OrienTel - Multilingual access to interactive communication services for the Mediterranean and the Middle East

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

OrienTel is a project funded within the European Commission's IST framework that focuses on collecting linguistic data for telephony-based IT applications across the Mediterranean and the Middle East. Languages covered in this *SpeechDat*-based project are Cypriote Greek, Turkish, Hebrew, different varieties of Arabic, French, English and German. Within the project's lifetime of 30 months, starting in September 2001, *OrienTel* will produce a set of 22 databases, develop dialect adaptation techniques, conduct research into multilingual acoustic modelling and deploy two demonstrators as a proof of concept.

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, Hebrew and Turkish, 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.

2. Commercial and technological settings

The *OrienTel* region presents a very heterogeneous picture both from a technological and a commercial point of view. In terms of purchasing power measured in US$ per capita, the potential *OrienTel* countries range from Yemen's 304 to the United Arab Emirates' 16,800. In terms of the telecommunications infrastructure, the fixed line teledensity ranges again from Yemen's 1.4 per cent coverage to Cyprus' 54 per cent, the number of mobile service subscribers from virtually zero for Syria to 7,871,500 in Egypt (CIT Publications 2000). State of the art technology in many of the Gulf States stands in sharp contrast to the outmoded systems installed in countries such as Yemen or Syria. 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,

---

[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*).

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 (split over two columns):

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

Country	1st language	2nd language	3rd language
Turkey	Turkish	-	German
Cyprus	Cypriote Greek	-	English

Table 3: *OrienTel* languages

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[3], and the simplifications made for the *SpeechDat*-predecessors of this project, and *SpeeCon*. The goal of the specification document that is currently being drafted 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).

8.4. 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.

---

[3] 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
Cyprus	1000	-	1000

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	$\geq 30\%$	$\geq 20\%$	$\geq 10\%$
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	$\geq 75\%$
	Public place/booth	
Mobile network 70% ± 5%	Home/office	$\geq 20\%$
	Public place/street	$\geq 20\%$
	Vehicle	$\geq 15\%$
	Hands-free car kit (optional)	$\geq 5\%$

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.

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.