

**DELIVERABLE IDENTIFICATION**

Identification number	LE2-4001—SD1.2.1
Type	Technical Report
Title	Environmental and speaker specific coverage for Fixed Networks
Status	Draft
Deliverable	SD1.2.1
Work Package	WP 1
Task	Task 1.2
Period covered	T01-T06
Date	26 February, 1997
Version	2.2
Number of pages	46
Author(s)	<p>           Francesco Senia, CSELT            Robrecht Comeyne, L&amp;H            Borge Linberg, AUC            Vesa Kuoppala, DMI            Aurelie Nataf, MATRA            Christoph Draxler, UMUNICH            Irene Chatzi, KNOWLEDGE            Finn Tore Johansen, Ingunn Amdal, TELENOR            Isabel Trancoso, INESC            Zdravko Kacic, UMARIB            Asunción Moreno, UPC            Johan Lindberg, KTH            Caloz Gilles, IDIAP            Louise Helliker, BT            Kamran Kordi, GEC         </p>
Work package (WP) / Task (T) responsible	<p>           WP 1 - Richard Winski /            Task 1.2 - Simon P.A. Ringland         </p>
Project contact point	<p>           Harald Höge,            Siemens AG, ZFE T SN 5, D-81730 München            Phone: +49 89 636 53374, Fax: +49 89 636 49802            E-mail: Harald.Hoege@zfe.siemens.de         </p>
CEC project officer	José Soler
Status	Public
Actual distribution	Consortium and CEC
Supplementary notes	

Key words	Speaker characteristics, dialects, telephone networks
Abstract	This report describes the main speaker and environmental factors that should be taken in account during a speech database collection. Dialect influences to the European languages are described too.
Status of abstract	Public

Received on	
Recipients catalogue number	

### **DOCUMENT EVOLUTION**

Version	Date	Status	Notes
1.0	10/04/96	First draft	To be discussed at the Athens meeting
1.1	03/09/96	Second draft	Reviewed internally by WP and Task managers
2.0	05/02/97	Third draft	All maps added and some section reviewed
2.1	24/02/97	Fourth draft	Some maps, Portugal, Switz. and UK reviewed.
2.2	26/2/97	Final report	Approved and delivered to CEC on

## **Table of Contents**

<b>INTRODUCTION</b>	<b>5</b>
<b>1. SPEECHDAT DATABASES SPECIFICATIONS</b>	<b>5</b>
<b>2. SPEAKER SPECIFIC CHARACTERISTICS</b>	<b>6</b>
2.1 Male/Female	6
2.2 Age	7
2.3 Weight and height	8
2.4 Smoking and drinking habits	8
2.5 Pathological speech	8
2.6 Professional vs. Untrained speakers	8
2.7 Socio-economic factors	9
<b>3. REGIONAL/DIALECTICAL FACTORS</b>	<b>9</b>
3.1 Belgium (Flemish and Belgian French)	11
3.2 Denmark (Danish)	12
3.3 Finland (Finnish and Finnish-Swedish)	13
3.4 France (French)	14
3.5 Germany (German)	15
3.6 Greece (Greek)	16
3.7 Italy (Italian)	19
3.8 Luxembourg (Luxembourgish French and Luxemburgish German)	21
3.9 Norway (Norwegian)	23
3.10 Portugal (Portuguese)	28
3.11 Slovenia (Slovenian)	29
3.12 Spain (Spanish)	31
3.13 Sweden (Swedish)	34
3.14 Switzerland (Swiss French and Swiss German)	38
3.15 United Kingdom (English and Welsh)	40

<b>4. GENERAL ENVIRONMENTAL SPECIFIC CHARACTERISTICS</b>	<b>42</b>
<b>4.1 Calling environment</b>	<b>42</b>
<b>4.2 Handsets</b>	<b>43</b>
<b>4.2 Network</b>	<b>44</b>
<b>BIBLIOGRAPHY</b>	<b>46</b>

## Introduction

When starting to design a Speech Database one should have in mind several questions and two of them could be:

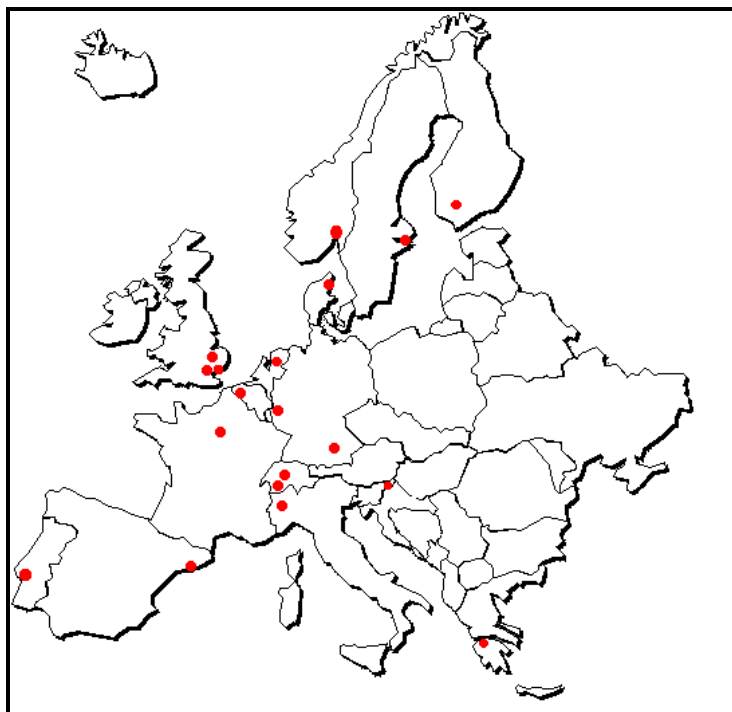
*“How are the speakers for a speech corpus selected?”*  
*“Which are the environmental characteristics I have to take into account?”*

For SpeechDat purposes, training and testing of recognition systems, the population of interest must be suitably random sampled. In general it is recommended to include all possible types of speakers in a speech corpus, unless there are imperative arguments to exclude specific speaker groups. Speaker characteristics, which are potentially important and should therefore be considered when selecting the speaker population, are well described in the EAGLES handbook. They will be briefly described in the first chapters of this report.

The environmental characteristics that can influence a speech database are originated from several factors, but mainly: environment where is a speaker, microphone used, connection between speaker terminal and the recording machine, and finally the recording machine itself. All these factors are described in the SpeechDat(M) report D3.1.2.1 and briefly below.

### 1. SpeechDat databases specifications

Within the LE2-4001 “SpeechDat” project it has been planned to collect the following speech databases<sup>1</sup> by using the country partner’s fixed telephone networks:



**Figure 1 - Countries involved in the Fixed Network Database Collection**

<sup>1</sup> The reported table has been extracted from the project’s Technical Annex (page 59, table 7.1). To that table has been added the Norwegian language, because Telenor joined the project after it was started.

Language (variant)	Country	Partner	Recorded Speakers
Danish	Denmark	AUC	+4000
Flemish	Belgium	L&H	1000
English	United Kingdom	GEC	+4000
Welsh	United Kingdom	BT	2000
Finnish	Finland	DMI	4000
French	France	MATRA	5000
French	France	PHILIPS <sup>2</sup>	0
Belgian French	Belgium	L&H	1000
Swiss French	Switzerland	IDIAP	+2000
			(PolyVar) 1000
Luxemb. French	Luxembourg	L&H	500
German	Germany	SIEMENS	+4000
Swiss German	Switzerland	IDIAP	1000
Luxemb. German	Luxembourg	L&H	500
Greek	Greece	KNOW & UPATRAS	5000
Italian	Italy	CSELT	+3000
Portuguese	Portugal	PT	+4000
Slovenian	Slovenia	SIEMENS	1000
Spanish	Spain	UPC	+4000
Swedish	Sweden	KTH	5000
Finnish Swedish	Finland	DMI	1000
Norwegian	Norway	TELENOR	1000

**Table 1 - Languages involved in the Fixed Network Database collection**

## 2. Speaker specific characteristics

Demographic factors form a very important set of relatively stable speaker characteristics which must be considered when designing sampling procedures for a corpus collection project. Each corpus should have sufficient demographic coverage.

### 2.1 Male/Female

Sex is known to have an enormous impact on speech quality. It is not well known at what age sex-related speech characteristics become prevalent; it appears that also cultural factors and stereotypes play an important role. Because no definitive recommendation can be given with respect to the age above which sexes should be distinguished and sampled individually, each corpus should comprise approximately equal numbers of speakers of both sexes. Children should be considered as a “third sex” and a small number of them should be recruited. In general speaker sex is suspected to affect at least four aspects of speech behaviour:

- pitch and intensity - women are known to have higher average pitch than man;

<sup>2</sup> PHILIPS will \*make available\* for the SpeechDat(II) consortium the SpeechDat(M) 1000 speakers FDB, i.e. no new FDB recordings by Philips.

- overall spectral slope - women are reported to tend towards breathy a voice quality than males;
- accuracy of pronunciation - women are reported to adhere more to standard pronunciation than men, at least for American English and Dutch, and it is not known whether this finding generalises to other languages;
- vocabulary and syntax - different on the level of vocabulary and syntax are only relevant when spontaneous speech is being recorded.

For these project purposes it has been decided to collect approximately equal numbers of speakers of both sexes with at most 5 % of difference from half of the population, according to the following table:

<b>Sex</b>	<b>Range %</b>
male	45 - 55
female	45 - 55

**Table 2 - Sex distribution**

⇒ Sex information should be reported in the database description files and these data will be checked in the validation phase. Databases that do not comply should be rejected.

## 2.2 Age

Age influences at least two aspects of speech behaviour:

- voice quality - it is not known the impact of the speaker's age on performance of automatic speech recognition but, if we are based on the capacity that people are moderately good at guessing age from speech signal characteristics, we could think to distinguish between the group under 20, the group between 20 and 60 and the group over 60.
- vocabulary and syntax - here the consideration described above on the impact of sex on speech behaviour apply in exactly the same way, but main differences between older generation and younger speakers can be mainly found in spontaneous corpus.

For SpeechDat project it has been decided to collect from speakers with the following age distribution:

<b>Age</b>			<b>Middle point</b>	<b>Minimum %</b>
0	..	15	12	1 recommended
16	..	30	23	20
31	..	45	38	20
46	..	60	53	15
61	..	∞	78	optional

**Table 3 - Age distribution**

⇒ Speaker's ages should be reported in the database description files, and when the exact ages are missing the middle points in the table can be used. These requirement will be checked in the validation phase and databases that do not comply will be rejected.

### **2.3 Weight and height**

It appears that people are moderately successful in estimate speaker weight or speaker height from speech recording alone, but the exact signal characteristics that enable people to guess them are unknown. It is unknown whether speaker weight or speaker height influences automatic speaker recognition, but in a sufficient large sample of speakers most weight/height groups will probably be represented. This is certainly true for databases with a large number of speakers. Particular attention should be paid when collecting databases with a few speakers involved

⇒ In any case, for SpeechDat purposes it has been decided not to investigate speaker weight or speaker height and these need not be reported.

