A Multilingual Corpus for Language Identification*

L.F. Lamel, G. Adda, M. Adda-Decker, C. Corredor-Ardoy†, J.J. Gangolf, J.L. Gauvain
Spoken Language Processing Group
LIMSI-CNRS, BP 133,91403 Orsay cedex, FRANCE
{lamel,gadda,madda,gangolf,gauvain}@limsi.fr
http://www.limsi.fr/TLP

ABSTRACT

In this paper we describe the design, recording, and transcription of a large, multilingual (French, English, German and Spanish) corpus of telephone speech for research in automatic language identification. The corpus contains over 250 calls from native speakers of each language from their home country, and an additional 50 calls per language from another country. Although the same recording protocol was used for all languages, slight modifications were necessary to account for language or country specificities. Issues in designing comparable corpora in different languages are addressed, including how to interact with callers so as to obtain the desired responses.

INTRODUCTION

We have designed and recorded a large, four-language corpus (French, English, German and Spanish) of telephone speech for research in automatic language identification (Corredor-Ardoy et al., 1997). The corpus is similar in style to the OGI multi-language corpus (Muthusamy, Cole & Oshika, 1992). The multilingual corpus contains speech from 250 native speakers of each language calling the LIMSI data collection system from their home country via a toll-free number, as well as 50 native speakers of each language calling from within France (or from Germany, Spain or the United Kingdom for native French speakers). The callers were recruited by a marketing survey company, who was responsible for balancing the subjects for sex, age, and dialect. The marketing survey company contacted the callers, gathered the necessary subject information and distributed calling scripts.

The scripts contained three types of data: general questions concerning the call and caller (code, sex, age, location, etc.); a series of items containing pre-defined texts to read (phonetically rich sentences, dates, times, spoken and spelled names) and fixed prompts ("what time is it now?"); and a set of questions aimed at obtaining spontaneous speech. The scripts were slightly modified to fit each language and country. For example, in France the first 4 digits of the phone number are used to locate the calling region, while in Germany callers were explicitly asked to specify the region.

Over 250 calls in each of the four languages were recorded and orthographically transcribed by native speakers of that language. Common protocols were used to carry out the transcriptions with regard to marking of spontaneous speech effects such as hesitations, word fragments and laughter, and non-speech events. We have found that it is essential that the transcribers are native speakers of the language, who have recently lived in the country, thus having up-to-date linguistic and pragmatic knowledge of the country and culture. The transcribers also participated in the definition of the scripts and questions assuring their naturalness.

CORPUS DESIGN AND ACQUISITION

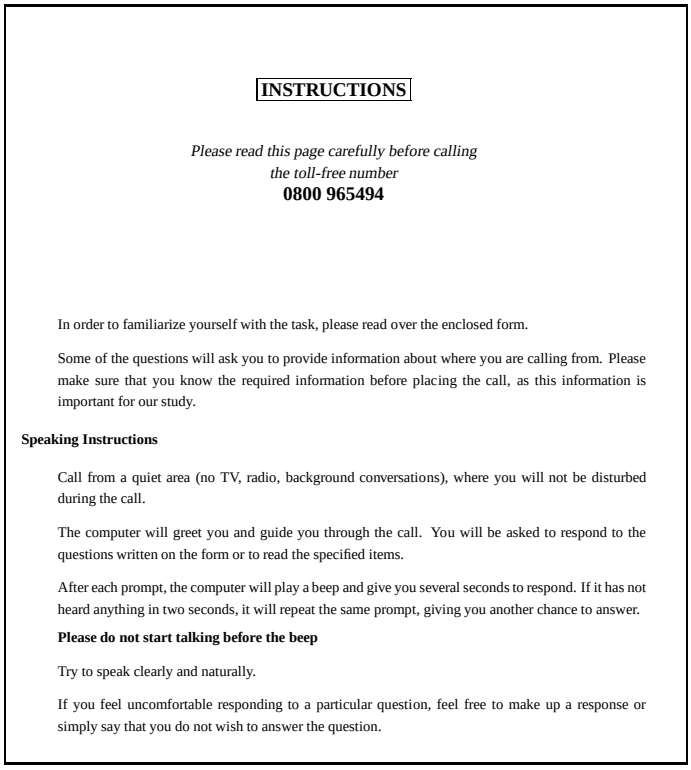
The multilingual corpus was designed to enable the development and testing of algorithms for automatic language identification. The corpus contains over 300 calls for each for 4 languages: British English, French, German, and Spanish. 250 of these are *matched calls* completed by native speakers of the language calling the LIMSI data collection system from their own country: that is native British English speakers calling from the United Kingdom, French speakers calling from France, etc. An additional 50 calls per language were made under "crossed" conditions: native French callers from the U.K., Germany and Spain; native British English, German and Spanish speakers calling the system from within France. The crossed calls were included for testing purposes to ensure that the language and not the telephone channel are being identified.

