ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Speech Translation for French within the C-STAR II Consortium and Future Perspectives

Hervé BLANCHON
GETA, CLIPS , IMAG, BP 53
385 rue de la Bibliothèque
38041 Grenoble Cedex 9, France
herve.blanchon@imag.fr

Christian BOITET
GETA, CLIPS , IMAG, BP 53
385 rue de la Bibliothèque
38041 Grenoble Cedex 9, France
christian.boitet@imag.fr

Abstract

Despite joining the C-STAR II consortium in late 1996, the CLIPS ++ group succeeded in building the French parts of a multilingual task-oriented spoken dialogue translation system and took part in multilingual, intercontinental demonstrations held on July 22nd 1999 by CLIPS (France), CMU (United States), ETRI (South Korea), ATR (Japan), IRST (Italy), and UKA (Germany). The challenge was to reach the minimum quality level adequate for handling specific tasks, which is quite higher than what is sufficient for casual chatting and can be achieved by putting together commercially available components.

After presenting the modules and the architecture of our C-STAR II demonstrator, we evaluate the results, both externally and internally. While the reactions to the final demonstrations were very positive, and many said that these prototypes should quickly lead to products, we feel that there is still much room for improving the overall quality in significant ways. In the last part, we focus on future avenues of research to further improve the quality of task-oriented speech translation, in particular by defining a more powerful and orthogonal task-oriented semantic pivot, using the linguistic and dialogic context, and generating information usable by speech synthesis to generate better prosody.

Keywords

Speech Translation, C-STAR, task-oriented semantic "IF" pivot

10.21437/ICSLP.2000-837

Introduction

The CLIPS ++ group joined the C-STAR II[1] consortium as a partner in September 1996. Led by the CLIPS-IMAG laboratory (Grenoble, France), our group was composed of three other laboratories: LATL (Genève, Switzerland), LAIP (Lausanne, Switzerland), LIRMM (Montpellier, France).

We adopted the interlingua approach already pursued by four C-STAR II partners, where the interlingua, called IF (Interchange Format) is a task-oriented (specialized) semantic pivot. We thus developed four modules: French speech recognizer, French to IF analyzer, IF to French generator, and French speech synthesis. These modules cooperate between each other and with other partners' demonstrators through an integrator module.

Low quality, but useful, speech translation can be achieved by putting together off the shelf speech recognizers, MT systems and synthesizers. This "quick and dirty approach" in the context of casual, episodic conversations has been demonstrated [Seligman & al. 98] to establish a "base line". However, there are situations where a far better quality is required, those involving some urgency and the participation of at least one professional "agent".

The task-oriented semantic pivot approach and the heuristic linguistic programming techniques used in the project allow adding new languages quite easily, and the quality is obviously better. In the context of "quick and dirty" translation, there may be four or five repetitions of an utterance before it is correctly recognized or manually edited, so that its complete processing takes 20 to 30 seconds. On the other hand, with the C-STAR II demonstrators, we rarely experienced more than one repetition of an utterance, and the translation took 4 to 7 seconds, with almost no feeling of waiting because of integration in a multimedia videoconference setting.

However, we think that there is still much room for improving the overall quality of task-oriented speech translation systems in significant ways, in particular by defining a more powerful and orthogonal task-oriented semantic pivot ("IF"), using the linguistic and dialogic context, and generating information usable by speech synthesis to generate better prosody.

[1] See also [Levin & al. 2000, Park & al. 98, Sugaya & al. 99, Sumita & al. 99] for partners' works.

