Developing Non-European Translation Pairs in a Medium-Vocabulary Medical Speech Translation System

Pierrette Bouillon[1], Sonia Halimi[1], Yukie Nakao[2], Kyoko Kanzaki[3], Hitoshi Isahara[3], Nikos Tsourakis[1], Marianne Starlander[1], Beth Ann Hockey[4], Manny Rayner[1]

(1) Geneva University, ISSCO/TIM, Boulevard du Pont-d'Arve, CH-1211 Genève 4, Switzerland

(2) LINA, Nantes University, 2, rue de la Houssinière, BP 92208 44322 Nantes Cedex 03, France

(3) NICT, 3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan 619-0289

(4) UCSC UARC, Mail Stop 19-26, NASA Ames Research Center, Moffet Field, CA 94035, USA

Abstract

We describe recent work on MedSLT, a medium-vocabulary interlingua-based medical speech translation system, focussing on issues that arise when handling languages of which the grammar engineer has little or no knowledge. We describe how we can systematically create and maintain multiple forms of grammars, lexica and interlingual representations, with some versions being used by language informants, and some by grammar engineers. In particular, we describe the advantages of structuring the interlingua definition as a simple semantic grammar, which includes a human-readable surface form. We show how this allows us to rationalise the process of evaluating translations between languages lacking common speakers. The grammar-based interlingua definition can also be used in other ways. We describe two applications: a simple generic tool for debugging to-interlingua translation rules, and a method for improving speech understanding performance by rescoring N-best speech hypothesis lists. Examples presented focus on the concrete case of translation between Japanese and Arabic in both directions.

1. Introduction

This paper presents recent work on MedSLT, a medium-vocabulary speech translation system intended to support medical diagnosis dialogues between a doctor and a patient who do not share a common language. The topic of conversation is assumed to be limited to a specific medical subdomain, defined by a related set of symptoms. Typical examples are headaches or chest pains. The architecture has been designed with the following key goals in mind:

1. Given the safety-critical nature of the task, precision is more important than recall.
2. It should be easy to adapt the core system to new languages and domains.
3. The user should be able to adapt to the limitations of the system's coverage with a minimum of training, and should not experience these limitations as arbitrary.

The first goal has disposed us towards an architecture that is primarily rule-based, and thus more readily predictable in terms of function, though we also use statistical tuning methods to increase efficiency. Speech recognition uses the Nuance platform, equipped with grammar-based language models. One of the system's distinguishing characteristics, compared to related work, is that all grammars used (for recognition, analysis and generation) are compiled from a small number of general linguistically motivated unification grammars, using the Open Source Regulus platform (Rayner et al, 2006). Early versions of the system used a single core grammar per language; more recent ones have gone further, and merged together grammars for closely related languages (Bouillon et al, 2007b). These core grammars are automatically specialized, using corpus-driven methods based on small corpora, to derive simpler grammars. Specialization will typically be along all of the following dimensions: task (recognition, analysis, generation), subdomain (headache, chest pain, etc), and context (doctor question, patient response). The specialization process uses the Explanation Based Learning algorithm. It starts with a parsed treebank derived from the training corpus, and then divides the parse tree created from each training example into a set of one or more subtrees, following a set of domain- and grammar-specific rules conventionally known in the Machine Learning literature as *operationality criteria*. The rules in each subtree are then combined, using the unification operation, into a single rule. The set of all such rules constitutes a specialized unification grammar.

Each of these specialized unification grammars is then subjected to a second compilation step, which converts it into its executable form. For analysis and generation, this form is a standard parser or generator. For recognition, it is a semantically annotated CFG grammar in the form required by the Nuance engine, which is then subjected to further Nuance-specific compilation steps to derive a speech recognition package. These final compilation steps include a second use of the training corpus to perform statistical tuning of the language model. The overall goal of the Regulus architecture is to simplify the normally very onerous task of writing and maintaining a large number of closely related grammars, retaining internal coherence between them. In particular, coherence between the recognition and analysis grammars guarantees that any spoken expression which is accepted by the recognizer can also be parsed.

