UNIVERSITÉ DE GENÈVE

Archive ouverte UNIGE
https://archive-ouverte.unige.ch

Chapitre d'actes | 2008 | Open Access

This version of the publication is provided by the author(s) and made available in accordance with the copyright holder(s).

Many-to-Many Multilingual Medical Speech Translation on a PDA

Bouillon, Pierrette; Flores, Glenn; Georgescul, Maria; Halimi Mallem, Ismahene Sonia; Hockey, Beth Ann; Isahara, Hitoshi; Kanzaki, Kyoko; Nakao, Yukie; Rayner, Emmanuel; Santaholma, Marianne Elina; Starlander, Marianne; Tsourakis, Nikolaos

How to cite

BOUILLON, Pierrette et al. Many-to-Many Multilingual Medical Speech Translation on a PDA. In: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas. Waikiki, Hawai'i (USA). [s.l.] : [s.n.], 2008. p. 314–323.

This publication URL: https://archive-ouverte.unige.ch/unige:3473

© This document is protected by copyright. Please refer to copyright holder(s) for terms of use.

Many-to-Many Multilingual Medical Speech Translation on a PDA

Pierrette Bouillon[1], **Glenn Flores**[2], **Maria Georgescul**[1], **Sonia Halimi**[1]
Beth Ann Hockey[3], **Hitoshi Isahara**[4], **Kyoko Kanzaki**[4], **Yukie Nakao**[5]
Manny Rayner[1], **Marianne Santaholma**[1], **Marianne Starlander**[1], **Nikos Tsourakis**[1]

[1] University of Geneva, TIM/ISSCO, 40 bvd du Pont-d'Arve, CH-1211 Geneva 4, Switzerland
{Emmanuel.Rayner,Pierrette.Bouillon,Nikolaos.Tsourakis}@unige.ch
{Sonia.Halimi,Marianne.Santaholma,Marianne.Starlander}@unige.ch

[2] UT Southwestern Medical Center, Children's Medical Center of Dallas
Glenn.Flores@utsouthwestern.edu

[3] Mail Stop 19-26, UCSC UARC, NASA Ames Research Center, Moffett Field, CA 94035–1000
bahockey@ucsc.edu

[4] NICT, 3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan 619-0289
{isahara,kanzaki}@nict.go.jp

[5] LINA, Nantes University, 2, rue de la Houssinière, BP 92208 44322 Nantes Cedex 03
yukie.nakao@univ-nantes.fr

Abstract

Particularly considering the requirement of high reliability, we argue that the most appropriate architecture for a medical speech translator that can be realised using today's technology combines unidirectional (doctor to patient) translation, medium-vocabulary controlled language coverage, interlingua-based translation, an embedded help component, and deployability on a hand-held hardware platform. We present an overview of the Open Source MedSLT prototype, which has been developed in accordance with these design principles. The system is implemented on top of the Regulus and Nuance 8.5 platforms, translates patient examination questions for all language pairs in the set {English, French, Japanese, Arabic, Catalan}, using vocabularies of about 400 to 1 100 words, and can be run in a distributed client/server environment, where the client application is hosted on a Nokia Internet Tablet device.

1 Introduction

There is an urgent need for medical speech translation systems. The world's current population of 6.6 billion speaks more than 6,000 languages (Graddol, 2004). Language barriers are associated with a wide variety of deleterious consequences in healthcare, including impaired health status, a lower likelihood of having a regular physician, lower rates of mammograms, pap smears, and other preventive services, non-adherence with medications, a greater likelihood of a diagnosis of more severe psychopathology and leaving the hospital against medical advice among psychiatric patients, a lower likelihood of being given a follow-up appointment after an emergency department visit, an increased risk of intubation among children with asthma, a greater risk of hospital admissions among adults, an increased risk of drug complications, longer medical visits, higher resource utilisation for diagnostic testing, lower patient satisfaction, impaired patient understanding of diagnoses, medications, and follow-up, and medical errors and injuries (Flores, 2005; Flores, 2006). Nevertheless, many patients who need medical interpreters do not get them. For example, in the United States, where, according to the 2006 American Community Survey, 52 million people speak a language other than English at home and 23 million people have limited English proficiency (LEP), one

study found that about half of LEP patients presenting to an emergency department were not provided with a medical interpreter (Baker et al., 1996). In Western Europe, although the problems are probably less acute than in the United States, it is still clearly the case that there are tens of millions of immigrants and refugees experiencing serious problems in this area. In short, there a substantial gap between the need for and availability of language services in health care, a gap that could be bridged through effective medical speech translation systems.

