COLLECTING AND SHARING BILINGUAL SPONTANEOUS SPEECH CORPORA: THE CHINFADIAL EXPERIMENT

Georges FAFIOTTE*, Christian BOITET*, Mark SELIGMAN*, ZONG Chengqing**

*GETA, CLIPS, IMAG-campus (UJF - Grenoble 1)
385 rue de la Bibliothèque, BP 53
F-38041 Grenoble Cedex 9 (France)

**National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
P.O.Box 2728 Beijing 100080 (China)

{georges.fafiotte, christian.boitet}@imag.fr, mark.seligman@spokentranslation.com, cqzong@nlpr.ia.ac.cn

ABSTRACT

We describe here the three main platforms in the ERIM family of Web-based environments for human interpreting, two of them in more details, ERIM-Interp and ERIM-Collect, then ERIM-Aid. Each platform supports an aspect of the collecting or study of spontaneous bilingual dialogues, translated by an interpreter. ERIM-Interp is the core environment, providing mediated communication between speakers and human interpreters over the network. Using ERIM-Collect, French-Chinese interpreting data have been collected within the 3-year "ChinFaDial" project supported by LIAMA, a French-Chinese laboratory in Beijing. These "raw" speech data will be made available in the spring of 2004 on an open-access basis, using the DistribDial server, on a CLIPS-GETA website. Our goal is to extend such corpora, on a collaborative scheme, to allow other research groups to contribute to the site whatever annotations they may have created, and to share them under the same conditions (GPL). An ERIM-Aid variant is intended to provide focused machine aids to Web-based human interpreters, or to monolingual distant speakers conversing in different languages.

KEYWORDS

data collection, spontaneous speech, dialogue, speech corpora, interpreter, interpreting, free distribution, freeware.

INTRODUCTION

One of our ultimate research goals is to build systems for automatic speech interpretation (translation of speech) over the Web. Much progress has been made in this area over the past ten years. NEC produced the first speech translation demo, within the tourist domain, in September 1992, but the most widely known coordinated research efforts to date include the C-STAR projects (international Consortium for Speech Translation Advanced Research) [6], the European NESPOLE! IST project [9], the German Verbmobil project [10], and the US DARPA Communicator program [8] with the Galaxy Communicator Software Infrastructure. All have demonstrated platforms enhancing spontaneous speech processing in multilingual person-person or person-system communication, always in restricted domains.

At the same time, we are convinced that human interpreters will remain vital, both as irreplaceable suppliers of subtle nuances and as models for automatic systems. Human interpreting, too, will inevitably be carried out through the Web or its successors. Thus we foresee a continuing need for research on Web-based interpreting, and for data collection of realistic Web-based interpreting sessions (see Furuse & al [4], for related data collection efforts).

We expect the collected data to be useful for training or tuning automatic speech translation systems. It can also be used to study dialogue phenomena in order to adapt software elements, lexicons, etc., to dialogue situations.

Unfortunately, the human resources required to collect such data are always scarce. Thus we recognize a need to recruit data contributors and processors from the world at large, following the open source model. We aim to induce volunteer interpreters or students of interpretation to translate bilingual dialogues online, by exchanging this on-line help for free use of our Web-based lab for *e*-learning of the interpretation trade.

The ERIM human and automatic speech translation platforms have been implemented in several variants.

In Section I, we describe the motivation and design for the ERIM-Interp base platform. In Section II, we present ERIM-Collect, an extension of ERIM-Interp dedicated to the collection of interpreting data. In Section III, we describe the ChinFaDial data collected so far, to be accessed soon on a free distribution web site. In Section IV, we sketch current developments, and plans to "consolidate" the platform and to add new plugins to extend it to the area of instruction or training for interpreters —an extension which we hope will in turn lead to the collection of new data. Conclusions follow.