Although performance of rule-based recognition systems is typically good on in-grammar coverage, a well-known problem is brittleness: users need to know what language the grammar supports. Our approach to this problem is to equip the system with an intelligent help module (Starlander et al, 2005) which after each utterance provides the user with in-coverage examples, chosen to be as close to the user's actual utterance as possible. The help module's output is based on a library of utterances which have already been evaluated as being within

grammar coverage and producing correct translations. At runtime, the system carries out a second round of recognition using a backup statistical recognizer, and uses the result to select examples from the library which are similar to the statistical recognizer's result in terms of a backed-off N-gram metric. (Chatzichrisafis et al, 2006) describes an experiment in which medical students with no previous exposure to MedSLT used it to perform a diagnosis task on simulated patients, acquiring all their knowledge of grammar coverage from the help module. Post-experiment debriefing showed that, even though the subjects often felt that they were unable to ask questions in the way they would ideally prefer, they also usually thought that the help functionality allowed them to find an alternate phrasing within grammar coverage.

The semantic representations used by all grammars are feature-value lists, flat except for one optional level of nesting used to represent subordinate clauses and similar constructions. Translation is interlingua-based, where interlingua representations are of the same form as those produced and consumed by the source and target language grammars; these interlingua representations are essentially canonical versions of English-language semantic representations. The rules which translate source language expressions into interlingua, and interlingua into target language expressions, are formulated as (optionally conditional) rewritings of lists of feature-value pairs to feature-value pairs.

To give a simple example showing processing flow, suppose that the system has been configured for translation from English to French, and that the user has spoken the sentence "is the pain occipital?" The source-language speech recognizer, which contains a compiled form of the English source-language grammar, decodes the input waveform into words, and simultaneous produces the semantic representation

```
[[utterance_type,ynq], [symptom,pain],
 [verb,be], [tense,present],
 [voice,active], [adj,occipital]]
```

This is next mapped into the interlingua representation

```
[[utterance_type,ynq], [symptom,pain],
 [verb,be], [tense,present],
 [voice,active],[prep,in_loc],
 [part,back], [body_part,head]]
```

which is identical to an English representation of the sentence "is the pain in the back of the head". This interlingua form is then mapped into the French target-language representation

```
[[utterance_type,sentence],
 [pronoun,vous],
 [path_proc,avoir], [tense,present],
 [voice,active], [symptom,mal],
 [locative_prep,à],[part,arrière],
 [body_part,tête]]
```

Finally, the French target-language grammar realizes this representation as the French surface form "avez-vous mal à l'arrière de la tête?" and passes it to a TTS engine for realisation in spoken form.

In previous work, we have presented initial results for several languages, including Japanese (Rayner et al 2005c; Nakao et al, 2006) and Arabic (Bouillon et al 2007a). The current paper focuses on enhancements recently added to the platform, which aim to simplify the task of developing functionality in these and other non-European languages. In Section 2, we describe how we have systematically defined "gloss" forms for grammars and semantic representations, to facilitate multiple views of these resources catering to the different requirements of native speaker informants and language engineers. The next three sections present concrete ways in which we have used these resources: Section 3 discusses rationalization of translation evaluation, Section 4 a generic tool we have developed which facilitates the construction of rules which translate from the source language into the interlingua, and Section 5 use of the interlingua to improve speech translation performance. Finally, Section 6 briefly describes the current MedSLT demo system, and Section 7 concludes.

2. Systematic use of gloss forms

Extending a multilingual system like MedSLT to a new language involves the construction of language resources such as grammars, lexica and translation rules. This normally requires collaboration between a language engineer and a native speaker informant, each of whom will possess knowledge that the other lacks. The language engineer will typically find it difficult to read sentences in the new language; the native speaker informant will find it difficult to understand data-structures they may need to examine, in particular expressions in the interlingua. This can often act as a major brake on development. In this section, we describe the solutions we have developed to address these problems, which are based on the idea of using grammars and macros to define multiple "gloss" forms. We begin with the case of non-European languages, and then consider the interlingua. Tables 3 and 4 show examples of all the gloss forms we will be discussing.

