# A way to integrate context processing in the MT component of spoken, task-oriented translation systems

Christian BOITET, Hervé BLANCHON & Jean-Philippe GUILBAUD
GETA, CLIPS, IMAG, BP 53
385 rue de la Bibliothèque
38041 Grenoble Cedex 9, France
Christian.Boitet@imag.fr Herve.Blanchon@imag.fr, Jean-Philippe.Guilbaud@imag.fr

## Abstract

Handling the context, be it global, dialogic or linguistic, seems necessary to increase translation quality in spoken dialogue translation systems aiming at casual users speaking spontaneously in task-oriented situations. Traditionally, the unit of translation is a text produced by speech recognition in analysis and some intermediate form in generation (as the IF in some CSTAR systems). We propose to enrich it by a structured textual representation of the context and sketch a way to use the context during analysis and generation, at the same time producing a modified context for handling subsequent utterances.

## Keywords

Spoken dialogue translation, context processing, compound text/context unit of translation

## Introduction

Rough quality MT of spoken dialogues is obtainable by simple linear composition of commercial speech recognizers, MT systems and speech synthesizers[1]. Turns are processed independently. However, a higher quality is needed in many situations implying some urgency, where professionals and naive users have to communicate in two or more languages.

To improve the overall quality of the system, several classes of improvements have to be considered: context processing, production of natural, system-initiated clarification dialogues, better use of prosodic cues in analysis, production of a more natural prosody, tighter components integration, multimodality processing, and better interfaces for user control and feed-back [Boitet 99]. As far as the MT component is concerned, using the context is the most promising way to improve quality.

We distinguish between three kinds of context, global, dialogic and linguistic, and show why they are crucial for translation, whether one uses a semantic pivot approach such as the CSTAR IF [Levin & al. 98)] or more language-dependent intermediate descriptors such as syntactic or semantic dependency trees of Example-Based MT or Transfer-Based MT systems.

We then study how they could be used in the MT parts of IF-based Spoken Language Translation systems (many of these ideas would also apply to transfer-based systems). It seems possible to rely on strongly simplified expressions of these contexts, e.g. lists for possible referents of anaphora or elisions (in analysis) or lists of previously used words or terms (to make lexical selection more coherent in generation), and speakers linguistic characteristics (male/female, name, title) for both analysis and generation.

Finally, we propose an XML format in which to pass to an analyzer both the context and the result of speech recognition, the latter being possibly a classical list of orthographic utterances or a list of word lattices, or some intermediate form. A similar format may be used to pass the context and the intermediate form to the generator.

This technique is inspired from MT systems for texts in which the unit of translation is not a sentence or phrase, but rather a whole paragraph, section or even chapter. We illustrate it by sketching how it could be implemented in Ariane-G5 [Boitet 97, Boitet & Guilbaud 2000].

## 1. Necessity of handling various types of context

It is generally admitted that context processing is necessary to achieve better "understanding" and translation of naturally produced utterances. However, when we looked for evidence in the CSTAR experiments logs, we actually found very few examples supporting

[1] Linguatech has launched the first commercial product, Talk & Translate™, in 2000. It uses IBM technology: Via voice for speech recognition and synthesis, and an improved version of LMT™ for translation. Dragon and Systran offer a less integrated product. NEC has been demonstrating its ST system since 1992, on a laptop at Telecom'99, but it was not yet commercial then.

this claim [Blanchon & Boitet 2000]. This is because the dialogues were heavily rehearsed, and the participants were not really producing spontaneous speech, but speech they knew would give good results.

For example, instead of answering: "Two" to "Do you want to stay one or two weeks?", the person playing the part of the client would answer "Two weeks", although it is less natural, in order to eliminate any problem resulting from not resolving the anaphoric reference. In this example, it is necessary to restore "day", "week" or "month" to translate correctly into Japanese (one cannot translate by 二 ni, but one should use 二日 futsuka, 二週間 nishuukan, or 二ッカ月 nikkagetsu).

The problem is also not very apparent in monolingual dialogues, because the participants solve most problems in an unconscious, form-based way. But, if we go through spontaneous bilingual translated dialogues [Fafiotte & Boitet 94, Park & al. 95] of if we restore the spontaneity in the above logs, we see that handling context-related problems appears quite important in order to raise the quality and naturalness of translation.

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

### 1.1. Global context

The global context contains at least:

- the general type of dialogue (reservation, enquiry, chatting, request for help…),
- the characteristics of the participants, in particular their names, sex, ages and relative politeness level,
- the roles of the participants (agent/client, doctor/patient, host/guest) and their current relation (unknown, friends, former teacher-student…),
- the names of their locations, beause they can be personified (as in "But Taejon has just told me that…").

That kind of information is also needed by human interpreters.

