The NESPOLE! VoIP Dialogue Database

Susanne Burger[1], Laurent Besacier[2], Paolo Coletti[3], Florian Metze[1], Céline Morel[1],

[1]Interactive Systems Laboratories, Carnegie Mellon University, Pittsburgh, USA
and University of Karlsruhe, Germany
[2]Laboratoire CLIPS, Equipe GEOD, Université Joseph Fourier, Grenoble, France
[3] Istituto Trentino di Cultura - Centro per la Ricerca Scientifica e Tecnologica, Trento, Italy
sburger@cs.cmu.edu

Abstract

This paper presents the status of the NESPOLE! data collection as of end of February, 2001. A multilingual VoIP (Voice over Internet Protocol networks) database consisting of 200 dialogues in 4 languages (English, German, Italian and French) was recorded and transcribed. Dialogue speakers were connected via a H323 video-conferencing terminal. We describe the task, the technical architecture, the recording procedure and the transcription process of the NESPOLE! data collection. We provide some statistics concerning the data and, finally, we address problems that arose during the collection and annotation process.

1. Introduction

The NESPOLE![1] Project (NEgociating through SPOken Language in E-commerce) is a joint EU NSF funded project exploring future applications of automatic speech-to-speech translation systems in e-commerce and e-service sectors [1]. The languages processed in this project are Italian, German, English and French. The scenario for the first showcase of NESPOLE! involves an Italian speaking agent, located in a tourism agency in Italy (APT) and an English-, German- or Italian-speaking client at an arbitrary location using a simple terminal (PC, sound and video cards, H323 video-conferencing software).

Human language technology modules used in the speech-to-speech translation chain need a definition of the task domain vocabulary involved in the showcase scenario. For this, five different data collection scenarios in a tourism domain were developed. Speech data were collected in each partner's language in summer 2000, fall 2000 and February 2001.

Chapter two shows the technical architecture at the recording sites. The recording procedure and the transcription process are described in chapter 3 and four. In chapter five, first statistics about speaker and transcriptions were provided. Finally, some problematic aspects of the first NESPOLE! data collection were addressed in chapter six.

2. Recording Set-up

2.1. Task

Each partner

- France: Université Joseph Fourier (CLIPS)
- USA: Carnegie Mellon University (CMU)
- Italy: Istituto Trentino di Cultura-Centro per la Ricerca Scientifica e Tecnologica (ITC-irst), AETHRA S.r.l , Azienda per la Promozione Turistica del Trentino (APT),
- Germany: University of Karlsruhe (UKA)

provides a Windows PC running H323 video-conferencing software and a human caller (=client). The caller/client first reads the scenario description, and then contacts APT's IP address and establishes an H323 connection. (H323 is a standard for transmitting voice and video over IP networks, VoIP).

The APT agent and the client communicate in the client's native language. The agent records the client's audio signal via H323 connection and the agent's signal via microphone. The client's site records vice versa (agent via H323, client via microphone). An audio card and recording software is used. During the conversation, the caller pretends to be the tourist described in the scenario, and asks for the information specified in the scenario. Each conversation was expected to last from five to fifteen minutes, depending on the caller and scenario.

2.2. Technical Set-up

After some experimental recordings in the summer of 2000, a final technical set-up was agreed upon. Table 1 contains details about the hardware and software that were used for the actual recordings.

Hardware:	PC Pentium 200 and up
Software:	Windows NT or Win 98 Total Recorder NetMeeting3.01
Microphone:	Headset or close microphone
Environment:	Quiet office

Table 1: Recording equipment.

Every recording site was equipped with a PC running Windows NT or Windows 98. The search for appropriate recording software resulted in the decision for a product called "Total Recorder". The TotalRecorder software is a recording program capable of capturing the audio directed to the soundcard. The set-up was selected to create a stereo file containing the client and agent's audio tracks. For the net connection, Netmeeting3.01™ was chosen.

Speech recognition systems benefit from parallel recording qualities. Therefore, it was decided having additional recordings of better quality from each site itself. Each recording site recorded the partner via H323 protocol and the own location via microphone. Thus, for example, CMU records the Italian agent via the H323 connection and

[1] see http://nespole.itc.it/

the English client via a headset (see figure 1). APT would record the same dialogue with the client on an H323 connection, and the agent via a headset. The results are two parallel recordings of the same dialogue; the Italian recording contains a H323 quality signal file for the client and a microphone quality signal file of the agent and vice versa in case of the English recording.