2.1 Gloss forms for non-European languages

During the development process, the linguistic informant is responsible for development of the corpus and the lexicon, and evaluation of translation quality. The language engineer uses these resources, and the informant’s linguistic intuitions, to construct the grammar and other more structured elements of the system. In our project, most of the language engineers are unable to read non-European scripts, and prefer to work exclusively in a Roman alphabet. In contrast, the informant often finds it unnatural to use a Romanized version of their language. The problem is especially acute for Arabic, where there is not even an accepted standard Romanized form.

Our approach to these problems has been to construct the system so that it can be easily be reconfigured to use different character sets. For languages with non-European scripts, we parameterize the lexicon, using macros, so that each word specifies both its native script and Romanized form; we also add a third, "gloss" form. For example, the entry for the Arabic word *tahus* (feel) is as follows (Regulus macro invocations are introduced with the @ operator):

```
v:[sem=[[state, tahus_bi],
        [tense, present]],
   subcat=pp, agr=2/\sing/\masc,
   vform=finite, subj_np_type=agent,
   obj_np_type=symptom]
  -->
  @a('تحس', tahus, feel).
```

By supplying different script-specific definitions of the macros, the base grammar can readily be compiled in three different versions, all strictly equivalent.

The non-trivial part of the scheme is arranging things so that the grammar specialization process is driven from a single corpus for one of the script alternatives, with the specialized grammars for the other versions derived from it in such a way that the specialized grammars are also strictly equivalent. The central idea is to tag the parses in the treebank with sufficient information that any parse in one script can be mapped onto a corresponding parse in the others; this is done by annotating each node with a unique label, which identifies the rule or lexical item attached to the node using the rule's source file and line number. Parses in the different variant grammars will attach the same file and line information to each node, and differ only with respect to surface lexical items.

The first step in the grammar specialization process is to parse the training corpus in one of the grammar variants; in order to make this corpus directly accessible to the grammar engineer, it is usually most convenient to use the Romanized variant. The result is a treebank of analyses. This treebank is then transformed into a reduced version, by removing all the specific surface lexical items from the trees and replacing them with uninstantiated variables. This reduced version, which identifies rules and lexical items only by their position in the source files, can then be fed into the EBL specialization process and used to train the different variants (Romanized script, native script and gloss) of the specialized grammar. For Japanese and Arabic, the two languages where we have so far implemented this scheme, building the two extra versions of the specialized grammar adds only a modest overhead of about 10% to the system build time.

2.2 Gloss forms for interlingua expressions

In Section 1, we saw examples of interlingua expressions. Language engineers, particularly ones conversant either with Prolog syntax or with formal semantics, generally have little difficulty reading these. Our experience, however, is that most native speaker informants find the interlingua unpleasant to deal with directly, and there is thus a strong motivation to develop a human-readable "gloss" form of the interlingua. On general grounds, it is also desirable to have a clear operational definition of what constitutes a well-formed interlingua expression. Taken together, these two considerations suggest the idea of casting the definition of well-formedness in the interlingua as a grammar. Given an interlingua expression *E*, we can say that *E* is well-formed if and only if the "interlingua grammar" can generate a surface string from *E*; if the grammar is designed with a little care, this string can moreover function as a human-readable gloss. Tables 3 and 4 show examples of these glosses.

The interlingua grammar is not obliged to take account of the complex surface syntax phenomena characteristic of real languages (movement, agreement, etc). There is moreover no reason to attempt to structure it in a general way consistent with any linguistic theory, since its central purpose is to define a semantics for a specific domain. It is thus possible for the interlingua to be defined by a small, tightly constrained semantic grammar. As we will see later in the paper, this also confers other computational advantages.

We anticipate that many readers may have an instinctive aversion to the idea of using a semantic grammar. Semantic grammars are indeed a widely abused notion, and it is important to consider why their use here is more principled than may at first appear. The main problem with most normal uses of semantic grammars is that they do not just describe semantic content, but also encode surface syntactic constraints; it is not surprising that this usually leads to difficulties, since the principle of autonomy of syntax strongly suggests that syntactic constraints are best captured independently of domain-specific semantic concepts. Our interlingua specification grammar, in contrast, is a *genuine* semantic grammar, which only needs to be associated with a simple artificial syntax.

