Relating recognition and translation quality with usability of two different versions of MedSLT

Marianne Starlander
ISSCO/TIM/ETI, University of Geneva
40, Bd du Pont d'Arve
1211 Geneva 4
Marianne.Starlander@unige.ch

Paula Estrella
ISSCO/TIM/ETI, University of Geneva
40, Bd du Pont d'Arve
1211 Geneva 4
Paula.Estrella@unige.ch

Abstract

The aim of this paper is to further compare two versions of our spoken language translator MedSLT which differ in terms of grammatical coverage; the first version is restricted to Yes-No answers and short elliptical sentences, while the second version allows all of the above but also full sentences. Previous work in this direction has focused on the relation between Word Error Rate (WER) and task performance (Starlander et al., 2008). In this paper this comparison is extended to study in more detail the task performance with respect to the number of successful interactions and to explain why some utterances could not be correctly recognized. Additionally this paper will explore how these results correlate with the quality of the translation produced by both versions of the systems.

1 Introduction

MedSLT is a speech-to-speech translation system that enables a health care provider to communicate with a foreign patient, by translating diagnosis questions and answers. Due to the safety-critical nature of our domain, our main goal is to achieve precision, which explains why we chose a linguistic recognition and translation based on unification grammars and specialization techniques, provided by the Regulus toolkit (Rayner et al., 2006). Example-based learning techniques enable us to extract domain-specific grammars from our unification grammar for English (Rayner et al., 2006) and for Romance languages (Bouillon et al., 2007b) to create two different English to Spanish versions of MedSLT varying according to the coverage, by modifying the operationality criteria (Starlander et al., 2008) and providing the system with a training corpus for each version of the system.

This faculty has enabled us to test different communication configurations appropriate to doctor-patient communication. In the past we have studied which system architecture is best suited for a system such as MedSLT undergoing different steps. First, through a Wizard of Oz experiment, we could confirm that the bidirectional version of our system outperformed the unidirectional version. We wanted to examine this point since a more open system could also lead to more errors and out-of coverage usage of the system, implying that, in the end, a more restricted but more predictable unidirectional system (only operated by the physician) could be more efficient. However, from the usability point of view, the bidirectional was preferred over a fixed-phrase or a unidirectional version of the same system (Starlander, 2007). The second step, described in the present paper, aims at determining which type of coverage is best suited for a bidirectional system intended to be used in an emergency setting. We have therefore developed two distinct versions of MedSLT: Version 1 (V.1) is more restricted and only allows yes/no and ellipsis, while Version 2 (V.2) also covers full sentences. The coverage is clearly different, since V.1 is based on a corpus containing only 141 entries, while V.2 has been specialized from a corpus of

453 entries, supporting dialogues like the fragment shown bellow:

Phy: Where is the pain?
Tran: ¿dónde le duele?
Pat: Me duele la garganta
Trans: I have a sore throat
Phy: How long have you had a sore throat?
Trans: ¿desde cuándo le duele la garganta?
Pat: Me duele desde hace dos días
Trans: I have experienced the pain for two days

Example 1. Dialogue sample with Version 2 (unrestricted)

The current paper is part of a wider study where the final goal is to see whether the traditional way of evaluating Spoken Language Translation (SLT) systems (i.e. by studying the resulting speech recognition on one hand and the quality of the machine translation with or without automatic metrics on the other hand) correlates with the results obtained when studying a SLT system from a usability point of view. We will address all three aspects in this paper by trying to find out if there is a correlation between the answers given to usability questionnaires, by measuring the quality of speech recognition (using the metrics WER and SER) and by assessing MT quality by means of both human and automatic metrics.

In section three we will focus on task completion aspects, studying in detail all the collected data in terms of number of successful utterances, number of interaction errors and number of recovered errors thanks to the help module integrated in MedSLT (Starlander et al., 2005). In section four we focus on the speech recognition (SR), comparing the surface measures WER and SER with speech understanding error rate (SemER) measured in terms of the percentage of sentences accepted by the users. In section five, we will focus on the quality of the machine translation using both automatic and human metrics. While studying these aspects, we are mainly interested to see whether the ranking for our two systems remain the same across the various types of evaluation described.