Subject Recruitment

The callers were recruited by a marketing survey company, who ensured a balance for sex and age (4 groups between 18 and 65 years of age) of the speakers. In order to represent different regional accents, subjects were recruited from 10 subareas in each country. Each participant received (via the marketing survey company) a set of general calling instructions and a script corresponding to the call. Each

* This work was partially financed by a CNET CTI project.

† Corredor-Ardoy was with the LIMSI-CNRS when this work was carried out. He is now working at Bouygues Telecom, Vélizy, France.



INSTRUCTIONS

Please read this page carefully before calling
the toll-free number
0800 965494

In order to familiarize yourself with the task, please read over the enclosed form.

Some of the questions will ask you to provide information about where you are calling from. Please make sure that you know the required information before placing the call, as this information is important for our study.

Speaking Instructions

Call from a quiet area (no TV, radio, background conversations), where you will not be disturbed during the call.

The computer will greet you and guide you through the call. You will be asked to respond to the questions written on the form or to read the specified items.

After each prompt, the computer will play a beep and give you several seconds to respond. If it has not heard anything in two seconds, it will repeat the same prompt, giving you another chance to answer.

Please do not start talking before the beep

Try to speak clearly and naturally.

If you feel uncomfortable responding to a particular question, feel free to make up a response or simply say that you do not wish to answer the question.

Figure 1: Caller instructions (English version).

script was unique and identified by a code, so as to associate caller information supplied by the marketing company with the call, and to provide feedback when calls were completed. Any payment for participation was handled directly by the marketing company.

The cover letter sent with each script explained that the LIMSI-CNRS was carrying out a study on telephone speech, for which it is necessary to record a large number of speakers from a wide geographic area. It also explained the modality of the call, the expected duration (about 5 minutes) and assured the caller that the recordings would remain anonymous and that they would not be recontacted as a result of the call. The instructions given to the callers are shown in Figure 1.

Caller script generation

A language-independent program was written to facilitate the generation of the caller scripts. The program makes use of language specific files to specify the system prompts and to complete any required fields. For each script a unique code is generated consisting of two words from the international alphabet (*alpha, bravo, charlie, ...*) and three digits, where the third digit is a checksum. Unique file identifiers that are linked to the user code, thus keeping track of all information needed to easily verify and transcribe each call.

The program randomly selects items from the prespecified language specific files, and generates a variety of presentation formats according to usage in the given language. Each script consists of a set of fixed questions to localize the call and to characterize the call/caller. These questions are the Speaker Code, Gender, Age range (older or younger than 25), Native language, Calling city, Zipcode, Telephone area code, Date, and Time. These questions are followed by a series of prompts asking the caller to read a written text or to provide a response to a specific questions (ellicited responses). For each language source files containing several thousand texts of different styles were created. These include texts extracted from newspapers, simple telephone introductory phrases, travel information queries, dates, times, credit card numbers, telephone numbers, spoken and spelled common words, spoken and spelled proper names, digit strings, money amounts, and complete names and addresses. Different presentation formulations were randomly selected according to the language. For example, a date in English could be presented in the following forms:

Friday, May 8th, 1998
Friday, May 8th
Friday, May 8, 1998
Friday, May 8
Friday, the 8th of May, 1998
Friday, the 8th of May,
Friday, 8 May, 1998
Friday, 8 May
the 8th of May, 1998

the 8th of May
8 May 1998
May 8, 1998
Friday, 8 / 5 / 98
8 / 5 / 98
08 / 05 / 98

The above different forms can all be derived from the following specification for the date:

$$\begin{aligned} \text{Date} \quad = \quad & [\,(\,\emptyset \mid \text{Friday}\,)\,(\,\text{May 8th} \mid \text{8 May} \mid \\ & \text{May 8} \mid \text{the 8th of May}\,)\,(\,\emptyset \mid 1998\,)\,] \mid \\ & [\,(\,\text{Friday}\,)\,(\,\text{8/5/98} \mid \text{08/05/98}\,)\,] \end{aligned}$$

A final series of questions aimed at collecting spontaneous monologues, were randomly selected at record time from a set of about 200 questions and were not written on the paper script. Example scripts for the four languages are shown in Figures 2 and 3.

