A Generic Multi-Lingual Open Source Platform for Limited-Domain Medical Speech Translation

Pierrette Bouillon[1], Manny Rayner[1,2], Nikos Chatzichrisafis[1], Beth Ann Hockey[2], Marianne Santaholma[1], Marianne Starlander[1], Yukie Nakao[3], Kyoko Kanzaki[3], Hitoshi Isahara[3]

(1) University of Geneva, TIM/ISSCO/ETI
{pierrette.bouillon, nikolaos.chatzichrisafis}@issco.unige.ch,
{marianne.santaholma, marianne.starlander}@eti.unige.ch

(2) ICSI/UCSC/NASA Ames Research Center
mrayner@riacs.edu, bahockey@email.arc.nasa.gov

(3) NICT, Kyoto, Japan
{isahara, kanzaki}@nict.go.jp, yukie-n@khn.nict.go.jp

Abstract. We present an overview of MedSLT, an Open Source platform for developing limited-domain medical speech translation systems. We focus in particular on the speech understanding architecture, which uses grammar-based language models derived using corpus-based specialisation methods from a single linguistically motivated grammar, and summarise the results of two evaluations which investigate the appropriateness of these design choices. Other sections describe the interlingua and its relationship with the recognition architecture, and the current demo system.

1. Introduction

Medical domains are an attractive area for spoken language translation systems. They offer not only potential for interesting applications that can be of real use, but can also be sufficiently constrained to permit reasonable performance. A recent high-profile example is the hand-held Phraselator translator[1], which is currently being used by the US military in Iraq.

Discussions with physicians suggest that doctor/patient examination dialogues are both useful and manageable as a task, and have several advantageous properties from the point of view of building a spoken language translation system. As interactions are highly constrained, the input to be recognised is limited. Examinations can also be divided into smaller subdomains based on symptom types, for example, headaches, chest pains, gastric complaints, and so on. This gives the possibility of further constraining the range of utterances that needs to be recognised at any specific moment in the dialogue.

Another important point is that the dialogue can be mostly initiated by the doctor, with the patient giving only non-verbal responses. Speech recognition in the doctor to patient direction is much easier than the reverse, since doctors who use the system regularly will have had time to acclimatise to it; patients, on the other hand, will in general have had no previous exposure to speech recognition technology, and may be reluctant to try it.

With this type of design, it is also possible to improve accuracy by displaying a back-translation into source language via interlingua, and to let the physician-user accept or reject the recognition before invoking translation. In this way, the doctor can check if the content of the sentence has been correctly understood, providing a safeguard against possible parsing and recognition errors.

The MedSLT project is currently engaged in developing a medical speech translation framework based on these general ideas. In the rest of

[1] http://www.phraselator.com

the paper, we describe the architecture of the system, summarise two recent evaluations we have carried out investigating our design choices, and describe the actual prototype that will be demoed at the conference.

2. The MedSLT architecture

MedSLT is an Open Source project[2] whose goal is to develop a generic platform for rapid construction of domain-specific multilingual speech translation systems. Translation is one-way in the doctor to patient direction, which means that most communication is in the form of yes-no questions that can be answered non-verbally. The system has a limited notion of dialogue context, so that it is possible to ask elliptical follow-on questions. For example, if the preceding question was "Is the pain sharp?", then "dull?" will be interpreted as "Is the pain dull?". Supporting ellipsis compensates to some extent for the restriction to yes-no questions. Instead of asking a single WH-question ("Where is the pain?"), the doctor can ask an initial yes-no question with a series of elliptical follow-ups ("Is the pain in the front of the head?"... "The back of the head?"... "The left side?"... "The right side?")[3].