1. The CLIPS ++ demonstrator

1.1 Components

The IF (interface format) [Levin & al. 98] relies on dialogue acts, concepts, and arguments. Dialogue acts describe speaker's intention, goal, need. Concepts define the focus of the dialogue act. Several concepts may appear in one IF. Arguments are values of discourse variables. For the sentence "The week of the twelfth we have single and double rooms available" pronounced by an agent, the following IF should be built: *a:give-information+availability+room (room-type=(single ; double), time= (week, md12)).*
The global architecture for speech translation using the IF approach is thus the following:

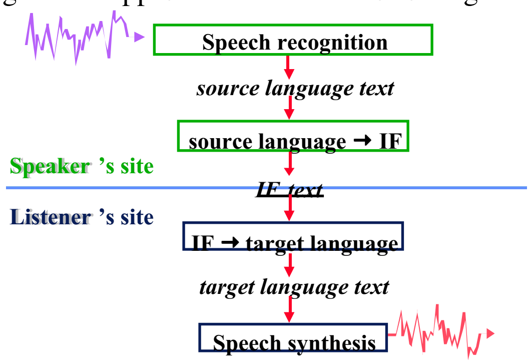


Figure 1: Overall components interaction

GEOD-CLIPS developed the French speech recognizer, GETA-CLIPS the French to IF analyzer, LATL the IF to French generator, and LAIP the French speech synthesis. LIRMM explored French-to-IF analysis with another methodology.

1.1.1 French speech recognizer

The module [Vaufreydaz & al. 99] is designed for speaker independent continuous speech recognition with a vocabulary specialized for the tourism domain of about 10k words. It is based on client-server architecture. A speech recognition server is used through "light" clients on the network. It is built with the JANUS-III toolbox of Carnegie Mellon University. It was implemented using:

- a context independent markovian acoustic model trained on 10 hours of continuous speech (BREF-80 corpus),
- a stochastic language model trained on a 140 million words corpus and optimized for the tourism task.

1.1.2 French-to-IF analyzer

This module [Blanchon & al. 2000, Boitet & Guilbaud 2000] is developed with Ariane-G5, a generator of machine translation systems supporting five specialized languages for linguistic programming, running under VM/ESA/CMS.
The input is an orthographic transcription of a spoken utterance. The following steps are performed one after the other:

- Morphological analysis and lemmatization of the words of the text,
- 1st access to transfer FR-IF dictionaries,
- Syntactical analysis for the recognition of semantically interesting structures: dates, quantity, numbers, prices, etc.
- 2nd access to transfer FR-IF dictionaries,
- Syntactic and morphological generation of the resulting IF.

1.1.3 IF-to-French generator

The IF-to-French module [Wehrli & Wehrle 98] was partly developed with GB-Gen, a broad lexical and syntactical coverage syntactic generation tool.
The transformation between an IF and a French text is made in three steps:

- Mapping of an IF into a GB-Gen semantic structure,
- Application of the GB-Gen generation procedures to produce a syntactic structure,
- Application of the GB-Gen morphological rules to produce a text in French.

1.1.4 French speech synthesizer

LAIPTTS [Keller 97, Keller & Zelner 98] is a "text to speech" rule-based synthesizer. Synthesis is made in three steps: text to phoneme mapping, prosody generation, signal generation.
Text to phoneme mapping uses general rules and specialized rules for numbers, abbreviations, fixed expressions, etc. It also uses 7000 general words and specialized dictionaries of proper nouns. Prosody generation uses psycholinguistic rules. Signal generation uses the MBROLA technique.

1.2 Architecture

1.2.1 Demonstrators integration

Two kinds of data are exchanged between the systems: video and sound for the videoconference, and data supporting the translation process.
Commercial products handle the videoconference. Data exchanged for the translation process itself are the IF structures (mandatory for the translation), and the recognition hypothesis and the generation from the locally produced IFs (for trace purposes). These exchanges are made through a communication server via the telnet protocol (cf. Figure 2).

1.2.2 Local components integration

All components of our demonstrator are servers. None of them communicates directly with another one, but they are all connected to a local communication server. We chose this architecture, because it is versatile and very convenient for distributed development.