We also stress that the grammars used to define the source and target languages are *not* semantic grammars. For example, the general English Regulus grammar (Rayner et al 2006, Chapter 8) is a complex, linguistically motivated feature grammar, whose non-terminals and features represent standard linguistic concepts such as S, NP, agreement, gapping and so on. A typical rule is the following (presented in simplified form; the details are not important),

```
np:[sem= @np-d-nbar-sem(Det, N),
    agr=3, agr=Agr, wh=Wh,
    sem_n_type=Type, conj=n,
    gapsin=GIn, gapsout=GIn,
    pronoun=n,
    @takes_pps(PPs)] -->
  d:[sem=Det, agr=Agr, wh=Wh],
  noun:[sem=N, agr=Agr,
        sem_n_type=Type,
        @takes_pps(PPs)].
```

which defines the general structure of a simple NP consisting of a determiner and a noun. In contrast, the interlingua grammar has a far simpler structure (in particular, it has far few features), and its non-terminals are semantic concepts such as SYMPTOM, LOCATION and BODY_PART. For example, consider the following

Interlingua rule, also presented in a slightly simplified form for expository purposes:

```
location:[sem=concat(BP, P, S)] -->
  body_part:[sem=BP],
  ?part:[sem=P],
  ?side:[sem=S].
```

This gives one possible realization of a LOCATION representation, as consisting of the concatenation of a BODY_PART (e.g. "head") with an optional PART (e.g. "front") and SIDE (e.g. "left").
In the following three sections, we will describe ways in which we have made concrete use of interlingua and interlingua gloss forms.

3. Evaluating translation quality

Systematic development of a rule-based translation requires frequent regression testing, which in turn implies an ability to judge correctness or otherwise of translations. When dealing with unusual language pairs, an important practical problem is the fact that it is hard to find informants who speak both languages and are able to evaluate translation quality. This has for example been the case with the Japanese/Arabic pair in MedSLT.
We have found in practice that the interlingua gloss form is adequate to support the task of translation evaluation; informants who are uncomfortable with direct inspection of the interlingua rapidly gain enough familiarity with the surface form to be able to use it with confidence. This means that it is possible to split the process of translation judging into two halves. A source language speaker judges translation from the source language to the interlingua gloss form, and a target language speaker judges translation from the interlingua gloss form to the target language.
Thus, if we for example want to judge whether *ago made hirogari masu ka* (Japanese) is a good translation of *hal yamtad al alam ila al fak* (Arabic), we present the Arabic-speaker with the Arabic→ Interlingua-gloss pair

هل يمتد الألم إلى الفك
→
YN-QUESTION pain radiate jaw PRESENT ACTIVE

and the Japanese-speaker with the Interlingua-gloss → Japanese pair

YN-QUESTION pain radiate jaw PRESENT ACTIVE
→
あごまで広がりますか

We judge the translation as correct if and only if both halves are judged correct.
Other things being equal, it is clear that being able to split judging into two independent pieces in this way wins in terms of efficiency and modularity. What is not *a priori* clear, however, is whether it really amounts to the same thing as direct judging of translation from source to target. For example, one might reasonably argue that splitting the evaluation into two pieces could yield misleading results, on the grounds that the to-interlingua half will in general be oriented too much towards the source language, and the from-interlingua half towards the target. We can advance various counter-arguments: for example, an important difference between MedSLT and many text translation systems is that we are only really interested in preserving correct meaning, and in fact in many cases intentionally choose a paraphrase rather than an exact translation.