The translation architecture is interlingua-based, and includes multiple recognition engines, back-translation, context-dependent translation and an intelligent help component. The flow of processing is as follows. Input speech is recognised using two different recognisers, both built on top of the Nuance platform (Nuance, 2005). The first recogniser uses a PCFG language model, which directly produces a semantic representation; this is described in more detail in Section 2.1. The second recogniser uses a class N-gram model built using the Nuance SayAnything© package; this produces a plain recognition string, from which a representation can be derived using a set of phrase-spotting rules. The phrase-spotting rules were developed on the training corpus used for both recognisers, and match the output of the grammar-based recogniser to a tolerance of 99% on this material.

The source language semantic representation is passed to a discourse processing module, which interprets it in the context of the previous dialogue, in order to resolve possible ellipsis. The resolved representation is then transformed into its interlingual counterpart. The inter-lingual form is first translated back into the source language and shown to the user, who has the option to abort further processing if they consider that the system has failed to understand what they said. If the user approves the back-translation, the interlingual form is transferred into a target language representation. This is then transformed into a target language surface string using a generation grammar, and finally passed to a speech synthesis unit.

The system optionally invokes a simple context-sensitive help module. This uses the result of robust SLM-based recognition to display a list of in-coverage example sentences. Examples are selected from a predefined list, using a heuristic that prioritises sentences maximizing the number of bigrams and unigrams shared with those extracted from the SLM recognition result.

The remainder of this section describes in more detail the speech understanding component (Section 2.1) and the translation component (Section 2.2).

2.1. Grammar-based recognition

At the start of the project, we felt that there was a case to be made for using grammar-based recognition methods. Initially, we had no training data for creating statistical language models. Also, the system is designed for expert users; an earlier study we had been involved in (Knight et al, 2001) suggested that grammar-based recognition gives better results for people familiar with the coverage of the system. Although these arguments favor the grammar-based approach, we wanted to be able to compare grammar-based speech understanding with a more standard architecture based on statistical language modeling and robust parsing, and have the option of reverting to the standard architecture if that seemed appropriate. In particular, this implied that source-language semantic representa-

[2] http://sourceforge.net/projects/medslt/

[3] The system does in fact also support WH-questions, since several doctors said they would like the option of using them as introductions to yes-no questions: "Where is the pain?"... "Is it in the front of the head?"

tions needed to be such that they could reasonably be produced using phrase-spotting techniques. In the end, we decided that the grammar-based recogniser did indeed offer better performance (cf. Section 3.1), but we retained the alternate statistical recogniser in order to provide input to the intelligent help component.

In the grammar-based recogniser, speech recognition uses a set of CFG language models, one per subdomain. The language models are compiled, using the Open Source Regulus toolkit[4], from a single linguistically motivated unification grammar. This makes it possible to support efficient structure-sharing between many similar subdomains with overlapping vocabulary and structure. Each subdomain-specific grammar is defined by a small training corpus, typically containing 400 to 800 examples. The same corpus material is also used to perform probabilistic tuning of the resulting CFG language model.

We have developed linguistically motivated unification grammars in Regulus format for several languages. The most mature one, for English, currently contains 180 unification grammar rules and 79 features, and also includes a function-word lexicon with about 450 entries; less elaborate grammars also exist for French, Japanese, Spanish and Finnish. The process used to transform a general unification grammar into a domain-specific recogniser goes through the following stages.

1. A set of domain-specific lexical entries is written and added to the function-word lexicon.
2. The training corpus is converted into a treebank of parsed representations, using a left-corner parser version of the grammar.
3. The treebank is used to produce a "raw" specialised grammar in Regulus format, using the Explanation Based Learning algorithm (van Harmelen and Bundy, 1988; Rayner, 1988). The granularity of the learned rules is determined by a user-supplied parameter.
4. The "raw" specialised grammar is post-processed into the final specialised grammar. This involves discarding duplicate rules, and carrying out a binarisation transform (Rayner et al 2002).
5. The post-processed specialised grammar is compiled into a CFG grammar in Nuance Grammar Specification Language (GSL) format.
6. The Nuance GSL grammar is compiled into a Nuance recognition package.