Data collection system

The data collection system consists of an SGI Indy workstation and a telephone interface ELAN BT8. The telephone interface is controlled by the workstation and the telephone inputs and outputs are directly connected to the audio channels of the Indy. In this way one workstation is able to simultaneously handle 4 telephone lines. Recordings were made on a numeric line with an 8kHz sampling rate (16 bits), and without automatic gain control. Although the telephone interface could handle DTMF codes, these were not used since digital telephones are not common in some of the geographic areas targeted.

Data recordings

The recordings were carried out sequentially, via toll-free numbers[1], starting with French, then English, followed by German and Spanish. In this way 3 of the 4 telephone lines were dedicated to the current language, with the fourth line available to handle any late calls from the previous langugae.

All calls on a given day (for a given language) are saved in the same subdirectory, which simplified archiving and verification of the calls.

The following different types of recording problems were handled:

- The caller hangs up unexpectedly before completing the call. After detecting the hangup, the system is reset and the call counter is incremented.
- No speech is detected. Either the caller did not respond, the caller spoke too softly, or the line/telephone quality is too poor. The system asks the caller to repeat and, if the situation continues, to speak louder. If still no speech is detected, the system asks the user to call back (preferably from a different telephone) and disconnects.
- An errorful response is detected: the caller is asked to repeat the last response. Several types of verification were employed depending upon the situation: none, speech detection, minimal duration for the speech, forced aligment of the speech with the known prompt text, speech recognition.

For each language a set of prompts were recorded by a native speaker of the language. The initial prompt informed the caller that they were connected to the data collection system and that the call was being recorded:

> *Thank you for calling the recording system at LIMSI-CNRS. Your voice will be recorded for the purposes of research and development in speech technology. This call is anonymous. If you do not wish to have your voice recorded, please hang up now. After each question, you will hear a beep. Please do not start talking until you hear the beep.*

The caller was then prompted to read the code, followed by the remainder of the items on the script.

The rate at which calls were received was quite variable, and highly dependent upon the efficiency of the market survey company. For English, French, and Spanish, it took about 4-6 weeks to obtain 250 acceptable calls with the appropriate age and gender distribution. Typically about 80% of the calls occurred over the 2nd and 3rd weeks, after which the calling rate was quite reduced. For German the call rate was much higher, and all the calls were recorded in a 2 week period.

Call verification

Calls were verified on a daily basis. This was found to be crucial for smooth functioning of the process during high rate periods. If too many calls accumulated it was hard to catch up. Unix scripts were written to simplify the verification procedure, providing a call summary and a means of listening to the data via an editor (Emacs). Verification, which took about 5 minutes per call, was used to decide if a call was acceptable. Transcription of the data was carried out only after the majority of recordings were obtained. The codes of the accepted calls were reported to the marketing company on a weekly basis.

[1] During the data collection period, there was a lapse in the agreement between France and Spain concerning international toll-free numbers. This forced us to delay the recordings from Spain. Eventually we got around the problem by having the subjects call the system "reversing the charges" (a collect call). The server simply detected speech on the part of the operator, and responded affirmatively.

Vous êtes connectés au système d'enregistrement du LIMSI-CNRS. Votre voix va être enregistrée et utilisée pour effectuer des travaux de recherche sur le traitement de la parole. Nous vous remercions pour votre participation. Si vous ne désirez pas être enregistré, vous pouvez raccrocher maintenant.

1 - Prononcez le code :

bravo zoulou 5 5 9

2 - Etes-vous un homme ou une femme ?

3 - Avez-vous plus ou moins de 25 ans ?

4 - Quelle est votre langue maternelle ?

5 - Prononcez et épelez le nom de la ville ou du village d'où vous appelez.