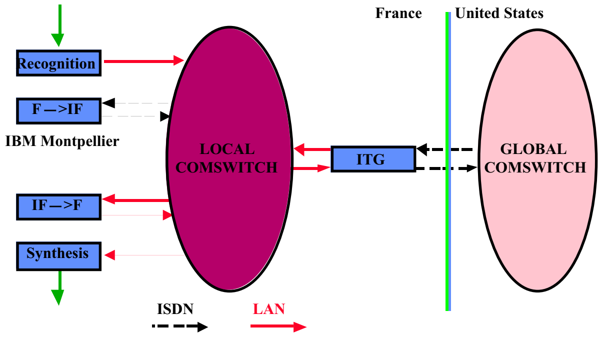


Figure 2: Global architecture of the demonstrator

We also believe that this architecture will be of interest for the foreseen commercial systems, It will be necessary to provide the customers with light clients interacting with very powerful servers which software will be updated for the benefit of all the users at the same time. This architecture fits also the needs of mobile applications.

1.3 Interface

The whole interface is distributed among two screens: one screen for the videoconference and one screen for the user interface.

In the client setting, the user interface screen is divided in two parts. On the left hand side there is the interface of the speech translation system. The right hand side is devoted to a web viewer. A picture of this interface is given in Annex.

In the agent setting, the user interface screen is also divided in two parts. On the left hand side there is the interface of the speech translation system. The right hand side is shared between a web viewer and a web selector allowing the travel agent to send web pages to the client.

2. Demonstrations

2.1 Settings

Several quadrilingual demonstrations were held on 22 July 1999 in the US, Japan, Germany, Korea, Italy and France. Our demonstration involved ETRI, UKA, and CMU. Our media coverage was quite good on the 22nd of July and after.

We also participated in 3 demonstrations with IRST hosted by ETRI who was present at Telecom'99 (Geneva). One of them was redirected live to IBM-France in Paris.

2.2 Scenario of a C-STAR II session

Playing a client, the following scenario was used in our July demo:

- Entering a virtual tourism agency with branches in the United-States, Korea and Germany, the client had first a greeting session with all the travel agents available,
- The client then planned successively trips to Taejon, New-York and Heidelberg booking a transportation (flight, or train) and a hotel, asking for tourist attraction and directions, paying with a credit card.
- For the demo purpose the client finally said thank you to the three travel agents.

For Telecom'99 we played the role of the agent with an enriched scenario.

2.3 Outcomes

The opening ceremony was well received and a very nice entry. Some mistakes were also entertaining. The demonstration lasted for about half an hour and the people in the public said that they did not see the time fly because of the variety of the situations.

The dialogue with the Korean agent was very appreciated (the Hangul script and the almost never heard language), picturing clearly the need for speech translation when there is a need for communication but no common language to support it.

In Europe, people are less sensitive to that matter as far as German and English are concerned. Most of the examples taken by the media for future application were about French-Korean and French-Japanese. We tried then to explain the need for that technology even if some communication is possible when the message has not to be distorted or misunderstood.

3. Evaluation and perspectives

Despite these successful experiments, quality, task-oriented speech translation is not yet ready to hit the market. In this section, we will comment on our demonstrator and discuss some ways to reach higher quality in the future.

3.1 Evaluation of our demonstrator

We implemented a purely sequential architecture with no shared information and very simple data structures exchanged between the components. This architecture is not better than the one used in the framework of the "quick and dirty" approach. In this sense, we do not take a real advantage of our deep knowledge of each module.

Also, no memory of the past is used, and no dialogue processing is integrated, even if we spent time designing dialogue models.

The IF was not fully covered both in analysis and generation. Because of its poor specification, it was learned using example databases. This was a time consuming task and slowed down our development.

The French-to-IF and IF-to-French modules run on remote machines. For speech recognition, a client process runs on the user's computer and the recognition server process runs on a remote machine. The speech signal is piped to the server, consuming a large bandwidth on the local network.

Higher quality can only be reached if we can design a more dialogue-oriented, integrated, interactive and tunable architecture.

3.2 Short terms goals

We envisage several ways to decisively improve the quality and usability of these systems and plan to work on some of them in the short and medium term in the framework of the NESPOLE! (NEgotiating though SPOken Language in E-commerce) European project[2], while the others are still long term goals.

Complete server-based architecture

For speech translation to be widely used, a more powerful server-based architecture should be used, so that the amount of specific software and hardware on the user side should be as small as possible.

In particular, acoustic, linguistic and task-related resources, which are very large and subject to frequent changes, should be stored only on the servers and not on each user's PC or NC. Users will then benefit of all updates on the fly.