I. ERIM-INTERP, FOR HUMAN INTERPRETATION ON THE WEB

From our previous work with the multimodal Wizard of Oz Speech Translation platform EMMI at ATR-ITL [5], and other work on monolingual multi-Wizard architectures (NEIMO [2]), and from experience gained in our lab with the C-STAR II and Nespole! projects, we concluded that, even with high quality automatic interpreting systems, there should be a real human "warm body" or "guardian angel" in the loop anyway. Thus a realistic design for online network-based interpretation should "integrate" both human and machine interpretation. ERIM platforms have been developed on this basis (ERIM in French stands for Network-based Environment for Multimodal Interpreting).

1. Motivation

Some companies have already developed proprietary network-oriented "interpreter's cubicles", which are the counterparts of existing fixed installations for interpreting

in multilingual meetings (for example at the UN or EU). However, the associated code is not available for research.
Furthermore, our typical scenario is quite different from that of classical interpreting, where interpreters are at hand for the entire duration of the conversations. Instead, we envisage two scenarios:

- "conference call": the interlocutors establish a schedule and book a time slot with an interpreter.
- "on demand interpretation": the speakers try to converse using whatever knowledge they may have of their interlocutor's language, or of a third common language. When the language barrier impedes communication, they ask an available interpreter to jump in to help.

Apart from these practical motivations, we also wish to conduct experimental studies on the effect of combining multimodal resources on bilingual or multilingual conversations. Thus facilities for recording are required.

2. Design

The design we have settled on to meet these criteria is shown in Figure 1.

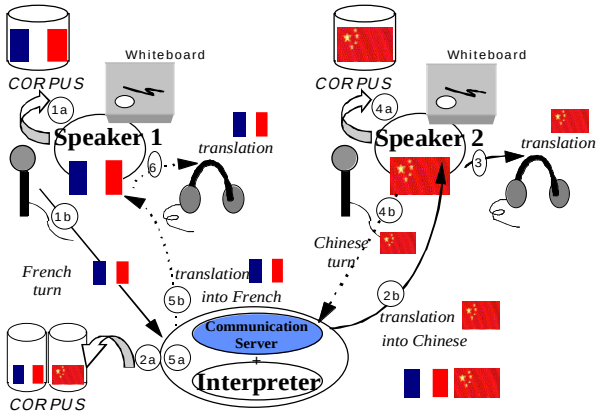

Figure 1: ERIM-Interp / ERIM-Collect

Our system consists of a central communication server, two speaker stations, one interpreter station, and one multimodality server (originally using MBone to handle video communication). It is also possible to exchange short typed messages, using an adapted version of the CommSwitch written by CMU for the CSTAR-II project.

3. Current status

The current implementation, in Tcl/Tk, is platform independent and runs on Windows, MacOS, soon Linux. It is flexible: a speaker and his/her visitor may share the same workstation, or the scenario can be extended to include more than two interlocutors, more than one interpreter (in "one-way interpreting" situations), and hence possibly more than two languages.
Voice transmission in streaming mode is necessary, as a "record-then-send" strategy of course leads to long waiting times. However, the supporting Voice/IP technology has not as yet proven entirely convincing over the Internet. Between Grenoble and Valence (110 Km), there were a few micro-gaps, well tolerable if participants spoke without overlapping. Such drawbacks are reducing rapidly. We may retain facilities for transmitting sound through phone lines. These facilities might be used in operational contexts by telephone operators, such as Prosodie in France: since this company is also an Internet service provider, it can merge both "tracks" into a single communication. These data are summarized in Figure 2.

Experiments: grades from 0 to 5	*text*	*voice: "record then send"*	*voice: "send and record" (streaming)*	*voice: same with overlapping*
Streaming	—	—	+	+
Connexion: Internet	100M bit	=	=	=
Reception quality	5	5	3	1
Speed of exchange	5	2	4	5
Reliability	5	5	4	1
Special problems / phenomena	None	User wary (too slow)	Some micro-cuts, good overall quality	Unusable, bandwidth too large

Figure 2: Oral communication over the web.