6 - Quel est le code postal de l'endroit d'où vous appelez ?

7 - Quels sont les 4 premiers chiffres de votre numéro de téléphone ?

8 - Quelle est la date d'aujourd'hui ?

9 - Quelle est l'heure actuelle ?

10 - Prononcez la phrase :

Il joue au héros des Jeux malgré lui.

11 - Prononcez la phrase :

Madame Delatour a renoncé au fourneau à charbon de sa jeunesse.

12 - Prononcez les jours de la semaine, en commençant par lundi.

13 - Prononcez les mois de l'année, en commençant par le mois de janvier.

14 - Pouvez-vous donner la date de naissance de quelqu'un que vous connaissez ?

15 - Prononcez la phrase :

Ils étaient aidés autrefois par de jeunes garçons, les mousses.

16 - Prononcez la phrase :

Bonjour chère madame, je voudrais vous demander un renseignement.

17 - Prononcez l'heure :

22h24

18 - Prononcez la date :

Mercredi 3/6/86

19 - Prononcez le mot :

presque

20 - Epelez le mot :

presque

21 - Prononcez la phrase :

On jurerait une affiche de rugby en division d'honneur.

22 - Prononcez le numéro de téléphone :

29 44 25 04

23 - Prononcez le numéro de carte bancaire :

4973 8346 7876 0004

24 - Prononcez le nom :

KAPPERT

25 - Epelez le nom :

KAPPERT

26 - Prononcez la phrase :

Quel est le vol le moins cher?

27 - Prononcez la suite de chiffres :

7 5 1 9 5

28 - Prononcez l'adresse :

M. A. GRUAZ
4, Route des Gardes
B.P. 72 92322 CHATILLON CEDEX France

29 - Prononcez la somme :

431F

30 - Prononcez la phrase :

Je voudrais avoir les prix.

Maintenant, nous allons vous poser quelques questions d'ordre général. Soyez assez aimable pour essayer de répondre en une ou plusieurs phrases. Si vous ne trouvez rien à répondre, vous pouvez inventer une réponse ou dire ce que vous voulez. Si vous ne voulez pas répondre, dites-le et passez à la question suivante.

Thank you for calling the recording system at LIMSI-CNRS. Your voice will be recorded for the purposes of research and development in speech technology. This call is anonymous. If you do not wish to have your voice recorded, please hang up now. After each question, you will hear a beep. Please do not start talking until you hear the beep.

1 - Read the code:

bravo whiskey 6 6 7

2 - Are you masculin or feminine?

3 - Are you older or younger than 25?

4 - What is your native language?

5 - Say and spell the name of the city or town you are calling from.

6 - What is the postal code in the city you are calling from?

7 - What is the 4 or 5 digit area code you are calling from?

8 - What is today's date?

9 - What time is it?

10 - Read the sentence:

I believe the ultimate reason a country exists is to benefit its citizens.

11 - Read the sentence:

The second point requires more extended comment.

12 - Say the days of the week, starting with Thursday.

13 - Say the months of the year, starting with June.

14 - What is the birthday of someone you know?

15 - Read the sentence:

They consider it simply a sign of our times.

16 - Read the sentence:

Hello. Yes. Good afternoon, Miss. Can you help me?

17 - Say the time:

2:51 am

18 - Say the date:

Sunday, the 6th of October 1981

19 - Say the word:

whole

20 - Spell the word:

whole

21 - Read the sentence:

Talks begin.

22 - Say the telephone number:

882 729

23 - Say the credit card number:

3749 437634 41807

24 - Say the name:

Jennifer

25 - Spell the name:

Jennifer

26 - Read the sentence:

Are there direct buses between Oxford and Cambridge?

27 - Read the sequence of digits:

4 0 8

28 - Say the name and address:

Mr H.S. YOUNG
36 New Bond Street
London W1A 4SF

29 - Read the price:

9798

30 - Read the sentence:

What is the earliest train from London to Manchester?

Now we will ask you some general questions. Please try to respond in one or a few sentences. If you do not have an immediate response, feel free to make one up. If you do not want to respond to a question, you may simply say that.

Figure 2: Example scripts in English and French

Se encuentra conectado al sistema de grabación del LIMSI-CNRS. Su voz va a ser grabada y utilizada para la investigación en el campo del tratamiento de la palabra. No comience a hablar hasta que no haya escuchado el bip sonoro. Le agradecemos de antemano su participación. Si en estos momentos no desea que su voz sea grabada, puede colgar.