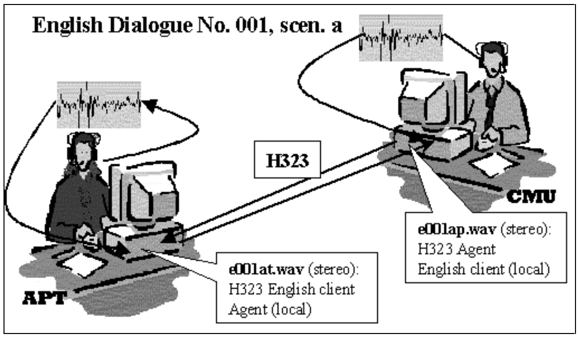


***Figure 1**:* Example for the recording scenario: APT (agent's site, Italy) records the English client via H323 connection and the Italian agent via headset (file name e001at.wav: e = English, a= scenario a, t=Trento), CMU (client's site, USA) records the Italian agent via H323 connection and the English client via headset (file name: e = English, a = scenario a, p = Pittsburgh)

2.3. Scenarios

Five scenarios were developed:

- Scenario a: Winter accommodation in Val-di-Fiemme
- Scenario b: All included tourist package
- Scenario c: Summer vacation in a park
- Scenario d: Castle and lake tours
- Scenario e: Looking for folklore and brochures

The main region was a region in northern Italy called Trentino. For every vacation package, tourist information regarding the location and available activities were prepared as handouts. Additionally, web pages containing links to professional tourist information centers and on-line forms for subject instructions were designed.

3. Recording Procedure

Subjects playing the role of clients were required to be fluent speakers of the language concerned.

Before the recording session, client subjects were given the description of at least two of the NESPOLE! scenarios and were asked to prepare for the experiment. This included information about topics, behavior (they had to act as a pre-informed tourist), etc. Additionally, each client was given access to the collection of web pages to obtain detailed background information about the scenario. Each subject was also asked to go through the on-line form, allowing her/him to learn more about her/his role, 'define' her/his family (marital status, how many children, of which age, etc.), destination of travel, special preferences, etc. This was to be taken by each subject to the session along with information about the part of Trentino (Italy) she/he was supposed to visit.

At the recording session, each participant was asked to sign a consent form, and provide information about factors possibly affecting her/his spoken language, e.g. parents' origin, education, etc., and/or affecting his/her voice, e.g. recent diseases or smoking habits.

Each participant sat in front of the computer, wearing a headset. The operator pressed the record button on "Total Recorder", and when the client felt ready, she/he pressed the call button on the Netmeeting window. To ensure synchronicity, "TotalRecorder" started to record only when receiving the pickup signal from Italy. After 10 min, the participant was urged to finish the call.

The Italian agents were professional agents working at Trentino tourist office APT.

4. Transcription

Transcriptions were produced for all "clean" microphone recordings. H323 recordings proved to be hard to work with even for human transcribers, because their quality is, in places, very poor. For seven English recordings, transcriptions of H323-quality audio-data were produced and aligned versus the corresponding "clean" transcriptions in the same style as speech recognition error rates are produced. Even though different noises were not counted as errors, the "error rate" was 3%, i.e. although the recordings were made of the same utterance, the transcriptions differ in three of one hundred words.

The NESPOLE! consortium did not establish a common set of transcription conventions. Each site chose their own kind of labeling method and set. The differences affect, in particular, the labeling of additional phenomena, such as breathing or filled pauses, turn segmentation and orthographical treatment of contractions.

4.1. French Transcriptions

The French partner delivered two kinds of transcriptions:

- A basic text transcription at the orthographical level without punctuation or additional labeling. Numbers are in digits
- A more detailed transcription of dialogue acts and spontaneous events

The turns correspond to speaker contributions and start with "A:" for agent and "C:" for client. No marker files or time stamps were segmented for the French transcriptions.

4.2. Italian Transcriptions

The transcriptions of the Italian dialogues are originally in XML format; they can be automatically translated into a readable text form.

The Italian site used "Transcriber"[1] as transcription tool.

A turn was identified by the presence of a pause longer than 1/2 sec, or by the intervening speech of the other party. Turn boundaries act also as time stamps in seconds; the XML transcriptions contain the time stamp information. The transcription of the agent's and the client's audio signals are mixed according to the time stamps. Each line is a turn; the agent's contributions starts with "A:" while the client's starts with "C:.".

