4 million lines, Turkey's teledensity of 26.6% places it sixth in the regional ranking table, yet its mobile sector has experienced nothing short of phenomenal growth: at the end of 1996 mobile penetration stood at just over 1.2%, but three years later it had increased ten-fold to 12%, and rose to just below 20% by the end of 2000.[4] Company-internal considerations of some *OrienTel* partners bear further evidence of the current interest in speech applications gradually extending from Europe towards some of the countries covered in the project. A survey of the needs for future language development undertaken by ELRA points into the same direction: particularly Arabic and Turkish are currently on the wish list of many companies active in the field of language and speech. Largely due to such infrastructural and commercial considerations, the *OrienTel* consortium chose nine out of a potential set of 19 countries between Morocco in the West and the Gulf States in the East. The countries treated in *OrienTel* so far are depicted in Table 1:

Country	Partner
UAE	Philips
Saudi Arabia	Lucent
Israel/Palestine	NSC
Egypt	IBM
Tunisia	UPC/ELDA
Morocco	ELDA/UPC
Turkey	Siemens
Cyprus	Knowledge/Patras Univ.

Table 1: *OrienTel* countries and partners

More countries may follow in case new partners decide to join the project.

3. Linguistic settings

From a linguistic point of view, too, the *OrienTel* region is far more diverse than any region covered in previous projects of a similar scope such as the various members of the *SpeechDat*-family or *SpeeCon* (Siemund et al. 2000, cf. also http://www.speecon.com). In order to treat the linguistic peculiarities of the area adequately, *OrienTel* follows a different strategy than previous *SpeechDat* projects. As Table 1 shows, each partner in the consortium is not responsible for a single language but for a whole country. The difference is an important one, since, as will be outlined below, in most *OrienTel* countries everyday-life is governed by more than a single language. One of the first project tasks was therefore to determine the various languages spoken in the *OrienTel* region, taking into account both linguistic and commercial criteria. From the consortium's point of view Arabic, Turkish, Hebrew and Cypriote Greek turned out to be of most immediate concern with Farsi being on the wish list for future stages of the project. Furthermore, English and French turned out to be of commercial interest as the dominant business languages in some *OrienTel* countries and because non-native varieties of European languages constitute a hitherto grossly neglected domain in linguistic research. This is also the reason why *OrienTel*'s language portfolio is complemented by German as spoken by Turks in Germany, who represent the largest linguistic minority of the country.

The most complex linguistic picture of the *OrienTel* region, however, is no doubt presented by Arabic and its variants. Arabic of the *OrienTel* area can be subdivided into four broad dialect regions, as outlined in Table 2:

Dialect region	Countries
Mahgreb Arabic	Morocco, Algeria, Tunisia, parts of Libya
Egyptian Arabic	Egypt, parts of Libya
Levantine Arabic	Syria, Lebanon, Israel + Palestine Authorities, Jordan
Gulf Arabic	Kuwait, Qatar, Bahrain, UAE, Saudi Arabia, Oman

Table 2: Dialect regions of Arabic

In order to represent all dialect regions adequately, all areas are attended to by at least one partner. In each country, the variety of languages spoken is rather large. In Morocco, for example, the official language is Modern Standard Arabic, the rather formal language of religion, the media and of public institutions. In everyday interaction though, people either tend to speak a local colloquial variant of Arabic that is only remotely related to the Standard (not to mention the various non-Arabic languages such as Berber) or, when it comes to commercial interaction, French as the language inherited from Morocco's colonial past. All three (or even more) languages have their place in everyday life and user-friendly applications have to take into account each country's linguistic diversity and its users' preferences. The databases produced in *OrienTel* are depicted in Table 3:

Country	1st language	2nd language	3rd language
UAE	Mod. Std. Arabic	Modern Coll. Arabic	English
Saudi Arabia	Mod. Std. Arabic	Modern Coll. Arabic	English
Israel/Pal. Auth.	Mod. Std. Arabic	Mod. Coll. Arabic	Hebrew
Egypt	Modern Std. Arabic	Modern Coll. Arabic	English
Tunisia	Mod. Std. Arabic	Modern Coll. Arabic	French
Morocco	Mod. Std. Arabic	Modern Coll. Arabic	French
Turkey	Turkish	-	German
Cyprus	Cypriote Greek	-	English

Table 3: *OrienTel* languages

[4] The market analysis is based on *Telecommunications Markets in the Middle East*. Exeter: CIT Publications, 2000.

As can be gathered from Table 3, the *OrienTel* consortium will produce a set of 22 databases in 8 countries, all of which will be made publicly available after the end of the project and a commercially reasonable quarantine period.