1 - Pronuncie el cdigo:

alpha tango 4 9 7

2 - ¿Es usted un hombre o una mujer?

3 - ¿Tiene usted más o menos de 25 años?

4 - ¿Cual es su lengua materna?

5 - Pronuncie y deletree el nombre de la ciudad desde donde usted llama.

6 - ¿Cual es el código postal del lugar desde donde usted llama?

7 - Diga el prefijo telefónico de su provincia seguido de los dos primeros números de su teléfono.

8 - Diga la fecha de hoy.

9 - Diga la hora actual.

10 - Lea la frase:

En Mijas se celebr ayer un festival en el que hubo buena entrada.

11 - Lea la frase:

Al parecer no hubo daos en el club.

12 - Diga los días de la semana comenzando por el Jueves.

13 - Diga los meses del año comenzando por Junio.

14 - Diga la fecha de nacimiento de alguien que conozca.

15 - Lea la frase:

Se espera que la ronda de negociaciones tendra un receso.

16 - Lea la frase:

Si no le importa, quisiera que usted me informase.

17 - Lea la hora:

19.20

18 - Lea la fecha:

Martes 13 de Diciembre

19 - Lea la palabra:

gracias

20 - Deletree la palabra:

gracias

21 - Lea la frase:

El cincuenta por ciento del capital de la empresa es de accionistas espaoles.

22 - Diga el número de teléfono:

(95) 609 02 75

23 - Lea el número de tarjeta bancaria:

6011 5952 0557 8998

24 - Lea el nombre:

Campuzano

25 - Deletree el nombre:

Campuzano

26 - Lea la frase:

¿Hay plazas libres en el vuelo Sevilla-Frankfurt para el pasado-maana por la tarde?

27 - Lea la serie de números:

4 0 8

28 - Lea la dirección:

Horacio Agudo Solé
C. Ferraz 18
34282 Palencia

29 - Pronuncie la suma de dinero:

9 558 493 pesetas

30 - Lea la frase:

¿Es muy largo el viaje entre Madrid y Mnaco?

A continuación le haremos algunas preguntas de orden general. Si no sabe que contestar, responda hablando sobre cualqier tema. Si no desea contestar, dígalo y pase a la singuinte pregunta.

Vielen Dank für Ihren Anruf bei der Aufnahmezentrale von LIMSI im Zentrum für Forschung und Wissenschaft (CNRS). Zur Forschung und Entwicklung in der automatischen Sprachverarbeitung wird Ihre Stimme aufgezeichnet. Der Anruf ist anonym. Falls Sie nicht möchten, daß Ihre Stimme aufgenommen wird, legen Sie bitte jetzt auf. Nach jeder Frage ertönt ein Signal. Bevor Sie zu sprechen beginnen, warten Sie bitte dieses Signal ab. Danke.

1 - Lesen Sie bitte folgende Codenummer:

Alpha Charlie 7 2 6

2 - Geben Sie Ihr Geschlecht an.

3 - Sind Sie älter als 25 Jahre?

4 - Welche ist Ihre Muttersprache?

5 - Von welchem Ort aus rufen Sie an? Geben Sie den Namen an und buchstabieren Sie ihn.

6 - In welchem Bundesland liegt dieser Ort?

7 - Geben Sie seine Postleitzahl an.

8 - Wie lautet die telefonische Vorwahl?

9 - Bitte geben Sie das heutige Datum an.

10 - Wieviel Uhr ist es jetzt?

11 - Lesen Sie bitte den folgenden Satz:

Erste Priorität ist für uns, zu informieren und den Gemeindemitgliedern Umweltprobleme näherzubringen.

12 - Lesen Sie bitte den folgenden Satz:

Die Auswahl kann deshalb nur einige wenige Restaurants berücksichtigen.

13 - Zählen Sie bitte die sieben Wochentage auf. Beginnen Sie mit Montag.

14 - Zählen Sie bitte die 12 Monate des Jahres auf. Beginnen Sie mit Januar.

15 - Bitte geben Sie das Geburtsdatum einer Ihnen bekannten Person an.

