The Italian NESPOLE! Corpus: a Multilingual Database with Interlingua Annotation in Tourism and Medical Domains

Nadia Mana[1], Roldano Cattoni[1], Emanuele Pianta[1], Franca Rossi[1], Fabio Pianesi[1] and Susanne Burger[2]

[1]ITC-irst
Centro per la Ricerca Scientifica e Tecnologica, Trento, Italy
{mana, cattoni, pianta, frarossi, pianesi}@itc.it

[2]Interactive Systems Laboratories
Carnegie Mellon University, Pittsburgh, USA
sburger@cs.cmu.edu

Abstract

This paper presents the Italian NESPOLE! Database. The database consists of three parts: The first two, called DB-1 and DB-2 concern the tourism domain, while the third part, DB-3, concentrates on the medical domain. The database includes audio files, transcriptions, Interlingua annotations in IF (Interchange Format) and translations into English, French and German. We describe how the database was built (data collection set-up, scenarios, recording procedure, data transcription and annotation) and statistically illustrates the corpus by providing a data analysis focused on language and spontaneous phenomena.

1. Introduction

The Italian NESPOLE! database is part of the multilingual VoIP (Voice over Internet Protocol networks) corpora collected for the NESPOLE! project[1].

NESPOLE! was a jointly EU/NSF funded project over three years and ended on February 2003. It aimed at supporting spontaneous multi-lingual human-human conversation in the context of advanced e-commerce by developing speech-to-speech translation (STST) technologies and multimodal interfaces. The NESPOLE! system (Lavie et al., 2001) uses a client-server architecture to allow an initially Internet-browsing user to connect seamlessly to a human agent who speaks another language, and provides speech-to-speech translation service. Commercially available PC video-conferencing technology are used to connect between the two parties in real-time. The languages addressed are Italian, German, English and French. The scenario for the first two showcases was the tourism domain and involved an Italian-speaking agent located in an Italian tourism agency, and an English-, German- or French-speaking customer at an arbitrary location. The third version was developed to evaluate the portability of the NESPOLE! STST system to new domains, and targeted first medical-aid assistance. A first monolingual data collection was conducted on the tourism domain for all four languages (Burger et al, 2001), followed by a second data collection on an extended tourism domain and a third collection on a medical domain (Mana et al., 2003).

The Italian database, developed within the NESPOLE! project, includes audio files, transcriptions, interlingual annotations in IF (Interchange Format) and translations in English, French and German. Here we describe how the database was built (data collection set-up, scenarios, recording procedure, data transcription and annotation) and illustrate the corpus statistically by providing a data analysis focused on language and spontaneous phenomena.

2. Data Collection

2.1 Scenarios

The Italian NESPOLE! database concerns 2 domains: tourism and medicine.

The first collected database, DB-1, consists of dialogues between a travel operator (Agent) and a tourist (Client) who wants to spend a holiday in Trentino, a region in northern Italy. The clients formulate questions following a predefined script, based on five scenarios:

- Scenario a: Winter accommodation in Val-di-Fiemme[2]
- Scenario b: All included tourist packages
- Scenario c: Summer vacation in a park
- Scenario d: Castle and lake tours
- Scenario e: Looking for folklore and brochures

Handouts containing information regarding the location and available activities for every vacation package were given to the agent.

DB-2 still concerns the tourism domain, but we focused on inclusive vacation packages, articulated in the following five new scenarios:

- Scenario f: All-inclusive summer package in a hotel or apartment.
- Scenario g: All-inclusive summer package in a campsite for a family.
- Scenario h: All-inclusive winter package in a hotel or apartment.
- Scenario i: All-inclusive summer package in a hotel or apartment for a family.
- Scenario j: All-inclusive summer package in a campsite.

1 NESPOLE! – NEgotiating through SPOken Language in E-commerce. For further details see the project web-site at http://nespole.itc.it

2 A resort in Trentino.

All scenarios regarded packages offering a vacation in Trentino during the summer or winter season. Vacation proposals varied in age and number of participants, duration of stay, accommodation type, cultural and recreational activities. Clients asked questions according to their plans, needs and constraints. Agents presented the packages, described prices, accommodation, possible activities and facilities.

DB-3 provides manufactured dialogues between a doctor and an ailing patient. The patient was feigning ill and contacted the doctor via the NESPOLE! system. The patient's task was to illustrate health problems and related symptoms, while the doctor formulated a diagnosis and suggested treatment. The medical scenarios focused on two possible health problems: chest pain and flu-like symptoms, each of them outlined to a less serious and a more serious situation.