Before entering the main body of the paper, we will in section two briefly sketch how MedSLT works, focusing in particular on aspects connected to Interlingua based translation and grammar specialization. We also describe the data used in the experiments, and how it was collected.

2 MedSLT versions and data collection

The two versions of our bidirectional medium-vocabulary SLT for the medical domain share most of the characteristics of the general MedSLT system, and we will briefly describe them in the following subsections.

2.1 Linguistic recognition and translation

A recent study (Rayner et al., 2009) suggests that current state of the art statistical machine translation systems do not provide the level of accuracy needed in order to produce a system usable in a hospital setting. This is why MedSLT is a controlled-language application, where users have to stay within the domain of diagnosis questions and answers. This choice entails that the coverage of the system is restricted to sentences covered by our grammar; however, the system is backed up with a help module (Starlander, 2005) that guides the users towards the grammar's coverage. The help module's output is based on a library of utterances which have already been evaluated as being within coverage and producing correct translations. If the user encounters a problem due to an out-of-grammar construction, he can choose a correct sentence from a list generated by the back-up statistical recognition. Although restriction to the coverage of the grammar is of course a problem, the upside is that it is also a warrant of good machine translation quality, since everything that is recognized can be parsed by the grammar and thus should be translatable. Another warrant for security is that recognized utterances are first presented to the user to be checked in the form of a back-translation before being sent to translation module. The back-translation is an utterance that has gone through all the steps of the MedSLT process: i.e. recognition and translation. In this paper we only report on the English to Spanish language combination, but MedSLT translates into six different languages through an Interlingua architecture (Bouillon et al., 2007a). The Interlingua constitutes the link between the representation of the original sentence and the target representation. So in this case a back-translation for a sentence uttered by a physician would be recognized and translated from English into our Interlingua representation and then back again into English, so that the physician can check the accuracy of the treated sentence.

In our medical context, we prioritize the coherence of our translations, by generating a single translation for all sentences sharing the same meaning, hereby avoiding the multiplication of translation rules. Another intended advantage of this architecture is that the translation focuses on the meaning of the sentence, thus producing less literal but more idiomatic translations. In section 5 we will study if this is really the case, and we will test automatic translation metrics on the data produced during the data collection that we will now describe.

2.2 Data collection

In order to compare the two versions of the system, a data collection was organized at the Children's Medical Center, Dallas, using physicians and standardized patients. The latter were professional medical Spanish interpreters from Children's Medical Center Dallas trained for a specific task. The eight physicians that participated had to determine whether the patient they were faced with suffered from a bacterial infection (streptococcal pharyngitis) or not. We asked the sixteen standardized patients to simulate a history of symptoms consistent with viral pharyngitis or streptococcal pharyngitis, using eight different carefully scripted, fixed scenarios. None of the participants had used the system before. Each standardized patient used both versions of the system with two different physicians. Half of the standardized patients started with the more constrained version (V.1) and half with the less restricted version (V.2). In the following sections we will analyze the collected data in detail.

3 Usability evaluation

We will start our evaluation by briefly giving the results of the questionnaires, which will be used as a baseline for the ranking of the two systems. Then we will go into a more detailed analysis of the systems' usability by applying the metrics usually employed in SLT evaluation exercises (Aberdeen et al., 2005; Stallard, 2000): task completion, length of dialogue in terms of time and number of interactions and metrics defined in terms of the relative proportion of successful and failed interactions.

3.1 Ranking by questionnaires

After having used both versions of the system, users answered a satisfaction questionnaire to indicate which one they preferred. The answers to the questionnaire are summarized in Table 1, which shows that both physicians and patients were clearly in favor of the less restricted version of MedSLT (V.2). It can also be noted that selecting one of these systems seems to be harder for physicians as shown by the larger number of users that did not specify the version of the system they preferred (column *Not sure* in Table 1).