16 - Lesen Sie bitte den folgenden Satz:

Mein Arzt empfahl dringend Bäder.

17 - Lesen Sie bitte den folgenden Satz:

Ähm, guten Tag, Frau Müller. Können Sie mir helfen , bitte?

18 - Geben Sie bitte folgende Zeit an:

10.22

19 - Geben Sie bitte folgendes Datum wieder:

Samstag, den 12. April 1967

20 - Bitte lesen Sie folgendes Wort:

Kanal

21 - Bitte buchstabieren Sie es nun:

Kanal

22 - Lesen Sie bitte den folgenden Satz:

Gestartet wurde am Uhrtürmchen in der Berger Straße.

23 - Sagen Sie bitte folgende Telefonnummer:

(0 2 71) 16 86 84

24 - Geben Sie bitte die folgende Kreditkartennummer wieder:

3749 159480 99931

25 - Bitte sagen Sie den folgenden Namen:

Buchner

26 - Bitte buchstabieren Sie ihn jetzt:

Buchner

27 - Lesen Sie bitte den folgenden Satz:

Guten Morgen, ich möchte mit dem Intercity von Prag nach Paris fahren.

28 - Lesen Sie folgende Ziffernreihe ab:

2 9 5

29 - Sagen Sie bitte folgende Namen und Adresse:

Frau Stud. rat. Rina Göpfert
Am Schwimmbad 19
D-41515 Grevenbroich

30 - Bitte lesen Sie folgenden Preis:

8428,32 DM

31 - Lesen Sie bitte den folgenden Satz:

Guten Morgen, ich möchte eine Auskunft für eine Flugverbindung von Oslo nach Dresden.

Nun stellen wir Ihnen einige allgemeine Fragen, die Sie so natürlich wie möglich beantworten sollten. Ihre Antworten können auch ohne Bezug auf die gestellten Fragen bleiben. Zweck ist lediglich, einige natürlich gesprochene Sätze aufzunehmen. Falls Sie eine Frage wirklich nicht beantworten möchten, so sagen Sie dies in Ihrer Antwort.

Figure 3: Example scripts in German and Spanish

#Calls	*Region*
1	Basse-Normandie
7	Bretagne
40	Centre
9	Champagne-Ardennes
15	Haute-Normandie
68	Ile-de-France
4	Limousin
14	Lorraine
31	Midi-Pyrenées
25	Nord-Pas-de-Calais
31	Pays de la Loire
14	Rhone-Alpes

Table 2: Call regions in France.

A set of criteria were used to determine if a call was acceptable or not. The first requirement was that the caller code be present and intelligible. Other important responses concerned the caller's native language, calling location (city, zipcode, area code). In order to ensure a minimal amount of speech data a subset of the read and ellicited items were mandatory, although some flexibility was permitted here if only one item was problematic. An additional minimal duration of 30s was required for the spontaneous speech. These criteria were provided to the marketing company before subjects were recruited. Calls were also refused if there was significant background noise not due to the telephone channel. Using these criteria about 12% of the calls were rejected.

CORPUS

The entire corpus contains over 70 hours of data, with at least 13 hours per language from the source country (matched data). Each call took about 5 minutes to complete, resulting in about 2 minutes of speech data.

Table 1 summarizes the matched data for the different languages. There are slighltly over 250 calls per language, with a male:female ratio of 50:50 for French; slighlty lower for Spanish (45:55) and with fewer males (42%) for English and German. The rejection rate varied slightly from language to language, with the 3% for German, 10% for French, 13% for English and 25% for Spanish (see Table 1. Rejections are mainly due to insufficient spontaneous data or the presence of significant extraneous background noise such as conversation, television or music, or children playing.

The complementary part of the corpus contains speech from native speakers of the four languages calling from outside of their home country. This data serves to ensure that any language identification algorithms are accurately identifying the language and not simply differences in the international telephone channels. Over 70 calls were recorded for each language. French speakers called from Germany,

#Calls	*Region*
25	Birmingham
25	Glasgow
20	Leeds
15	Liverpool
22	London
17	Manchester
15	Newcastle-upon-tyne
22	Norwich
23	Nottingham
29	Plymouth
26	Southampton
13	Wales

Table 3: Call regions in the United Kingdom.