### **2.4 Smoking and drinking habits**

Smoking and drinking habits certainly influence the voice quality.

⇒ SpeechDat partners are free to report them, if they decided to collect these information. In any case these factors will not be taken into account during the validation stage.

### **2.5 Pathological speech**

The boundary that divides pathological speech from non-pathological speech is very difficult to draw. Hoarseness due to smoking can be regarded as a very mild speech disorder, whereas more severe speech disorders include, for instance, paralysis of the vocal cords and aphasia. The EAGLES handbook clearly describe five different levels of speech disorders, but for the purposes of this project it is not desirable to include speakers with severely pathological speech. On the other hand, speakers with mild pathological disorders, such as hoarseness, can be included in this corpus designed for recognition. Of course, researches that focus specifically on pathological speech, for instance when a recogniser is developed for the use as an environmental control device for handicapped persons, should amply include such type of speakers.

⇒ It is not expected to report these information in the SpeechDat databases and they will not be further investigated.

### **2.6 Professional vs. Untrained speakers**

Professional speakers will not be taken into account because they are not really representative of the 'normal' speech behaviour in the community; in fact they are selected mainly to develop text-to-speech systems. The only exception may be the SDB database which has the possibility of employing a professional mimic for impostor testing.

⇒ In the SpeechDat project professional speakers will not be recruited

## 2.7 Socio-economic factors

A clear distinction between different social classes exists, for instance, in India, where each member of the society belongs to a specific caste. However, in most cultures it is very difficult to distinguish between social classes. The division into three categories lower-class, middle-class and upper-class seems to be most widely accepted for Western cultures but in the SpeechDat project socio-economic class can be documented in the partner's preferred way. However the following classification, where social grades are defined by occupation, is used widely in the UK:

grade	social class	occupation
A	Upper Middle	Class Higher Managerial (Administrative or Professional)
B	Middle Class	Intermediate Managerial (Administrative or Professional)
C1	Lower Middle Class	Junior Managerial(Administrative or Professional Supervisory Grades), Clerical Workers
C2	Skilled Working Class	Skilled Manual Workers, Craftsmen, Specialised Workers
D	Working Class	Semi and Unskilled manual Workers, Apprentices, Labourers and Mates to C2 occupations
E	Those at the lowest level of subsistence	Widows Living on state benefit alone, Retired Person living on state pension alone, Invalids living on state pension alone, Unemployed for at least 6 months and living on state benefit alone with no chief wage earner in household.

**Table 4 - Social classes**

Code social class according to their previous occupation for:

- Persons sick for less than 6 months
- Retired persons other than those living in state pension alone
- Persons unemployed for less than 6 months

⇒ In the SpeechDat project it has been decided to optionally report social status, possibly by using the previous table. If other means are used they should be well documented.

## 3. Regional/Dialectical factors

Last but not least the regional background of speakers can have a large effects on their speech. People speak differently dependent on the specific region in which they were grown up and also on factors such as the linguistic background of the parents. It is widely assumed that the high-school period is most decisive for the regional or dialectal colouring in one's speech.

⇒ In the SpeechDat project is recommended to obtain information about the high-school period when collecting data about the speaker's background.

Although there is an enormous amount of literature on Dialectology the impact of dialects on standard speech is not well understood. Linguistics and dialectologists appear to disagree about the number of major dialects in a language area, and about the boundaries between the areas where a specific dialect is spoken. In collecting SpeechDat databases the factor *regional/dialectical colouring* should be properly accounted for. However, since the basic data to determine number of dialects and dialect boundaries are difficult to obtain and

probably not always reliable, it is recommended that *dialect* is operationalised by *geographic region*.

There is considerable uncertainty as to how to treat dialects in corpora collected to develop speech recognition systems for use in telephone information systems, but a common guideline arising from the attached language specific choice, is that speakers should be selected proportionally to their presence in a certain region. This appears to be appropriate for general purpose corpora such as the SpeechDat one.

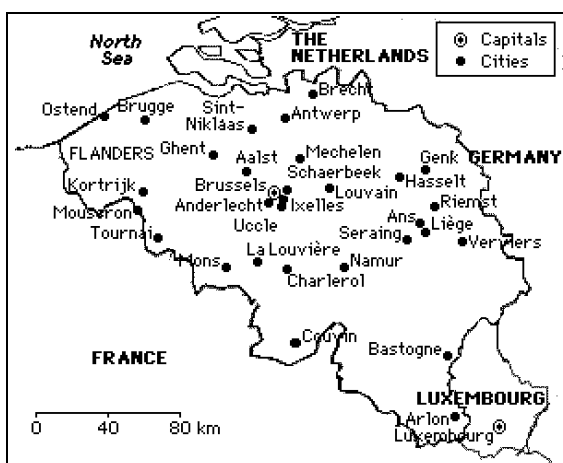
⇒ In the SpeechDat project it is mandatory to select and report dialect and regional origin of a call. It is mandatory supply at least 1 % of speakers each region and this figure will be checked during the validation stage<sup>3</sup>.

In the next paragraph each partner involved in the Fixed Network speech collection has reported the characteristics of their countries, the number and boundaries of their dialects and the number of speakers selected each in region proportionally to the region's population.

---

<sup>3</sup> Actually the 0.5% will be considered acceptable (agreed at the Athens meeting)

### 3.1 Belgium (Flemish and Belgian French)



The Belgium (10.068.000 people in 1993) is administratively split in 9 provinces and there are two official languages: Flemish and Belgian French. In this country L&H will collect two 1000 speakers databases, one for each language.

Geographical regions for dialectal variants of Flemish in Belgium - "Vlaanderen" are:

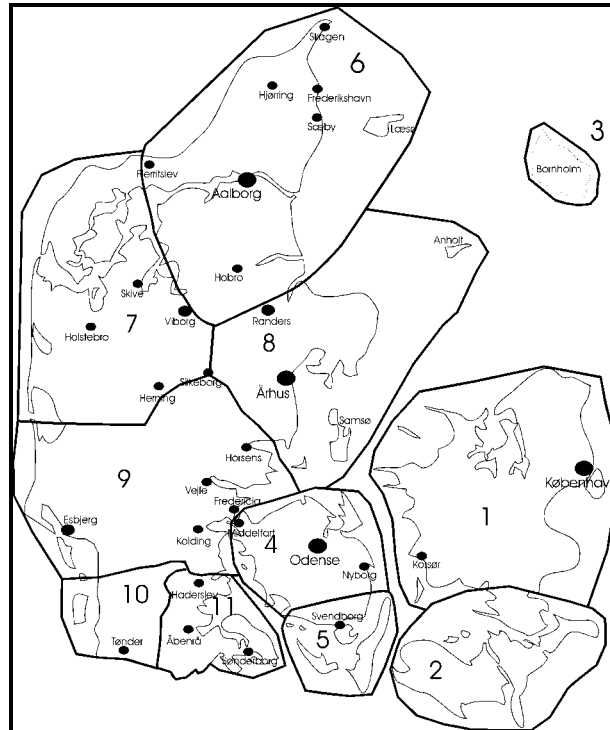
No.	Region	Description
1	Westvlaams	This is a dialect typically spoken in Bruges, Ostend, Diksmuide, Roeselare, Kortrijk, Tielt, Ypres which covers the province of West-Vlaanderen.
2	Oostvlaams	This is a dialect typically spoken in Eeklo, Ghent, St-Niklaas, Dendermonde, Oudenaarde, Aalst which covers the province Oost-Vlaanderen.
3	Brabants	This is a dialect typically spoken in Brussel, Antwerp, Mechelen, Turnhout, Halle, Vilvoorde, Leuven which covers the province Antwerp and the province Vlaams Brabant.
4	Limburgs	This is a dialect typically spoken in Hasselt, Maaseik, Tongeren which covers the province of Limburg.

The geographical regions for dialectal variants of French - "la Wallonie" are:

No.	Region	Description
1	l'Ouest-Wallon	This is a dialect typically spoken in La Louvière, Charleroi, Philippeville which covers the eastern part of the province Hainaut and the south-eastern part of the province Namur.
2	le Namurois	This is a dialect typically spoken in Namur, Dinant which covers the province of Namur and the southern part of the province Brabant.
3	l'Est-Wallon	This is a dialect typically spoken in Waremme, Huy, Liège, which covers the province Liège (without the Oostkantons: German dialect).
4	le Wallo-Lorrain	This is a dialect typically spoken in Marche-en-Famenne, Bastogne, Neufchâteau which covers the province of Luxembourg (without Arlon: German dialect).
5	le Picard	This is a dialect typically spoken in Tournai, Mons, Beaumont which covers the western part of the province Hainaut (this is a French dialect, whereas the other dialects are typically wallon).

### 3.2 Denmark (Danish)

Denmark had a population of 5.181.000 people in 1993. The only official language in Denmark is Danish. CPK, Aalborg University, will organise the recording of 4000 speakers selected from 11 geographical regions as follows.



**Figure 2 - Denmark linguistic regions**

No.	Region	Population	Target	
			N.	%
1	Northern Zealand	-	720	18
2	Southern Zealand and Lolland Falster	-	240	6
3	Bornholm	-	240	6
4	Northern Funen	-	400	10
5	Southern Funen and islands	-	240	6
6	Northern Jutland	-	400	10
7	Western Jutland	-	320	8
8	Eastern Jutland	-	560	14
9	Southern and central Jutland	-	400	10
10	Western part of southern Jutland	-	240	6
11	Eastern part of southern Jutland	-	240	6
Total		-	4000	100

The specified proportions are equivalent to those applied in the SpeechDat(M) project. Although the threshold of 6% possibly could have been lowered within the SpeechDat(II) project when recording 4000 speakers, it was decided to maintain this in order to ensure representative samples for all regions.

### 3.3 Finland (Finnish and Finnish-Swedish)

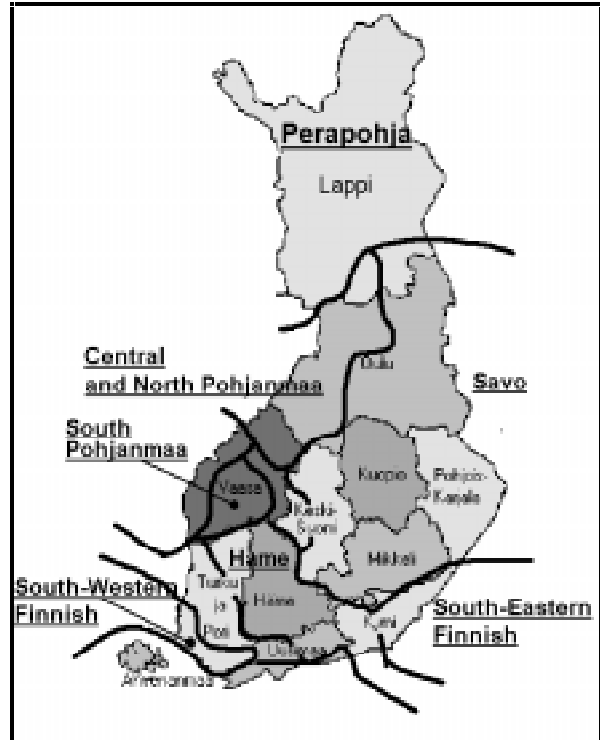
Finland (5.068.000 people in 1993) is split into 12 provinces (coloured in the map) and there are two official languages: Finnish and Finnish-Swedish. This country will be addressed by DMI collecting a 4000 speakers Finnish database, and a 1000 speakers Finnish-Swedish database.