An ideal system would be able to interpret accurately and flexibly between patients and health care professionals, using unrestricted language and a large vocabulary. Providing functionality of this kind is, unfortunately, well beyond the current state of the art. Although it goes without saying that we are in favour of continued research towards this highly desirable goal, it is also worth thinking about what can be achieved in terms of building systems now, or in the immediate future, which will already be concretely useful. This paper describes the current version of MedSLT (Bouillon et al., 2005), an Open Source medical speech translation system for doctor-patient examination dialogues aimed at short-term deployment. We start by reviewing the basic goals and constraints.

The fundamental issue is reliability. In conversations with medical professionals, it has been made clear to us on numerous occasions that doctors are only interested in translation systems which are extremely reliable. At the recent workshop on medical and safety-critical translation (Bouillon et al., 2008a), the question of evaluation metrics for doctor/patient communication applications was considered during the panel discussion; the consensus view reached was that a metric which conforms to most physician's intuitions should give a negative score for seriously incorrect translations somewhere around 25 to 100 times the positive score for a correct translation. Given the usual tradeoff between precision and recall, this implies that the dial needs to be moved heavily towards the "precision" end.

This top-level design decision has several implications. Although data-driven architectures for speech translation are currently more popular, their key strengths are in the direction of robustness, high recall, and suitability for inexperienced users. This is reflected in the development methodologies, which typically involve optimising BLEU, or a similar surface-oriented metric. It is not at all clear that this kind of approach is appropriate for the kind of high-precision translation required here; the comparative studies that have been carried out suggest that rule-based architectures offer considerable higher accuracy, especially when the system is designed for use by experts, such as doctors (Knight et al., 2001; Rayner et al., 2005a; Lee and Seneff, 2005). These users can both learn the system's intended coverage, and also adapt to an interface which gives them feedback on what the system has understood, allowing them to abort incorrect speech processing without producing a translation. Thus our first conclusion is that rule-based architectures are, at least at the moment, the most suitable ones for the doctor-patient examination dialogue task.

For similar reasons, it appears right now difficult to build a useful bidirectional translation system for this kind of domain. It is in principle highly desirable to support two-way dialogue, and several quantitative medical studies (Stewart, 1995; Michie et al., 2003) have demonstrated the advantages of patient-centered communication and shared decision-making. This must however be balanced against the fact that the patient is at a grave disadvantage compared to the doctor; they will in general have had no previous opportunity to familiarise themselves with the system and its correct mode of operation, implying that communication in the patient-to-doctor direction will be far less reliable than in the doctor-to-patient direction. These problems are not insurmountable, and we are actively addressing them. None the less, in the short-term, it seems reasonable to say that a useful system is much easier to build if it only translates in one direction.

Our basic architecture, then, is one-way controlled-language rule-based translation in the doctor to patient direction, and it is not unreasonable to hope that physician users, once they have learned the controlled language, will get good performance. This raises the next question: how they will learn the boundaries of the controlled language's coverage? Evidently, the system will be more useful if it is in some way able to help the user acquire this knowledge; several studies with controlled-language spoken dialogue systems have

shown that inclusion of an online help component often makes a critical difference to usability (Gorrell et al., 2002; Hockey et al., 2003; Rayner et al., 2005a).

Another important set of issues are concerned with portability. Even though medical examination questions in any given subdomain (headaches, chest pains etc) are reasonably constrained, there are a large number of such subdomains. Developing a separate grammar for each subdomain is unattractive; it is much better to be able to share structure between the various subdomain grammars, working within a uniform general framework. Similar problems arise with respect to support of multiple language pairs. In most Western countries, it is relatively easy to find someone who can interpret for a French or Spanish-speaking patient, considerably harder to find a Japanese or Tamil interpreter, and extremely challenging to locate anyone who can speak Albanian or Mongolian. In other words, translation adds most value for rare languages, and it is consequently important to be able to add new patient languages quickly. At least in Europe, it is also the case that a system will be more useful if it can be deployed in many countries, which implies that it is also important to be able to cover multiple doctor languages. Supporting the cross-product of multiple source and target languages ("the N^2 problem") is difficult if translation is performed directly from source to target; as has been argued many times in the literature, requirements of this kind strongly dispose towards an interlingua-based translation architecture.