- Scenario k: chest pain 1
- Scenario l: chest pain 2
- Scenario m: flu-like syndrome 1
- Scenario n: flu-like syndrome 2

2.2 Recordings

For the tourism domain the role of the agent was played by professional tourism agents working at APT (tourism board of Trento), whereas for the client role subjects speaking Italian fluently were recruited and instructed to act as clients. For the medical domain subjects playing the role of doctor and patient were both recruited from a pool of physicians of the medicine faculty in Florence, Italy. Similarly to the tourism data collection, speakers were required to familiarize themselves with their role, the dialogue scenario description and the use of the NESPOLE! system interface. All speakers involved in the data collection communicated through the Internet, using thin terminals (PCs with sound and video cards, and H323[3] video-conferences software). The network configuration of the NESPOLE! machines (agent, mediator and client machines) was different in the two domains. For the tourism domain, the agent and mediator machines were set at APT, while the client's location was at ITC-irst. The quality of the H323 remote speech depended on the internet conditions during the recording time. For the medical domain the data collection was conducted at the medicine faculty in Florence. The doctors and patients were at the same site, but in separate rooms. In this case 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.

In DB-2 and DB-3 the NESPOLE! system provided for a multimodal communication (Taddei et al., 2002) too, allowing speakers to share images, maps and web-pages displayed on the monitors of both parties and to perform drawing gestures on maps, simultaneously visible to both. In the tourism domain, the use of maps facilitated the agent's task of giving directions, providing information about position of relevant objects, highlighting locations, etc. In the medical domain anatomical maps were mainly used by patients to mark pain location, but by request of the doctor. Additional details on the technical set-up and the recording procedure can be found in (Burger et al., 2001) for DB-1 and in (Mana et al., 2003) for DB-2 and DB-3.

3 H323 is a standard for transmitting voice and video over IP networks).

3 Transcription, Annotation and Translation

3.1 Transcription

All the collected dialogues were transcribed, using the Transcriber tool[4] for DB-1 and the TransEdit tool[5] for DB-2 and DB-3. Transcriptions include segmentations in turns corresponding to speaker contributions. In addition to the orthographic representation of words, the transcriptions provide labeled spontaneous phenomena, such as filled pauses, repetitions or corrections, false starts, aborted words, etc. according to a predefined convention set[6].
An example is reported here:

> i102h_2_0013_NAD_00: <hes> -/io vole=/- sono interessata +/a una/+ a una vacanza in Trentino in inverno per due persone . [lit. I'd be interested in a vacation in Trentino, in winter, for 2 persons]
>
> i102h_1_0014_CRI_00: bene . <B> <Laugh> due persone <P> in inverno . [lit. fine. 2 persons. In winter]

3.2 Interlingua Annotation

The transcribed dialogues were manually tagged on the basis of the Interchange Format (IF), a task-oriented, language independent, meaning representation formalism, aiming at representing the communication intentions of the speaker more than the literal expression of such intentions. An IF representation corresponds roughly to a clause (or fragment of it) called a *Semantic Dialogue Unit* (SDU). The representation consists of four components:

1. the *speaker tag*, where `c:` indicates the client (in our dialogues the traveler or the patient), and `a:` the agent (in our dialogues the travel agent, or the doctor);
2. the *speech act*, e.g. `thank`, `give-information`;
3. a possibly empty sequence of *concepts*, describing the conceptual focus the utterance, e.g. `+hotel`, `+pain`;
4. a possibly empty list of *arguments* as name-value pairs, specifying details of the intended SDU meaning. Arguments are licensed by concepts.

The following are three examples of utterances tagged with their corresponding IF labels:

1. *Thank you very much*
   c:thank
2. *And we'll see you on February twelfth*
   a:closing (time=(february, md=12))
3. *There is an hotel in the town*
   a:give-information+existence+accommodation (accommodation-spec=hotel, location=town)

4 http://www.ldc.upenn.edu/mirror/Transcriber/

5 Burger S., Meier U., "TransEdit. A New Way to Transcribe Speech Data." Manual by Helman J.

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

Let us look at the IF in Example 3 in more detail. The first element is the speaker tag `c:`, identifying the travel agent. The second component is the `give-information` speech act, which describe the communication intention of passing some information to the hearer. The speech act is followed by the concepts `+existence` and `+accommodation`, which are the two main concepts of the SDU. The combination of a speech act with one or more concepts results in what is called a *domain action*. In our example the domain action can be paraphrased as "communicating information about the existence of some accommodation". The domain action licenses a set of *arguments* that are semantically related to the concepts of the domain action. Here the concept `+accommodation` licenses the `accommodation-spec=` argument, specifying the type of accommodation the speaker is referring to, whereas `+existence` licenses the `location=` argument, specifying that the hotel can be found in the town.

