The NESPOLE! VoIP Multilingual Corpora in Tourism and Medical Domains

Nadia Mana[1], Susanne Burger[2], Roldano Cattoni[1], Laurent Besacier[3], Victoria MacLaren[2], John McDonough[4], Florian Metze[4]

[1]Itc-irst, Centro per la Ricerca Scientifica e Tecnologica, Trento, Italy
[2] Interactive Systems Laboratories, Carnegie Mellon University, Pittsburgh, USA
[3] Laboratoire CLIPS, Equipe GEOD, Universitè Joseph Fourier, Grenoble, France
[4] Interactive Systems Laboratories, Universität Karlsruhe, Karlsruhe, Germany

mana@itc.it

10.21437/Eurospeech.2003-464

Abstract

In this paper we present the multilingual VoIP (Voice over Internet Protocol networks) corpora collected for the second showcase of the Nespole! project in the tourism and medical domains. The corpora comprise over 20 hours of human-to-human monolingual dialogues in English, French, German and Italian: 66 dialogues in the tourism domain and 49 in the medical domain. We describe in detail the data collection (technical set-up, scenarios for each domain, recording procedure and data transcription), as well as statistically illustrated corpora and a preliminary data analysis.

1. Introduction

NESPOLE![1] is a jointly EU/NSF funded project, designed to provide fully functional speech-to-speech capabilities within real-world settings of common users involved in e-commerce applications. The design principles of the NESPOLE! system are described in [1]. The NESPOLE! system uses a client-server architecture to allow a common user, who is initially browsing through the web pages of a service provider on the Internet, to connect seamlessly to a human agent who speaks another language, and provides STST (speech-to-speech translation) service. Commercially available PC video-conferencing technology is used to connect between the two parties in real-time. The languages addressed are Italian, German, English and French. The scenarios for the first two showcases (1 and 2A) are in the tourism domain and involve an Italian-speaking agent located in an Italian tourism agency (APT), and an English-, German- or French-speaking customer at an arbitrary location. The third version (showcase 2B) was developed to evaluate the portability of the NESPOLE! STST system to new domains. Within the selected medical domain, the scenario was limited to a first aid medical assistance service.

In order to develop the human language technology modules used in the STST chain for the first showcase [2], a task domain vocabulary was needed for each involved language. For this reason, a first data collection was conducted on the tourism scenario during 2000 [3].

In the system's context, one of the goals of the second showcase was to demonstrate project scalability and portability. This demonstration was exhibited in two distinct NESPOLE! domains, as extended tourism and as medical domain. However developing the translation modules required a task domain vocabulary for each language involved; hence new data collection in the tourism and medical domains was designed and conducted in 2002. Differing from the previous data collection, the principle goal was to define common protocols, procedures and conventions for the building of monolingual corpora by each partner (CMU, Carnegie Mellon University, USA – UKA, University of Karlsruhe, Germany – CLIPS, Université Joseph Fourier, France – ITC-irst, Trento, Italy). The technical set-up and use of the NESPOLE! system interface [4] were also novel with respect to the previous data collection. Such an interface may load and share images, web pages, etc., and includes drawing functions that are useful for performing gestures on the images during dialogues. In this respect, the collected dialogues are multimodal.

The following sections describe in detail the data collection (technical set-up, scenarios for each domain, recording procedure and data transcription), as well as statistically illustrated corpora, a preliminary data analysis, and concluding remarks.

2. Data Collection

The goal of the 2002 data collection was to record monolingual dialogues between human subjects, using the tourism and medical domains. For the tourism domain, the dialogues were held between a travel agent and a prospective tourist (client), whereas the medical domain dialogues were conducted between a doctor and a patient.

Each partner provided human subjects for the data collection[2]. These subjects were native speakers of each site's respective language, English, German, French, or Italian.