For desktop applications, the most pressing problem is speech recognition. With the current networks, the speech signal consumes too much bandwidth. For connections over LAN preprocessed or compressed is probably the solution. We will try those ideas within NESPOLE!

For mobile application, over cellular phone, it will be necessary to handle low quality data. We will follow this direction within C-STAR III.

Context processing

Three points are targeted here: global context, dialogue context and linguistic context.

The global context contains at least the type of dialogue, the characteristics of the participants, in particular their names, sex[3],, ages and relative politeness level, their intentions if available, and perhaps the names of their locations, because they can be personified[4]. Human interpreters also need that kind of information.

The dialogue context should contain a representation of the past dialogue, the present stage of the dialogue if it follows some known script, and some predictions about the future. In the short term, much could already be achieved if the analyzer could access a sorted list of speech acts predicted by a suitable dialogue model.

The most necessary part of the linguistic context is the list of possible "centers", that is, possible referents for anaphoric elements or ellipses. Here is an example from French to German which illustrates this point:

(1a) Nous avons deux chambres, une sur cour avec WC et l'autre sur rue avec douche et WC[5].

…2 Zimmer,…

(1b) Pour aller à la gare, ne prenez pas la première rue à droite, mais la seconde[6].

…die erste Straße…

(2) D'accord, je prends la seconde[7].

Einverstanden, ich werde das/die zweite nehmen.

When translating (2), the gender will be neutral in case of (1a) and feminine in the case of (1b).

Because Ariane-G5 can use a relatively long fragments of text as a unit of translation, a practical solution to use these contexts is to let the integrator module send to the analyzer (resp. to the generator) a text containing the contexts and the result of speech recognition (resp. the IF).

Example of a possible input to analysis[8]:

```
<ctxt_glob> <speaker> client <client> Madame
Durand 70 years <agent> Herr Biedemeyer 52
years <firme> NTG <topic> hotel reservation
</ctxt_glob>
<ctxt_dial> <stage> central episode <past_sp
_acts> question-info info request <future_
sp_acts> yes no question-info </ctxt_dial>
<ctxt_ling> pension-hotel_NF réserver_VT
chambre_NF cour_NF réserver_VT pension-
régime_NF prendre_VT rue_NF </ctxt_ling>
<utterance> <alt> d'accord_/_encore je
prends_/_rends la seconde <alt> la cour prend
la seconde </utterance>
```

2 http://nespole.itc.it/

3 In German or Japanese, proper names must be used in greetings. "Bonjour, Monsieur!" is possible in French, but we cannot say "Guten Tag, (mein) Herr!" in German. "Guten Tag, Herr Müller" is necessary. And if a Japanese says "Smith-san", we must choose between "Mr Smith", "Mrs Smith", and "Ms Smith".

4 "But Taejon has just told me that…"

5 We have 2 rooms, one on the back with WC and the other on the street with shower and WC.

6 To go to the station, don't take the first street on the right, but the second.

7 OK, I'll take the second one.

8 We anticipate here on the section concerning tighter integration of components.

3.3 Medium-term goals

Better & richer IF

There is really a need for a more structured IF with a cleaner specification and more expressive power (cf. 3.2.1.). This will be a major task in the NESPOLE! project.

Prosody processing

We also would like the speech recognizer to generate prosodic marks which could be used by the analyzers and then be encoded in the IF.

Conversely, the prosodic marks contained in the IF expressions could be used by the generators in conjunction with other semantic and pragmatic (speaker's intention) features to produce outputs better suited to the situation, and containing marks or tags usable by the speech synthesizers to generate adequate prosody.

More user-system interaction & feedback

On the ergonomic side, two complementary approaches should be pursued:

- Adjustment to the user by the system (automatic tuning to user's profile).
- Adjustment to the system by the system (learnability, advice on how to make the speech easier to recognize and translate).