II. THE ERIM-COLLECT PLATFORM

1. Motivations and goals

It is widely recognized that realistic and large corpora are necessary resources for building Speech Recognition (SR) and Machine Translation (MT) systems. If the Web has recently been put to use as the largest possible corpus (in order to enhance monolingual Speaker-independent SR and to train acoustic parameters on many speakers), modeling casual spontaneous spoken language requires transcribed speech corpora of hundreds of hours.
Speech translation systems likewise need large parallel translation corpora of transcribed and aligned spontaneous utterances in dialogue context, ideally with complete sets of parse trees. However, few such corpora have been developed (by NEC, ATR and a few others), and these are not publicly available. Why not? Because these corpora are extremely expensive to transcribe, once collected, and annotate. After so much has been spent in compiling a corpus, giving it away seems unreasonable.
With these considerations in mind, we have developed the ERIM-Collect variant of our platform (cf. Figure 1) to enhance collaborative generation and use of bilingual speech corpora, namely to:

- collect only raw data, as multimodal as possible,
- use volunteers to produce the data,
- induce volunteering by offering free service (one of the ERIM variants described here) in exchange for free data (users should agree to "donate their chat to science"),
- distribute the data as freeware (via GPL licensing) on the Web, in a "replayable" form: for each dialogue, descriptors indicate essential facts about the participants (but no names!), along with the list of turns, indications of files, speakers, and time stamps for each turn,
- make it possible for other researchers to enrich the corpora by adding annotations in parallel files, again sharable through the web,
- develop the collection platform so that it can itself be offered as freeware on the Web.

Our own research objective is to use collected corpora for studying and modeling real life spontaneous spoken

language and dialogues. We hope to validate our hypothesis that, depending on specific dialogue situations, translation process settings, and other modalities, specific linguistic traits can be expected.

For instance, two speakers in a bilingual dialogue may hear one another's original speech or not, may use video or fixed images, etc. Their linguistic behavior is expected to vary accordingly: the number of clarification sub-dialogues may vary; third person use or indirect speech may be used more in the presence of a speech translation system than with a human interpreter; the use of deictic and anaphoric elements may turn out to depend on the use of visible markable objects on whiteboards, maps, images.

2. Design

ERIM-Collect is seen as an extension of ERIM-Interp:

- Recording is carried out locally during the conversation. Speech files are in PCM 16kHz-16bit-mono format.
- After the conversation, local descriptors and speech files are sent to the collection server, where the central descriptors are consolidated.
- Everything possible should be recorded: speech, short texts, whiteboard events, use of buttons or menus, video, objects which the speakers refer to (e.g. file paths, urls).

3. Current status

In the current version 3 of ERIM-Collect, only voice and short texts are collected, whiteboard actions will be next.

Figure 3 shows the screen which is currently presented to a conversational partner. This version (400 Kbytes of code in Tcl/Tk) is composed of three main modules (CommServer, Interpreter, and Speaker).

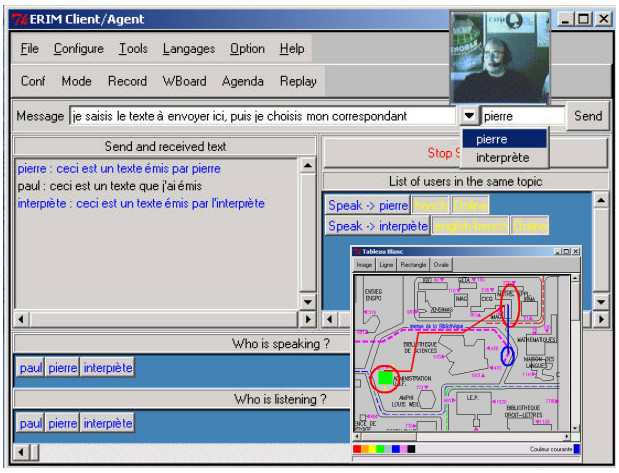

Figure 3: Speaker (Client or Agent) screen

The CommServer module and an Interpreter module may run on the same workstation. Two speakers can converse using the same station, in "office visit" mode.

Other modules include a MultiModServer, to handle whiteboard sharing, and a CollectServer, to merge session sub-corpora. As for playback of a previously recorded bilingual dialogue, a full reconstruction is available, with simplified visual tracking. One can extract monolingual versions of the dialogues. A first version of the DistribDial / Replay component, a web site for accessing dialogues with such replays, is now completed.