In order to evaluate these competing arguments in a concrete setting, we thought that a reasonable test would be to perform a direct evaluation of the quality of translation in French → Japanese, a language-pair where we did have access to a few informants who could speak both languages. We could then compare the result of this evaluation with the implicit evaluation resulting from composition of evaluations for French → Interlingua and Interlingua → Japanese. We consequently ran the system in offline mode for the French → Japanese pair, to produce translations for 507 French sentences. The relevant French → Interlingua translations were judged by a French native speaker, and the French → Japanese and Interlingua → Japanese translations by a Japanese native speaker fluent in French, who had not previously been involved in the project. As in previous evaluations, translations were judged as "good", "ok" (acceptable but not perfect) and "bad". Finally, we compared the direct and composed translations.
Of the 496 utterances which produced translations, 69 received different judgements on the two methods. However, all but 9 of these judgements represented divergences between "good" and "ok", which in general tends to be a fairly subjective choice. Of the 9 seriously divergent sentences, one was determined to be a simple clerical error, and the other 8 represented issues involving granularity of the level of definition in the interlingua. The most important case (three occurrences) was caused by the fact that the French expressions *depuis combien de temps* ("since when"), and *pendant combien de temps* ("for how long") had incorrectly been mapped onto the same Interlingua expression. This was adequate for translation into English, since both can be translated as "how long"; however, Japanese, like French, requires the two concepts to be realized differently. Our overall conclusion was that the composed evaluation essentially gave almost the same results as the direct one, and that the few problem cases resulted from minor shortcomings in the interlingua which were easy to correct. This is comforting, since it is extremely difficult to find evaluators for at least half of the language-pairs covered by the system.

4. Debugging translation into interlingua

As already described, translation in MedSLT proceeds in two stages: source language to interlingua, and interlingua to target language. In practice, the second of these is by far the easier to debug, as the target language

informant can usually determine without difficulty whether a proposed translation into the target language is correct. Translation into the interlingua is more problematic. In earlier versions of the system, most decisions on correctness of translation into interlingua required subsequent translation into a target language. If this failed, it was often difficult to determine which of the two translation steps contained the problem.

Introduction of a well-defined grammar-based interlingua has greatly improved this situation. It is now trivial to decide whether or not an interlingua expression is well-formed. Rather to our surprise, we have also discovered that it is easy to provide automatic help for many cases where ill-formed interlingual expressions are generated by the source-to-interlingua translation step. The generator derived from the interlingua grammar turns out to be simple enough that it is computationally feasible to attempt generation from partially instantiated semantic forms, something that is definitely not the case with normal generation grammars. Our solution is based on the idea of exploiting this fact.

When in debugging mode, the system can be set to react to production of ill-formed interlingua by creating multiple variants of the bad interlingua form, and then attempting to generate surface interlingua expressions from them. Currently, we create three types of variants, formed by deletion of each element of the interlingua expression in turn, addition of an extra uninstantiated element, and substitution of each element in turn with an uninstantiated element; uninstantiated elements are instantiated during generation. The following is a typical example. Suppose that a French-to-Interlingua rule incorrectly maps the French source-language elements

```
[[duration_prep, pendant],
 [temporal, nuit]]
```

(representing *pendant la nuit*) into the Interlingua elements

```
[[prep, at_time],
 [time, night]]
```

This is a plausible mapping, but, as it happens, the current interlingua definition normalizes temporal prepositions, and the correct target should actually be

```
[[prep, in_time],
 [time, night]]
```

The result is that translation of an example like *avez-vous mal pendant la nuit* ("do you have pain at night") produces the ill-formed Interlingua expression.

```
[[utterance_type,ynq], [pronoun,you],
 [state,have_symptom], [symptom,pain],
 [tense,present], [voice,active],
 [prep,at_time], [time,night]]
```

When Interlingua debugging is turned on, the translation engine systematically goes through all possible deletions, additions and substitutions of single elements, using uninstantiated elements for additions and substitutions, and then attempts to generate interlingua forms from each of these. Since the Interlingua grammar is so tightly constrained, the process is very fast; it takes less than a second. The result is the following warning:

```
INTERLINGUA REPRESENTATION FAILED
STRUCTURE CHECK. APPLYING MODIFICATION
([prep, at_time] --> [prep, after_time])
GIVES
"YN-QUESTION you have pain after night
PRESENT ACTIVE"
```

This warning is not sufficiently informative that it tells the rule-writer exactly which change to make in order to fix the bug. It does however strongly suggest that the bug is in the rule producing the element `[prep, at_time]`, as opposed to one of the five other rules used in the translation. In practice, we have found that the interlingua debugging tool provides useful feedback on interlingua problems in over 75% of all cases, and has greatly speeded up the process of developing translation rules.