3.4 Long-term goals

Tighter integration between components

Direct integration of components is extremely difficult, and contradicts the quest for modularity and server-based architecture. The possibilities for improvement here are to:

- use richer interface data structures between components, such as tree lattices or tree charts,
- use common primary linguistic resources (lexical and grammatical data bases to generate the linguistic data for each component,
- improve the system architecture (pipe-line, agents, blackboard, whiteboard).

Real multimodality

Finally, a perhaps elusive goal is to develop true multimodal systems, perhaps by writing unimodal grammars for each channel, and multimodal rule packages on layers organized in a hierarchy, as in the following diagram. We do hope to propose some first answers within NESPOLE! on this point

Multimodal user-system interaction could then also be used to alleviate cognitive load, for example: light visual interactive disambiguation.

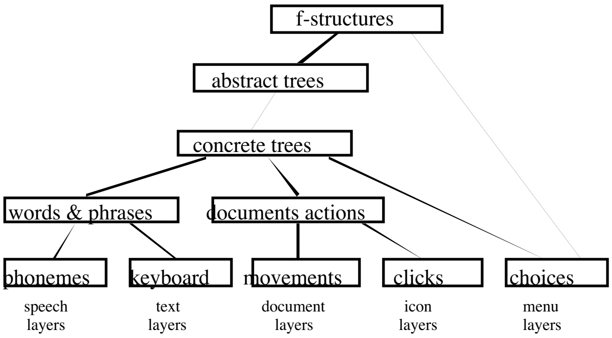


Figure 3: Layers' hierarchy

Conclusion

Thanks to the snowball effect of the consortium activity, we have been able built a reasonable demonstrator within a 2 years period. Apart form the technological and scientific goals we championed in the last part of this article, we would like to spend time on finding the niches for speech translation.

On top of letting us progress on the speech translation techniques, through its users group the NESPOLE! project is very good context to highlight good potential for commercial applications.

Acknowledgments

Thanks to many people at IBM France, who made it possible to reach acceptable response time for the French-to-IF module.

We are also very grateful to the IMAG institute, to our university, and to the CLIPS laboratory, for supporting this project through grants and direct help.

References

Blanchon H. & al. (1999) *Participation Francophone au Consortium C-STAR II*. La tribune des industries de la langue de du multimédia vol. 31-32 August-December 1999, pp. 15-23.

Boitet C. (1997) *GETA's MT methodology and its current development towards personal networking communication and speech translation in the context of the UNL and C-STAR projects*. Proc. PACLING-97, Ohme, 2-5 September 1997, Meisei University, pp. 23-57.

Boitet C. & Guilbaud J.-P. (2000) *Analysis into a formal task-oriented pivot without clear abstract semantic is best handled as usual translation*. Proc. ICSLP-2000.

Keller E. (1997) *Simplification of TTS architecture vs. Operational quality*. Proc. EUROSPEECH'97, September 1997, Rhodes, Greece.

Keller E. & Zellner B. (1998) *Motivations for the prosodic predictive chain*. Proc. ESCA Symposium on Speech Synthesis, Jenolan Caves, Australia, vol. 1/1, pp. 137-141.

Lavie & al. (2000) *The Janus-III translation system*. To appear in Machine Translation.

Levin & al. (1998) *An Interlingua Based on Domaine Actions for Machine Translation of Task-Oriented Dialogues*. Proc. ICSLP'98, 30th November - 4th December 1998, Sydney, Australia, vol. 4/7, pp. 1155-1158.

Park J. & al. (1998) Spontaneous Speech Translation System Development (in Korean). KSCSP vol. 15/1, pp. 281-286.

Seligman M. & al. (1998) Dictated Input for Broad-coverage Speech Translation. Proc. Association for Machine Translation in the Americas (AMTA-98), October 28, 1998, Langhorne, PA, USA.