The most complex part of this process is stage 5, conversion of the specialised unification grammar to CFG form. The basic algorithm is described in (Rayner et al, 2001). The central idea is simply to perform an enumerative expansion of the unification grammar by non-deterministically instantiating each feature to all of its possible values; the resulting grammar is then filtered to remove non-reachable rules.

As described in (Rayner et al, 2001), the efficiency of the naive algorithm can be greatly improved by adding a pre-processing step which performs a suitable factoring of the grammar. The current version of Regulus further refines the naive method by iteratively alternating the expansion and filtering stages, non-deterministically expanding each feature in turn and then filtering the result before proceeding to the next feature. On large grammars, this "iterative expansion" technique can reduce the time and space requirements of the compilation algorithm by several orders of magnitude. Use of iterative expansion has allowed Regulus successfully to compile several grammars which exceeded resource bounds for the Gemini compiler (Moore et al, 1997; Moore, 1998).

In Section 3, we summarise the results of two recent evaluations, which investigate the design choices made in the recognition component. We focus now on the translation component.

2.2. Translation component

Since one of the aims of the project is to be able to support rapid implementation of new language-pairs, translation is interlingua-based and is done in four main stages: (1) parsing of the source sentence, (2) mapping from the source representation to interlingua, (3) mapping from interlingua to the target representation and (4) generation, using a suitably compiled Regulus grammar for the target language. These stages are illustrated in Fig. 1 for the sentence "do you

[4] http://sourceforge.net/projects/regulus/

ever have headaches in the morning?" → "avez-vous déjà eu vos maux de tête le matin?"

```
source_representation =
[[utterance_type,ynq],    [pro-
noun,you],[voice,active],
[tense,present],[freq,ever],
[state,have_symptom],
[prep,in_time],[time,morning],
[secondary_symptom,headache]]
                ↓
interlingua =
[[utterance_type,ynq],
[pronoun,you],[voice,active],
[tense,present],[freq,ever],
[state,have_symptom],
[prep,in_time],[time,morning],
[symptom,headache]]
                ↓
target_representation =
[[utterance_type,sentence], [pro-
noun,vous],[voice,active],
[tense,passé_composé],[freq,déjà],
[path_proc,avoir],
[temporal,matin],
[symptom,mal_de_tête]]
```

Figure 1: Translation flow for the sentence: "do you ever have headaches in the morning?"

In accordance with the generally minimalist design philosophy of the project, source, target and interlingua representations have been kept as simple as possible, and in many cases consist of a flat list of unordered attribute-value pairs. One level of nesting can optionally be added to represent subordinating clause constructions, like "do you have headaches when you drink red wine?" or "do you have pain when you watch TV?":

```
[[utterance_type,ynq],
[pronoun,you],[voice,active],
[tense,present],
[state,have_symptom],
[secondary_symptom,headache],
[sc,when],
[clause,
  [[utterance_type,dcl],
   [pronoun,you],
   [voice,active],
   [action,drink],
   [cause,red_wine]]]]
```

Figure 2: Source representation for "do you have headaches when you drink wine?"

In many applications, the underspecified nature of this minimalist representation would pose serious problems. In this project however, it seems to be sufficient: the domain is very restricted, and both analysis and generation grammars are automatically specialised to our specific application with a corpus and a typed lexicon, limiting the ambiguity of the general grammar.

```
transfer_rule(
[[tense,present],
 [freq, ever]],
[[tense,passé_composé],
 [freq,déjà]]).
```

Figure 3: Example of mapping rules from interlingua to target representation in French

The minimalist representation language offers several advantages. In general, it is compatible with the statistical version of the system and can easily be produced by a robust parser based on phrase spotting rules. In the context of translation, the minimalist representation greatly simplifies the construction of mapping rules from and to interlingua. In most cases, these simply have to transform atoms or list of atoms. For example, in order to map the present tense in the English sentence from Fig. 1, "do you ever have headaches in the morning" to the past tense in its French counterpart "avez-vous déjà eu ces maux de tête le matin", we only need the rule in Fig. 3, which changes the English present tense to a French past tense in the presence of *ever*.