Despite being task-oriented, the IF has been conceived with the goal of accommodating as many domains as possible, by clearly distinguishing the IF parts (speech acts, concepts, etc.) that are domain-independent, from those that are domain-specific. This has positively contributed to the portability of STST systems, resulting in the current version of the IF, which covers two very different domains: tourism and medical assistance[7]. The table below shows a breakdown of speech acts, concepts, and arguments by domain.

	Travel	*Medical*	*General*	*Total*
Speech Acts	0	4	75	79
Concepts	49	16	79	144
Arguments	43	~50	~310	403

Table 1: Speech Acts, Concepts and Arguments

3.3 Translation

In order to have a database completely aligned in four languages (English, Italian, German and French) all the textual transcriptions of the Italian dialogues DB-1, DB-2 and DB-3 were translated into the other three languages by professional translators. The translators were appropriately instructed about the structure of the database and it's content to obtain results that respect the original text as much as possible.

The transcripts were translated closer towards their conceptual structure rather than providing a literal representation; moreover, dialectal and idiomatical expressions were translated respecting the translation language.

All the symbols indicating spontaneous phenomena and aborted words were respectively replaced in translations by '...' and +w. Repetitions and corrections, instead, were translated and remained labelled, as in the source text, by the symbols +/.../+ e -/.../-. Parts enclosed in square brackets, and not annotated by means of the IF, were translated too and left within square brackets. Finally foreign words were labelled as in the source text.

4. Results and Discussion

4.1 Collected Dialogues

73 dialogues (39 DB-1; 17 DB-2; 17 DB-3) were collected, for a total of 8 hours, 58 minutes and 24 seconds.

	Dialogue Number	*Total Recording Time*	*Average Dialogue Duration*
DB-1	39	3h 43' 00"	5' 43"
DB-2	17	3h 51' 43"	13' 38"
DB-3	17	1h 24' 03"	4' 57"
TOTAL	73	8h 58' 24"	7' 22"

Table 2: Dialogue Number, Recording Time and Average Duration

On average, dialogues lasted seven minutes and 22 seconds, but were significantly longer in DB-2 (more than 13 minutes instead of 4-6 minutes in DB-1 and DB-3). There are several reasons for the difference: First, the speaking domain changed from the DB-2 to DB-3 dialogues: the physicians in DB-3's first aid medical scenario tended to ask simple and concise questions; patients typically replied with specific and short answers, even if the replies were not always focused. From the DB-1 to the DB-2 dialogues, where the domain, tourism, remains the same, the task became more challenging: finding an all-inclusive package meeting the requirements led clients to ask specific, detailed questions and agents to provide rather lengthy detailed descriptions.

4.2 Turns, Tokens, Types

For each dialogue, the total number of spoken turns, word-tokens and word-types was counted. Turns are, generally, a speaker's contribution to the conversation, and are identified by the presence of a pause longer than 0.5 seconds or by interruption of the speaker. Word-tokens are occurrences of given word-types – e.g. the sentences "*The hotel is called Bellavista"* and *"Bellavista is the hotel name"* together contain 10 word-tokens and seven word-types.

	DB-1	*DB-2*	*DB-3*
turns per dialogue	a: 20.95 c: 20.64 a+c: 41.59	a: 76.47 c: 77.35 a+c: 153.82	a: 14.82 c: 14.59 a+c: 29.41
word-tokens per dialogue	701.7	1568.1	496.7
types per dialogue	64.4	133.05	58.89
word-tokens per turn	a: 22.1 c: 11.6 a+c: 16.9	a: 14.2 c: 6.1 a+c: 10.1	a: 16.1 c: 17.6 a+c: 16.8
token/type ratio	10.8	11.7	8.4

Table 3: Average number of turns, tokens and types per dialogue, tokens per turn and token/type ratio

7 For reference: http://www.is.cs.cmu.edu/nespole/db/

As shown in Table 3, the average number of turns per dialogue is 41.59 in DB-1, 153.82 in DB-2 and 29.41 in DB-3. Clients and agents contributed about the same number of turns to each dialogue (DB-1: 20.95 agent vs 20.64 client; DB-2 76.47 agent vs. 77.35 client; DB-3 14.82 agent vs14.59 client).