Finally, there is the question of hardware platforms. Laptop-based systems are less than ideally portable; anecdotally, at least, a system which could be deployed on a PDA would be more well-received by and feasible for physicians. In some cases, this is just a question of convenience (the doctor will prefer to carry her translation device with her when going on ward rounds). It is also easy to imagine situations where platform issues could become more critical.

The rest of the paper describes the architecture of the MedSLT prototype, which has evolved in response to the requirements we have outlined above. The basic functionality offered is one-way rule-based speech translation in the doctor to patient direction, and supports translation from any language to any language in the set {English, French, Japanese, Arabic, Catalan}. Recognition uses a separate grammar-based language model for each language and subdomain; all subdomain grammars are derived from general, domain-independent language resources (Section 2). Translation is interlingua-based, and completely decouples translation between source language and interlingua from translation between interlingua and target language (Section 3). An intelligent help facility interactively guides users towards the system's coverage (Section 4). The various top-level components — recognition, translation, interactive help, speech output — are combined into a speech-to-speech translation system which can be deployed on a hand-help mobile platform (Section 5), using a client/server model which runs recognition processes on a remote machine. Recognition performance for the distributed client/server system is as good as that which would be obtained on a high-end laptop.

2 Recognition

Speech recognition uses the Nuance 8.5 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). 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 recogniser can also be parsed. Regulus has also been used to build several other large speech-enabled systems, a prominent example being NASA's Clarissa (Rayner et al., 2005b).

Early versions of Regulus used a single core grammar per language; more recent ones have gone further, and merged together grammars for closely related languages (Bouillon et al., 2007). These core grammars are automatically specialised, using corpus-driven methods based on small corpora, to

derive simpler grammars. Specialisation is both with respect to task (recognition, analysis, generation) and to subdomain (headache, chest pain, etc). The specialisation process uses the Explanation Based Learning algorithm (Rayner, 1988). 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 specialised unification grammar. Each of these specialised 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.

Table 1 gives examples of the coverage of the English-input headache-domain version, and Table 2 summarises recognition performance in this domain for the three input languages where we have so far performed serious evaluations. Differences in the sizes of the recognition vocabularies are primarily due to differences in use of inflection.

3 Interlingua-centred translation

Translation in MedSLT uses a rule-based interlingua architecture. Source language semantic representations are translated into interlingual representations by one set of rules, and these interlingua representations are then translated into target language representations by a second set of rules. Finally, the target language representations are converted into surface words using a target language grammar. We will focus here on two specific aspects of the translation architecture: first, our representation language, Almost Flat Functional Semantics, and second, the role of the interlingua.

3.1 Almost Flat Functional semantics

The representation language used for source, interlingua and target forms has gone through several iterations of redesign. In the earlier Spoken Language Translator (SLT) project (Rayner et al., 2000), which in many ways was a precursor to MedSLT, the language used was Quasi Logical Form (QLF), an unscoped logic-based representation. For example, a QLF representation of "Does coffee give you headaches?" would have been something like

```
[dcl,
 form(verb(present,no,no,no,yes),E,
      [[give1,
        E,
        term(q(bare,sing),X,
             [coffee1, X])
        term(ref(pro,you,sing),Y,
             [person,Y]),
        term(q(bare,plur),Z,
             [headache1, Z])]])]
```

Although QLF is wonderfully expressive, the complex nature of the representations meant that translation rules were also highly complex, with the usual implications for development and maintenance costs. In MedSLT, our first inclination was to move to a minimal formalism: early versions of the system used a language which consisted only of simple feature-value lists, with one optional level of nesting used to represent subordinate clauses and similar constructions. Determiners were not in general included in the representation, both because they are often difficult to translate, and also because recognition is not usually able to distinguish them reliably. Thus, continuing the example, "Does coffee give you headaches?" was represented as

```
[[utterance_type,ynq],
 [action,give],
 [cause,coffee],
 [pronoun,you],
 [symptom,headache],
 [tense,present], [voice,active]]
```