2.1. Task description

The task for the 2002 data collection on the tourism domain was similar to the task of the previous data collection on the same domain. The scenario designed for the tourism domain was as follows: While navigating on the Internet page of the Tourism Board of Trentino (a region in northern Italy), an English, French or German user, intending to have vacation in Trentino, is seeking additional information not printed on the

[1] NESPOLE! – NEgotiation through SPOken Language in E-commerce. For further details, see the project web-site at htpp://nespole.itc.it

[2] According to the project's Technical Annex, the French partner was not involved in the medical data collection.

web pages. Using a link from the provider web page, the user can establish a live real-time interactive connection with an Italian human operator from the Tourism Board. Through the NESPOLE! system, interactive multimodal applications are then automatically activated and support the communication between the two parties. According to this scenario, the clients' task was to ask pertinent information in order to plan a vacation, while the agent provided clear details.

For medical data collection, the doctor and patient used the NESPOLE! system to communicate in a simulated long-distance medical consultation. In the scenario, the patient is allegedly ill and contacts the doctor via the NESPOLE! system, abstaining from a traditional phone call. The patient's task was to illustrate health problems and related symptoms, while the doctor formulated a diagnosis and suggested a cure.

2.2. Technical set-up

For the technical set-up, the data collection process was guided by a principle of uniformity. Both domains used the same software for all sites: (1) Windows® 2000 (NT, in one case) as Operating System, (2) NetMeeting® 3.01 with the G711 audio codec for the net connections, (3) the NESPOLE! software developed for the H323 layer (the User-Interface [2] and the Mediator) and (4) TotalRecorder® 3.3 as the audio recording tool. All sites recorded in signed u-law PCM 16 bit stereo on separated channels (respectively, for local clean speech and for remote speech through the H323 connection). The sampling frequency was 16 kHz[3].

There were some unavoidable differences in the hardware employed, depending on the site: different PC set-up (desktops or laptops), different soundcards, and microphones (headset or close microphones, with or without pre-amplifier). However such differences did not affect the software layer.

The network configuration of the NESPOLE! machines was different in the two domains. For the tourism domain, the agent and Mediator machines were set in fixed locations (at the Tourism Board in Trento, Italy) while the client's location varied according to the language: CMU (USA) for English, UKA (Germany) for German, IRST (Italy) for Italian, CLIPS (France) for French. Therefore the quality of the H323 remote speech depended on the internet conditions between Italy and the client's country during the recording time - for a discussion of the issues concerning H323, VoIP and speech recognition see [5]. For the medical domain, each of the three collected languages (English, German and Italian) had doctors and patients in the same site; the NESPOLE! machines run on a LAN, and thus even the H323 remote speech was considered of a quality comparable to the local clean speech.

Typical environmental conditions included separated rooms and a quiet office (one exception was the English data collection in the medical domain due to the physical constraints of the 5th US-Army military Hospital).

2.3. Tourism and Medical Scenarios

For the tourism domain, the first showcase has been extended to include five new scenarios, which summarize all packages available for vacations in Trentino, either during the summer or winter. Packages were addressed to couples, families or small groups of two to three persons, and included several possible activities (different kinds of excursions, visits to natural parks, castles or museums, access to a swimming pool, sauna, etc.), and three different possible accommodations (hotel, apartment or campsite) in two areas (mountain or lake). The proposed vacation could last one-week or a few days. The five scenarios are: *All-inclusive summer package in a hotel or apartment* (Scen. F); *All-inclusive summer package in a campsite for a family* (Scen. G); *All-inclusive winter package in a hotel or apartment* (Scen. H); *All-inclusive summer package in a hotel or apartment for a family* (Scen. I); *All-inclusive summer package in a campsite* (Scen. J). For each scenario, several web pages were designed to contain information about holiday resorts, local attractions and forthcoming events. Prepared electronic documents also summarized information relevant to the packages and maps of the holiday resorts.