The main Finnish dialects [4] (underlined in the figure) are in the following table.

<b>Finnish dialects</b>
South-western Finnish
Hame (Tavast)
South Pohjanmaa (Etelä-Pohja)
Central and North Pohjanmaa (Pohjois- ja Keski-Pohja)
Perapohja
Savo (Savolax)
South-eastern Finnish (Finnish Karjala, Finnish Karelian)

For the regions of call will be used the provinces of Finland as reported in the following table.

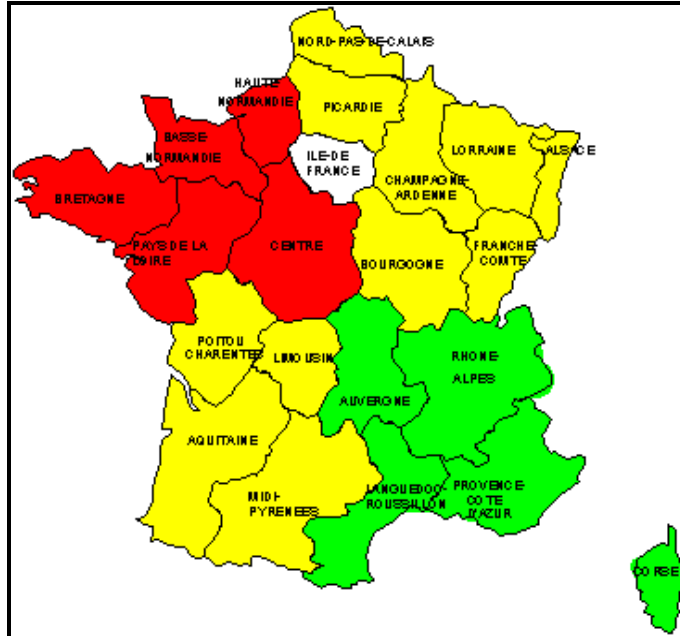
No.	Region
1	Uusimaa
2	Turku and Pori
3	Hame
4	Kymi
5	Mikkeli
6	Pohjois-Karjala
7	Kuopio
8	Keski-Suomi
9	Vaasa
10	Oulu
11	Lappi



Finnish-Swedish is spoken in Ahvenanmaa and in the coastal area of Finland. Additional information are in the Sweden section.

### 3.4 France (French)

In France, there are 22 administrative regions and in each region we can identify a different accent. Otherwise, in some regions, there is an important population and in other ones there are less people. For this reason, it's certainly difficult to have exactly the same number of speakers in each region. However, all the accents have to be represented so we can't use a "proportional criteria" to decide how many speakers will be recorded in each region.

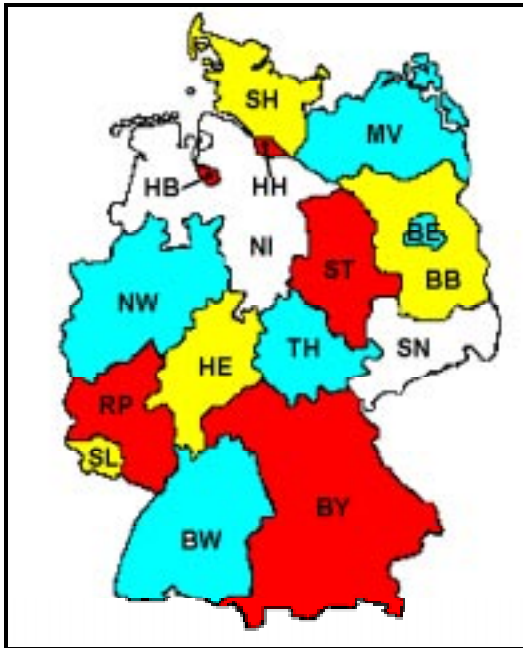


Matra has decided to divide France in five main part: l'Ile-de-France, the north-east, the north-west, the south-east and the south-west (these are the five regions chosen by France Telecom for the new French dialling set-up in October 96). In each "big" region, the objective is to record 1000 speakers with a minimum of 500 and a maximum of 1500. Of course, in each "big" region, we will distribute the speakers over the 22 French administrative regions.

A 5000 speakers speech database will be collected by MATRA. Philips will make available to all SpeechDat partners its 1000 speakers France speech database collected in the framework of the previous SpeechDat(M) project.

No.	Region	Description	Population
1	Ile de France	Ile-de-France.	10.660.000
2	North-East	Nord-Pas-de-Calais, Picardie, Champagne-Ardenne, Lorraine, Alsace, Bourgogne, Franche-Comte.	13.774.000
3	North-West	Haute-Normandie, Basse-Normandie, Bretagne, Centre, Pays de la Loire.	11.363.000
4	South-East	Auvergne, Rhone-Alpes, Languedoc-Roussillon, Provence-Cote d'Azur, Corse.	13.303.000
5	South-West	Poitou-Charente, Limousin, Aquitaine, Midi-Pyrenees.	7.551.000
Total			56.651.000

### 3.5 Germany (German)



Germany is split in 16 administrative regions (federal states) and had a total population of 80.400.000 people in 1993.

Siemens will collect a 4000 speaker speech database and the linguistic regions chosen are simply the federal states of Germany.

Code	Region	Population	Target	
			N.	%
BW	Baden-Württemberg	10.000.000	488	12.2
BY	Bayern	11.600.000	577	14.4
BE	Berlin	3.500.000	174	4.4
BB	Brandenburg	2.600.000	129	3.2
HB	Bremen	600.000	40	1.0
HH	Hamburg	1.600.000	80	2.0
HE	Hessen	5.900.000	294	7.3
MV	Mecklenburg-Vorpommern	1.900.000	95	2.4
NI	Niedersachsen	7.500.000	373	9.3
NW	Nordrhein-Westfalen	17.700.000	880	22.0
RP	Rheinland-Pfalz	3.900.000	194	4.9
SL	Saarland	1.100.000	55	1.4
SN	Sachsen	4.600.000	229	5.7
ST	Sachsen-Anhalt	2.800.000	139	3.5
SH	Schleswig-Holstein	2.600.000	129	3.2
TH	Thüringen	2.500.000	124	3.1
Total		80.400.000	4000	100

### 3.6 Greece (Greek)

#### 3.6.1 Historical Background

Studies on the development of the Greek language maintain [5] [6] that the ancient Greek dialects, by the end of the Hellenistic period, had been displaced by the commonly used language Kioni, which was based on ancient Greek. Through the Kioni traces only of the local ancient dialects have survived until the present time. Hence, strictly speaking, the term “dialect” can apply to very few cases, one of which may be the Cretan. For all other local forms of Greek covered in this project (Vlachs and Tsakonians are not included), the term “idiom” is preferred.

In 1922, in all of continental Greece, and all the islands close to it, the Aegean islands, the greatest part of Asia Minor and in Constantinople, the Greeks spoke a number of idioms which shared many characteristics. These idioms and especially those of the Peloponnese constitute the basis of the contemporary common oral Greek. The final classification of the various idioms is to a large extent still an open question.

Taking into consideration the vast urbanisation of recent years, the geographical limits of any remaining dialectical differences become even more problematic. Especially so in the present study where it is not possible to characterise as separate areas those with very small population.



### 3.6.2 Definition of Areas According to Pronunciation Variations

The definition of the areas is concerned only with major pronunciation variations, not with vocabulary or syntax, and the distinction into four large areas is loosely made in that, in one large area smaller areas with differences in pronunciation are incorporated. These include for instance areas that are isolated geographically, as e.g. Mani in southern Peloponnese.

In the case of the Ionian Islands, otherwise known as the Heptanese, the original plan was to place them in a separate area, because as a whole they are characterised by a vocabulary with many words of Italian origins and a melodious intonation of a southern type. However, their pronunciations is not different today from Standard Modern Greek (Kontosopoulos p. 68), except for the articulation of /I/, particularly in Corfu, and the intonation is not the same in all of these islands (e.g. differences between Zante, Corfu, Lefkas). In addition, heptanesian pronunciation is evidenced in the Peloponnese, as e.g. in a large community in Patras. For these reasons and because they constitute a small part of the Greek population they are placed along with the Peloponnese and those areas where standard Greek is spoken.

### 3.6.3 Characteristics of each idiom/dialect

#### 1. Standard (Urban) Modern Greek.

It is the language spoken and written by most Greeks, in both formal and informal discourse.

#### 2. Northern and Semi-Northern Modern Greek.

The most striking characteristic of the northern idiom has to do with the pronunciation of the unstressed vowels, where unstressed /I/ and /u/ disappear, /e/ and /o/ become /I/ and /u/ respectively, and /a/ remains unchanged. For instance, /cer'etisa/ becomes /cir'etsa/. This also results to the formation of consonant clusters and consonant word endings that are not allowed in standard Greek: /'anèropi/ becomes /'anèrup/, .../m'iti/ becomes /m'it/-/peä'eftika/ becomes /piä'eftika/.

#### 3. Cretan Modern Greek

It is spoken in Crete and in other places where the Cretans of the Diaspora retain elements of their local pronunciation throughout their lives. The most characteristic element of the Cretan pronunciation is that the consonants /k,y,x,g/ become rough before /e/ and /I/. Other typical characteristics are, the change of /t/ to /è/ before the semivowel /I/, the dropping of /n/ before /è/ and frequently of /s/ before /m/. These are also important morphological and lexical differences from the standard language.

#### 4. Aegean Modern Greek

The Aegean is not homogeneous in terms of dialects. The Dodecanesian, the Cycladic, the Cycladocretan, the Northern and Seminothen Greek, coexist in part in various places.

In the Dodecanesian pronunciation, the loose articulation of consonants is a dominant characteristic. This makes their speech difficult for the average Greek to understand. The phenomenon of "tsitakismos" (i.e. the pronunciation of /ce, ci/ becomes /tse, tsi/ occurs in many islands of this area. In Chios, they add or retain a final /n/ to nouns and verbs that the other Greeks do not. In the Cyclades one meets almost all Modern Greek idioms. Here, too, tsitakismos is common.