The payoff was that translation rules could be made much simpler, since they only had to map lists of pairs to lists of pairs. It is evident, however, that a flat representation like the one above is underconstrained. Unlike the QLF representation, there is

Where?	Is the pain above your eye?
When?	Have you had the pain for more than a month?
How long?	Does the pain typically last a few minutes?
How often?	Do you get headaches several times a week?
How?	Is it a stabbing pain?
Associated symptoms?	Do you vomit when you get the headaches?
Why?	Does bright light make the pain worse?
What helps?	Does sleep make the pain better?
Background?	Do you have a history of sinus disease?

Table 1: Examples of English MedSLT coverage

nothing here which says which pieces of structure occupy the different argument positions of "give". Thus, *a priori*, it might equally well mean "Do headaches give you coffee?", which could for example give rise to an unintended ambiguity if the English grammar were used to generate a surface string.

In nearly all cases, problems of this kind did not actually arise in practice, once natural sortal constraints had been added to the grammar; here, the object of "give" is constrained to be a symptom, blocking "Do headaches give you coffee?". None the less, each language had a few awkward constructions. For instance, in English there is no very convincing way to deal with the verb "precede": the representation

```
[[utterance_type,ynq],
 [event,precede],
 [symptom,headache],
 [symptom,nausea],
 [tense,present],
 [voice,passive]]
```

could mean either "Are the headaches preceded by nausea?" or "Is the nausea preceded by headaches?". Examples like these caused enough difficulties for translation rule writers that we decided to move to a slightly more expressive formalism. This new formalism, which we call Almost Flat Functional semantics (AFF), combines elements of the SLT project's QLF and the original flat feature-value list semantics. It is essentially a version of the flat semantics, enhanced by adding functional markings to the feature-value pairs; continuing the running example, "Does coffee give you headaches?" is represented in AFF as

```
[null=[utterance_type,ynq],
 null=[action,give],
 agent=[cause,coffee],
 indobj=[pronoun,you],
 obj=[symptom,headache],
 null=[tense,present],
 null=[voice,active]]
```

AFF, which we describe in detail in (Rayner et al., 2008), appears to be a good compromise between the competing goals of tractability and representational adequancy. It solves all the ambiguity problems that arose with the original flat representation language, but only makes translation rules slightly more complex; over five-sixths of the rules for the flat representations could be carried over directly to AFF. Processing times for parsing and generation are also very similar for the two formalisms.

Language	Vocab	WER	SemER
English	447	6%	11%
French	1025	8%	10%
Japanese	422	3%	4%

Table 2: Recognition performance for English, French and Japanese headache-domain recognisers. "Vocab" = number of surface words in source language recogniser vocabulary; "WER" = Word Error Rate for source language recogniser, on in-coverage material; "SemER" = semantic error rate (proportion of utterances failing to produce correct interlingua) for source language recogniser, on in-coverage material. Tests were carried out on a laptop-based version of the system; as described in Section 5, performance on the mobile version is essentially the same.

English	Is the pain above your eye?
Interlingua	YN-QUESTION pain be above-loc eye PRESENT
English	Have you had the pain for more than a month?
Interlingua	YN-QUESTION you have pain duration more-than one month PRESENT-PERFECT
English	Do you get headaches several times a week?
Interlingua	YN-QUESTION you have headache several times per week PRESENT
English	Is it a stabbing pain?
Interlingua	YN-QUESTION stabbing pain be PRESENT
English	Do you vomit when you get the headaches?
Interlingua	YN-QUESTION you vomit sc-when [you experience headache PRESENT] PRESENT
English	Do you have a history of sinus disease?
Interlingua	YN-QUESTION you have history of sinus-disease PRESENT

Table 3: Interlingua surface forms corresponding to some of the English sentences from Table 1

3.2 Grammar-based interlingua

Apart from the representation formalism, the other main novelty in MedSLT's approach to interlingual translation is the definition of the interlingua itself. Early versions of the systems only enforced loose constraints on interlingua representations. Over the last year, however, we have completely reorganised this part of the system, and introduced a new treatment, where the interlingua is conceptualised as a language in its own right, specified by a Regulus grammar (Bouillon et al., 2008b). Given an interlingua representation 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; Table 3 gives some examples. The interlingua grammar is not obliged to take account of the complex surface syntax phenomena characteristic of real languages (movement, agreement, etc), and 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, which in turn means that generation, and hence checking of validity for an interlingua expression, is extremely efficient.