For the medical domain, four scenarios were developed. Given the immense diversity of ailments, the scenarios focused on two possible health problems: chest pain (Scen. K, L); flu-like syndrome (Scen. M, N). Each health problem was outlined according to two possible situations: a more serious situation (e.g. referring to a very old person, with permanent diseases and allergies) and another one less serious (e.g. referring to a young person with a history of good health). According to information by physicians, scenarios were described in terms of personal data (sex, age, occupation, weight, height, allergies, diseases, etc.) and symptom description (listing, when started, duration, frequency, medicine taken, etc). Additionally, images of the human body or body parts (e.g. head, chest, back) were provided.

2.4. Recording Procedure

For the data collection in the Showcase 2A tourism domain fluent speakers of the concerned language (English, French, Italian and German) were recruited and instructed to act as clients. These subjects received instructions about the scenario, their task, the system functionalities and interaction modalities prior to their scheduled recordings. By the end of the tutorial, they could understand their task and how to perform it, while the agents, who were Italian operators of the Trentino Tourism Board, were trained and instructed to use the NESPOLE! interface.

The session began when the client used the *Audio Tuning Wizard* feature of NetMeeting to properly tune the audio settings for his/her voice. The client then called the agent via NetMeeting. Recording began once the agent answered. Clients formulated questions according to their plans, needs and constraints (e.g. preferred accommodation, available budget, interests for specific activities), following a predefined script. Agents presented the packages, describing prices, accommodation, activities and facilities. Unlike the 2000 data collection [3], the agent used NetMeeting's file sharing ability to load maps and web pages, which improved answers to the clients' questions and supplemented conversation. The Whiteboard function of the NESPOLE! Interface [4] allowed both parties to make marks, or gestures, which were simultaneously visible to both parties. Agents used maps to provide directions or to show the location of hotels, castles, museums, etc., and all the agents' gestures were saved. Agents also used online tourism sites to supplement the information they provided to the clients.

[3] In some cases constraints on soundcards forced higher sampling frequencies (22kHz or 44kHz): in such cases the dialogues have been later down-sampled to 16kHz.

For medical data collection, subjects playing the role of patients and doctors were recruited from a pool of physicians[4]. Similar as for the tourism data collection, patients and doctors were required to familiarize themselves with their roles and a scenario description, according to specific instructions. Patients received a script containing the main information about their health status and problems. Sessions were conducted similarly to the tourism domain. However, while agents were the main users of maps in the tourism data collection, here anatomical maps were mainly used by the patients to mark pain location.

2.5. Transcription

Common transcription conventions allow for consistency and eases future experiments, especially between languages. We agreed upon a subset based on the transcription conventions for spontaneous speech developed under the German project VERBMOBIL[5]. This transcription system was chosen because it can be easily converted into other formats, which meets all of NESPOLE!'s grammar and speech recognition requirements and conforms to pre-existing processing tools[6].

Transcription was performed with *TransEdit*, a Windows-based tool for transcription and segmentation. Using automatic turn numbering and format management, *TransEdit* supported consistent transcription in all four partner sites. It also provided an audio application allowing multiple audio signals to be displayed concurrently, so transcribers could view the audio of both speakers simultaneously. Users created turn segmentations by highlighting and designating turns. *TransEdit* retained the segmentations in a separate file known as a marker file, and the transcription in a transliteration file.

3. Results

3.1. Showcase 2A – Tourism Domain

In total, 66 dialogues were recorded for Showcase 2A. Table 1 breaks down the distribution by language and scenario.

	F	G	H	I	J	TOT
English	4	3	3	3	3	16
French	4	3	3	3	4	17
German	4	3	3	3	3	16
Italian	5	3	3	3	3	17
TOT	17	12	12	12	13	**66**

Table 1: Collected tourism dialogues

The data corpus comprises 16.5 hours of dialogue; because audio data was recorded on both ends of the NetMeeting® connection, a total of 33 hours of audio data was collected. The Italian dialogues were the shortest, averaging 13.6 minutes, while the French were the longest, averaging 16 minutes. The German and English dialogues averaged 15.3 and 15.2 minutes, respectively.