The average number of word-tokens per dialogue was 701.7 for DB-1, 1568.1 for DB-2 and 496.7 for DB-3. The number of word-types per dialogue was 64.4 for DB-1, 133.05 for DB-2, and 58.89 for DB-3. Dividing the number of word-tokens by the number of words-types then yields the average token-to-type rate (DB-1 10.8; DB-2 11.7; DB-3 8.4). These values indicate how many words were uttered between introductions of new words to the dialogue; they show that in the medical domain, people tend to produce more new words per dialogue than in the tourism domains.

DB-2 dialogues were longer and richer; i.e., they contained more word-tokens and word-types than the other sets (nearly twice as many as DB-1).

Finally, the number of word-tokens per turn highlights a significant difference between agents' and clients' contributions to the dialogues: in the tourism domains agents produced twice as many word-tokens per turns than clients (DB-1: 22.1 vs. 11.6; DB-2: 14.2 vs. 6.1), whereas in the medical domain, clients produce more word tokens per turn (DB-3 17.6 vs. 16.1). Again, this difference would seem to be due to the differences in scenario and task: in the tourism domain most of information is provided by agents who have to satisfy their clients' requests for information, while in the medical domain the roles are inverted. Clients, as the patients, have to speak more in order to describe their specific health problems and symptoms.

4.3 Spontaneous Phenomena

As mentioned in section 3.1, some classes of spontaneous phenomena were annotated in transcription files. In order to estimate the impact of these phenomena on the dialogues, we first clustered them into two classes: (1) empty and filled pauses and noises, and (2) non-grammatical phrases (e.g. false starts, repetitions, interrupted words). For each class the ratio of the phenomena to the total number of word tokens was calculated.

	Pauses & noises	***%***	***a-grammatical phrases***	***%***
DB-1	4042	14.8	-	-
DB-2	3172	11.9	600	2.3
DB-3	863	10.2	225	2.7

Table 4: Some spontaneous phenomena

The results, reported in Table 4, show percentages over 10% for pauses and noises in each domain (14.8% for DB-1, 11.9% for DB-2 and 10.2% for DB-3); it is worth noticing here that this class of phenomena are expected to affect dialogues only very slightly. The impact of non-grammatical phrases on dialogues is more critical; for this class, the percentages were significantly lower: 2.3% for DB-2 and 2.7% for DB-3 (DB-1 was omitted from this calculation).

5. Conclusion

The collected corpora have been used to train the HLT modules of the NESPOLE! system and to develop the IF.

These corpora may be useful for developing and training other systems, such as speech recognizers, natural language generators, and machine translation systems, among others. Furthermore, the collected data may be used for linguistic and pragmatic analysis of spontaneous speech in human-to-human communication mediated by machines. For these reasons, the authors plan to make them publicly available through ELRA organization soon.

6. Acknowledgements

The work described in this paper has been supported by the European Union (1999-11562) and by the National Science Foundation (99822227) as part of the joint EU/NSF MLIAM research initiative.

The authors thank all participants in data collection, transcription and annotation. Many thanks also to Timothy Notari and Zachari Slogane for their help.

References

Burger S., Besacier L., Coletti P., Metze F., and Morel C. (2001), "The NESPOLE VoIP Dialogue Database", in Eurospeech 2001 Proceedings, Aalborg, Denmark.

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

Levin L., Gates D., Lave A., Waibel A. (1998), "An Interlingua Based on Domain Actions for Machine Translation of Task-Oriented Dialogues", in ICSLP98 Proceedings, Sydney, Australian.

Mana N., Burger S., Cattoni R., et al. (2003), "The NESPOLE! VoIP Multilingual Corpora in Tourism and Medical Domains", in Eurospeech 2003 Proceedings, Geneva, Switzerland.

Florian Metze, J. McDonough, H. Soltau, A. Lavie, L. Levin, C., Langley, T. Schultz, A. Waibel, R. Cattoni, G. Lazzari, N. Mana, F. Pianesi, E. Pianta (2002): "Enhancing the Usability and Performance of NESPOLE!: a Real-World Speech-to-Speech Translation System", *HLT 2002*, San Diego, California U.S.

Taddei L., Costantini E., Lavie A. (2002): "The NESPOLE! Multimodal Interface for Cross-Lingual Communication - Experience and Lessons learned", in ICMI 2002 Proceedings, Pittsburgh, U.S.