In earlier versions of the system, translation rules could only be tested in the context of a specific source/target pair; when problems arose, it was often difficult to know whether the source or target rule-set was at fault. The interlingua surface form, which is designed as a highly simplified version of English, has allowed us to effect a complete decoupling of development work into monolingual components, since it is now possible to evaluate translation from source to interlingua, and from interlingua to target, independently of each other. We were initially apprehensive that composition of translation judgements involving the artifical interlingua language would not necessarily agree with judgements of translations from a real source language to a real target language; thus, for example, it was not clear whether a judgement of correct translation from French to Interlingua, together with a second judgement of correct translation from Interlingua to Japanese, would always correspond to a direct judgement of correct translation from French to Japanese. In fact, at least for the domains we are working with in MedSLT, agreement between the two ways of evaluating appears to be in excess of 98% (Bouillon et al., 2008b). This is good enough that we have felt comfortable in moving all our development over to the new decoupled methodology.

Interlingua-based development is organised around a set of "combined interlingua corpora", with one corpus per subdomain. Each corpus is created in three stages. First, all source-language development corpora for the given subdomain are translated into the interlingua; second, the results are sorted, to group each unique interlingua form together with the source-language examples that mapped into it; finally, each interlingua form is

translated into each target language, with the results again attached to it.

Organising development in this way has several advantages. There is no duplication of effort during multilingual regression testing, since each parsing, translation and generation step is performed exactly once. It is thus possible to run regression tests more frequently, and very easy to write scripts which analyse the corpus to identify lack of uniformity in coverage. The combined corpus also turns out to be useful for constructing resources for the intelligent help component, as we will show in the next section.

4 Intelligent help

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 covers. Our approach to this problem is to equip the system with an intelligent help module (Starlander et al., 2005; Chatzichrisafis et al., 2006) 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, during development, as being within grammar coverage and producing correct translations. The libraries of help examples are specific to each subdomain and language pair, and are extracted from the combined interlingua corpora described in Section 3.2. The extraction process is trivial: for a given subdomain D, source language S, and target language T, the help corpus simply walks through the combined corpus for D, collecting all from-language examples tagged with an S which are attached to an interlingual form which also has a to-language example tagged with a T. This guarantees that help suggestions will always be appropriate to the subdomain and target language currently loaded in the system.

At runtime, the system carries out a second round of recognition using a backup statistical recogniser, and uses the result to select examples from the library which are similar to the statistical recogniser's result in terms of a backed-off N-gram metric; the back-off classes are defined in terms of their syntactic and semantic properties (e.g. "cardinal number" or "singular symptom noun"), and are extracted from the lexicon during the compilation process. (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.

5 MedSLT on a PDA

The top-level components we have described in preceding sections are combined as follows. Input speech is passed to two versions of the source-language Nuance 8.5 recogniser, one using a Regulus-derived grammar-based language model (GLM), and one using a class N-gram statistical language model (SLM). The GLM recogniser is configured to produce N-best output; currently, N is set to 6. The Regulus grammar's semantic representations are compiled down into the Nuance platform's GSL representation language (Rayner et al., 2006, Chapter 8), and recognition hypotheses consequently contain both a word string and an associated source language semantic form.

Each source language semantic form is first subjected to a simple form of surface-oriented ellipsis processing (Rayner et al., 2006, §6.5); for example, in a context where the previous question was "Does coffee give you headaches?", the follow-up question "Chocolate?" would be resolved to a semantic form representing "Does chocolate give you headaches?". The resolved semantic form is then processed through the translation component, using the source-language $\rightarrow$ interlingua rule-set, and the resulting interlingua representation is passed to the "interlingua grammar" (cf. Section 3.2) to check well-formedness. The first hypothesis in the list which produces well-formed interlingua is selected. N-best processing of this type results in a proportional reduction in semantic error rate of about 15–20% on in-coverage material (Bouillon et al., 2008b).