65 participants acted as clients. The English had seven male and nine female speakers. Both the German and French groups had 10 males and six females, the Italians 11 males and six females. 12 participants were agents. Of these, the agents for the French, German and English groups were all female, while the Italian group had three females and two males. The only agent who was not a native Italian speaker was a native French speaker and appeared in eight of the French dialogues.

3.2. Showcase 2B – Medical Domain

For Showcase 2B, the medical domain, a total of 49 dialogues was recorded; the distribution of these recordings by language and scenario is reported in Table 2.

	K	L	M	N	TOT
English	5	4	3	4	16
German	3	5	5	3	16
Italian	5	4	4	4	17
TOT	13	13	12	11	**49**

Table 2: Collected medical dialogues

The medical recordings make up eight hours and 25 minutes of audio data, again including one set of recordings from each side of the dialogues. On average, German dialogues lasted about 5.5 minutes, English dialogues about 5.3 minutes, and Italian dialogues 4.6 minutes.

The 36 participants included 12 females and 24 males acting either as physicians or patients. 10 of the English, six of the German and all Italian participants were professional doctors by trade; the other English participants included two registered nurses; the rest of the German participants were six medical students and one registered nurse.

3.3. Preliminary Data Analysis for Turn Duration

Different scenarios, even in the same domain, may have a strong impact on the dialogues. In order to investigate such an impact, we calculated some preliminary statistics regarding turn length and talk_time on the three collected corpora.

3.3.1. Talk_time

The overall *talk_time* a single speaker talks during a dialogue is the sum of the durations of all segmented turns spoken by this speaker; in other words, it is the actual time when the speaker was talking during the dialogue but was not listening to or waiting for an answer.

Comparing the data collected for Showcase 1, Showcase 2A and 2B in terms of the duration of talk_time of agents versus clients (doctor vs. patient, respectively) shows in case of Showcase 1 data that agents talked 5.4 minutes per dialogue in average while clients' talk_time was not even the half of this time, 2.4 minutes (see figure 1). Showcase 2A showed a more balanced talk_time for both dialogue partners with a slight longer duration of agents' talk-time. Also the doctors and patients of the Showcase 2B dialogues talked for similar durations.

[4] Physicians of the 5th US-Army Hospital in Heidelberg (Germany) and of the Uniklinikum in Mannheim (Germany) were recruited for English and German, respectively. Italian subjects were recruited among physicians of the Medicine Faculty in Florence, Italy.

[5] See http://verbmobil.dfki.de/

[6] See http://www.is.cs.cmu.edu/trl_conventions

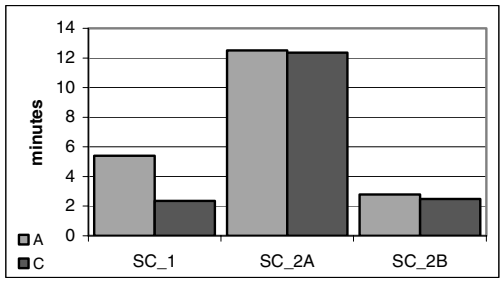

Figure 1: Talk_time in minutes, A: agent/doctor, C: client/patient

3.3.2. Long turns – short turns

In our second analysis, we calculated how many speaker turns lasted (a) 10 seconds or longer and (b) 0.7 seconds or briefer. These data may indicate a lopsided relationship between speakers; in such cases one speaker would record a high number of long turns, while the other speaker's turns would be short and mostly confirmations and interjections.

In Showcase 1, the dialogues show many long turns for agents and a few long turns for clients (see figure 2). Both clients and agents exhibited a large number of short turns. Thus, Showcase 1 dialogues showed a lopsided effect in favour of the agent. Looking at the result of Showcase 2B, in which the conversation was supported by the use of drawings functions, 27% of agents turns and nearly 26% of client turns were long. The frequency of short turns was also balanced for agents and clients, but short turns occurred three times less often than in Showcase 1 data. The doctor-patient dialogues of Showcase 2B are also nearly balanced in terms of long and short turns; contrary to the other corpora, these dialogues also show doctors with slightly fewer long turns but more frequent short turns.