### 6.3.5 Database collected

Knowledge will collect a 5000 speakers speech database according to the following table.

No.	Region	Description	Population	Target	
				N.	%
1	Standard	Athens (Attica), South Euboea, Thessalonike, the Peloponnese, Kythera, Ionian islands	7.750.000	4050	81
2	Northern and Semi-Northern	From the northern shore of the Corinthian Gulf up to the northern Greek frontiers, Lefkas, Sterea Hellas, Hepeiros, Thessay, Macedonia, Thrace, North Sporades, Thasos, Lemnos, Invros, Lesbos, Samos, Tinos	1.000.000	500	10
3	Cretan	Crete	600.000	300	6
4	Aegean	The Dodecanese, the Cyclades, south Sporades, Chios	300.000	150	3
Total			9.650.000	5000	100

### 3.7 Italy (Italian)



Italy with its global population of 57.250.000 inhabitants in 1992 is a republic geographically split into 20 administrative regions. CSELT will collect a 3000 speakers database in addition to the 1000 speakers collected within the SpeechDat(M) project.

CSELT decided to regroup the Italian administrative regions into five<sup>4</sup> macro-regions based on some linguistic criteria [7] [8]:

a) vowels distribution:

- regions in the North have a 5 vowels system with /E/ - /e/ and /O/ - /o/ collapsing in two intermediate sounds;
- In the South vowels are slightly diphthongized in internal stressed syllables: e.g.
  - faro = /"f { : e r o/ (Puglia)
  - l'una = /"l U: u n 6/ (Puglia)
  - l'una = /"5 (w) u: n a/ (Sicilia)<sup>5</sup>
- People from the Centre use a system based on seven vowels, that distinguish open and close timbers, like
  - péscā (= fishing) /"p e s k a/ vs. pèsca (= peach) /"p E s k a/
  - bótte (= barrel) /"b o t: e/ vs. bòtte (= hits) /"b O t: e/

b) consonant distribution:

- people in the North regions pronounce inter-vocalic /s/ as voiced /z/, while in the Centre and South the sound is unvoiced.

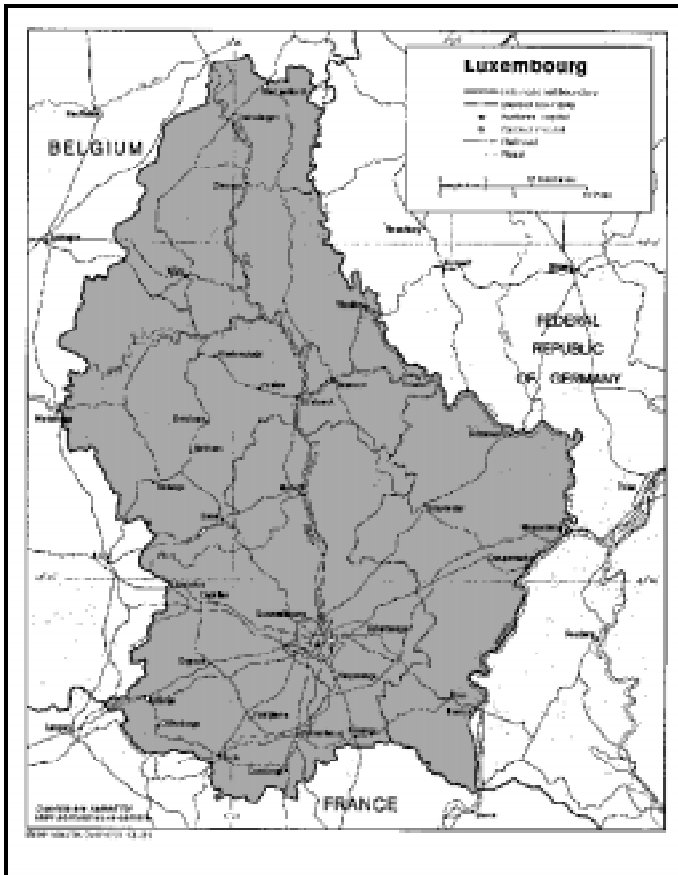
The included figure shows all dialect regions defined, mapped in the correspondent administrative regions. People will be selected proportionally to the real populations of the regions as reported in the following table:

<sup>4</sup> In the SpeechDat(M) project we used a four-regions subdivision, but the current one appears to be more correct.

<sup>5</sup> Reported in extended SAMPA as described in <http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm>

No.	Region	Description	Population	Target	
				N.	%
1	North	Piemonte, Val d'Aosta, Liguria, Lombardia, Trentino-Alto Adige, Friuli-Venezia Giulia, Veneto, Emilia-Romagna	25.600.000	1350	45.0
2	Sardinia	Sardegna	1.650.000	100	3.3
3	Centre	Toscana, Umbria, Marche, Abruzzo, Molise, Lazio,	12.500.000	650	21.6
4	Center-South	Basilicata, Campania, Calabria	8.400.000	450	15.0
5	South	Puglia, Sicilia	9.100.000	450	15.0
Total			57.250.000	3000	100.0

### 3.8 Luxembourg (Luxembourgish French and Luxemburgish German)



The Grand Duchy of Luxembourg, a constitutional monarchy, is an independent sovereign state, tucked between Belgium, France and Germany. It is 51 miles long and 36 miles wide, encompassing an area of 999 square miles with a population of 400,900 inhabitants (in 1994).

The country is divided into two clearly defined regions:

- The “Eisléck” or “Oesling” in the north, which is part of the Ardennes, on the western rim of the Eifel, and covers one-third of the territory. It is a wooded country of great scenic beauty. Highest point: 1823 feet.
- The “Good country” in the centre and the south, covering the remaining 68 % of the territory, is

mainly rolling farmland and woods. Average height: 900 feet. Culminating point 1400 feet. It is bordered in the east by the wine-producing valley of the Moselle, and in the extreme south west by a narrow strip of red earth which forms the Luxembourg iron- ore basin.

Of the country's 400,900 inhabitants, some 85,000 live in Luxembourg- city and its immediate surroundings. The average population density is 145 people per square kilometre. The number of foreign residents in Luxembourg has already exceeded 32 % of the population and is slightly higher than 50 % in the capital. It is the highest proportion of foreigners of any EU country. An intelligent policy of integration has thus far helped to avoid any friction between the various communities. The Belgians, Germans, and French constitute the largest groups among these, closely followed by the Portuguese and the Italians.

“Lëtzebuergesch” is the everyday spoken language of the people, and the main symbol of the Luxembourgers' national identity. Since the creation of a dictionary, and a grammar in the 1950s, this former West- Moselle-Frankish dialect is now recognised as the national language (since 1984), while both French and German remain administrative languages. “Lëtzebuergesch” or Luxembourgish is taught in schools and in language courses mostly addressed to the resident foreigners. Although of Germanic origin (the earliest written attestation is in glosses of the 11th century from the monastery of Echternach), Luxembourgish has sufficiently differentiated itself from its parent language, so as no longer to be understood by most Germans. Indeed many French words have been adopted into the language and were transformed, sometimes beyond recognition.

Both German and French culture meet in Luxembourg. Franco-German bilingualism, without predetermined preference for either language, is a typical aspect of the country's social

structure. If both German and French are used in the press, in political and in religious life, French is nevertheless the official language of the administration, jurisdiction, parliament, education, and of some literary circles. Public offices though are held to answer wherever possible in the language they are addressed in.

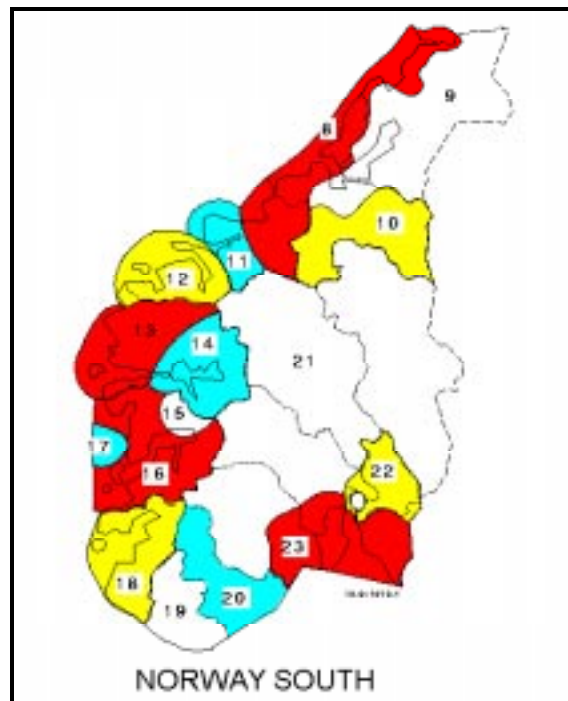
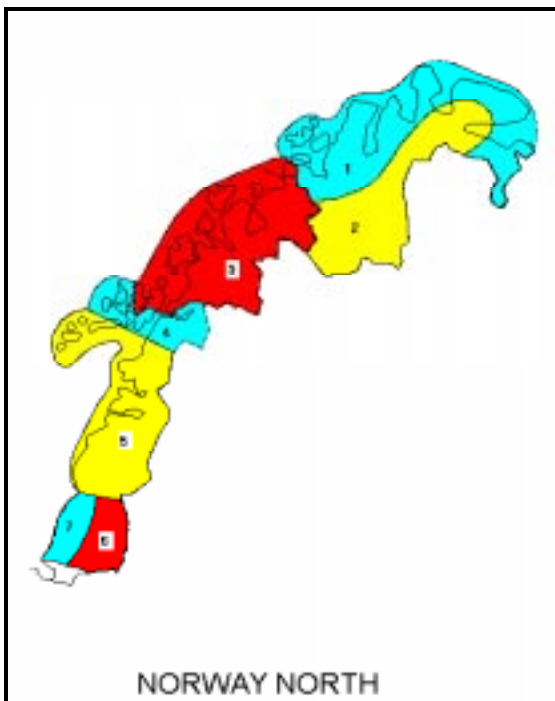
This peculiar language situation is a direct result of the size of the country, and its historic associations with both France and Germany. When going abroad which after all, is not very far the Luxembourgers have to speak other languages, simply because their own is not understood elsewhere. Thus it comes as no surprise that many Luxembourgers speak English too. This is obviously more often the case in the capital and in other centres, than in rural areas, where there is hardly a need for more than two foreign languages.

L&H will collect two databases of 500 speakers each one: the first one in “Français luxembourgeois”, i.e. the typical dialect spoken in Luxembourg. The second one in “Deutsch”, that is the second typical dialect spoken in Luxembourg.

### 3.9 Norway (Norwegian)

Norway has a population of 4.370.000 (01.01.96). It is a constitutional monarchy and it is split in 19 counties. The official language is Norwegian, with two written forms, “bokmål” and “nynorsk”. Telenor will collect a 1000 speakers database with manuscript sheets and annotation in both forms. It is expected that 15% of the speakers will select “hynorsk”.