5. Using interlingua to improve speech understanding performance

The fact that the interlingua grammar gives us a tight and efficient operational definition of well-formedness for interlingua expressions can also be used to improve speech understanding performance. Even though the source-language grammars have been carefully tuned, it is extremely difficult to constrain them to the point where all utterances within grammar coverage are also meaningful in the context of the medical translation task. This means that it is quite often possible that recognized utterances can fail to produce any translation.

The interlingua grammar, however, allows us to add additional constraints, thus in effect improving the language model for the source language. We set the source language recognizer to produce multiple recognition hypotheses, which are first ranked in order of plausibility according to the confidence score assigned by the recognizer. We then select the highest-ranked hypothesis in this list which produces a well-formed interlingua representation. By construction, this can only improve speech translation performance compared to the simpler strategy of always choosing the hypothesis with the best recognizer score; the cases where N-best processing produces different results from 1-best are precisely those in which 1-best processing would give rise to ill-formed interlingua, and hence to no translation. Figure 1 shows a typical example where N-best processing was able to improve performance. It is important to note that the practical utility of the scheme depends on the fact that well-formedness of an interlingua representation can be checked quickly. It

would have been possible to achieve the same result by processing each hypothesis all the way up to generation of a target translation, and picking the first one which produced a non-trivial result; the overhead, however, compared to the interlingua-based approach, would have been more than an order of magnitude greater.

Actual utterance:
"avez vous mal au niveau des yeux"
N-best hypotheses:
1: "a elle mal au niveau des yeux"
2: "avez vous mal au niveau des yeux"
3: "elle a mal au niveau des yeux"
4: "a elle mal au dessus des yeux"
5: "une fois au niveau des yeux"
6: "un mal"

Figure 1. French example showing how the combination of N-best recognition and interlingua filtering can improve speech translation performance. All hypotheses are within grammar coverage, but the second one is selected, since the first fails to produce well-formed interlingua.

In a preliminary experiment designed to estimate the improvement in performance resulting from interlingua-based N-best rescoring, we ran recorded and transcribed speech data for three langages through offline versions of the system, using the French data from (Chatzichrisafis et al 2006), the English data from (Rayner et al 2005b) and the Japanese data from (Rayner et al 2005c). In each case, we judged examples by hand to determine which ones produced correct interlingua representations. Table 2 presents the results. As usual for a grammar-based speech application, we give figures both for the full set of utterances and for the subset consisting only of utterances inside grammar coverage; in view of the fact that all versions of MedSLT include an intelligent help component which actively steers users towards the grammar coverage, performance on this subset is arguably the figure which most accurately reflects the user's intuitive estimation of system performance. It is encouraging to see that the results across the three languages are very similar. In each case, addition of N-best rescoring yields a relative decrease in the proportion of utterances producing incorrect interlingua by about 9% for all the data, and about 20% for the in-coverage data.

6. The MedSLT demo system

The current version of MedSLT covers 6 languages (English, French, Spanish, Catalan, Japanese and Arabic), and uses vocabularies of about 400 to a little more than 1000 surface words per language, with the variation mainly depending on the degree to which the language uses inflection. Word error rates for speech recognition on in-coverage examples range from 3% to 8%, again depending on the language.

7. Summary and conclusion

In the context of the Regulus grammar-based MedSLT spoken translation system, this paper has discussed the advantages for development in non-European languages of: 1) having the ability to create and maintain parallel native script, gloss and Romanized forms; and 2) structuring the interlingua definition as a simple semantic grammar with a human-readable gloss. We describe how processing of the script alternatives is supported in MedSLT, including insuring that specialized grammars in variant scripts are kept strictly equivalent. With this capability, language engineers have Romanized and/or gloss versions for grammar development, and native speaker informants can work in their normal scripts for building corpora and lexica, and judging translations.

