Practicing Controlled Language through a Help System integrated into the Medical Speech Translation System (MedSLT)

Marianne Starlander, Pierrette Bouillon,
Nikos Chatzichrisafis, Marianne Santaholma
University of Geneva, TIM/ISSCO,
40, Bd. Du Pont d'Arve
CH-1211 Geneva 4, Switzerland
{Marianne.Starlander, Marianne.Santaholma}@eti.unige.ch
{Pierrette.Bouillon, Nikos.Chatzichrisafis}@issco.unige.ch

Manny Rayner, Beth Ann Hockey
UCSC/NASA Ames Research Center,
Mail Stop T-27A, Moffett Field
CA 94035-1000, USA
mrayner@riacs.edu
bahockey@email.arc.nasa.gov

Hitoshi Isahara, Kyoko Kanzaki, Yukie Nakao
National Institute for Communications Technology
NICT, 3-5 Hikaridai Seika-cho, Soraku-gun,
Kyoto, 619-0289 Japan
{isahara,kanzaki}@nict.go.jp
Yukie-n@khn.nict.go.jp

Abstract

In this paper, we present evidence that providing users of a speech to speech translation system for emergency diagnosis (MedSLT) with a tool that helps them to learn the coverage greatly improves their success in using the system. In MedSLT, the system uses a grammar-based recogniser that provides more predictable results to the translation component. The help module aims at addressing the lack of robustness inherent in this type of approach. It takes as input the result of a robust statistical recogniser that performs better for out-of-coverage data and produces a list of in-coverage example sentences. These examples are selected from a defined list using a heuristic that prioritises sentences maximising the number of N-grams shared with those extracted from the recognition result.

1 Introduction

The aim of this paper is to show how to use the idea of controlled language in the medical spoken language translation system MedSLT. The goal of MedSLT is to translate diagnosis questions (such as: "where does it hurt?", "In the front of the head?", "in the back", etc.) asked by a doctor in an emergency setting. It is thus a very limited domain, where questions usually follow a fixed scenario or guidelines[1]. For example, standard examination questions about chest pain always include intensity, location, duration, quality of pain, and the factors that increase or decrease the pain. The answers to these questions can be successfully communicated by a limited number of one or two word responses (e.g. yes/no, left/right, numbers) or even gestures (e.g. nodding or shaking the head, pointing to an area of the body). Translation can thus be unidirectional.

In order to obtain an accurate translation, the system uses a grammar-based speech recogniser. For this type of application a grammar-based approach appears to give better results than a statistical-based recognition (Knight et al., 2001, Rayner et al. 2004, Bouillon et al. 2005). Diagnosis seems to be a very convergent sublanguage, where it is possible to guess the syntactic structures that a doctor will use, and thus to describe them in a grammar.

The advantage of this approach is that the grammar enforces more global constraints on the recognized utterance than the simple bigrams or trigrams of a statistical language model: more complete sentences are thus well recognized, which improves the translation. The drawback is the lack of robustness: if the sentence structure is not in the grammar or if a word is not in the lexicon, the recognition completely fails. Helping the user to find the in-coverage structures is thus the only practical way of using such a system. A possible solution would be to provide the user with precise formulation guidelines, similar to the existing writing rules of the conventional controlled languages, e.g. the Simplified English writing rules (AECMA, 2001), but would the user be able to use those while he speaks? In this paper

[1] For USA, www.guidelines.gov, National Guideline Clearinghouse, as of Monday 9th of May 2005.

we will present the solution we adopted, following previous work implemented by (Gorell et al. 2002) and (Hockey et al., 2003). We provide the user with a simple help feature that guides him through the controlled language used by the system, by listing related in-coverage sentences in a help window. In the rest of the paper we evaluate the relevance of this approach for this specific application. We will first describe the MedSLT architecture in more detail; in section 3, we will explain how the help system works, and finally we will evaluate its efficiency and its effect on the user in sections 4 and 5.

2 The MedSLT system

MedSLT[2] is an Open Source project which is developing a generic platform for building medical speech translation systems; early versions are described in (Rayner, 2002). The basic philosophy behind the MedSLT system architecture is to attempt an intelligent compromise between fixed-phrase translation on one hand, like (Phraselator, 2005), and linguistically motivated grammar-based processing on the other, like Verbmobil, (Wahlster, 2000), or NESPOLE! (Metze et al., 2002ab) and Spoken Language Translator (Rayner et al., 2000).