[1] http://www.ldc.upenn.edu/mirror/Transcriber

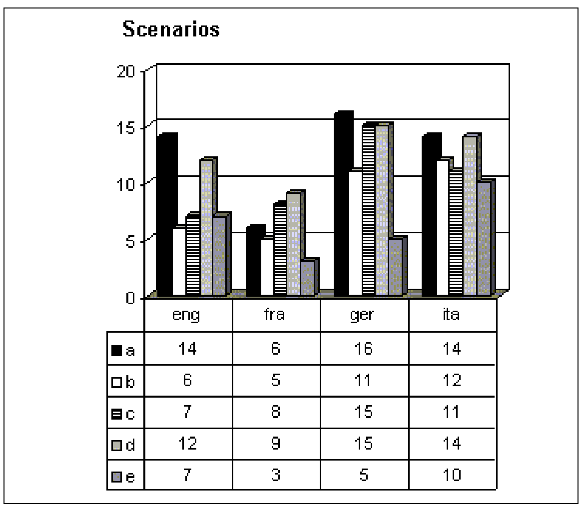
4.3. German and English Transcriptions

The German and English Transcriptions follow a subset of the conventions for the transcription of spontaneous speech established for the European project VERBMOBIL[1]. CMU and UKA made the decision to follow VERBMOBIL[2] conventions, because software applications, processing tools and transcribers trained in these conventions previously existed at these two sites. Additionally, this convention system covers all needs of grammar and speech recognition development and is easy to convert into other formats.

The TransEdit[3] application was used for transcription. The tool provides for click able labels, and automatic support for turn numbering and format management. TransEdit's audio application allows several signal file tracks to be displayed, so that a complete dialogue can be processed in one step. Turn segmentation is realized by moving the selected sections into the segmentation row. The result is a so called marker file, containing 'begin' and 'end' time stamps at sample points and automatically created identifiers for turns according to their transcriptions. Turns correspond to speaker contributions.

5. Results/Status

5.1. Recorded Dialogues



	eng	fra	ger	ita
a	14	6	16	14
b	6	5	11	12
c	7	8	15	11
d	12	9	15	14
e	7	3	5	10

Figure 2: Scenario distribution; a: winter vacation in Val di Fiemme, b: all inclusive package, c: summer in the park, d: castles and lakes, e: folklore and brochures

200 dialogues have been recorded so far. Figure 2 shows the distribution of recordings of the five scenarios for all four recording sites.

Most of the recordings concern scenario a (50), d (50) and c (41). There are 62 German dialogues recorded, 61 Italian, 46 English and 31 French. All French and English recordings are completely transcribed. 27 German dialogues are completely transcribed. Due to technical problems, for 31 recordings only the client part was transcribed. 54 Italian dialogues are completely transcribed, 6 have transcriptions of the Italian agent's part, one of the client's part.

5.2. Speaker Distribution

Figure 3 shows the speakers' distribution for each recording site. The total number of speakers is 94: 25 females and 69 males (there were no special restriction as to participation frequency for subjects).

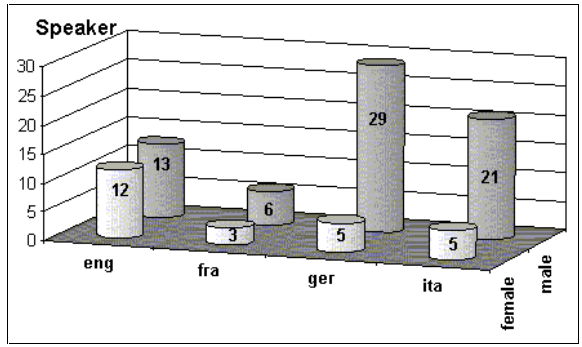


Figure 3: Speakers' distribution

5.3. Verbosity Rates

For the current paper, the original transcriptions were converted to VERBMOBIL II [2] format and filtered so that only the orthographic version of the text was left.

For all German transcriptions, the vocabulary list contains 2447 different types at a total word count of 33044 spoken tokens, English 1610 different types at a total word count of 43825 spoken words, French 2052 types at a word count of 37243 spoken words and all Italian transcriptions result in a vocabulary of 3216 types for 45710 tokens.

The participants of the experiment were asked to produce rather verbose contributions. Figure 4 shows how they fulfilled this task.

The French conversations resulted in the longest dialogues, with an average of 132 turns per dialogue; the German-speaking subjects follow with 97 turns per dialogue, then the English with 95 turns. Italian dialogues were the shortest, with an average of 70 turns. Turns were counted for complete transcriptions of dialogues (client and agent part).

We counted tokens per clients' turns, because for English, French, and German dialogues the clients were the native speakers, but the agents' task generally involved more speaking. To avoid this additional issue, also for Italian transcripts only the clients' turns were taken. All languages showed very similar results. The German turns were the shortest with slightly fewer than seven words per turn on average. Italian and French speakers produced approximately seven words per turn on average and English speakers over eight.