	V.1	**V.2**	**Not sure**
Patients	18.8	68.8	12.5
Physicians	12.5	62.5	25.0

Table 1. Percentage of users in favor of each system

This can easily be explained by the fact that the differences between the versions are not directly affecting them, since the physician-side coverage remains the same. The choice between V.1 and V.2 only affects the physician if the patient appears to be having problems answering his questions.

3.2 Task completion

In this experiment, a task is considered successfully accomplished when a doctor correctly identifies the type of infection the patient was asked to simulate. As shown in Table 2, V.2 clearly outperforms V.1 in this part of the evaluation, obtaining 62.5% of correct diagnosis compared to 43.8%. Intuitively, this result suggests that it is more difficult to obtain the necessary information from patients by asking questions that must be answered only by yes-no or elliptical sentences.

	V.1	**V.2**
No diagnosis	18.8	12.5
Correct diagnosis	43.8	62.5
Incorrect diagnosis	37.5	25.0

Table 2. Task completion in % of achieved diagnosis

In the cases where no diagnosis was given, it was due to the fact that the physicians, despite the instructions to do so, were reluctant to make a diagnosis without a direct strep test or examination. Some diagnosis errors are most easily explained by the fact that the eight scenarios predetermined with our medical advisor and partner Glenn Flores were

of different levels of difficulty. Two cases out of the eight were particularly difficult to diagnose correctly. Moreover, in the more difficult scenarios, the fact that some questions could not be asked because of coverage holes such as "Where is the rash" and "Do you have diarrhea" contributed to mislead the physicians. Finally, in at least one case, it was clear that the patient did not follow the set scenario, which also led to a diagnosis error.

Notwithstanding these observations, it should be kept in mind that, as a whole, the less restricted version ranked first on the task completion. We will now study if we find a similar result when looking more closely at the interactions performed during the data collection.

3.3 Interaction analysis

In order to analyze in more detail all the interactions with the system, we classified interactions into 8 categories, where we differentiate the *accepted utterance* (AC) from the accepted utterances coming from the *help system* (AC_help), which were not taken into account in the previous studies. Interactions that were not accepted are classified into: *bad recognition* (NA_BR) and *lack of translation* (NA_NT), according to the reason for not being accepted by the user. Additionally, we distinguish *interaction errors* in both accepted and unaccepted interactions (AC_IE and NA_IE).

User	**Phy**	**Phy**	**Pat**	**Pat**
System	**V.1**	**V.2**	**v.1**	**v.2**
AC	42.8	42.7	59.1	59.1
AC_help	15.3	17.3	6.0	6.4
AC_IE	0.7	0.8	0.8	3.6
REP	2.4	1.5	0.2	0.0
Total AC	61.2	62.3	66.0	69.1
NA	2.9	1.9	1.7	0.6
NA_BR	29.2	31.3	25.9	25.6
NA_NT	3.1	3.1	2.8	1.1
NA_IE	2.9	0.9	1.5	0.8
NA_help	0.7	0.6	2.0	2.8
Total NA	38.8	37.8	34.0	30.9

Table 3. Interaction analysis by users for the restricted version (V.1) and the less restricted version (V.2)

As shown in Table 3, we also counted the number of sentences that were replayed by the users (REP) to make an utterance heard/recognized again. This option to re-launch a dialogue was primarily used by physicians (around 2% of times) with slightly higher figures for the more restricted version V.1. This can be explained by the fact that the ellipsis treatment only works if the context sentence (i.e. the information about the original question) is still available for treatment, as explained in (Bouillon et al., 2007c). Thus, repetition of utterances occurred when an interaction failed several times and the context was lost. This situation is less frequent when using V.2 of the system, where users could also generate full sentences, which do not need the context.