At run-time, the system behaves essentially like a phrasal translator which allows some variation in the input language. This is close in spirit to the approach used in most normal phrase-books, which typically allow "slots" in at least some phrases ("How do I get to ---?"). However, in order to minimize the overhead associated with defining and maintaining large sets of phrasal patterns, these patterns are derived from a single large linguistically motivated unification grammar, using the Open Source Regulus platform (Rayner, 2003, Regulus, 2005), which implements an example-based specialisation method driven by small corpora of examples.

The linguistically motivated compile-time architecture makes the system easy to extend and modify. In particular, it makes it easy to port the grammar between different medical sub-domains, which seem to be quite convergent.

The translation module is implemented in SICStus Prolog, and is interlingua-based. Translation consists of three main stages illustrated in Figure 1: (1) mapping from the source representation to the interlingua, which may include ellipsis processing; (2) mapping from the interlingua to the target representation and (3) generation, using a suitably compiled Regulus grammar for the target language. In accordance with the generally minimalist design philosophy of the project, semantic representations have been kept as simple as possible, namely a flat list of attribute-value pairs.

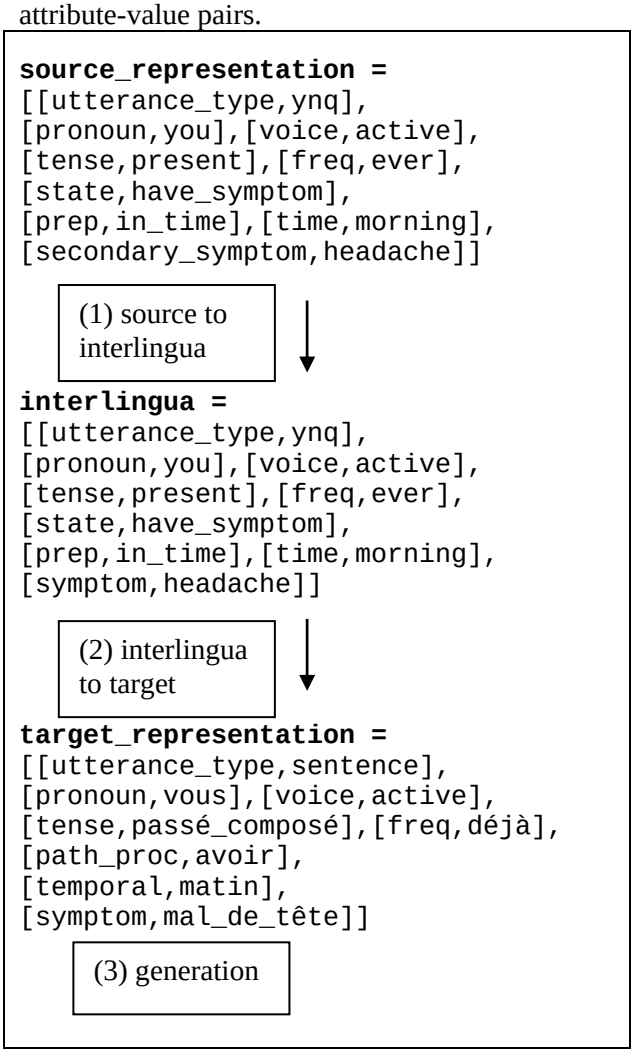

Figure 1: Translation flow for the sentence: "do you ever have headaches in the morning?"

The run time system provides a graphical user interface, which allows the user to select the input and output languages and the sub-domain. The user initiates speech recognition through a push-to-talk interface. The processing of a translation is as follows: first the input speech is recognised by using a recogniser built on top of the Nuance platform (Nuance, 2005). The acquired 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. To increase the reliability of the translations, the interlingua form is first translated back into the source language and shown to the user. The user thus has the possibility of aborting further processing when he

[2] MedSLT: http://sourceforge.net/projects/medslt/ and http://www.issco.unige.ch/projects/medslt/

believes that the system has failed to understand what was originally 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, and finally passed to a speech synthesis unit.

MedSLT currently translates from English into French, Japanese, Finnish and Spanish, and from Japanese and French into English. The covered medical sub-domains are symptom based headaches, chest pain and abdominal pain.