The simplicity of the representation also makes it easy to define conditional rules that test for the presence or absence of partially specified elements. For example the following rule translates "have pain" to "avoir mal" in the context of a subordinating conjunction element, annotated with the tag "sc".

```
% have pain when-> avez-vous mal
quand
transfer_rule(
[[state,have_symptom],
 [symptom,pain]],
[[path_proc,avoir],
 [symptom,mal]])
```

```
:- context([sc,_]).
```

Figure 4: Example of conditional rule

Another advantage of the flat representation for translation is that the interlingua itself is easy to read and to validate automatically and/or manually. As shown in Fig. 1, interlingua structures are for now essentially canonical versions of English representations. Apart for simple transformations like adding/deleting one atom in the representation (for example, absorbing a support verb or adding a canonical preposition), the main systematic transformations carried out when moving from source to interlingua level are concerned with temporal and causal concepts, which are central in our medical domains. Thus for example the source representation of "are your headaches caused by red wine" is represented at the interlingua level as in Fig. 2 and can be paraphrased as: "do you have headaches when you drink red wine". In the same way, "are the headaches accompanied by nausea" receives an interlingual representation that can be paraphrased as "do you experience nausea when you have headaches". In this way, all the different temporal and causal constructions are mapped to one of the following interlingua schemas: (1) Clause1 WHEN Clause2; (2) Clause1 BEFORE Clause2 and (3) Clause1 AFTER Clause2.

Of course, this would not be a suitable solution in a general domain. It is however motivated in this application, where the translations are not required to be literal. We have found that the non-literal translation is often clearer than the literal one since it is more explicit, especially for divergent languages. It also adds robustness at the level of translation, since we are not forced to translate all the variants that can be found in all languages. Evaluation of the quality of the translation is described in the next section.

3. Summary of evaluations

This section describes two recent evaluations, which investigate the following questions:

Evaluation 1:
How does a grammar-based speech understanding architecture compare against a statistical/robust one for the MedSLT task?

Evaluation 2:
Do recognisers based on specialised unification grammars offer better performance than ones based on the original general grammars?

3.1. Evaluation 1

Our first evaluation used the English-to-French and English-to-Japanese versions of the system, with the headache subdomain. Both versions of the recogniser were trained from the same corpus of 575 standard examination questions supplied by a medical professional[5].

We collected data from 12 native speakers of English. Each subject was first given a short acclimatisation session, where they used a prepared list of ten in-coverage sentences to learn how to use the microphone and the push-to-talk interface. They were then encouraged to play the part of a doctor, and conduct an examination interview, through the system, on a team member who simulated a patient suffering from a specific type of headache. The subject's task was to identify the type correctly out of a list of eight possibilities. Half of the subjects used the grammar-based version of the system, and half used the SLM based version. We collected a total of 870 recorded utterances.

The recorded data was first transcribed, and then processed through offline versions of both the grammar-based (GLM) and statistical (SLM) processing paths in the system. We first set the system to translate from English into English, via the interlingua, and then had an English-speaking judge evaluate each back-translation. Utterances for which the back-translation was judged acceptable were regarded as correctly recognised. Results for speech understanding performance are shown in Table 1; here, SemER refers to the proportion of utterances producing an unacceptable back-translation.

	In coverage		**Out of coverage**	
	GLM	SLM	GLM	SLM
WER	5.7%	12.7%	57.5%	47.8%
SER	19.4%	36.7%	99.8%	91.4%
SemER	18.5%	28.1%	87.9%	89.0%

[5] Dr. Vol Van Dalsem III, El Camino Hospital, Mountain View, California.