For example, 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.

If a Japanese says "Smith-san", in English we must choose between "Mr. Smith", "Mrs. Smith", and "Ms Smith". Hence the need for knowing the name and sex.

The age is also important to choose between a direct and a polite form (tu/vous in French): in France, "tu" will certainly be used in addressing a little child (under 10-11), and "vous" is obligatory for adults. (The convention is different for French Canadian, but this information belongs perhaps more to the linguistic context.).

The relative politeness level is very important in many languages, not only in Japanese and Korean. For example, English has many such forms although it has no tu/vous distinction:

- Give me a room with bath
- A room with bath, please
- Please give me a room with bath
- Could you give me a room with bath?
- Would you please leave a message for Mr Smith?
- Would you be so kind as to take a message?
- …

The participants roles and relation is often used to generate proper address forms: "Doctor, is this serious?", "I am sorry, my friend", "But, Mister X, I don't find it."…

### 1.2. Dialogue context

Producing the correct speech act and translating appropriately the dialogue connectors ("yes", "but", "I understand", "right"…) is very important [Tomokiyo 00].

If we use the IF approach, where the speech act is contained in the IF representation, the generators have the necessary information.

With other approaches, the linguistic structures generally don't contain it, so that generators should be given a pair (speech act, structure). With all approaches, the main problem is, during analysis, to compute the speech act from the utterance and the previous context.

In the long term, analyzers integrating linguistic and dialogue processing in a tight way might be developed. For the moment, it seems more realistic to develop separately the dialogue processor and to pass a representation of the dialogue context to the linguistic analyzer, which then must be extended to the use of the available context.

Ideally, 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.

In this framework, analyzers should produce not only their usual output (IF or linguistic structure), but also the identified speech act, if not explicit in the output.

### 1.3. Linguistic context

Utterances in natural dialogues contain many instances of anaphora and ellipsis. This considerably limits the output quality of the current analyzers, which handle utterances without information about the previous utterances even if the missing elements are present in a previous utterance of the same turn.

Context is also important for lexical disambiguation (e.g. "Je prendrai un express" -> "I will take an expresso/an express train") and for consistent lexical selection from an utterance to the next.

The most necessary part of the linguistic context seems to be the list of possible "centers", that is, possible referents for anaphoric elements or ellipses (main context words such as nouns and verbs). 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[2].

…2 Zimmer,…

(1b) Pour aller à la gare, ne prenez pas la première rue à droite, mais la seconde[3].

…die erste Straße…

(2) D'accord, je prends la seconde[4].

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).

## 2. A simple idea: unit of translation = context +utterance or IF

### 2.1. General schema

The idea we propose is simple: we now define the unit of translation as composed of a context, plus an utterance to be analyzed into one or more IF, or a (list of) IF to be generated. A dialogue processor (DP) is added for each language used in the multiparty conversation.

Every dialogue processor has access to the context of each translation unit. Figure 1 shows the general organization envisaged to handle the context.

When a speaker speaks in L1, the speech recognizer for L1 produces one or more hypotheses in a textual form (1, Fig. 1). The integrator component asks the dialogue processor of L1 to produce a context representation (Context-$A^{I}_{L1}$) and combines it with the output of SR to form the input to the analyzer of L1 (2, Fig. 2).

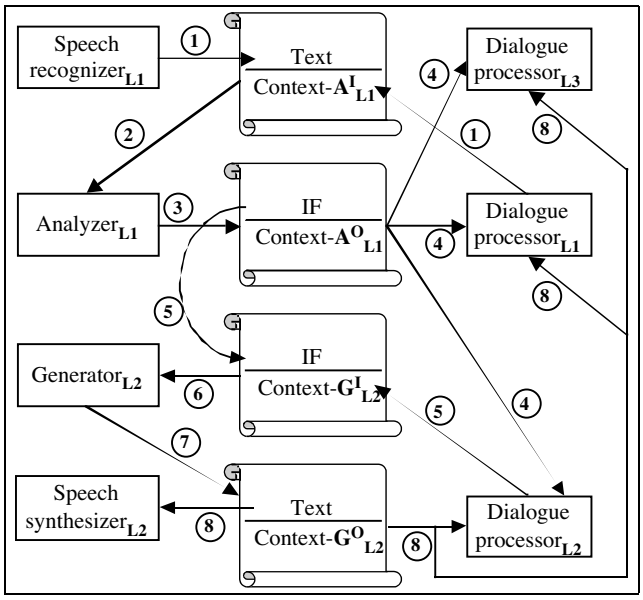


*Figure 1: General process organization*

The analyzer of L1 then produces both its usual output (here, an IF or a list of IF) and the new context (Context-$A^{O}_{L1}$), where the computed speech act(s) replace the list of predicted speech acts, and the list of possible centers has been updated by adding restored elided elements (3, Fig. 1).