In the following section we will describe the MedSLT help module in more detail.

3 Help Module

The help module's function in this application is to accelerate the learning effect. By providing the user with a list of sentences related to the one that has been uttered, the help module improves recognition success in two ways. First, it educates novice users about possible ways to express questions and demonstrates the system's coverage. Second, it deals with misrecognitions by allowing the user to simply select the intended question out of this help list.

The help module relies on three different general observations made during this project: 1) statistical-based recognisers obtain better WER on out-of-coverage data than grammar-based recognisers; 2) statistical recognisers are more likely to produce a recognition result when users utter out-of-coverage sentences and 3) the users generally improve their performance as they use the system.

In brief, the approach is the following: the help module uses the result of a robust statistical recogniser built with Nuance SayAnything for this application (See (Rayner et. al., 2004) for a description) to display a list of in-coverage example sentences. These examples are selected from a defined list using a heuristic that prioritises sentences maximising the number of N-grams shared with those extracted from the recognition result.

The screenshot in Figure 2 shows how the help system is integrated in the MedSLT system described above. The recognition is performed using both the grammar-based recogniser (primary recognition) and the statistical one (secondary recognition). As before, the result of the grammar-based recognition is then translated back into the source language, and both the recognition result (*Raw recognition result*) and the back translation (*What the system understood*) are shown to the user. The new feature is that the recognition result of the statistical language model is passed to the help module which produces a list of relevant sentences which are then displayed in the lower part of the application window (*Recognition help*). At this point, the user can accept the recognition, and the system's interpretation, by pressing the *Translate button*. The sentence is then translated and synthesized in the target language. Alternatively, the user has the option of looking at the list of help sentences and, on this basis, rephrasing the input sentence or selecting a sentence out of the list of help sentences.

A more precise description of the help module is outlined below. During initialisation, the help module reads in a reference file, which consists of sample sentences, relevant to the active sub-domain (headaches, chest pain, abdominal pain, etc).

Reference sentences are then placed into groups according to their N-grams. For example, the sentence "is the pain deep" will be placed into following bi-gram groups:

- is pain
- pain [deep / gradual]

The sentence "is the pain in the front", will similarly be placed in the first bi-gram group, and will also form two new groups with:

- pain in
- in front

This process is repeated for all reference sentences. At the beginning of each system-user interaction round the recognition result of the statistical recogniser is passed to the help system and is then itself analysed into its N-grams. Each *n*-gram which occurs in both the input and reference sentences contributes n^2 to the relevance score.

Once all N-grams of the input sentence have been processed, the reference sentences are sorted by their relevance score. Up to ten highly relevant sentences are displayed in the application window. As we can see from the example above, the help system takes into consideration stop words ("the", etc.) by excluding them from the scoring process. It also uses syntactic and semantic word classes as shown in Table 1 for grouping related sentences into the same bigram groups. For example, "is the pain gradual" will be put in the same bi-gram class as "is the pain deep since "deep" and "gradual" belong to the same class.