Table 1: Recognition performance

Utterances that were judged as acceptably recognised were then translated further into the target languages French and Japanese. These translations were judged by native-speaker judges for each language; there were six judges for French, and three for Japanese. Judges were asked to categorise translations as "good", "ok" or "bad". For each target language, and each processing method (GLM or SLM), we consolidated the results using a majority voting scheme. If two-thirds of the judges (i.e. four for French, or two for Japanese) agreed that the translation was clearly "good" or "bad", we counted the translation as belonging to the appropriate category. Otherwise, we counted it as "ok" (labeled Acceptable Translation). Table 2 shows performance results for speech translation.

	French		Japanese	
	GLM	**SLM**	**GLM**	**SLM**
Bad Recognition	54.6%	59.8%	54.6%	59.8%
Good Translation	34.4%	30.8%	36.4%	32.8%
Acceptable Translation	8.7%	7.7%	3.6%	3.3%
Bad Translation	0.3%	0.2%	0.5%	0.5%
No Translation	2.0%	1.5%	4.9%	3.7%

Table 2: Translation performance

The most interesting aspect of these results is the striking difference in performance between the two recognisers on in-coverage data, as shown in Table 1; the GLM scores much better than the SLM, especially on WER (5.7% versus 12.7%) and SER (19.4% versus 36.7%). Robust processing lets the SLM recover somewhat on SemER (18.5% versus 28.1%), but the GLM still comes out a clear winner. Interestingly enough, although the SLM scores better than the GLM on the out-of-coverage data in terms of WER (47.8% versus 57.5%) and SER (91.4% versus 99.8%), the two systems score about equally even here in terms of semantic error rate (87.9% versus 89.0%).

In summary, the grammar-based recogniser is a great deal better than the statistical one on the in-coverage data and about the same on the out-of-coverage data, when the evaluation metric is the semantic error rate. This reflects the "all-or-nothing" nature of the medical speech translation task, where partial translations are worse than useless. The extra robustness offered by the statistical recogniser does indeed result in a lower word error rate on the out-of-coverage data, but both recognisers perform miserably here. What counts is dependable performance on the in-coverage data, and this is why the grammar-based system wins comfortably.

3.2. Evaluation 2

Our second evaluation was designed to investigate the impact on recognition performance resulting from the grammar specialisation process described in Section 2.1. This time, we used the Japanese to English version of the system, again with the headache subdomain. We compiled both the unspecialised and the specialised versions of the Japanese grammar into Nuance recognition packages, and evaluated them on a corpus of 544 spoken Japanese utterances, collected from four different Japanese native speakers using a protocol similar to that used for Evaluation 1. Once again, we present the results separately for the 314 in-coverage utterances and the 230 out-of-coverage utterances. (In this context, "in-coverage" refers to the coverage of the *unspecialised* grammar). Table 3 shows figures for recognition performance (WER, SER, speed as a multiple of real time[6]).

Recognition			
Version	**WER**	**SER**	**xRT**
	In-coverage (319 utterances)		
Unspecialised	20.9%	29.8%	1.62
Specialised	7.1%	14.7%	0.06
	Out-of-coverage (225 utter.)		
Unspecialised	83.0%	100.0%	1.71
Specialised	68.5%	100.0%	0.08

Table 3: Effect of grammar specialisation on recognition performance

Table 4 contains the figures for performance on the Japanese to French speech translation task. The column headings "BadRec", "Good", "OK", "Bad" and "None" refer respectively to the proportion of utterances which are incorrectly recognised, translated completely correctly, translated acceptably, translated badly, and which fail to produce a translation.

Translation					
Version	**Bad**	**Good**	**Ok**	**Bad**	**Non**

[6] As measured on a 3.2GHz mobile Intel P4 processor

	Rec				e
In-coverage (in % out of 319 utterances)					
Unspec.	26.3	52.4	12.2	2.8	6.3
Spec.	8.2	69.3	13.8	3.4	5.3
Out-of-coverage (in % out of 225 utterances)					
Unspec.	96.4	1.8	1.3	0.4	0.00
Spec.	85.8	7.1	3.1	0.0	4.0

Table 4: Effect of grammar specialisation on translation performance

On the in-coverage part of the corpus, the specialised version outperforms the non-specialised one on WER and proportion of misrecognised utterances by a factor of about 3, and on recognition speed by a factor of about 25. This clearly illustrates the performance gains resulting from the grammar specialisation process, which result in a much more efficient recognition grammar.