Speakers will be recruited from 23 dialect regions. The regions are designed by a linguist in order to cover the main phonetic variation in Norwegian dialects. A target distribution is specified to be approximately proportional to the population in each region. A minimum number of speakers are however allocated for the smallest regions in order to obtain adequate coverage of all dialect-specific allophones. The number of speakers from each dialect region in the final database should not be less than 20% below the target. Region 24 is included to cover callers with foreign accents.



## North regions

No.	Region	Description (municipalities)	Population	Target	
				N.	%
1	Finnmark nord	From Finnmark county: Vardø, Vadsø, Hammerfest, Alta, Hasvik, Kvalsund, Måsøy, Nordkapp, Porsanger, Lebesby, Gamvik, Berlevåg, Båtsfjord, Sør-Varanger.	64.737	20	2
2	Finnmark sør	From Finnmark county: Kautokeino, Loppa, Karasjok, Tana, Nesseby. From Troms county: Lyngen, Storfjord, Kåfjord, Skjervøy, Nordreisa, Kvænangen.	29.282	20	2
3	Troms	Troms county except Lyngen, Storfjord, Kåfjord, Skjervøy, Nordreisa, Kvænangen.	133.602	40	4
4	Narvik-området	From Nordland county: Narvik, Tysfjord, Lødingen, Tjeldsund, Evenes, Ballangen, Bø, Øksnes, Sortland, Andøy.	53.138	20	2
5	Bodø-området	From Nordland county: Bodø, Gildeskål, Beiarn, Saltdal, Fauske, Skjerstad, Sørfold, Steigen, Hamarøy, Røst, Værøy, Flakstad, Vestvågøy, Vågan, Hadsel, Moskenes.	101.003	30	3
6	Mo i Rana-området	From Nordland county: Dønna, Nesna, Hemnes, Rana, Lurøy, Træna, Rødøy, Meløy.	44.848	20	2
7	Brønnøysund-området	From Nordland county: Sømna, Brønnøy, Vega, Vevelstad, Herøy, Alstadhaug, Leirfjord, Vefsn, Grane, Hattfjelldal.	40.136	20	2
Total			466.746	170	17

## Middle regions

No.	Region	Description (municipalities)	Population	Target	
				N.	%
8	Ytre Trønderlag	From Nordland county: Bindal. From North Trøndelag county: Namsos, Fosnes, Flatanger, Vikna, Nærøy, Leka. From South Trøndelag county: Hemne, Snillfjord, Hitra, Frøya, Ørland, Agdenes, Rissa, Bjugn, Åfjord, Roan, Osen. From Møre og Romsdal county: Kristansund, Averøy, Frei, Gjernes, Tingvoll, Sunndal, Surnadal, Rindal, Aure, Halså, Tustna, Smøla.	122.193	30	3
9	Indre Trønderlag	From North Trøndelag county: Steinkjer, Meråker, Stjørdal, Frosta, Leksvik, Levanger, Verdal, Mosvik, Verran, Namdalseid, Inderøy, Snåsa, Lierne, Røyrvik, Namsskogan, Grong, Høylandet, Overhalla. From South Trøndelag county: Trondheim, Orkdal, Melhus, Skaun, Klæbu, Malvik.	290.664	70	7
10	Søndre Trønderlag	From South Trøndelag county: Oppdal, Rennebu, Meldal, Røros, Holtålen, Midtre Gauldal, Selbu, Tydal.	31.845	20	2
11	Molde-området	Molde, Rauma, Nesset, Midsund, Aukra, Fræna, Eide.	51.303	20	2
12	Ålesund-området	From Sogn og Fjordane county: Selje. From Møre og Romsdal county: Ålesund, Vanylven, Sande, Herøy, Ulstein, Hareid, Volda, Ørsta, Ørskog, Norddal, Stranda, Stordal, Sykkylven, Skodje, Sula, Giske, Haram, Vestnes, Sandøy.	134.730	30	3
Total			630.735	170	17

### West regions

No.	Region	Description (municipalities)	Population	Target	
				N.	%
13	Ytre Sogn og Fjordane	Sogn og Fjordane county except Selje, Sogndal, Balestrand, Leikanger, Aurland, Vik, Lærdal, Årdal, Luster.	76.020	20	2
14	Indre Sogn og Fjordane	From Sogn og Fjordane county: Sogndal, Balestrand, Leikanger, Aurland, Vik, Lærdal, Årdal, Luster.	28.558	20	2
15	Voss	From Hordaland county: Voss, Granvik, Ulvik	16.168	20	2
16	Hordaland	Hordaland except Bergen, Askøy, Fjell, Sund, Os, Voss, Granvin, Ulvik.	131.792	30	3
17	Bergens-området	From Hordaland county: Bergen, Askøy, Fjell, Sund, Os.	277.287	70	7
Total			529.825	160	16

### South-West regions

No.	Region	Description (municipalities)	Population	Target	
				N.	%
18	Rogaland	Rogaland county.	357.027	90	9
19	Vest-Agder	Vest-Agder county.	150.426	40	4
20	Aust-Agder	Aust-Agder county.	100.211	30	3
Total			607.664	160	16

**South-East regions**

No.	Region	Description (municipalities)	Population	Target	
				N.	%
21	Indre østlandet	Oppland and Hedmark counties. Telemark county except Porsgrunn, Skien, Bamble, Kragerø, Siljan, Nome. Buskerud county except Kongsberg, Øvre Eiker, Nedre Eiker, Drammen, Lier, Røyken, Hurum.	495.583	100	10
22	Oslo-området	Oslo and Akershus county. From Østfold county: Rømskog.	929.246	140	14
23	Øst- og Vestfold	Østfold county except Rømskog. Vestfold county. From Telemark county: Porsgrunn, Skien, Bamble, Kragerø, Siljan, Nome. From Buskerud county: Kongsberg, Øvre Eiker, Nedre Eiker, Drammen, Lier, Røyken, Hurum	710.158	100	10
24	Foreign background	Those that do not match any of the regions above.	-	-	-
Total			2.134.987	340	34
Grand Total			4/370.000	1000	100

### 3.10 Portugal (Portuguese)



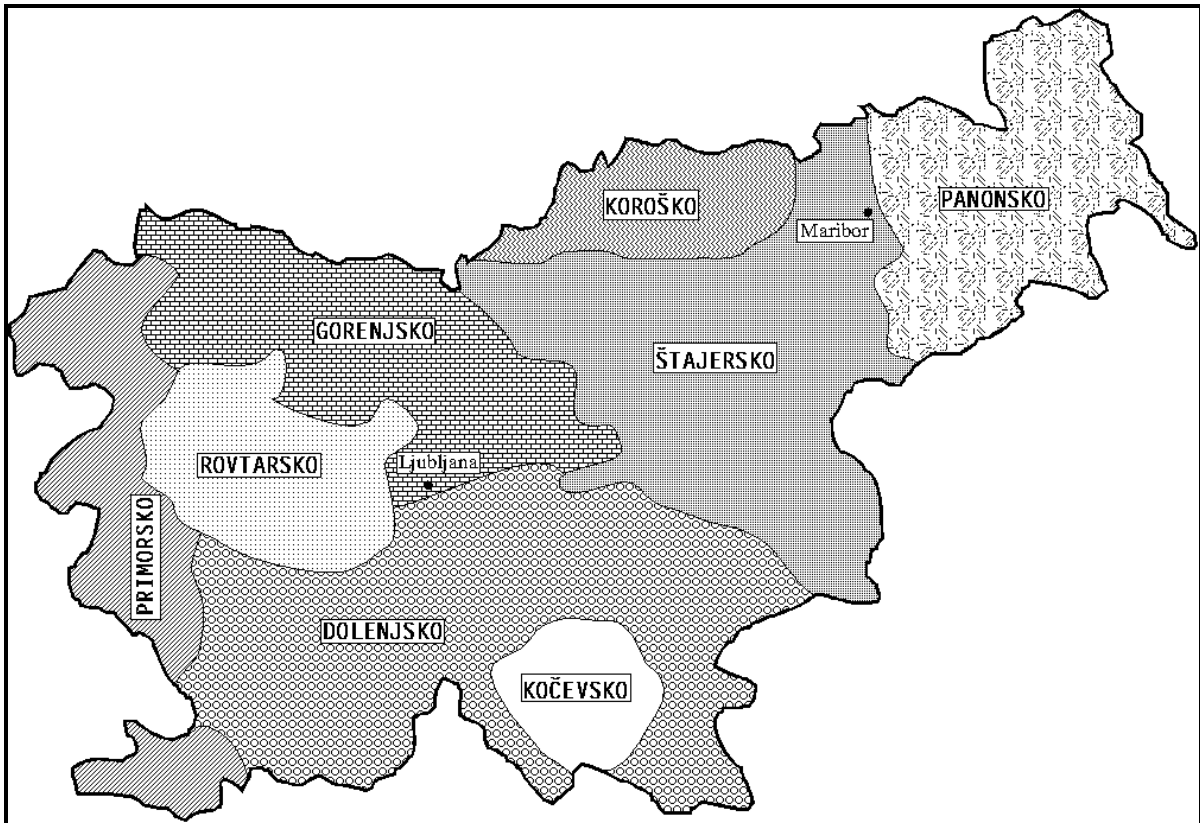
Portugal is a republic of 9.919.000 people (1994) and the country is split into 20 districts. The official language is Portuguese.

PT will collect a 4000 speaker speech database. The approach adopted for speaker recruitment involves selecting speakers among the employees of Portugal Telecom (about 20,000). The company has a wide geographical coverage, thus guaranteeing a good representation of many regional accents.

Regions 6 and 7 are out of the reported map.

No.	Region	Description	Population	Target	
				N.	%
1	North	Minho, Douro-Litoral, Trás-on-Montes	3.531.000	1440	36.0
2	Centre	Beira-Leitoral, Beira-Alta, Beira-Baixa	1.711.000	680	17.0
3	Tagus Valley	Estremadura, Ribantejo	3.310.000	1320	33.0
4	Alentejo	Alto-Alentejo, Baixo-Alentejo	524.000	200	5.0
5	Algarve	Algarve	345.000	160	4.0
6	Açores	Azores islands	241.000	80	2.0
7	Madeira	Madeira islands	257.000	120	3.0
Total			9.919.000	4000	100

### 3.11 Slovenia (Slovenian)



Slovenia is a republic (1.965.986 people in 1993) and is split in 7 regions. The official language is Slovenian. There are more than 40 dialects, however 10 main dialects were selected for database recording in the framework of the SpeechDat II project. On a geographical map 8 dialect regions are denoted, whereas two dialects are the dialects as spoken in the cities Maribor and Ljubljana. The number of calls for each dialectical region is proportional to the population.

The speaker recruitment will involve speakers among the employees of Slovenian Poste. It has a good geographical coverage, thus guaranteeing a good regional coverage. The recorded database will be balanced according to the project specifications with the recruitment of employees and scholars at primary and secondary schools in Slovenia.