4. Linguistic research and dialect adaptation

The rather complicated linguistic situation in the *OrienTel* countries calls for innovative approaches to speech recognition techniques. Thorough research will be conducted into multilingual acoustic modelling and the development of multilingual lexicons, including descriptions of phonetic inventories. An important goal will also be the development of phonetic and orthographic transcription strategies. By default written Arabic and Hebrew orthography depict consonants only. Even though it is possible to render vowels by supra- and supersegmental markers, fluent reading of such "annotated" words is awkward even for native speakers. Strategies will therefore be developed to prompt speakers reliably even if the meaning of, for example, a single command word cannot be gathered from the context of whole sentences. The problem of vowels is of particular importance especially since it is largely the vowels on which current Hidden Markov Modelling heavily relies (cf. Rabiner/Juang 1993). Once parts of the various databases become available, it will therefore be one of the main research tasks to assess the linguistic features of dialect clusters and develop techniques of dialect adaptation across the Arabic-speaking world.

5. Foreign accent adaptation

Apart from the databases representing Standard varieties and local dialects, a separate set of data will be produced for foreign accent adaptation. Due to the *OrienTel* countries' colonial, protectorate or migration history, the most prominent foreign languages in the region are French, English and, for different reasons, German. On the one hand, collecting data of this kind will ensure true multilinguality of applications in the *OrienTel* countries. On the other hand, French, English and German services already under operation in the EU can be adapted to foreign accent variation.

6. Demonstrator development

In order to show that the multilinguality approach taken in *OrienTel* is feasible, the project will produce two demonstrator applications. The exact kind of services will be specified at a later stage of the project. Considerations will, however, take into account the convergence of internet, WAP and voice for service portals, unified messaging, customer care applications, directory assistance and banking. The two demonstrators will reflect two different types of services and will account for two different linguistic regions.

7. Dissemination of information and results

In order to keep the speech recognition community informed about the *OrienTel* efforts, the project will contribute to scientific discussions concerning the languages of the *OrienTel* region at conferences, in publications and through relevant mailing lists. It will furthermore continuously update the project's website with information on *OrienTel* activities and publish the results (cf. section 9 below). The 22 databases will be made publicly available through the European Language Resources Association (ELRA) in due course after the project has ended.

8. Database specification

Due to the linguistic heterogeneity of the region, questions of database specification such as corpus composition, orthographic and phonetic transcription strategies constitute a crucial part of the project. Particularly Arabic and Hebrew pose interesting problems for speech recognition that were never tackled in projects of the *OrienTel* scale before. Cases in point are the rendering of vowels, the right-to-left writing system and the transcription of oral or colloquial speaking styles. While at the time of writing the present paper quite a few design details are still under discussion (the design phase is due for completion before the LREC2002 conference starts), some of the cornerstones can already be reported at the present stage.

8.1. Recording scenarios and platforms

All *OrienTel* databases will be recorded from fixed and mobile networks via ISDN lines and multiple channels, i.e. either through a Basic Rate Interface or a Primary Rate Interface (cf. Senia 1998). A dialogue will be implemented by the application driving the recordings. The dialogues will be designed to make the caller speak and act comfortably.

8.2. Corpus and vocabulary

Data collections will rely on three separate sets of prompt sheets, namely one each for

- the 'foreign' languages in Arabic-speaking countries, i.e. English and French, including Turkish, Greek, Hebrew and German
- Modern Standard Arabic
- Modern Colloquial Arabic

While the specifications for English, French, Greek and German are largely based on previous *SpeechDat* projects and *SpeeCon*, the design for Arabic and Turkish presents a novelty. All three sets of prompt sheets, however, contain the following items, though in varying quantities with at least 47 items per sheet:

- isolated digits
- digit and number strings
- natural numbers
- currency amounts
- yes/no questions
- dates
- times
- application keywords and phrases
- word spotting phrase using embedded application words

- directory assistances names (proper names, place names, company names)
- spellings
- phonetically rich words and sentences
- spontaneous utterances

8.3. Transcription and annotation

The *OrienTel* transcription and annotation conventions are largely based on conventions used by the Linguistic Data Consortium and ARPA in producing the ATIS CD-ROMs[5], and the simplifications made for the *SpeechDat*-predecessors of this project, and *SpeeCon*. The goal of the specification document that should be finished by the time of the present workshop is to define a coarse transcription that can be performed quickly, but covers adequately the acoustic events most important for the training and testing of automatic speech recognisers. The transcription is orthographic (cf. the lexicon section below for phonetic renderings) and includes a few markers representing audible acoustic events (speech and non-speech) present in the corresponding waveform files. The phoneme symbol set aims at the localisation of the main acoustic events according to a coarse categorisation rather than a full description of all possible sounds that may appear during a recording. Extra marks contained in the transcription aid in interpreting the text form of the utterance; markers for non-speech acoustic events and distortions have been chosen such that they can be automatically removed or modified to yield the base transcription. The overall aim is to keep as much speech in the corpus as possible and to avoid the need for deleting recordings from the corpus due to some extra noises, disfluencies, etc. All items for all languages covered will be transcribed in standard orthography and will be Romanized in the label files. A Sampa transliteration will be generated and discussed with the Department of Phonetics and Linguistics at UCL (cf. http://www.phon.ucl.ac.uk/home/sampa/home.htm) if need be. Administrative information on speech files and their properties will be stored in SAM files (cf. http://www.icp.grenet.fr/Relator/standsam.html).