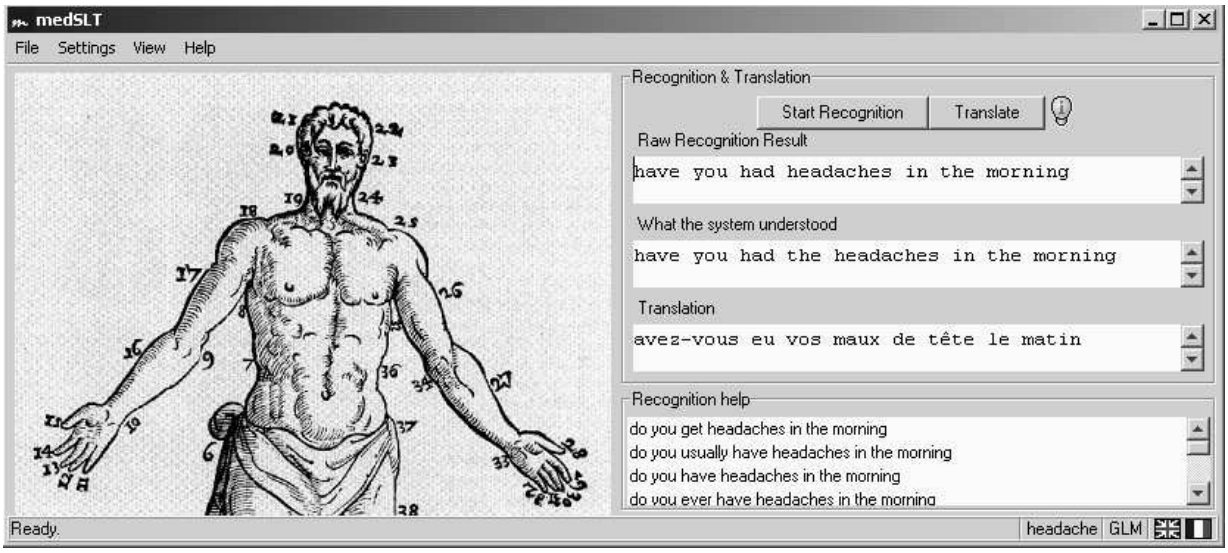

Figure 2: Screenshot of the MedSLT system, for Eng–Fre, using the GLM version and providing recognition help

Class 1	Class 2	Class 3
radiate	sudden	bursting
extend	gradual	deep
spread		sharp

Table 1 Sample Relevance Classes

Classes with synonyms as in Class 1 are useful for showing alternate ways of expressing a certain topic. Classes with antonyms (Class 2) help to group reference sentences with contrasting meaning and are helpful for demonstrating system coverage. Classes containing words of the same category, for example adjectives in Class 3, are useful for displaying the wide range of topics that the system is able to process

In the following sections we will evaluate if this kind of simple help system is efficient to help a doctor to make a correct diagnosis of his patient. We will first describe the evaluation setting of the help module, and then the results will be presented and discussed in section 5.

4 Evaluation

We want to evaluate the efficiency of the help-module described on the medical examination task.

We collected data from six native speakers of English who used the headache version, English to French, of the Open Source MedSLT system. 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 with the help-module switched **on**, and half used the same version with the help-module switched **off**. Most of the subjects needed around 70 utterances to complete their task and diagnose the correct headache type. The length of sessions ranged from 30 to 170 utterances.

We first transcribed the recorded data, and then compared the transcribed sentences with the data available in the session logs for all subjects. These logs contain the following information:

- <primary recognition>: the output of the grammar-based system (GLM).
- <secondary recognition>: the output of the statistical language model (SLM).
- <back translation>: translation of the primary recognition from SL (English) to SL (English) via the interlingua.
- <translation request>: sentence sent to translation, most of the time equal to primary recognition; if this is not the case, this means the user has picked a new sentence from the help module.
- <translation result>: output sentence in target language.

This information enabled us to trace back when the user has picked a sentence from the help module window. We also reconstructed the help sentences that were shown in the window, in order to study these.

We annotated the logs and extracted the following information:

- Utterances for which the back-translation was judged acceptable in comparison to the original transcript were regarded as correctly recognized (W).
- Utterances that were not sent to translation were annotated as interrupted (and thus badly recognized). In the case of a successful processing of the data, we studied whether the subject used the help, when it was available, or not.
- Utterances taken from the help module, making the difference between exact matches (H), similar help sentence (HS) and entirely new help sentence (HN).
- We also marked whether the subject repeated exactly the same sentence several times or if he repeated the same kind of sentence before finding the "good" sentence.

We will now present the results from the previously described data.

5 Results

We will first discuss the effect of the help by comparing the performance of the subjects using help with those not using help. Then we will examine in more detail the way the help module was used and finally the effect of the module on the user, to determine how it helped them to progress towards a more effective use of the MedSLT system.

5.1 Comparison help – no help

Table 2 shows that there is a learning effect showing for both subject classes. But this trend is clearly more marked for subjects using help, as there is a progression of 20.61 percentage points between the first and the last quarter of the session, against 10.66 percentage points for those using MedSLT without help. This shows that the help module is successful in accelerating user training.

Part session	Well-recognised Help ON	Well-recognised Help OFF
All data	**42.17 %**	**46.39 %**
First quarter	**30.70 %**	**38.12 %**
Last quarter	**51.32 %**	**48.78 %**
Difference	20.61	10.66

Table 2: Comparing overall performance of GLM measured by proportion of **well-recognised utterances**, with or without access to help. Figures are presented for all data, just the first and last sessions, and the difference between first and last quarters

A surprising result in Table 2 is, however, that the subjects without help achieve a slightly better overall performance (46. 39%) than those using help (42.17%). This difference is probably due to the fact that in the category of subjects using help, one subject performs particularly below average, while in the category not using help one subject has particularly high scores, as shown in Table 3 below.