University of Maribor will collect a 1000 speakers database.

The table below indicates the dialect regions (identified with a number), main cities in the region, population of the region, number of callers for 1000 speakers database and number of callers in case of necessary oversampling, as well as the percentage of the population in the region.

No.	Region	Description	Population	Target	
				N.	%
1	Panonsko	Murska Sobota, Ptuj, Lendava, Ljutomer	233.486	119	11.9
2	Stajersko	Celje, Krsko, Slovenska Bistrica, Velenje	440.931	224	22.4
3	Korosko	Dravograd, Ravne na Koroskem, Slovenj Gradec	73.789	38	3.8
4	Dolenjsko	Grosuplje, Postojna, Ajdovscina, Novo Mesto	259.330	132	13.2
5	Kocevsko	Kocevje	18.523	10	1.0
6	Rovtarsko	Idrija, Skofja Loka, Vrhnika, Logatec	84.733	43	4.3
7	Gorenjsko	Domzale, Jesenice, Kranj, Radovljica	226.336	115	11.5
8	Primorsko	Izola, Koper, Nova Gorica, Tolmin	156.030	79	7.9
9	Maribor		151.221	77	7.7
10	Ljubljana		321.607	163	16.3
Total			1.965.986	1000	100



No.	Region	Description	Population	Target	
				N.	%
1	North West	Galicia, Asturias	3.100.000	400	10
2	North	País Vasco, Navarra	2.170.000	280	7
3	Central	Aragon, Cantabria, Castilla_La_Mancha, Castilla_León, La_Rioja, Madrid, Extremadura (North)	9.300.000	1200	30
4	South	Andalucía, Canarias, Extremadura (South), Murcia	7.750.000	1000	25
5	East	Cataluña, Valencia, Baleares	8.680.000	1120	28
Total			31.000.000	4000	100

The phonetic dialectal variations can be summarised as follows [9] [10]:

1) North West:

- Vowels [e,o] in word final position can be pronounced as [i, u]
- Voiced alveolar nasal [n] in word final position is pronounced as voiced velar nasal [ŋ]
- Occasionally, [t] is pronounced as [s]
- Some consonant groups are reduced (akto: ato, rrepuxnante: rrepunante, taGsi: tasi, eGsakto: esato).

2) North:

- The main difference is the intonation.
- Occasionally [t] is pronounced as [s].

3) Central:

- Correspond to the standard Spanish or Castiliano. However in Madrid, usually the phoneme [d] in word final position is pronounced as [t]

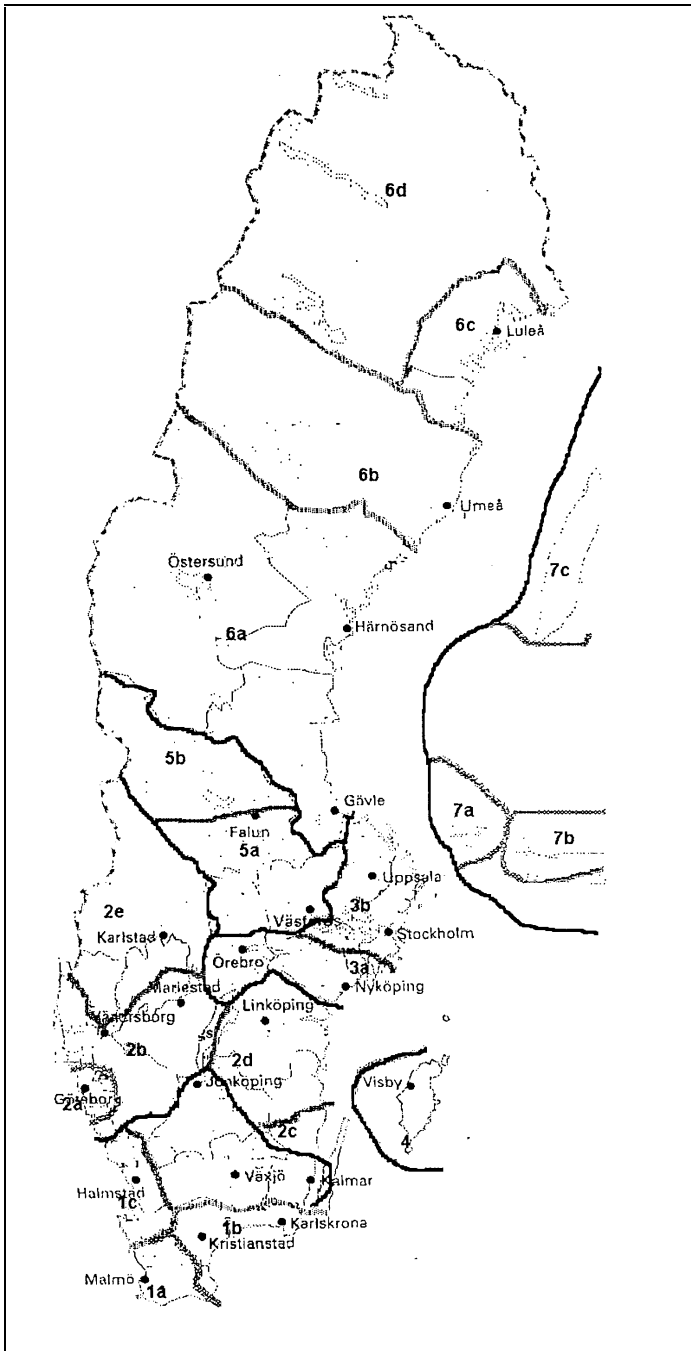
4) South:

- Phoneme [t] is pronounced as [s] in Sevilla, Córdoba, Huelva y Málaga.
- Phoneme [s] is pronounced as [t] in wide areas of Jaén, Cádiz, Granada y Almería, otherwise [t] is pronounced as [s] as in the former paragraph.
- Phoneme [s] in word final position is aspirated or lost.
- Phonemes [s, z] in syllable final position are aspirated, assimilated or lost (mizmo: mi(h)mo, mimmo, mimo).
- A voiced consonant after phoneme [s] becomes voiceless (desBjar: defjar).
- Phoneme [x] voiceless velar fricative is aspirated (muxer: mu(h)er).
- Commonly, there is no distinction between [j] and [L].
- Voiced alveolar nasal [n] in word final position is pronounced as voiced velar nasal [ŋ] or disappears.
- Voiced alveolar lateral [l] and voiced alveolar tap [r] in syllable final position usually are lost.
- It is common to pronounce phonemes [l] in syllable final position before a plosive as [r].
- It is also common a neutralization of phonemes [l] and [r] (alta: arta, olor: olol)
- Phoneme [r] followed by [l] can be assimilated (perla: pella)
- Phoneme [tʃ] voiceless palatal affricate, can lose the affricate property.

5) East:

- Phoneme [d] in word final position is pronounced as [t]
- Phoneme [s] between vowels is pronounced as [z]
- Occasionally [T] is pronounced as [s]
- Clear distinction between [j] and [L]

### 3.13 Sweden (Swedish)



The population of Sweden was 8.837.496 people at the end of 1995. Sweden is a constitutional monarchy and is divided into 24 counties. Swedish is the only official language. KTH will collect a 5000 speaker database in Swedish as spoken in Sweden. DMI in Finland will collect a 1000 speaker Finnish Swedish database. In order to facilitate a comparison between “standard” Swedish and Finnish Swedish, some information about the latter is included below.

The population counts reported in the tables below are based on the national registration of people living in Sweden in 1995.

The geographical division of Sweden into the dialectal areas shown on the map below has been performed by Professor Emeritus Claes-Christian Elert, a Swedish expert in this field. Since the spoken language changes gradually when moving between different geographical areas it is impossible to define definite borders between dialects, and one should rather speak of central areas of a dialect than of exact borders between them.

The dialectal division of Sweden reflects an effort to divide the spoken standard Swedish language into different geographical variants. It does not regard genuine dialects, confined to rather small and specific areas, but rather the spoken language used by most people in most situations in the areas defined. This principle results in seven major Swedish dialect regions, six of which cover Sweden and one Finland. The regions can be further divided into smaller areas, where the dialect in each one differs less from the dialects within its own region than from the dialects of the others. Still the difference is considered large enough to motivate the subdivision, resulting in a total of 18 dialect areas in Sweden and 3 in Finland. KTH will record speakers from the 18 areas within Sweden, while the Finnish Swedish dialects will be recorded in Finland.

Below we give short characteristics of the 7 major Swedish dialect regions, but first we give an overview of the 18 Swedish vowels:

### Swedish vowels

Orthography	i	e	ä	y	u	ö	o	å	a
SAMPA, short vowel	I	e	E	Y	u0	2	U	O	a
SAMPA, long vowel	i:	e:	E:	y:	}:	2:	u:	o:	A:

### Swedish dialect areas

#### 1. *South Swedish*

South Swedish diphthongization (raising of the tongue, late beset rounding of the long vowels), retracted pronunciation of r, no supradentals, retracted pronunciation of the fricative sje-sound. A tense, creaky voice quality can be found in large parts of Småland.

#### 2. *Gothenburg, west, and middle Swedish*

Open long and short ä and (sometimes) ö vowels (no extra opening before r), retracted pronunciation of the fricative sje-sound, open å-sound, thick l-sound.

#### 3. *East, middle Swedish*

Diphthongization into e/ä in long vowels (possibly with a laryngeal gesture), short e and ä collapses into a single vowel, open variants of ä and ö before r (i, i: and 9, 9: in SAMPA).

#### 4. *Swedish as spoken on Gotland*

Secondary Gotland diphthongization, long o-vowel pronounced as å.

#### 5. *Swedish as spoken in Bergslagen*

u pronounced as central vowel, acute accent in many connected words.

#### 6. *Swedish as spoken in Norrland*

No diphthongization of long vowels, some parts have a short u pronounced with a retracted pronunciation, thick l-sound, sometimes the main emphasis of connected words is moved to the right.

#### 7. *Swedish as spoken in Finland*

Special pronunciation of u-vowels and long a, special sje and tje-sounds, r is pronounced before dentals, no gravis accent.

The tables below contains the geographical regions associated with the different Swedish dialect areas and also the population in the respective areas as well as the number of speakers to be recorded within SpeechDat (II). The targeted number of speakers is directly proportional to the relative number of speakers in each area.