4. The prototype

The current version of the MedSLT prototype covers three subdomains (headaches, chest pains, and abdominal pain) with a vocabulary of about 300-550 words per domain.

The run time system is accessed though a GUI illustrated in Figure 5. By clicking on suitable areas in the picture of the human figure, the user can change subdomain. The Settings menu allows the user to select the input language (currently English or Japanese) and the output language (currently English, French, Japanese or Finnish), the sub-domain and the type of recognition package (statistical language model, grammar-based language model derived from the general grammar or grammar-based language model derived from a specialised grammar). This information appears at the bottom of the window. The user initiates speech recognition using the *Start Recognition* button. The *Raw Recognition Result* box shows the actual words returned by the recogniser, while the *What the system understood* box displays the back-translation from the interlingua. The physician-user can then accept or reject the utterance before invoking translation by pushing the *Translate* button.

The project team will exercise the system across its range of coverage, and include side-by-side comparison, highlighting the differences in speech recognition and in end-to-end performance using recognisers based on statistical language models, general grammars, and specialised grammars. Demo attendees will also be invited to try out the various versions of the system themselves. Examples of domain sentences and translations are given in Appendix 1.

5. References

KNIGHT, Sylvia, GORRELL, Genevieve, RAYNER, Manny, MILWARD, David, KOELING, Rob, LEWIN, Ian (2001). Comparing Grammar-Based and Robust Approaches to Speech Understanding: A Case Study. In Proceedings of Eurospeech, Aalborg, Denmark, pp. 1179-1782.

MOORE, Robert C., DOWDING, John, BRATT, Harry, GAWRON, Jean Mark, GORFU, Yonael, CHEYER, Adam (1997). 'CommandTalk: A Spoken-Language Interface for Battlefield Simulations'. In Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 1-7.

MEDSLT (2005). https://sourceforge.net/projects/medslt/. As of 31 January 2005.

MOORE, Robert, (1998). 'Using natural language knowledge sources in speech recognition'. In Proceedings of the NATO Advanced Studies Institute.

NUANCE (2005). http://www.nuance.com. As of 15 January 2005.

PHRASELATOR (2004). http://www.phraselator.com. As of 8 Dec 2004.

RAYNER, Manny (1988). 'Applying explanation-based generalization to natural-language processing'. In Proceedings of the International Conference on Fifth Generation Computer Systems, Tokyo, Japan, pp.1267-1274.

RAYNER, Manny, DOWDING, John, HOCKEY, Beth Ann (2001). 'A baseline method for compiling typed unification grammars into context free language models'. In Proceedings of Eurospeech, Aalborg, Denmark, pp. 729-732.

RAYNER, Manny, BOUILLON, Pierrette (2002). 'A flexible Speech to Speech Phrasebook Translator'. In Proceedings of ACL-02 Workshop on Speech-to-Speech Translation: Algorithms and Systems, Philadelphia, pp. 69-76.

RAYNER, Manny, BOUILLON, Pierrette, VAN DALSEM III, Vol, HOCKEY, Beth Ann, ISAHARA, Hitoshi, KANZAKI, Kyoko (2003). 'A limited-domain English to Japanese medical speech translator build using REGULUS 2'. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (demo track), Sapporo, Japan, pp. 137-140.

RAYNER, Manny, BOUILLON, Pierrette, HOCKEY, Beth Ann, CHATZICHRISAFIS, Nikos, STARLAND-

ER Marianne (2004). 'Comparing Rule-Based and Statistical Approaches to Speech Understanding in a Limited Domain Speech Translation System'. In Proceedings of TMI 2004, Baltimore, MD USA, 2004, pp. 21-29.

VAN HARMELEN, Frank, BUNDY, Alan (1988). 'Explanation-Based Generalisation = Partial Evaluation'. In Artificial. Intelligence 36(3), pp. 401-412.