#Calls	*Region*
44	Baden-Württemberg
36	Bayern
10	Berlin
8	Bremen
7	Hamburg
8	Hessen
20	Mecklenburg-Vorpommern
3	Niedersachsen
79	Nordrhein-Westfalen
23	Sachsen-Anhalt
7	Schleswig-Holstein
4	Thüringen

Table 4: Call regions in Germany.

#Calls	*Region*
29	Paìs Vasco
14	Murcia
44	Madrid
20	Galicia
30	Extremadura
31	Cataluña
20	Cantabria
23	Aragón
18	Andalucìa Or.
24	Andalucìa Occ.

Table 5: Call regions in Spain.

Language	*#Calls*	*%Reject*	*#Male*	*#Female*	*Total data*
English (UK)	258	13%	109	149	14.8 hours
French (France)	259	10%	129	130	13.1 hours
German (Germany)	257	3%	109	148	15.8 hours
Spanish (Spain)	253	25%	114	139	17.9 hours

Table 1: Summary of data under matched language/country conditions.

French from Germany	30
French from UK	50
French from Spain	33
English from France	72
German from France	77
Spanish from France	82

Table 6: Summary of crossed calls.

Spain and the UK, while British English, German, and Spanish speakers called from within France. A summary of the crossed calls are given in Table 6. For each language there are over 2 hours of data recorded with the crossed condition. It should be pointed out that collecting the crossed condition calls was much more difficult than collecting the matched condition calls. The data collected depends also upon the subject population, whether they are tourists visiting for a short period of time, students living abroad for a few months, or long-term residents with their private and professional life in the foreign country. The latter groups may interject words in the native language from where they are calling, pronouncing words for places and names according to local conventions.

TRANSCRIPTION

The entire corpus has been orthographically transcribed by native speakers of the languages. The basic philosophy is to represent as accurately as possible the graphemic forms that were produced by the speaker, while marking important noises produced by the speaker (breath noise, coughing, etc.) or externally (background noise such as TV, radio, traffic, conversations, etc.). The former noise sources generally occur inbetween words, whereas the latter ones are typically superimposed on the speech signal. There also can be different noise sources coming from the microphone or the telephone line.

Unix scripts were written to transcribe all the calls for a given day. These scripts kept track of whether or not the call has already been verified, or if a (partial) transcription already exists. In this case a backup version of the file is created. For new calls, the code is used to create an initial transcription file containing one line per utterance, and reformatted prompt texts for all read-speech items. This file was then presented to the user in an Emacs buffer, with control sequences allowing transcribers to play the utterances and display the signal. Transcribers found it useful to play the prompt question when transcribing the spontaneous responses to questions.

A set of common transcription conventions were used for all languages (M. Adda-Decker et al., 1995). These conventions are similar to those used by the Linguistic Data Consortium (LDC) for similar style data. The main conventions are summarized here.

- The graphemic form of a word is that found in a standard dictionary. Accented characters are represented using ISO Latin (iso-accents-mode in Emacs). This underlying form is used even when elisions are produced by the talker *goo(d) morn'* for *good morning*.

- Mispronounced words are marked by an asterisks at the start of the word: ***correct-word**. If the word is unknown, the is transcribed as ******.

- Word fragments are transcribed. If the word is known, then the missing part is included in parentheses: **tr(ee)**, **(s)treet**, **tr()**.

- Filler words are marked in the transcripts. A list of filler words and their spellings was defined for each language.
  Some example filler words for French: *ah, aye, ben, bon, euh, oh, ohlala, ouf*
  Some example filler words for German: *ähm, ach, mmh, buh, na, naja, och, puh*

- Capitalizations are used only when distinctive, that is they are used to mark proper names (and substantives in German).

- The transcriptions do not contain punctuation markers, except where these have been verbalized by the talker. In this case they are written with the marker attached to the word: **.period**, **,comma**.

- When non-ambiguous, numbers are written as a sequence of digits. A standard form is used by default, for example, in French the date *1980* is assumed to be *mille neuf cent quatre-vingts*. If the caller said *dix-neuf cent quatre-vingts*, the transcription is written as: *19 100 80*.
  Telephone numbers or credit card numbers are also read in different manners: *4045 0800 2324 4170* can be read as a sequence of digits, numbers, or a combination of them: "four zero four five zero eight hundred...",

"forty forty-five zero eight double zero ...". The written form corresponds unambiguously to what the caller said.