The interlingua grammar with gloss supports a translation evaluation process with a separate evaluator for each language, rather than requiring an evaluator who knows both source and target languages. We presented a comparison of the two methods for Japanese to French, done by an evaluator unfamiliar with the system, and found the results nearly the same. This result shows promise for evaluating language pairs for which bilingual speakers of the two languages are scarce.

Defining the interlingua as a semantic grammar has also proven useful in other contexts. We have shown how it permitted implementation of a debugging mode for translation from target language into interlingua. The interlingua grammar makes it trivial to identify ill-formed interlingua, and we were able to provide automatic help in many cases. We have also described preliminary experiments showing how a tight operational definition of the interlingua was able to effect a non-trivial improvement in speech understanding performance at the cost of only a small overhead.

8. References

Bouillon, P., Rayner, M., Chatzichrisafis, N., Hockey, B. A., Santaholma, M., Starlander, M., Isahara, H., Kanzaki, K., Nakao, Y., A (2005). *Generic Multi-Lingual Open Source Platform for Limited-Domain Medical Speech Translation*, in: Proceedings of the Tenth Conference on European Association of Machine Translation, 30-31, May, Budapest, Hungary, pp. 51-58.

Bouillon P., Halimi S., Rayner M., Hockey B.A, (2007a). *Adapting a Medical Speech to Speech Translation System (MedSLT) to Arabic*, in: Proceedings of the Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, ACL 2007, June 27, Prague, Czech Republic, pp. 41-48.

Bouillon P., Rayner M., Novellas Vall Br., Starlander M., Santaholma M., Nakao Y. and Chatzichrisafis N., (2007b) . *Une grammaire partagée multi-tâche pour le traitement de la parole : application aux langues romanes, in: TAL (Traitement Automatique des Langues)*, Volume 47, 3/2006, Hermes & Lavoisier.

Chatzichrisafis, N., Bouillon, P., Rayner, M., Santaholma, M., Starlander, M. and Hockey, B.A., (2006). *Evaluating Task Performance for a*

Unidirectional Controlled Language Medical Speech Translation System, in: Proceedings of the First International Workshop on Medical Speech Translation, HLT-NAACL, June 9, New York.

Nakao, Y., Rayner, M., Chatzichrisafis, N., Kanzaki, K., Bouillon, P., Hockey, B.A. and Isahara, H., (2006) . *Making MedSLT Easier to Use: Back-Translation and the Help System*, in: Proceedings of NLP 2006, 16, March, 2006, Yokohama, Japan (in Japanese).

Rayner, M., Bouillon, P., Santaholma, M. and Nakao, Y., (2005a). *Representational and architectural issues in a limited-domain medical speech translator*, in: Proceedings of TALN/RECITAL, 6-10, June, Dourdan, France, pp.163-172.

Rayner, M., Bouillon, P., Chatzichrisafis, N., Hockey, B.A., Santaholma, M., Starlander, M., Isahara, H., Kanzaki, K. and Nakao, Y. (2005b) *A Methodology for Comparing Grammar-Based and Robust Approaches to Speech Understanding*, in: Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP), Lisboa, Portugal, pp. 1103-1107.

Rayner, M., Chatzichrisafis, N., Bouillon, P., Nakao, Y. Isahara, H., Kanzaki, K., Hockey, B.A., Santaholma, M. and Starlander, M. (2005c). *Japanese Speech Understanding Using Grammar Specialization*, in: Proceedings of HLT-EMNLP, 6-8, October, Vancouver, British Columbia.

Rayner, M., Hockey, B.A. and Bouillon, P., (2006) . *Putting Linguistics into Speech Recognition*, CSLI, Stanford.

Starlander, M., Bouillon, P., Chatzichrisafis, N., Santaholma, M., Rayner, M., Hockey, B.A., Isahara, H., Kanzaki, K and Nakao, Y., (2005). *Practising Controlled Language through a Help System integrated into the Medical Speech Translation System (MedSLT)*, in: Proceedings of the MT Summit X, 12-16 September, Phuket, Thailand.