8.3.1. Strategies for recording colloquial Arabic

Dialectal Arabic is exclusively a spoken language and can very rarely be found in the written form. This fact imposes constraints on the recording procedure. There are a number of possibilities with regard to collecting colloquial Arabic speech:

1. recording spontaneous speech only;
2. presenting audio prompts to the speaker who then only needs to repeat what has been prompted;
3. presenting written prompt sheets to the speaker that he/she needs to read. The prompts can be
    a. written in vowelized Arabic script or
    b. transcribed using Latin alphabet.

The first option is likely to provide the most natural results. However, transcription of such spontaneous material would be very difficult as well as time- and money-consuming. That is why it has been decided to include a limited number of spontaneous items in the recordings. An example of such a spontaneous item is asking the speaker to answer a question on, e.g., the sightseeing sites of his/her town. Furthermore, spontaneous speech will be recorded as response to questions concerning dates, natural numbers and proper names. An example is enquiring the speaker's date and place of birth. Such questions are expected to produce short and simple utterances.

Typically, a *SpeechDat*-like database contains a number of items whose content has been defined in advance, such as phonetically rich sentences and application words. For this type of recordings option 2 or 3 need to be taken into account (the audio prompts or prompt sheets). In order to determine the best approach a number of basic experiments have been carried out with Moroccan speakers. These have enabled us to dismiss option 3b (transcription with Latin characters). During the experiments using this prompting approach none of the speakers was able to pronounce words naturally. On the other hand, the best results were obtained using audio prompts. However, this option has been dismissed too since it poses too many practical problems with such a high number of speakers and items which need to be recorded. Colloquial Arabic is difficult to read for most of the speakers since they are not used to reading it. Nevertheless, the majority of the speakers manage to come up with a correct pronunciation after having analysed the prompted text for a moment. Because of that, however, it is important to grant the speakers some extra time to become acquainted with the script before the recording starts.

8.3.2. Strategies for recording standard Arabic

Conventionally, the orthographic representation of standard Arabic relies on consonants. Although it is possible to represent vowels using diacritics (located super- and suprasegmentally), fluent reading of such scripts remains a challenge; most of the speakers are not used to reading texts containing vowel diacritics. In complete sentences it is possible for the speaker to deduce the vowels from the context. However, for isolated words it is necessary to mark the vowels in order to disambiguate between the different options. This situation is similar to that of colloquial Arabic in which the speakers need extra preparation time to be able to read vowelised scripts without hesitation and pauses.

8.4. Specification of speakers

The number of speakers to be recorded is 2000 per country. This number is distributed between the set of databases to be collected. Table 4 on the following page shows the minimum number of speakers per country and recorded language. A maximum overlap of 15% in the total number of speakers per country between the different databases is allowed.

[5] Cf. http://www.ldc.upenn.edu/, http://www.arpa.gov/, and http://www.atis.org , respectively.

Country	Colloquial	Standard	Business
Morocco	1000	500	500
Tunisia	1000	500	500
Egypt	1000	500	500
UAE	1000	500	500
Saudi Arabia	1000	500	500
Turkey	-	1700	300
Israel	500	500	1000[6]
Cyprus	1000	-	1000[7]

Table 4: Number of speakers per database

8.4.1. Gender

The distribution of male and female speakers should be 50% each per database, with an allowed deviation of 5% for the whole database per language. There is no gender restriction for "Age" and "Dialect". For "Environment", the gender distribution must be 30-70% for each sub-category.

8.4.2. Age

Table 5 presents the distribution of speaker age:

Age	16-30	31-45	46-60
Proportion	≥ 30%	≥ 20%	≥ 10%
Requirement	Mandatory	Mandatory	Mandatory

Table 5: Distribution of speaker age

Naïve speakers should be recorded rather than experienced or trained speakers to guarantee more natural speaking styles, voices and dialects.

8.4.3. Dialect

Many (though not all) of the languages spoken in the *OrienTel* regions are not the speaker's actual mother tongue. In such cases, we consider a person who spent most of his/her childhood, or who grew up in the concerned region, as having no foreign accent. Language-specific cases should be documented in the LSPs.

The specific number of dialects relevant to each country should be discussed in the LSP documentation. The distribution of speakers over dialect regions refers only to the colloquial varieties of the language. Speech should be collected from a minimum of three different dialect regions (if possible), with at least 20 speakers recorded for each defined dialect.