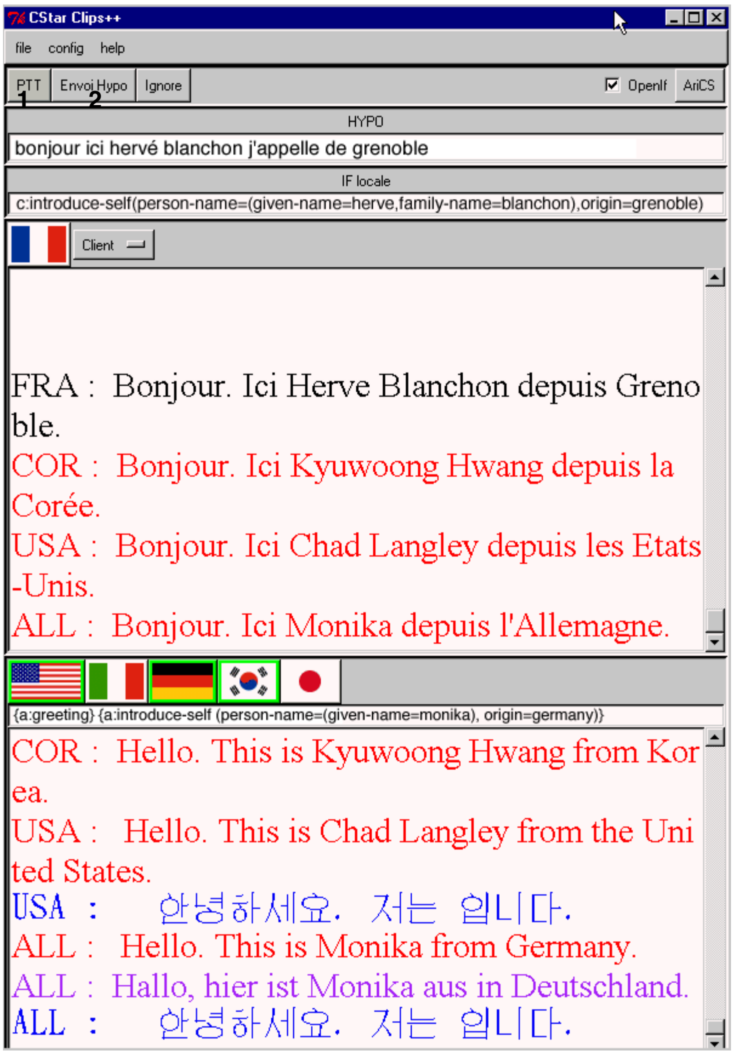
Sugaya & al. (1999) *End-to-End Evaluation in ATR-MATRIX Speech Translation Sytem between English and Japanese*. Proc. EUROSPEECH'99, September 5-9, 1999, Budapest, Hungary, vol. 6/6, pp. 2431-2434.

Sumita & al. (1999) *Solutions to Problems Problems Inherent in Spoken-language Translation: The ATR-MATRIX Approach*. Proc. MT-Summit VII, Singapore, September 13-15, 1999, vol. 1/1.

Vaufreydaz & al. (1999) *A Network Architecture for Building Application that Use Speech Recognition and/or Synthesis*. Proc. EUROSPEECH'99, Budapest, Hungary, September 5-9, 1999, vol. 5/6, pp. 2159-2162.

Wehrli E. & Wehrle E. (1998) *Overview of GBGen*. Proc. 9th International Workshop on Natural Language Generation, Niagara-on-the-lake, Canada, August 1998.

Annex: Interface of the demonstrator



1 2 3 4 5 6 7 8 9

Legend:

1) Put the speech recognizer into the wait for an spoken utterance

3) Speech recognition result *(Hello, here is Hervé Blanchon, I am calling from Grenoble)*

2) Send the speech recognition result to the French–IF module

4) Produced IF text `(c:introduce-self(person=(name=(given-name=herve, family-name=blanchon), origine=grenoble)`

5) Retrogeneration into French for the control *(Hello, here is Hervé Blanchon from Grenoble)*

6) Answer of the other participants prefixed by the origin of the answer *(Hello, here is Kyuwoong Hwang from Korea ; Hello, here is Chad langley from the US ; Hello, here is Monika from Germany)*

7) Available languages to be displayed

8) Last received IF reçue *(here from Germany:* `{a:greeting} {a:introduce-self(person-name=(givenname=monika),origine=germany)}`*)*

9) Generation into the other selected languages