The interlingua representation belonging to the selected hypothesis is then processed through the translation component, using the interlingua → source-language rule-set, to produce a "back-translation" in the source language. Particularly when non-trivial ellipsis processing has been used, the back-translation can be very different from the recognition string, and gives the user a more accurate picture of what the system's main rule-based processing has understood.

In parallel with the grammar-based processing path, the speech input is also submitted to the SLM-based recogniser, which produces a second recognition result. This is passed to the intelligent help component (Section 4), which produces a set of in-coverage help examples. The user is now shown both the back-translation and the help examples, and can make one of three choices: accept the recognition hypothesis corresponding to the back-translation; choose a help example to translate instead (we have found that this is useful about 10% of the time (Starlander et al., 2005)); or abort processing.

If the user does not abort, the selected interlingua representation is translated though the interlingua → target-language rule-set, to produce a target-language representation. Finally, the target-language generation grammar is used to convert this representation into a surface string, which is realised in spoken form either using a TTS engine, where one is available, or by concatenating recorded speech files.

Until recently, all our development work on MedSLT assumed deployment on a high-end laptop. Over the last year, we have ported the application so that it can also be used on a mobile platform. Although it is possible to deploy simple speech applications on stand-alone hand-held platforms (Waibel et al., 2003), it is currently very challenging to run the complex recognition grammars used by MedSLT on embedded devices; experiments using the current version of Nuance VoCon, currently the world's leading embedded recognition platform, showed that even heavily simplified versions of the MedSLT grammars could not be run under VoCon. In view of these problems, we have elected to use the generic distributed solution described in (Tsourakis et al., 2008). The client, which contains the user interface, runs on a Nokia Linux N800 Internet Tablet; most of the heavy processing, including in particular speech recognition, is hosted on the remote server, with the nodes communicating over a wireless network. A picture of the tablet, showing the current (very simple) user interface for the mobile version of MedSLT, is presented in Figure 1. The top row in the display is the back-translation; the sentences appearing under it are produced by the help component.

(Tsourakis et al., 2008) presented performance results for an evaluation carried out on another Regulus application, and showed that recognition performance in the client/server environment was no worse than on a laptop, as long as a similar microphone was used. We used this study as a model, to carry out a simple evaluation comparing the laptop and PDA versions of the English MedSLT recogniser. Six subjects were each given the same 50 MedSLT sentences to read to the system, using three different configurations. In the first, the system ran on a laptop, and the subject used a normal headset with a noise-cancelling microphone; in the second, they used the PDA version, with a comparable headset and microphone; and in the third, they used the PDA, with the machines's built-in microphone. The order in which the subjects used the different versions of the system was varied, so that each system was used equally often in first, second and third place. Table 4 presents the results, which again suggest that performance on the PDA is not worse than on the laptop, as long as a close-talking microphone is used. The somewhat higher error rates, compared to those in the first line of Table 2, are due to the fact that most of the subjects in the comparison test were not native English speakers.

6 Conclusions and further directions

We have presented an overview of MedSLT, a unidirectional controlled-language medical speech translator which can be used to carry out doctor-patient examination dialogues. The architecture's primary motivation is high reliability; other considerations are ease of supporting multiple source and target languages, provision of intelligent help to alleviate the brittleness of a controlled language solution, and deployability on a hand-held device. Tests with med-