- The spelled words and names are transcribed by a sequence of letter, for example, **L O N D O N**. Certain words can be spelled in alternative ways such as **B A L L** or **B A double L** for the word *ball*.
- Acronyms are written as a single word with all capital letters, independent of whether they are pronounced as a word *LIMSI, NATO* or are spelled *IBM, TGV*.
- Extraneous noises are marked with square brackets [b], using a set of predefined codes. If the noise covers several words, the start is marked with [b-] and then end with [-b]. Some of the more common markers are for noises produced by the talker: breath noises (inspiration, expiration), sniffling, throat clearing, cough, whispering, laughter; as well as environmental noises: background conversation, TV, radio, paper rustling; or system noises: microphone, telephone beep, telephone hang-up.

A large lexicon is used to help locate typographical errors, as these appear as out-of-vocabulary words. After creating pronunciation lexicons, the transcriptions were verified by carrying out a forced alignment of the text with the speech signal.

DISCUSSION AND CONCLUSIONS

In this paper we have described the design, recording and transcription of a large, multilingual corpus for automatic language identificiation. Our experience is that it is important to involve native speakers of language in all steps of the corpus constitution: design of scripts, prompts, call verification, and transcription.

A marketing survey company was used to recruit callers according to specified geographic, age and gender distributions. In spite of this it was still difficult to obtain an equal balance of male and female callers (typically it is easier to get female callers). This was particularly true for the crossed language/country calls. The market survey companies took different approaches to locating subjects. Some recruited in one geographical region at a time (in Spain), where others (in France and Germany) covered all areas in parallel. With the first approach it is easier to track the number of callers per region, thus obtaining a more even distribution, however the data collection process is longer.

Concerning the caller behavoir, language or country related differences were observed. For example, the British callers were relatively reluctant to speak spontaneously, and tended to give short responses to most questions. In contrast, the Spanish callers where quite outspoken particularly, when asked to describe their families or their homes. This is reflected by the total amount of data for Spanish (17.9 hours) compared to 14.8 hours for British.

The most effective questions for the British speakers, ie .those which inspired the callers the most, concerned dream vacations and music. The shortest responses concerned shopping and participation in sports. For the French callers, the longest responses were given when asked to describe their dream house and where they prefer to shop (by mail, big supermarkets or small shops). German speakers gave the longest responses to questions about public transportation and how they spend their evenings after work. They did not respond to more personal questions about preferences or desires.

The total corpus contains calls from over 320 native speakers of each language. 250 of these calls were obtained under matched conditions, ie. native speakers calling the system from their home country, and 70 calls with crossed language/country of origin. This corpus has been used to carry out experiments in automatic language identification (Corredor-Ardoy et al., 1997) and phone recognition in multiple languages (Corredor-Ardoy et al., 1998).

ACKNOWLEDGEMENTS

Other contributors in the corpus constitution and transcription were F. Connerade, S. Foukia, M. Neumann, C. Ulrich, and H. Visser.

REFERENCES

"Identification Automatique de la Langue à travers le réseau téléphonique," Contract report CNET no. 94 1B 089, nos. 1-7.

M. Adda-Decker, J.L. Gauvain, G. Adda, L. Lamel (1995), "Identification Automatique de la Langue à travers le réseau téléphonique: Conventions de transcriptions d'enregistrements téléphoniques," Internal contract working document, CNET no 94 1B 089, October, 1995.

C. Corredor-Ardoy, L. Lamel, M. Adda-Decker, J.L. Gauvain (1998) "Multilingual Phone Recognition of Spontaneous Telephone Speech", to appear in the *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-98*, Seattle, Washington, May 1998.

Y.K. Muthusamy, R.A. Cole, B.T. Oshika (1992),"The OGI Multi- Language Telephone Speech Corpus," *Proceedings of the International Conference on Spoken Language Processing, ICSLP-92*, Banff, Canada, October 1992, pp. 895-898.

C. Corredor-Ardoy, J.L. Gauvain, M. Adda-Decker, L. Lamel (1997), "Language Identification with Language-Independent Acoustic Models," *Proceedings of the European Conference on Speech Technology, ESCA EuroSpeech'97*, Rhodes, Greece, September 1998, pp. 55-58.