	Entire session	**First quarter**	**Last quarter**	**Diff.**
Sub.1, H+, 77 utt.	53.3 %	42.1 %	78.9 %	+ 36
Sub.2, H+, 62 utt.	33.9 %	**37.5%**	**25.0%**	**- 12**
Sub.3, H+, 33 utt.	39.4 %	12.5 %	50.0 %	+ 37
Sub.4, H-, 71 utt.	**62.0%**	38.9 %	61.1 %	+ 22
Sub.5, H-, 80 utt.	38.8 %	35.0 %	40.0%	+ 5
Sub.6, H-, 170 utt.	38.5 %	40.5 %	45.2%	+ 4.7

Table 3: Performance of all subjects by proportion of well-recognised utterances. H+ stands for help ON and H- for help OFF. Figures are presented for all data, just the first and last sessions, and the difference between first and last quarters

Table 3 also shows that two of the subjects using help actually improve their performance by more than 36%, while subject 2 shows the contrary, performing worse in the last quarter of the session. This "counter performance" is explained by the fact that the subject was trying the same sentence very often (10 times, out of a total of 62 utterances on the entire session), for which there was unfortunately no help available, because the quality of the secondary recognition produced by the SLM was too bad.

This brings us to our next observation: how does the help system prevent the doctor from stagnating? If a sentence cannot be recognised, the user should find an alternative, thanks to the help. The following Table 4 shows indeed that subjects using help have a far lower repetition rate (1.94 %). The subjects without help tend to repeat a badly recognised sentence (11.47 %) because they have no access to the help examples.

	% of repeated sentence
Subject 1	2.60 %
Subject 2	3.23 %
Subject 3	0 %
Average H+	**1.94 %**
Subject 4	8.45 %
Subject 5	11.25 %
Subject 6	14.70 %
Average H-	**11.47 %**

Table 4: Percentage of repeated sentences, comparing subjects with and without help.

We will now look more closely at the influence of the help module on the users.

5.2 Using help

We will now focus on the three subjects that used the help module. How often did they effectively use it, and how did they proceed? Table 5 shows that the help is used in three different ways. The subject can either find an exact match to the intended question in the help window, or pick a sentence with similar meaning, or even select an entirely new sentence from the help. As a whole, 10.55 % more sentences could be processed by the system thanks to the help module. This percentage amounts to 19.2% of all badly recognised sentences that could be "recovered" this way.

	Subj1	**Subj2**	**Subj3**	**Average**
Help used on total number of sentences	14.3 %	11.3 %	6.1 %	**10.55** %
Help provides exact match (H)	5.6 %	2.4 %	5 %	4.3 %
Help gives a similar sentence (same meaning) (HS)	8.3 %	0 %	5 %	4. 4 %
Help providing entirely new sentence (HN)	16.7 %	14.6 %	0 %	10.4 %
Total help usage on badly recognised sentences	30.6 %	17.1 %	10 %	**19.2** %

Table 5: Proportion of help use by the different subjects

The question here is: does the help influence the user too much and make them deviate from the original sentence? Most of the time, when the user cannot pick an exact match from the help window, he would choose a similar sentence, that has almost exactly the same meaning (4.4 %). In that case, the mission of the help module has been accomplished, as the communication between the doctor and the patient can continue. But we noticed, that quite often, as shown in the table above, the user would select an entirely new sentence if a similar one was not available (10.4 %), for example: "Does sleep usually relieve your headache", instead of the intended sentence: "Does coughing relieve your pain" the intended sentence. In this case, the help sentence is still related to the intended one. In that type of example the user is not changing his strategy entirely, but he is probably only changing slightly the order of questioning. Some examples show a more radical change like picking the sentence: “is your headache usually caused by bright light”, when the original utterance was: “what do you usually take for your headache”. This practice of picking entirely new sentences that are not even related semantically from the help window helps the user to progress with the diagnosis, where insisting on one sentence that is not parsing blocks progress.