The dialogue processors store the new context (4, Fig. 1). Of course, only $DP_{L1}$ stores the linguistic context (of L1).

For each target language, L2 for example, $DP_{L2}$ produces the context for the generation of L2 (Context-$G^{I}_{L2}$), keeping the same global and dialogic contexts as in Context-$A^{O}_{L1}$ (5, Fig.1). This new context is combined with the IF(s) to form the input to the generator of L2 (6, Fig. 1).

The generator of L2 produces both its usual output (or more, see below) and the new context (Context-$G^{O}_{L2}$), where the list of possible centers has been updated (7, Fig. 1).

The text output is sent to the speech synthesizer, while the DPs store the new context (only $DP_{L2}$ stores the linguistic context) (8, Fig. 1).

### 2.2. Possible representation as a tagged text

Rather than to exchange compiled data structures, we propose to combine the representation of the context and that of the utterance (output of SR, IF, or linguistic tree), as a tagged text, which be the real unit of translation submitted to the MT component, in analysis or generation.

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

[3] To go to the station, don't take the first street on the right, but the second one.

[4] OK, I'll take the second one.

Here is an example of a possible input to analysis for the utterance "D'accord, je prends la seconde", in a preliminary XML format showing the context and a hypothetical word lattice[5]. One could of course prefer to use more specific tags, or attributes.

```
<ctxt_glob> <speaker/> client
<client/> Madame Durand 70 years
<agent/> Herr Biedemeyer 52 years
<firm/> NTG
<topic/> hotel reservation
</ctxt_glob>
<ctxt_dial> <stage/> central episode
<past_sp_acts/> request-information
give-information request-action
<future_sp_acts/> accept reject
request-information
</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>
```

Graphical view of the utterance:

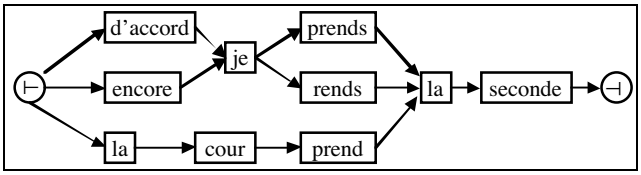


*Figure 2: Utterance word lattice*

## 3. Sketch of context processing in analysis and generation

### 3.1. Analysis

Let us sketch how an analyzer written in Ariane-G5 could handle the context passed in this way.

First, morphological analysis has to transform the whole tagged text into a "flat" decorated tree. This is quite easy: the XML tags and attributes will be put in the dictionary with appropriate attributes, and the linguistic elements will be analyzed as usual.

In the case of French, the current analyzer is built to use the morphosyntactic tags attached by the language model of the recognizer, if any, to reduce the ambiguities: for example, it will analyze TOUR_NM only as a masculine noun, although "tour" may be a masculine or feminine noun (a tour/a tower).

This tree will then be passed to the structural analysis phase, which will first structure the context in a subtree, and transform the lattice into a second subtree with a convenient form, e.g.

```
'TEXT' ('CTEXT' ('CTXTGLOB' (…),…), 'UTTERANCE' (
  'SOLUT' ('ALT' ('D"ACCORD_ADV' [attributes], '
    ENCORE_ADV' […], 'JE' […],
    'ALT' ('PRENDRE_VT' […], 'RENDRE_VT' […]),'
    ALT' ('IL' [pron, fem, sing dirobj…], 'LE' [art, fem, sing]),
    'ALT' ('SECOND_ADJ' […], 'SECONDE_NF' […])),
  'SOLUT' ('LE_ART' [def, fem, sing…], 'COUR_NF' […],
    'PRENDRE_VT' […],'ALT' ('IL' [pron, fem, sing dirobj…],
    'LE' [art, fem, sing…]), 'ALT' ('SECOND_ADJ' […],
    'SECONDE_NF […]))))).
```

Structural construction interleaved with disambiguation can then be performed almost as in the analysis of a connected text: local conditions apply first, then local ambiguities such as "prendre/rendre" may be solved using the context (PRENDRE_VT). The two (or more) interpretations can be constructed in parallel (as sister subtrees).

Suppose SECONDE_ADJ would be preferred over SECONDE_NF because SECONDE_NF is not a likely object of "prendre" (take). The same rule may be used (in the same grammar application) to find in the context subtree the most probable elided noun for all subtrees where SECONDE_ADJ appears: looking from right to left for a feminine noun, one would find RUE_NF (street) before CHAMBRE_NF (room) and restore the head of the nominal phrase accordingly.

For each possible interpretation, one has then to build a structure conformant with the IF (or with the conventions used for intermediate linguistic trees). The global and the dialogic contexts may be used in the process.