Language	Subset	#Utts	Bad interlingua		Improvement	
			w/o N-best	with N-best	Absolute	Relative
French	All	2130	30.5%	27.6%	2.9%	9.5%
English	All	870	42.0%	38.0%	4.0%	9.4%
Japanese	All	544	39.2%	37.7%	3.5%	8.9%
French	In coverage	1583	12.7%	9.7%	3.0%	23.6%
English	In coverage	515	11.2%	9.1%	2.0%	19.0%
Japanese	In coverage	331	10.6%	7.8%	2.8%	26.5%

Table 2. Improvement in speech understanding performance resulting from addition of N-best rescoring using the interlingua, in three languages. "Subset" = "all" or "only in-coverage sentences"; "#Utts" = number of utterances processed; "Bad interlingua" = proportion of examples producing incorrect interlingua, with and without N-best rescoring; "Improvement" = absolute and relative reductions in proportion of utterances producing bad interlingua.

Example 1:	1) هل يمتد الألم إلى الكتفين 2) hal yamtad al alam ila al katifayn 3) YN-QUESTION pain radiate shoulder PRESENT ACTIVE 4) 肩まで痛みが広がりますか 5) kata made itami wa hirogari masu ka 6) shoulder UNTIL pain SUBJECT expand POLITE-PRESENT Q
Example 2	1) كم مرة أحسست بنوبات ألم 2) kam marra ahsasta bi nawbat alam 3) WH-QUESTION you have attacks of pain how-often PRESENT ACTIVE 4) どれくらいの頻度で痛みは起こりますか 5) dorekurai no hindo de itami wa okori masu ka 6) how GEN degree BY pain TOPIC occur POLITE-PRESENT Q
Example 3	1) هل تحس بالألم لأكثر من خمس ساعات 2) hal tahus bi al alam li akthar min khams saat 3) YN-QUESTION you have pain duration more-than five hour PRESENT ACTIVE 4) 五時間少なくとも痛みますか 5) go jikan sukunakutomo itami masu ka 6) five hour at_least hurt POLITE-PRESENT Q

Table 3. Examples of different types of gloss forms for translation between Arabic and Japanese (for each example: 1= Source Arabic, 2= Source Arabic Romanized, 3= Interlingua Gloss, 4= Target Japanese, 5= Target Japanese Romanized, 6= Target Japanese Gloss

Example 1:	1) チーズを食べると痛みはひどくなりますか 2) chizu wo taberu to itami wa hidoku nari masu ka 3) YN-QUESTION pain become-worse sc-when [you eat cheese PRESENT ACTIVE] PRESENT ACTIVE 4) هل يشتد الألم عندما تأكل الجبن 5) hal yachtaddou al alam indama takoul al joubn 6) Y-N_QUESTION intensify-sing3-PRESENT DEF pain-masc-NOUN when-TIME eat-sing2-PRESENT DEF cheese-masc-NOUN
Example 2	1) きりきりした痛みですか 2) kirikirisuru itami desu ka 3) YN-QUESTION boring pain be PRESENT ACTIVE 4) هل الألم واخز 5) hal al alam wakhiz 6) Y-N_QUESTION DEF pain-masc-NOUN piercing-masc
Example 3	1) コーヒーを飲むと頭痛はひどくなりますか 2) kouhii wo nomu to zutsu wa hidoku nari masu ka 3) YN-QUESTION headache become-worse sc-when [you drink coffee PRESENT ACTIVE] PRESENT ACTIVE 4) هل يشتد الألم عندما تشرب القهوة 5) hal yachtaddou al alam indama tachroub al qahwa 6) Y-N_QUESTION intensify-sing3-PRESENT DEF pain-masc-NOUN when-TIME drink-sing2-PRESENT DEF coffee-fem-NOUN

Table 4. Examples of different types of gloss forms for translation between Japanese and Arabic (for each example: 1= Source Japanese, 2= Source Japanese Romanized, 3= Interlingua gloss, 4= Target Arabic, 5= Target Arabic Romanized, 6= Target Arabic Gloss)