Interestingly, a last observation is that the help was less used in the second half of the session, with an average over the subjects of 19.59 % for the first half and 1.71 % for the second half. This would also tend to reinforce the idea that the user feels more familiar with the coverage of the system; the training task of the help module is thus fulfilled.

6 Conclusion

The results of our evaluation demonstrate that even such a simple help module based on a robust recogniser helps to improve the translation produced by MedSLT in two different ways: through lowering the amount of badly recognised sentences and through training the user to produce in-coverage sentence structures. This preliminary work shows promise for experimenting with more sophisticated help strategies and suggests many interesting directions for further work. One obvious step would be to use our help corpus to see if we could associate help messages with badly recognised sentences, in the same way as a controlled language checker. Messages could address words outside the coverage of the lexicon used by the recogniser or bad collocation pairs, for example wrong preposition after a verb.

References

AECMA: The European Association of Aerospace Industries. 2001. *AECMA SIMPLIFIED ENGLISH: A guide for the Preparation of Aircraft Maintenance Documentation in the International Aerospace Maintenance Language*.

P. Bouillon, M. Rayner, N. Chatzichrisafis, B.A. Hockey, M. Santaholma, M. Starlander, H. Isahara, K. Kanzaki, Y. Nakao 2005. *A 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.

G. Gorrell, I. Lewin, M. Rayner. 2002. *Adding Intelligent Help to Mixed-initiative Spoken Dialogue System.* In "Proceedings of the Seventh International Conference on Spoken Language Processing", Denver, CO, USA.

B. A. Hockey, O. Lemon, E. Campana, L. Hiatt, G. Aist, J. Hieronymys, A. Gruenstein, J. Downing. 2003. *Targeted Help for Spoken Dialogue Systems: intelligent feedback improves naïve users performance.* In "Proceedings of the 10th conference of the European Chapter of Association for Computational Linguistics", Budapest, Hungary.

S. Knight, G. Gorrell, M. Rayner, D. Millward, R. Koeling, I. Lewin. 2001. Comparing Grammar-*Based and Robust Approaches to Speech Understanding: A Case Study.* In "Proceedings of Eurospeech", Aalborg, Denmark.

MedSLT, 2005. http://sourceforge.net/projects/medslt/ and http://www.issco.unige.ch/projects/medslt/. As of 10 May 2005.

F. Metze, J. McDonough, H. Soltau, A. Lavie, L. Levin, C. Langley, T. Schulzt, A. Waibel, R. Cattoni, G. Lazzari, N. Mana, F. Pianesi, E. Pianta. 2002a *Enhancing the Usability and Performance of NESPOLE!: a Real-World Speech-to-Speech Translation System,* In "Proceedings of HLT 2002", San Diego, California U.S., March 2002.

F. Metze, J. McDonough, H. Soltau, C. Langley, A. Lavie, L. Levin, T. Schultz, A. Waible, R. Cattoni, G. Lazzari, N. Mana, F. Pianesi, E. Pianta. 2002b. *The NESPOLE! Speech-to-Speech Translation System*. In "Proceedings of HLT 2002, San Diego, California U.S, March 2002.

Nuance, 2005. http://www.nuance.com. As of 10 May 2005.

Phraselator (2005). http://www.phraselator.com. As of 10 May 2005.

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

M. Rayner, B. A Hockey and J. Dowding. 2003. *An Open Source Environment for Compiling Typed Unification Grammars into Speech Recognisers*. In "Proceedings of the 10th Conference of the European Chapter of Association for Computational Linguistics (demo track)", Budapest, Hungary.

M. Rayner, P. Bouillon. 2002. *A flexible Speech to Speech Phrasebook Translator*. In "Proceedings of ACL-02 Workshop on Speech-to-Speech Translation: Algorithms and Systems", Philadelphia, USA.

M. Rayner, P. Bouillon, B. A. Hockey, N. Chatzichrisafis, M. Starlander. 2004, *Comparing Rule-Based and Statistical Approaches to Speech Understanding in a Limited Domain Speech Translation System*, in "Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation – TMI 2004", Baltimore, MD USA.

Regulus, 2005. http://sourceforge.net/projects/regulus/. As of 10 May 2005.

W. Wahlster (Ed.) .2000. *Verbmobil: Foundations of Speech-to-speech Translation*, Berlin, Heidelberg, New York, Springer-Verlag.