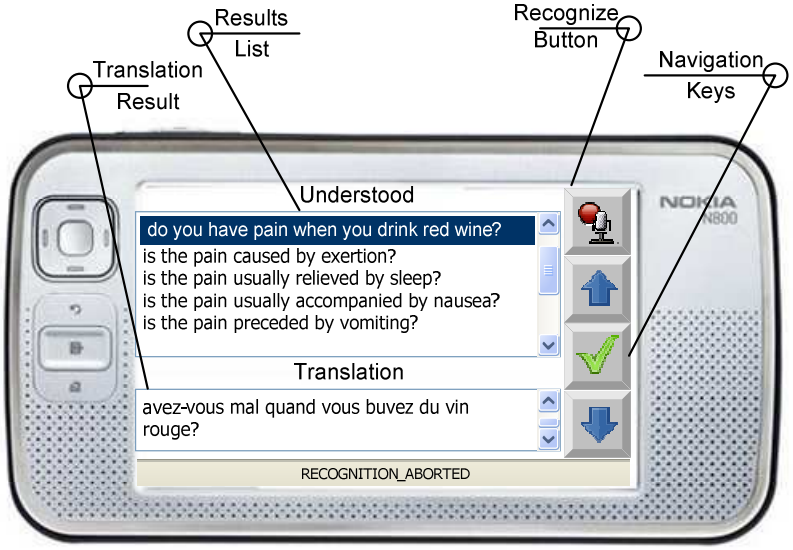

Figure 1: Mobile version of the MedSLT system, running on a Nokia tablet. The first line in the upper pane is the back-translation; the others show output from the help component. The input sentence was "Is the pain caused by red wine?"

Platform	Laptop	PDA	
Microphone	Close	Close	Onboard
WER	7.3%	6.3%	14.7%
SER	23.6%	22.6%	35.7%
SemER	18.5%	16.5%	28.7%

Table 4: Comparison of English language speech understanding performance for three hardware configurations of the MedSLT system, constrasting 1) laptop with close-talking microphone "laptop/close", 2) PDA with close-talking microphone ("PDA/close") and 3) PDA with onboard microphone ("PDA/onboard")

ical students (Chatzichrisafis et al., 2006) show that the system can be used successfully in simulated interviews with standardised patients.

As we have already said, our belief is that MedSLT's architecture is about as ambitious as is feasible, given the limitations of today's technology, if the goal is to build something that health care professionals would actually consider using in a real medical situation. It is certainly possible to build systems that offer far larger coverage (Gao et al., 2006; Ehsani et al., 2006); the question is whether the increased recall is sufficient to motivate the concomitant reduction in precision.

In the long-term, we think that the most promising line of attack is to attempt to combine controlled-language and broad-coverage strategies. If the doctor can verify uncertain information gleaned through a broad-coverage translator by asking questions through a high-precision controlled-language channel, then it may be possible to get the best of both worlds. We hope to begin concrete investigation of these ideas in a later project.

References

D.W. Baker, R.M. Parker, M.V. Williams, W.C. Coates, and Kathryn Pitkin. 1996. Use and effectiveness of interpreters in an emergency department. *Journal of the American Medical Association*, 275:783–8.

P. Bouillon, M. Rayner, N. Chatzichrisafis, B.A. Hockey, M. Santaholma, M. Starlander, Y. Nakao, K. Kanzaki, and H. Isahara. 2005. A generic multi-lingual open source platform for limited-domain medical speech translation. In *Proceedings of the 10th Conference of the European Association for Machine Translation (EAMT)*, pages 50–58, Budapest, Hungary.

P. Bouillon, M. Rayner, B. Novellas, M. Starlander, M. Santaholma, Y. Nakao, and N. Chatzichrisafis. 2007. Une grammaire partagée multi-tâche pour le

traitement de la parole: application aux langues romanes. *TAL*.

P. Bouillon, F. Ehsani, R. Frederking M. McTear, and M. Rayner, editors. 2008a. *Proceedings of the Workshop on Speech Processing for Safety Critical Translation and Pervsaive Applications*, Manchester, England.

P. Bouillon, S. Halimi, Y. Nakao, K. Kanzaki, H. Isahara, N. Tsourakis, M. Starlander, B.A. Hockey, and M. Rayner. 2008b. Developing non-european translation pairs in a medium-vocabulary medical speech translation system. In *Proceedings of LREC 2008*, Marrakesh, Morocco.

N. Chatzichrisafis, P. Bouillon, M. Rayner, M. Santaholma, M. Starlander, and B.A. Hockey. 2006. Evaluating task performance for a unidirectional controlled language medical speech translation system. In *Proceedings of the HLT-NAACL International Workshop on Medical Speech Translation*, pages 9–16, New York.

F. Ehsani, J. Kinzey, D. Master, K. Lesea, and H. Park. 2006. Speech to speech translation for medical triage in Korean. In *Proceedings of the HLT-NAACL International Workshop on Medical Speech Translation*, pages 17–23, New York.