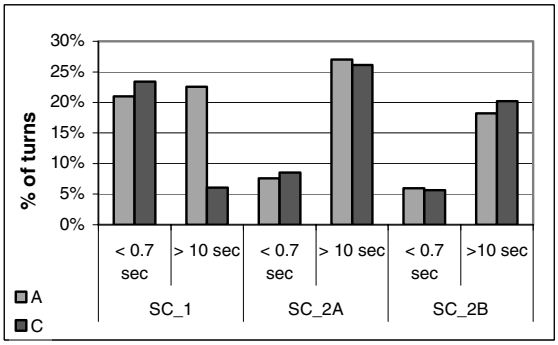

Figure 2: Long (>10 sec) and short (<0.7 sec) turns, A: agent/doctor, C: client/patient

3.3.3. Discussion

These statistics highlight some changes between the 2000 and 2002 data collections. First, speaker roles (agent vs. client) seemed more balanced in both of the 2002 domains: talk_time and the number of long and short turns of agents/doctors are close to that of clients/patients. We have several explanations for this phenomenon: (a) the defined tasks of Showcase 2A and 2B were more demanding than Showcase 1, forcing clients/patients to speak more to obtain the required information; (b) the use of the NESPOLE! interface, newly equipped with multimodal functions, let speakers use shorter and more effective sentences while giving instructions – this particularly influenced agents in Showcase 2A. The second change highlighted by the statistics is that only the medical domain showed clients (actual patients) recording slightly more long turns. We credit the task itself: doctors gathering information use succinct questions. Patients receiving a diagnosis tend to provide answers with details and long explanations.

We plan to investigate further into these issues, taking into account the linguistic content of the dialogues as well.

4. Conclusion

In this paper we presented the NESPOLE! VoIP multilingual corpora collected in tourism and medical domains. Preliminary analysis highlights that the corpora represent a valuable resource, from both an acoustic and a linguistic point of view. The corpora are scheduled to be available through the main linguistic organizations, LDC (US) and ELRA (EU).

5. Acknowledgements

The work described in this paper has been partially supported by the National Science Foundation under Grant number 9982227, and by the European Union under Contract number 1999-11562 as part of the joint EU/NSF MLIAM research initiative.

We thank all participants in data collection and transcription. In particular, Elisabetta Fauri, Franca Rossi, Erica Costantini, Francesca Guerzoni, Juliet Brown, Denise Hill, Robert Isenberg, Zachary Sloane, Anne-Claire Descalle, the APT operators in Trento, the physicians of the 5th US-Army Hospital, Heidelberg, of the Uniklinikum in Mannheim, Germany, and of the Medicine Faculty in Florence, Italy.

6. References

[1] Lavie A., Langley C., Waibel A., Pianesi F., Lazzari G., Coletti P., Taddei L., Balducci F.: "Architecture and Design Considerations in NESPOLE!: a Speech Translation System for E-commerce Application", in *HLT 2001 Proceedings*, San Diego, U.S.A. , 2001.

[2] Lavie, A., Metze, Pianesi F., et al: "Enhancing the Usability and Performance of NESPOLE!: a Real-World Speech-to-Speech Translation System", in *HLT 2002 Proceedings*, San Diego, California U.S., March 2002

[3] Burger S., Besacier L., Coletti P., Metze F., and Morel C. "The NESPOLE! VoIP Dialogue Database", *EuroSpeech 2001 Proceedings, pp. 2043-2046,* 2001.

[4] Taddei L., Costantini E., Lavie A. "The NESPOLE! Multimodal Interface for Cross-lingual Communication - Experience and Lessons Learned", *Proceedings of ICMI 2002*, Pittsburgh, USA, 14-16 October 2002.

[5] Metze F., McDonough J. and Soltau H., "Speech Recognition over NetMeeting Connections", *EuroSpeech 2001 Proceedings,* 2001.