Tokens-per-type is to be read as: how many words have been spoken on average, theoretically, before a new type was introduced. We counted the token-per-type rate for the clients' contributions. The rates show that the Italian clients added at every ninth word spoken a new type to the

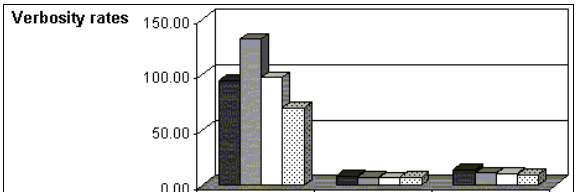
---

[1] http://verbmobil.dfki.de/

[2] http://www.is.cs.cmu.edu/trl_conventions/transcription.html

[3] further information about Trans Edit: sburger@cs.cmu.edu

vocabulary; native speakers of a French and German produced a new type at every eleventh spoken word, and native speakers of English after about fourteen words.



	turn/dia	token/turn	token/type
eng	94.64	8.48	13.67
fra	132.42	7.14	11.32
ger	97.19	6.86	10.68
ita	70.20	7.00	9.37

Figure 4: Average numbers for turns per dialogue, words per turn for all clients' turns, types per tokens for all clients' turns

5.4. Vocabulary Growth

In Figure 5, we compare the vocabulary growth for all four languages. The rates are based on all dialogues, client and agent contributions. The figure shows less growth for the English vocabulary, very similar curves for French and German vocabulary and a high growing rate for the Italian recording. These results have to be seen as characteristic for the NESPOLE! data base but may not be representative for growth curves of the involved languages in general. Only the Italian growth rate contains pure native speaker data. The rates of the other three languages contain non-native speaker contributions of only a few Italian agents and depend on their individual fluency in English, French or German.

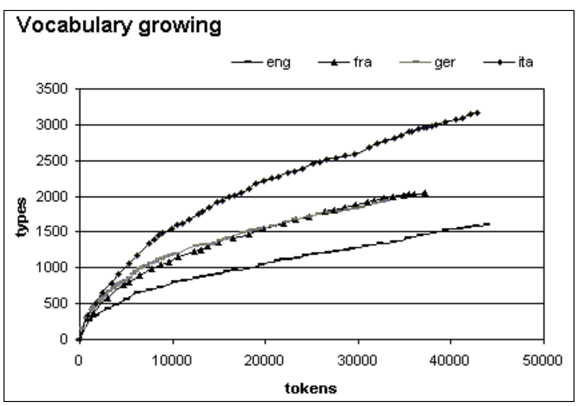


Figure 5: *Vocabulary growing graphs for the English, French, German and Italian dialogues*

6. Discussion

The main problems that arose during data collection and annotation were due to the original idea that data collection were to be conducted in a decentralized fashion, with each site taking care of the data it collected in a fairly independent way from the other sites. Thus, detailed agreements as to common names, formats and procedures seemed unnecessary, and each site was left free to rely on its already established data collection procedures and conventions. However, in the course of the data collection, it turned out that a greater degree of coordination among the sites would have been extremely useful. Consistent naming convention system turned out to be needed to facilitate data sharing among organizations, archiving, and to conduct multilingual experiments involving data from different partner languages.

The decision was then taken to develop tools to convert the transcriptions of the data already collected into a common standard, a task we are currently accomplishing.

The following list summarizes still pending problems that must be taken into account in future data collection, the archiving of the collected NESPOLE! data, and any analysis of the data collected from the different sites.

The H323 transfer line is often of very low quality. The corresponding transcriptions have the same low quality, for obvious reasons.

Data transmission introduces delays. Thus, when the client wanted to interrupt the agent, there was about 1s during which the agent kept on speaking. This sometimes resulted in strange dialogues, where client and agent don't know whose the turn is.

At the moment, the transcriptions are compatible only at the word level. In some cases, dialogue turns may have not been marked consistently across sites. It has also to be taken in account, that the four languages follow slightly different orthographic rules for contractions, so that two to three words become one word on the written word level. After converting all transcriptions to the same format and conventions, vocabulary statistics will become more interesting and reliable.

Acknowledgements

The data collection and transcription at the different sites was done by Christina Barbero, Francesca Guerzoni and Paolo Coletti, Yannick Fouquet, Solange Hollard and Laurent Besacier, John Mc Donough, Christiane Reihl, Hagen Soltau, and Florian Metze, Courtney Conrad, Victoria Maclaren, Celine Morel, Kay Peterson and Susanne Burger.

This material is based upon work supported by the EU under Grant No. IST1999-11562 and by the US National Science Foundation under Grant No. 9982227. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Lazzari G., Spoken translation: challenges and opportunities, ICSLP'2000, Beijing, China.

[2] Burger, S., "Transliteration spontan-sprachlicher Daten - Lexikon der Transliterationskonventionen - VERBMOBIL II", Verbmobil TechDok-56-97, München, 1997.