III. CHINFADIAL: FRENCH-CHINESE SPOKEN DIALOGUE CORPORA

1. Collecting spontaneous bilingual dialogues

ERIM-Collect has been used in the ChinFaDial project for collecting French-Chinese interpreted spontaneous spoken dialogues in the hotel reservation domain. This project has been funded by LIAMA, a joint French-Chinese laboratory under both French INRIA, CAS and Chinese MOST supervision. Our partner in this project is the Chinese Information Processing group at the National Laboratory for Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CAS-IA).

About 12 hours of spontaneous translated spoken dialogues on "hotel information and reservation" in French and Chinese have been recorded thus far (size = 30Kbyte per second, about 1,3 Mbytes in total).

Participants to this collection effort are at this time:

Participants	Chinese	French	Total
Fr-Ch Interpreters	2	2	4
Interlocutors	3	3	6

There are 65 recorded dialogues with these characteristics:

Characteristics	Minimum	Average	Maximum
Duration (sec)	457	635	874
Number of turns	28	52	78
Turn length (sec)	4	12	57

2. Collecting and sharing multilingual dialogue corpora

A website and a small DistribDial server have been set up to freely distribute the sound files, their descriptors, and a Replay module. Our goal is to extend it to allow other groups to contribute to the site whatever annotations they may have created, and to share them under the same conditions (GPL).

Other data collection using ERIM-Collect will begin soon, thanks to AUF (the Association of Universities using French) who is funding a "VTH-FRA.Dial" project, to collect dialogues between French and Vietnamese, Tamil, and Hindi. We are also planning to distribute an ERIM-Collect "hardened" version on DistribDial, so that others can use it to do their own spoken dialogue collection.

IV. CURRENT DEVELOPMENT AND PROSPECTS

1. ERIM-Aid: machine aids to interpreting, to conversing in two languages, hence to collecting

In our "on demand interpretation" scenario, intermittent Web-posted interpreters may be asked to jump from one conversation to another, and thus from one topic to another. This conversation switching is likely to be quite difficult and stressful. Thus machine aids could be welcome. We also envisage providing machine aids for the conversational partners, to help them do without interpreters so far as possible, if necessary.

Currently implemented "communication aids" include facilities to:

- See and hear others (participants and interpreters).
- Share whiteboard data, possibly markable, "pointable".
- Access an agenda for scheduling rendezvous.

Possible "language aids," to both the human interpreter and the speakers, are of three kinds:

- Access to dictionaries via typed or voiced requests, and via automatic word spotting followed by filtering, dictionary look-up, and display in a dedicated window.
- Speech recognition, to alleviate difficulties of oral understanding when not using the interpreter, and to produce a log of the dialogue (which can additionally help an interpreter jump in), after possible reduction.
- Fully or partially automatic speech translation.

Most communication aids have been implemented at this time. The current scheduling agenda is global for an ERIM site, but each user handles it through a personalized view. Language aids are the next step. An interface to existing free dictionary resources on the PAPILLON site [11] should be available soon. A speech recognizer has been connected to the platform in another ERIM variant – the automatic interpretation setup ERIM-SpTra (automated Speech Translation), not presented here. This Speech-To-Text facility could here help as well to issue draft transcripts during the dialogue.

2. Use of the platform by volunteer interpreters

Data collection being time-expensive, our own goal is not to do too much of it for its own sake, but to get it as byproduct of some "mutualized" use of the platform, in the open source non-profit paradigm.

Professional interpreters are unlikely to help on a non-profit basis, since interpreting is their livelihood. Student interpreters, however, may find cooperation to be a good way of learning their trade.

At the 2008 Olympic Games in Beijing, for example, student interpreters could well be asked to aid bilingual communication in exchange for free tickets. Assume, for example, that a French speaker and a Chinese speaker want to converse. They could then go to a PC, activate ERIM-Interp or ERIM-Aid for French-Chinese, click on the icon of an available interpreter, and begin a conversation, which would be recorded by ERIM-Collect if participants agree and use the service free of charge.