### South Swedish

No.	Region	Description (major cities in the area)	Population	Target	
				N.	%
1a	Southwest Skåne	Malmö,Lund	678.203	384	7.6
1b	Northeast Skåne & Blekinge	Kristianstad, Karlskrona	322.619	191	3.8
1c	Northwest Skåne & south Halland	Helsingborg, Halmstad	457.008	259	5.2
1d	(South) East Småland	Växjö, Jönköping, Kalmar	586.063	344	6.9
Total			2.081.471	1.178	23.5

### Gothenburg, west, and middle Swedish

No.	Region	Description (major cities in the area)	Population	Target	
				N.	%
2a	Gothenburg area	Gothenburg, Mölndal, Kungsbacka	707.627	400	8.0
2b	Västergötland	Skövde, Skara, Mariestad, Vänersborg, Borås	804.643	455	9.1
2c	East Småland	Västervik, Oskarshamn, Borgholm	132.131	75	1.5
2d	Östergötland	Linköping, Norrköping	412.167	233	4.6
2e	Värmland, Dalsland	Karlstad, Karlskoga	332.066	188	3.7
Total			2.388.634	1.351	26.9

### East, middle Swedish

No.	Region	Description (major cities in the area)	Population	Target	
				N.	%
3a	South Södermanland	Nyköping, Eskilstuna	411.498	233	4.6
3b	Stockholm-Uppsala area	Stockholm, Uppsala	2.042.900	1.156	23.0
Total			2.454.398	1.389	27.6

### Swedish as spoken on Gotland

No.	Region	Description (major cities in the area)	Population	Target	
				N.	%
4	Gotland	Visby	58.120	50	1.0
Total			58.120	50	1.0

### Swedish as spoken in Bergslagen

No.	Region	Description (major cities in the area)	Population	Target	
				N.	%
5a	Bergslagen	Västerås, Falun	571.319	323	6.4
5b	Upper Dalarna		74.688	50	1.0
Total			646.007	373	7.4

### Swedish as spoken in Norrland

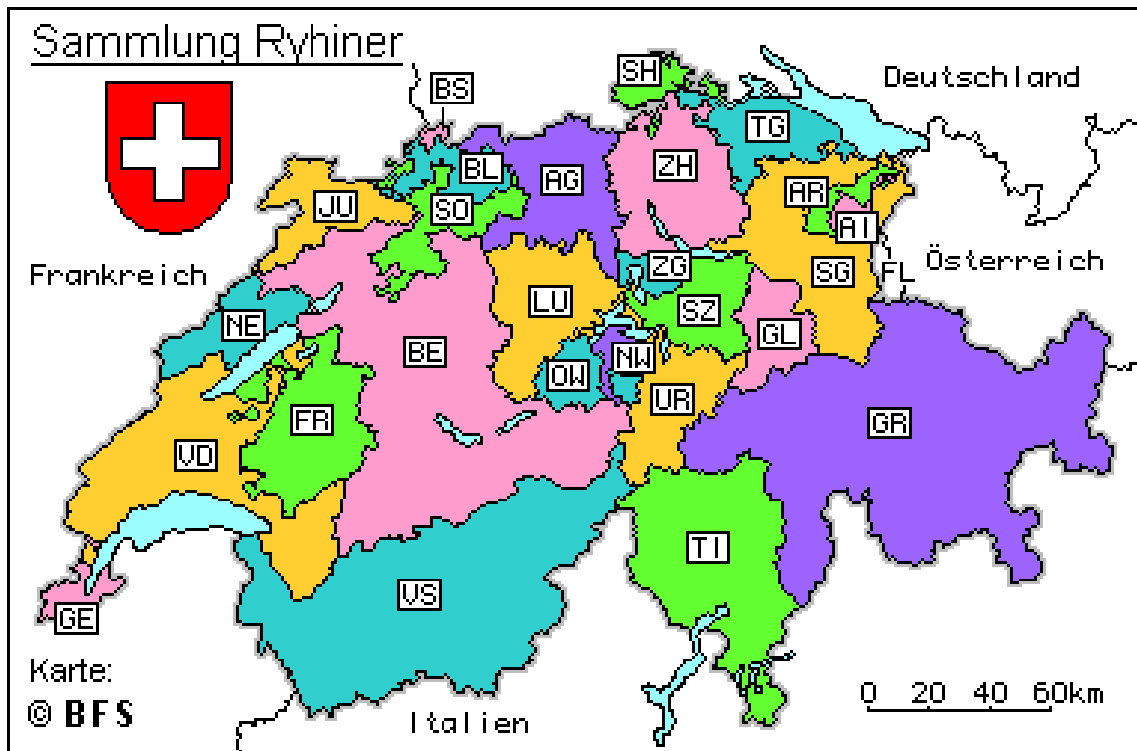
No.	Region	Description (major cities in the area)	Population	Target	
				N.	%
6a	Middle norrland & Jämtland	Gävle, Bollnäs, Sundsvall, Härnösand, Östersund	682.383	386	7.7
6b	Västerbotten (the province, i.e. incl. South Lappland)	Umeå	260.472	147	2.9
6c	Norrbotten (the county, i.e. coastal area)	Skellefteå, Piteå, Luleå	175.105	99	2.0
6d	North Lappland	Malmberget, Kiruna	90.906	51	1.0
Total			1.208.866	683	13.6

Grand Total	8.837.496	5024	100
-------------	-----------	------	-----

### Swedish as spoken in Finland

No.	Region	Description (major cities in the area)	Population	Target	
				N.	%
7a	Åland	Mariehamn	-	-	-
7b	Åboland- Nyland	Åbo, Helsingfors	-	-	-
7c	Österbotten	Vasa, Kristinestad	-	-	-
Total			-	1000	100

### 3.14 Switzerland (Swiss French and Swiss German)



The Switzerland with a population of 6.670.000 people in 1993 is a federal republic split in 26 cantons. There are four official languages: German, French, Italian (spoken by about 300.000 people) and Romanche (spoken by about 50.000 - 100.000 people).

Concerning the Swiss French and the Swiss German Polyphone, the different regions correspond to the different districts from Switzerland. So the number of calls for each district is proportional to the population.

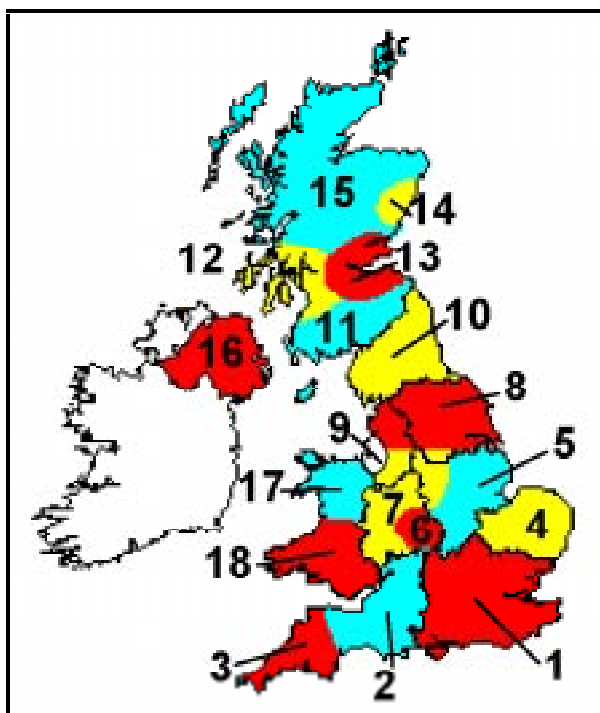
For the Swiss French Polyphone, we have:

No.	Region	Description	Population	Target	
				N.	%
1	Vaud		596.700	740	37.0
2	Genève		387.600	480	24.0
3	Fribourg		218.700	272	13.6
4	Valais		177.800	220	11.0
5	Neuchâtel		163.900	204	10.2
6	Jura		68.600	84	4.2
Total			1.613.300	2000	100.0

For the Swiss German Polyphone, we have:

No.	Region	Description	Population	Target	
				N.	%
1	Zurich	Zurich	1.162.100	230	23.0
2	Basel	Basel, Solothurn	684.200	135	13.5
3	Bern	Berner Seeland, Mitteland, Freiburg	627.400	124	12.4
4	Aarau	Aargau	518.900	103	10.3
5	St. Gallen	St. Gallen, Appenzell	505.800	100	10.0
6	Luzern	Luzern, Zug	424.000	84	8.4
7	Schwytz	Uri, Schwytz, Unterwald, Glaris, Oberwallis	348.400	69	6.9
8	Berner Oberland	Berner Oberland	313.700	62	6.2
9	Schaffhausen	Schaffhausen, Thurgau	290.700	57	5.7
10	Graubunden	Graubunden	182.000	36	3.6
Total			5.057.200	1000	100.0

### 3.15 United Kingdom (English and Welsh)



This country (58.099.000 people in 1993) is composed of four main regions (England, Wales, Scotland and Northern Ireland) split in several smaller units. Units are known as counties in England and Wales and as districts in Northern Ireland. Scotland has nine regions and three island areas. The Isle of Man is a British crown dependency.

The official language is the English but there are several dialectical influences.

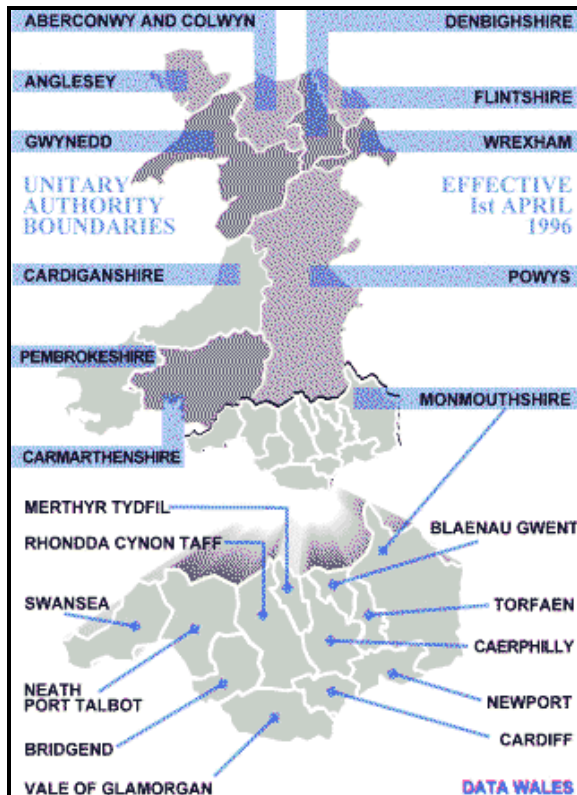
There are two separate FDB collections: one for English language - Marconi-Hirst is collecting 4000 speakers from England, Scotland, Wales and Northern Ireland; and one for Welsh - BT is recording 2000 Welsh speakers, all resident in Wales.

#### English regions

The following regional accents have been identified for British English. Each region has been identified with a number and the main cities and areas it covers.