The speaker's dialectal region is determined by asking the question *"in which district did you grow up" or "where did you spend most of your childhood"*, not the question *"where do you live"*. The allocation of city/district names to the corresponding dialect region can be determined according to the information provided by each partner in the LSP documentation.

8.4.4. Distribution of environments

The speaker distribution for the mobile network should be between 65 to 75% of the total number of speakers in the database; e.g., if there are 1000 speakers in the database, between 650 and 750 of them should be recorded through the mobile network. At least 30% of each gender must be recorded in each environment.

Both the fixed and mobile networks are further divided into specific environments. Speaker distribution over each environment is shown in Table 6:

	Environment	Speaker distribution
Fixed network 30% ± 5%	Home/office	≥ 75%
	Public place/booth	
Mobile network 70% ± 5%	Home/office	≥ 20%
	Public place/street	≥ 20%
	Vehicle	≥ 15%
	Hands-free car kit (optional)	≥ 5%

Table 6: Distribution of recording environments

8.5. Specification of the lexicon

The lexicon is an alphabetically ordered table of distinct lexical items that occur in the corpus with the corresponding pronunciation information. Each distinct word should have a separate entry, which will be laid down in the order orthography ⇨ frequency ⇨ transliteration (for Arabic and Hebrew) ⇨ phonetic transcription ⇨ variants (optional).

The lexicon is derived from the annotated database and is set up as follows:

- Standard Language: Arabic & Hebrew script, both vocalized and not vocalized.
- European languages: Latin script
- Colloquial Language: Region specific Arabic script (same as in orthographic annotations)
- Acronyms such as *IBM* should appear as complete words in the lexicon, i.e. as letters with no spaces in between. The reason is that there are often different ways of pronouncing them (spelled and expanded).

The phonetic alphabet used will be SAMPA, and is thus case-sensitive. While a Hebrew SAMPA alphabet is currently under negotiation for standardization as part of the *SpeeCon* project, *OrienTel* will make an effort to further standardize Arabic and Turkish SAMPA alphabets. Sampa symbols for each language are defined in the language-specific documents accompanying each database and are considered as a standard set of phonemes for that language.

9. Disclaimer and Contact

Since the specifications outlined in this document are still being discussed at present and are thus still subject to revision, the latest state of the *OrienTel* art can always be gathered from the continuously updated *OrienTel* website at http://www.orientel.org. The co-ordinators of the project can be contacted either via the internet pages or through rainer.siemund@philips.com.

[6] Hebrew.
[7] Greek.

10. References

CIT Publications (2000). *Telecommunications markets in the Middle East*. Exeter: CIT Publications.

Höge, H., C. Draxler, H. van den Heuvel, F.T. Johansen, E. Sanders, H. Tropf (1999). Speechdat multilingual speech databases for teleservices: Across the finish line. In *Proceedings of EUROSPEECH '99*, vol. 6 (pp. 2699–2702). Budapest: ESCA.

Höge, H., H. Tropf (1996). SpeechDat (M) Final Report (D06/D07). Available from http://www.speechdat.org.

Moreno, A., B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, J. Allen (2000a). SpeechDat-Car. A large speech database for automotive environments. In *Second International Conference on Language Resources and Evaluation. Proceedings vol. II* (pp. 895--900). Athens: ELRA.

Moreno, A. R. Comeyne, K. Haslam, H. v. d. Heuvel, H. Höge, S. Horbach, G. Micca (2000b). SALA: SpeechDat across Latin America. Results of the first phase. In *Second International Conference on Language Resources and Evaluation. Proceedings vol. II* (pp. 877–882). Athens: ELRA.

Pollak, P., J. Cernocky, J. Boudy, K. Choukri, H. v.d. Heuvel, K. Vicsi, A. Virag, R. Siemund, W. Majewski, J. Sadowksi, P. Staroniewicz, H. Tropf, J. Kochanina, A. Ostrouchov, M. Rusko, M. Trnka (2000). SpeechDat(E) - Eastern European Telephone Speech Databases. In *Proceedings LREC'2000 Satellite workshop XLDB - Very large Telephone Speech Databases, 29 May 2000* (pp. 20–25). Athens: ELRA.

Rabiner L.R., and B. Juang (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *IEEE Transactions on Acoustics, Speech and Signal Processing* 77(2), 257–286.

Senia, F., I. Chatzi (1997), Installation of the recording device and documentation. *Deliverable SD2.1 of the SpeechDat II project LE2-4001-SD2.1*. Available from http://www.speechdat.org/speechdat/deliverables/public/SD21V22.doc .

Siemund R., H. Höge, S. Kunzmann, and Marasek K., (2000). *SpeeCon* - speech data for consumer devices. *Second International Conference on Language Resources and Evaluation. Proceedings vol.* II (pp. 883—886). Athens: ELRA.