We will also develop an ERIM-Learn variant platform (not described here) to help student interpreters to practice, while at the same time fostering large-scale data collection and improvement in the open source mode. Language students may also be interested in the training.

3. Unification of platform variants

We have started to integrate the three ERIM-platforms presented here, as well as the two other (ERIM-SpTra and ERIM-Learn) just mentioned. The platform independence and plug-and-play architecture of ERIM-Interp make this integration effort quite realistic.

CONCLUSION

We have presented several platforms in the ERIM family. Each platform can aid in the study of spontaneous cross-lingual communication on the Web. The core platform is ERIM-Interp for Web-based human interpretation. ERIM-Collect is used – and intended to be used – to alleviate the current scarcity of data, particularly open data, which can support the construction of speech translation systems.

We have then described ChinFaDial, a collection of spontaneous bilingual interpreted spoken dialogues for French-Chinese. This data, then the collecting software itself, will be distributed as shareware (GPL) on the DistribDial web site.

ERIM-Aid will add various machine aids for interpreters and conversational partners. We mentioned ERIM-Learn, a further extension of ERIM-Interp, which can serve as a Web-based language lab for student interpreters, while also providing valuable facilities for language learners.

We plan to continue research in the ERIM framework by collecting and distributing more data concerning more languages (Vietnamese, Tamil, Hindi to French). This data collection will be made possible by a unified version of the ERIM environment, offering all the functionalities of the ERIM variants. More specifically, we hope that sudent interpreters will volunteer to interpret and to train with ERIM, while users agree to "give their dialogues to science" in exchange of using ERIM-Interp for free.

ACKNOWLEDGEMENTS

This work has been supported by CLIPS-IMAG (UJF University Grenoble 1, CNRS, INPG) and funded by the LIAMA French-Chinese Laboratory (within the ChinFaDial project), by the Rhône-Alpes Region (ERIM project), and by Spoken Translation, Inc. (Berkeley) for ERIM-SpTra. Thanks to Zhai JianShe (Nanjing University) for early prototyping and to Julien Lamboley (INSA, Lyons) for steady recent development, to Brigitte Meillon (at CLIPS-MultiCom), and to members of the GETA and NLPR-CASIA-Beijing teams, for volunteering to participate in data collection and related experiments.

REFERENCES

[1] Blanchon H. (1994). Perspectives of DBMT for monolingual authors on the basis of LIDIA-1, an implemented mockup. COLING-94. Kyoto, 5-9 Aug. 1994, Y. Wilks ed., vol 1/2, pp. 115-119.

[2] Coutaz J., Salber D., Carraux E. & Portolan N. (1996). NEIMO, a Multiworkstation Utilisability Lab for Observing and Analyzing Multimodal Interaction. CHI'96 companion.

[3] Fafiotte G. & Zhai J. (1999). A Network-based Simulator for Speech Translation. NLPRS'99. Beijing, 5-7/11/99, B. Yuan, T. Huang & X. Tang ed., pp. 511-514.

[4] Furuse O., Sobashima Y., Takezama T. & Uratani N. (1994). Bilingual Corpus for Speech Translation. AAAI-94 Workshop on Integration of Natural Language and Speech Processing. Seattle, 31/7-1/8/94, ATR Interpreting Telecommunications.

[5] Loken-Kim K.-H., Yato F. & Morimoto T. (1994). A Simulation Environment for Multimodal Interpreting Telecommunications. IPSJ-AV Wokshop, March 94, 5p.

[6] Morimoto T., Takezawa T., Yato F., Sagayama S., Tashiro T., Nagata M. & al. (1993). ATR's Speech Translation System: ASURA. EuroSpeech'93. Berlin, 21-23/9/83, 4p.

[7] C-STAR web site: http://www.c-star.org

[8] DARPA: http://www.darpa.mil/ito/research/com/index.html http://fofoca.mitre.org/doc.html and GALAXY web site: http://www.sls.lcs.mit.edu/sls/whatwedo/architecture.html

[9] NESPOLE! web site: http://nespole.itc.it

[10] VERBMOBIL web site: http://verbmobil.dfki.de

[11] PAPILLON web site: http://www.papillon-dictionary.org