G. Flores. 2005. The impact of medical interpreter services on the quality of health care: A systematic review. *Medical Care Research and Review*, 62:255–299.

G. Flores. 2006. Language barriers to health care in the united states. *New England Journal of Medicine*, 355:229–231.

Y. Gao, B. Zhou, R. Sarikaya, M. Afify, H.-K. Kuo, W.-Z. Zhu, Y. Deng, C. Prosser, W. Zhang, and L. Besacier. 2006. IBM MASTOR SYSTEM: Multilingual automatic speech-to-speech translator. In *Proceedings of the HLT-NAACL International Workshop on Medical Speech Translation*, pages 57–60, New York.

G. Gorrell, I. Lewin, and M. Rayner. 2002. Adding intelligent help to mixed-initiative spoken dialogue systems. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, Denver, CO.

D. Graddol. 2004. The future of language. *Science*, 303:1329–1331.

B.A. Hockey, O. Lemon, E. Campana, L. Hiatt, G. Aist, J. Hieronymus, A. Gruenstein, and J. Dowding. 2003. Targeted help for spoken dialogue systems: Intelligent feedback improves naive user's performance. In *Proceedings of the 10th EACL*, Budapest, Hungary.

S. Knight, G. Gorrell, M. Rayner, D. Milward, R. Koeling, and I. Lewin. 2001. Comparing grammar-based and robust approaches to speech understanding: a case study. In *Proceedings of Eurospeech 2001*, pages 1779–1782, Aalborg, Denmark.

J. Lee and S. Seneff. 2005. Interlingua-based translation for language learning systems. In *Proceedings of ASRU-2005*, San Juan, Puerto Rico.

S. Michie, J. Miles, and J. Weinman. 2003. Patient-centeredness in chronic illness: what is it and does it matter? *Patient Education and Counseling*, 51:197–206.

M. Rayner, D. Carter, P. Bouillon, V. Digalakis, and M. Wirén, editors. 2000. *The Spoken Language Translator*. Cambridge University Press.

M. Rayner, P. Bouillon, N. Chatzichrisafis, B.A. Hockey, M. Santaholma, M. Starlander, H. Isahara, K. Kanzaki, and Y. Nakao. 2005a. A methodology for comparing grammar-based and robust approaches to speech understanding. In *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP)*, pages 1103–1107, Lisboa, Portugal.

M. Rayner, B.A. Hockey, J.M. Renders, N. Chatzichrisafis, and K. Farrell. 2005b. A voice enabled procedure browser for the International Space Station. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (interactive poster and demo track)*, Ann Arbor, MI.

M. Rayner, B.A. Hockey, and P. Bouillon. 2006. *Putting Linguistics into Speech Recognition: The Regulus Grammar Compiler*. CSLI Press, Chicago.

M. Rayner, P. Bouillon, B.A. Hockey, and Y. Nakao. 2008. Almost flat functional semantics for speech translation. In *Proceedings of COLING-2008*, Manchester, England.

M. Rayner. 1988. Applying explanation-based generalization to natural-language processing. In *Proceedings of the International Conference on Fifth Generation Computer Systems*, pages 1267–1274, Tokyo, Japan.

M. Starlander, P. Bouillon, N. Chatzichrisafis, M. Santaholma, M. Rayner, B.A. Hockey, H. Isahara, K. Kanzaki, and Y. Nakao. 2005. Practising controlled language through a help system integrated into the medical speech translation system (MedSLT). In *Proceedings of MT Summit X*, Phuket, Thailand.

M.A. Stewart. 1995. Effective physician-patient communication and health outcomes: a review. *Canadian Medical Association Journal*, 152:1423–1433.

N. Tsourakis, M. Georghescul, P. Bouillon, and M. Rayner. 2008. Building mobile spoken dialogue applications using Regulus. In *Proceedings of LREC 2008*, Marrakesh, Morocco.

A. Waibel, A. Badran, A.W. Black, R. Frederking, D. Gates, A. Lavie, L. Levin, K. Lenzo, L. Mayfield Tomokiyo, J. Reichert, T Schultz, D. Wallace, M. Woszczyna, and J. Zhang. 2003. Speechalator: two-way speech-to-speech translation on a consumer PDA. In *Proceedings of Interspeech 2003*, Geneva, Switzerland.