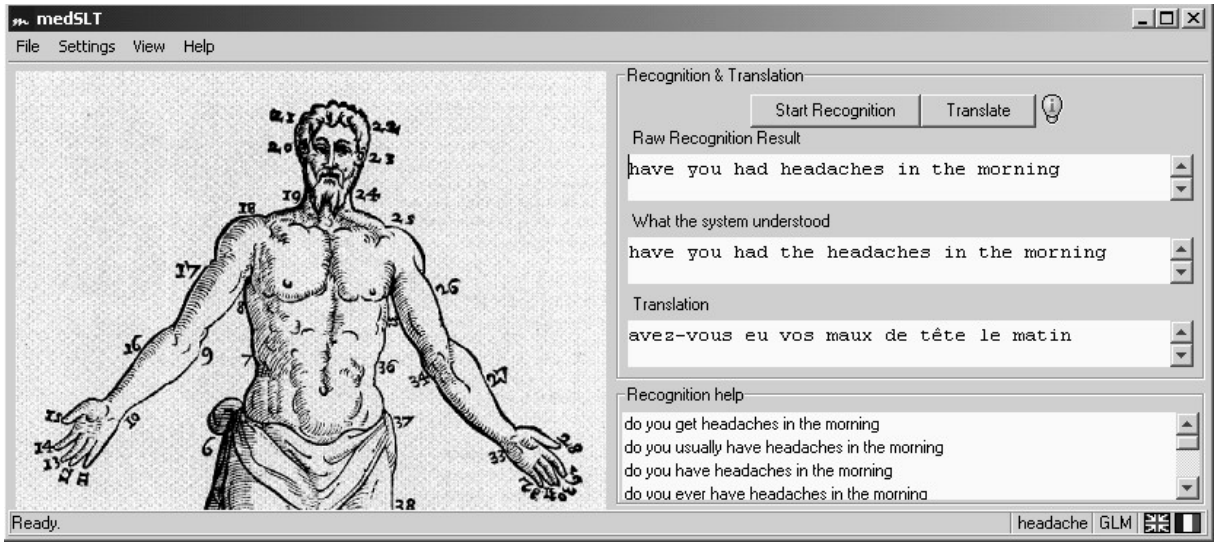

Figure 5: Screenshot of the MedSLT system, for Eng–Fre, using the GLM version and providing recognition help

Japanese-French
Source: massaaji wo suru to itami ga yawaragi masu ka
Target: la douleur est-elle soulagée par des massages
Source: kouhii de zutsu ga okori masu ka
Target: avez-vous vos maux de tête quand vous buvez du cafe
Source: hon wo yomu to itami ga okori masu ka
Target: avez-vous mal quand vous lisez
Source: kurai heya ni iru to itami wa osamari masu ka
Target: la douleur est-elle soulagée par l'obscurité
Source: akarui hikari de zutsu ga okori masu ka
Target: vos maux de tete sont-ils causés par une lumiere forte
English-Finnish
Source: does the pain go down to your neck
Target: säteileekö särky kaulalle
Source: how long does the headache last
Target: kuinka kauan päänsärky kestää
Source: where does it hurt
Target: missä särky tuntuu
Source: is the pain on the side of your head
Target: tuntuuko särky pään sivulla
Source: do you have the pain every day
Target: esiintyykö särkyä joka päivä
English-French
Source: does it hurt when you move your head
Target: avez-vous mal quand vous bougez la tête
Source: do you suffer from anxiety
Target: êtes-vous anxieux
Source: are you thirsty before you have the pain
Target: la douleur est-elle précédée d'une sensation de soif
Source: are the headaches occurring less frequently
Target: la fréquence de vos maux de tête diminue-t-elle
Source: are your headaches usually accompanied by drowsiness
Target: êtes-vous habituellement somnolent quand vous avez vos maux de tête

Appendix 1: Examples of domain sentences and translations