The analyzer then chooses the "best" interpretation, and erases the others. When this is done, it can modify the context, by adding to the linguistic "centers" the new words confirmed (from initial the word lattice), in order, and appending the computed speech act to the list of past speech acts.

If the analyzer is not able to choose an interpretation because several possible ones remain, there may be two solutions to solve the ambiguity: disambiguation by the speaker or disambiguation by the listener.

[5] This illustrates the idea to increase quality through tighter integration of recognition and analysis.

With the first approach, a question may be asked to the speaker of the turn to let him choose the good interpretation [Blanchon & Fais 97]. This is what a human interpreter would do. Actually, a previous study by [Oviatt & Cohen 1991] has shown that up to 30% of turns in bilingual dialogues interpreted by professional interpreters are clarification dialogues between the interpreter and a participant.

In the second approach, one interpretation may be randomly selected and synthesized for the listener with a warning that other interpretations of the speaker's turn are available. If the listener is not happy with the synthesized message, s/he may (iteratively) ask to hear the next one ("next one please!") before asking the speaker to clarify its turn. The practicality of this idea is demonstrated by the use of oral commands in dictation systems.

### 3.2. Generation

As said earlier, the speech act of the utterance to be generated is directly available in the input in the case of the IF approach. Otherwise, the generator should retrieve it from the context available in the translation unite.

Other parts of the context may then be used to generate the correct address forms, the adequate politeness level, and possibly auxiliary nodes or attributes later transformed into prosodic marks used by the speech synthesizer in order to produce a more natural prosody.

The linguistic context might be used as a kind of memory, to guide lexical selection. For example, the client may speak of a "pension" and the agent may answer with an utterance containing "hotel". Both words map to the same concept in the IF. Knowing from the linguistic context that PENSION_NF has appeared more recently than HOTEL_NF, and from the dialogue context that the agent is speaking, the French generator might generate PENSION_NF, thereby politely using the same term as the client.

Near the end of the syntactic generation, the generator should update the linguistic context by appending to it the list of used content words and removing duplicates to the left if deemed useful.

Then, morphological generation can produce a result in the form of a tagged text in the same format as the input to analysis (context, then utterance).

One could actually generate more than one textual form, simply using new tags to delimit them. For example, one form could be the normal orthographic form, and another could contain prosodic marks to enter the speech synthesizer at a more internal level.

## Conclusion

We have shown the importance of handling context for building higher quality speech translation systems for situations where participants produce truly spontaneous speech. We have also outlined a possible architecture that necessitates no modification of the speech recognizers, but the addition of a separate dialogue processor, and the extension of the analyzers and generators to handle units composed of XML-tagged textual representations of the utterance and of the context.

Although the current funded projects in Spoken Language Translation (such as Nespole!) are really more oriented towards development than research, the most important questions for them being robustness, speed and extensibility, we hope to experiment with these new ideas in parallel "research tracks", possibly in the context of the CSTAR-III consortium.

## References


[Blanchon & Boitet 2000] Speech Translation for French within the C-STAR II Consortium and Possible Enhancements, ICSLP'2000, Beijing.

[Blanchon & Fais 1997] Asking Users About what They Mean: Two experiments and Results. Proc. HCI'97, San Francisco, 24-29 August 1997, Elsevier, pp. 609-612.

[Boitet & Guilbaud 2000] Analysis into a formal task-oriented pivot without clear abstract semantics is best handled as "usual" translation, ICSLP'2000, Beijing.

[Boitet 1997] GETA's 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 Sept. 1997, pp. 23-57.

[Boitet 1999] On the need for different architectures for chat translation and for task-oriented translation of spoken dialogues, CSTAR-II Schwetzingen seminar, Sept. 1999.

[Fafiotte & Boitet 1996] An Analysis of the First EMMI-Based Experiments on Interactive Disambiguation in the Context of Automated Interpreting Telecommunications. Proc. MIDDIM-96, le Col de Porte, 12-14 August 1996, GETA, pp. 224-237.

[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.

[Oviatt & Cohen 1991] Discourse structure and performance efficiency in interactive and non interactive spoken modalities. In Computer Speech and Language 5(4), pp. 297-326.

[Park & al. 1995] Transcription of Collected Dialogues in a Telephone and Multimedia/Multimodal WOZ Experiment. TR-IT-0090, ATR-ITL, Kyoto, February 95, 123 p.

[Tomokiyo 2000] Discursive analysis of task-oriented spoken dialogues in Japanese, French and English dialogues — Analyse discursive de dialogues oraux finalisés en français, japonais et anglais : élaboration et validation par analyse de corpus réels, doctoral thesis, Paris VII, June 2000, 275p.