No.	Region Description
1	London and South East: covering London, Sussex East.
2	South West: Hampshire, Wilts and Gloucestershire.
3	Devon and Cornwall.
4	Suffolk and Norfolk
5	Northampton, Oxford, Cambridge and Lincolnshire
6	Birmingham, Worcester, Coventry and Lichfield
7	Stafford, Derby, Nottingham, Hereford, Crewe, Stoke-on-Trent, Chesterfield
8	Manchester, Sheffield, Huddersfield, Humberside, York, Lancashire
9	Merseyside: Liverpool, Wigan, Southport, Preston
10	Darlington, Cleveland, Stockton, Cumbria, Tyne and Wear, Northumberland
11	Dumfries and Galloway, Borders
12	Glasgow, Dumbarton, Paisley, Greenock
13	Edinburgh, Dunfermline, Kinross, Fife, Dundee
14	Aberdeen, Bachory, Stonehaven
15	Highland, Stirling, Dufftown, Elgin, Brechin
16	Northern Ireland
17	Gwynedd, Clwyd
18	Dyfed, Powys, Swansea, Cardiff, Newport

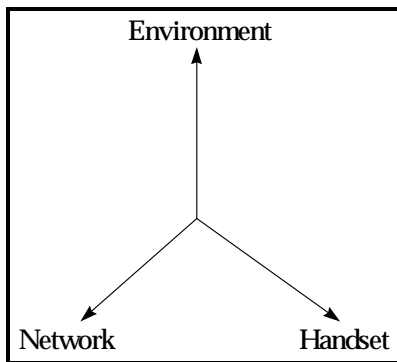
## Welsh regions



There are between half a million and one million people in Wales who speak the Welsh language. They are all bilingual and also speak English. Welsh is a Celtic language, whereas English is a Germanic language.

Code	Region	Description	Population	Target	
				N.	%
SWW	South West Wales	Blaenau Gwent, Monmouthshire, Tor-faen, Newport, Mid and South Powys, Ceredigion, Carmarthenshire, Pembrokeshire, Swansea	-	820	41
SEW	South East Wales	Neath and Port Talbot, Bridgend, Vale of Glamorgan, Cardiff, Caerphilly, Rhondda Cynon Taf, Merthyr Tydfil	-	320	16
NWW	North West Wales	Caernarfon and Meirionydd, Aberconwy and Colwyn, Anglesey	-	580	29
NEW	North East Wales	Denbighshire, Flintshire, Wrexham, North Powys (Welshpool Area)	-	280	14
Total			-	2000	100

## 4. General environmental specific characteristics



**Figure 3 - Environmental characterisation**

In speech corpus collections over the Public Switched Telephone Network a number of specific issues must be considered but in the following we will address only the ones involved in the Fixed Networks:

- calling environment,
- handset and
- telephone network.

Generally speaking these three factors can be represented in a three-dimensional space, as shown in fig. 3. They will be closely examined in the next sections<sup>6</sup>. Most of the following has been extracted from [1] that has been extended and updated to cover the new SpeechDat requirements. Network interface and signal processing are addressed in [2].

For each call received these three factors will be checked and stored in the appropriate files by using some defined code words [3]. The “domain” of each factor, i.e. the set of possible values can be used, can be increased by adding new values to the currently proposed lists, but the new ones should be agreed between all partners in advance in order to avoid an explosion of values with the same meaning but with different coding. The actual complete list will be reported in the final documentation for each database also if not all values have been used. The most recent set will be stored on the SpeechDat WWW server; it will be regularly updated when new values will be agreed.

Some values can be furthermore specialised by using a few country-specific “*modifiers*”, by adding a short string after a leading slash “/”, e.g. “GSM/TIM and GSM/OMNITEL” for Italian MDB. The most recent set of modifiers and their values will be stored on SpeechDat’s WWW server too.

### 4.1 Calling environment

Scientists with a speech technology background stress the importance of making recordings under conditions which are as close to the application conditions as possible. The main advantage of real-life recordings is that they contain the effects on the speaker’s behaviour due to the environmental conditions; for example, it is well known that talkers raise their voice level in noisy conditions (“Lombard effect”). Other conditions may have different effects on the talker’s behaviour, but there is little or no documentation.

In general, to be really useful calling environments should be reported by selection an appropriate keyword from a defined list. The SpeechDat agreed list is:

---

<sup>6</sup> In the following tables, the shaded rows are reported only for completeness, as they are involved only in the Mobile Database Speech collection.

<b>Environment</b>	<b>Description (used to describe ...)</b>
BOOTH	calls made from a telephone booth
PAYPHONE	calls made from a payphone
HOME	calls originated at home; they can be affected by background homely noises such as TV, children, ...
OFFICE	calls made from office, where usual background noises are background talking, telephone rings, typewriting noises, ...
FACTORY	calls coming from very noisy environments, where the background noise is generated by machines.
PUBLIC	calls generated by using telephones located in public places, with high background talking, such as stations, airports, bars, ...
STREET	calls generated by using telephones from sites with high background traffic emission noise
VEHICLE	calls generated by using cellular phones from inside moving vehicles
OTHER	calls coming from places not included in the previous classification

⇒ For SpeechDat purposes it has been decided to record at least 2 percent of calls from public places. Calling environment have to be reported in the database description files, and the figures obtained will be checked in the validation stage. When a simple Yes/No question is raised to the caller, such as “Are you calling from a public place?”, not public calls can be reported by using the code word “NOT\_PUBLIC”.

## 4.2 Handsets

Depending on the recording protocol and the selection of speakers it may or may not be possible to obtain reliable information about the type of handset a speaker is using.

In the modern, increasingly digital, fixed PSTNs the handset is one of the most important sources of variation. Often the make and type of a handset are not clearly specified on the equipment, and even if they are, it is often difficult to find out what they actually mean, because of the very large number of different makes that are used out there. What people can reliably report about a handset is whether it is cordless or not, and whether it has a rotary or push buttons. The cordless condition is, of course highly relevant for the transmission characteristics of the local loop. Moreover, it is very likely that a cordless handset has an electret microphone. The type of microphone cannot reliably be estimated with sufficient accuracy from the difference between rotary dials and push buttons, also if it appears that rotary dials telephones have carbon buttons microphones while push-button telephones have electret microphones. In any case, even when the type of microphone is known, there remains still a lot to be guessed when it comes to the actual transfer function of the handset.

In conclusion: reliable knowledge about the transfer function of the handsets used by the speakers in a corpus can only be obtained if the speakers are visited in person by an experimenter who is present during the recordings or if the speakers can be made to use handsets provided by the corpus collectors. This, of course, is only possible under some very restricted and expensive recording protocols.

⇒ For the SpeechDat corpus collection it is recommended to include different types of handsets in the corpus design and, if it is possible, to report them in the database description files. In any case this issue will not be checked during the validation stage.

The following table lists the possible keywords that can be used to report handset information.

<b>Handset</b>	<b>Description</b>
ROTARY	Rotary telephones that usually generate pulses
TOUCHTONE	Push buttons telephones which usually generate DTMF
CORDLESS	Cordless telephones
OTHER	Calls made with telephones not included in the previous classification

## 4.2 Network

Presently, in Europe there are three classes of networks operational, i.e., the fixed (wired) network, the analogue (TACS) cellular network and the digital (GSM) network. Recently the DECT standard has entered the market as new emerging system.

In the fixed network a distinction must be made between the local loop (i.e., the connection between the handset and the first central office switch through which the signal is routed) on the one hand and the transmission channels connecting central office switches on the other.

The latter are quickly migrating to all digital bulk transport channels. Consequently, the signals are virtually immune to added noise after having been digitised in the first switch they pass, until they are converted to analogue form for transmission over the receiver's local loop (if that is at all necessary). One corollary is that there is no real difference in signal quality between local and long distance calls. In the countries where are still analogue switches and analogue trunk lines in operation, one must be prepared to find considerable differences between local and long distance calls.

The local loops, on the other hand, will very likely remain analogue for the majority of the subscribers well into the next century. There will, however, be an increasing share of digital local loops, especially for business applications. In these cases the analogue part of the transmission is limited to the microphone and the AD-converter in the handset. All ISDN phones in use in Europe should produce standard A-law signals. It is assumed that essentially all digital handsets connected to PABXs also use A-law codecs, even if the signalling between handset and PABX does not conform to the ISDN protocols.

In every telephone call there are always two local loops, namely at the sender and at the receiver side.

More importantly, the local loop at the side of the recording station should not be analogue. There is now a large number of recording workstation on offer that connect directly to ISDN lines. Some of them are reported in the SpeechDat deliverable SD2.1

⇒ While no restrictions are defined for the local loop at the caller side, it is mandatory for SpeechDat partners to connect the database collection machines to all-digital local loops. As consequence of this choice in the whole SpeechDat databases the signals should be stored in 8 bit, 8 KHz A-law format.

The following keywords can be used to report in the databases description files information about the network used. The shaded keyword are useful for the mobile databases.

<b>Network</b>	<b>Description</b>
PABX	Private telephone network
FIXED	Public switched telephone network
TACS	Analogue cellular telephone network
GSM	Digital cellular telephone network
DECT	Digital Enhanced Cordless Telecommunication network

Recently new digital mobile networks using the DECT standard are being employed. DECT, Digital Enhanced Cordless Telecommunication, is an access method using a common air interface. It has a micro cellular structure (outdoors 250 m; indoors 20-50 m), using both Frequency (10 bands) and Time (24 time slots) Domain Multiplexing. This results in a high traffic capacity, needed in office environments, and dens populated areas. Besides all kinds of data communication, also telephony is supported for home environment, office and wireless local loop. The speech is coded with 32 Kbit ADPCM. The handset automatically selects the proper frequency and time frame from the strongest radio base station. Seamless handover is ensured, and all features of a fixed office line can be supported. Because the speech quality of a DECT system is equivalent to a digital fixed line office system, and few systems are operational in this moment, no special effort is taken to collect mobile telephone speech using DECT.

DECT systems for office and home are already on the market. Also mobile handsets using alternatively DECT and GSM are announced and one should be already on the market.

## **Bibliography**

- [1] R. Winski et al, "Recording platforms for telephone speech corpus collection", SpeechDat(M) doc. ref. LRE-63314-D3.1.2.1, 6 March 1996
- [2] F. Senia et al, "Installation of the recording device and documentation", SpeechDat doc. ref. LE2-4001-SD2.1
- [3] F. Senia, "Specification of speech database interchange format", SpeechDat doc. ref. LE2-4001-SD1.3.1, 23 January 1997

## **Finland**

- [4] Johdatus Suomen murteisiin, Martti Rapola, Suomalaisen Kirjallisuuden seura.

## **Greek**

- [5] Robert Browning: The Greek Language, Medieval and Modern. Ed. Papademas, Athens, 1988
- [6] N.G. Kontosopoulos: Dialects and Idioms of Modern Greek. Athens 1994

## **Italy**

- [7] Zarko Muljacic, "Fonologia della lingua italiana", Società editrice in Mulino, Bologna
- [8] L. Canepari, "Introduzione alla fonetica", Einaudi, Torino, 1896

## **Spain**

- [9] A. Zamora Vicente, "Dialectología española", Madrid, ed. Gredos; 1972
- [10] García Mouton, "Lenguas y Dialectos de España", Cuadernos de Lengua Española, Ed. Arco/libros, s.l., 1994