Regarding the usage of the help system, it is striking to observe that it was used twice as much by the physicians as by the patients, with a slightly higher usage for the less restricted version (V.2). This trend is explained by the fact that the proportion of utterances accepted straight away is also much higher for the physicians that have to learn the systems' coverage in terms of content, while the patients essentially have to find the correct way to formulate an answer according to which version of the system is being used. In this part of the analysis, the difference between V.1 and V.2 is almost nonexistent,compared to the number of interaction errors (row AC_IE of Table 3). Patients had 3.6% interaction errors with V.2 (less restricted version) and all the other figures remain under 1%. Most of these interaction errors on the patient side are due to the fact that the version used for the experiment was not able to recognize the negative form because of a coverage hole. The users of V.2 had no other choice than to answer to negative sentences by using a short "no" answer. An example is presented below, where the produced back-translation could easily be misread, explaining the high proportion of interaction errors in V.2. This problem could not occur in V.1, since in that case the only possible answer was "no" and not a full sentence.

The following examples illustrate the importance of the back translation, since in these cases only checking the recognition results would not have been enough to detect some errors. The first example of error prone utterances in Table 4 presents the answer of a patient to the question whether he/she has a cough ("*no, I don't have a cough*") that is translated as "*no, I cough*"; this type of error is easily detectable when comparing the recognition and back-translation outputs (rows

in bold of Table 4); similarly, the same error was detected in the second example.

	Ex. 1	Ex. 2
Transcript	no, no tengo tos	no muy fuerte
SR	**no tengo tos**	**no muy fuerte**
Backtransl.	**tengo tos**	**el dolor es muy fuerte**
Transl.	No, I cough	the pain is very severe

Table 4. Error prone utterances using the less restricted version (V.2) of the system

It is difficult to rank the two versions according to this detailed interaction analysis; V.2 produced a higher percentage of accepted interactions, but this this figure includes a relatively large number of interaction errors. However, V.2 scored over V.1, in that a smaller number of accepted utterances lacked translations; in this category, V.2 obtained 1.1% for the patients vs. 2.8% obtained by V.1.

The next section will study the performance of both systems in terms of recognition quality.

4 Recognition quality

(Wang et al., 2003) argue that the classical way to evaluate speech recognition quality by measuring Word Error Rate (WER) and Sentence Error Rate (SER) do not really reflect the quality of SL systems. Therefore, we will evaluate the two versions of MedSLT with a usability metric, namely the Semantic Error Rate (SemER) that we measured as the acceptance of a sentence by a group of users. Additionally, results will be compared to those of the classical WER and SER metrics, in order to explore a possible relation between the classical metrics for SLT evaluation and the real usability of a system.

4.1 WER and SER

Table 5 shows the WER/SER results calculated as in (Jurafsky and Martin, 2009). It can be noted that results differ considerably for doctors and patients using both systems: WER and SER are much lower for patients. This is an expected result given the nature of utterances generated by each participant in the communication process. Although the SER for patients using V.1 seems to be salient (as it is < 0.1%) it can be easily understood when analyzing the types of sentences in this category: patients using V.1 generated around 85% short sentences (53% ellipsis + 32% yes/no + 15% full) that do not present major errors while V.2 produced mostly full sentences (45% full + 26% ellipsis + 29% yes/no).

At the system level, Table 5 shows that WER/SER scores are higher for V.2 (less restricted) than for V.1 (restricted). Intuitively, this shows that offering the user a less constrained interface also induces more errors in the current setup; we found that more utterances in V.2 were out of coverage. The contribution to this result of the erroneous treatment of negations (discussed Section 3) will be further studied in future evaluations using a fixed version of the system.

	V.1	**V.2**
WER Physicians	29.4 %	32.7 %
WER Patients	14.1 %	25.2 %
SER Physicians	36.0 %	39.6 %
SER Patients	**0.1%**	37.6 %

Table 5. Results for WER and SER metrics

To follow the current literature about the actual impact of WER and SER on the real communication quality with a SLT system, we will now compare those figures with what we call Semantic Error Rate.

4.2 SemER

Since users are shown a back-translation of the treated sentence, it is possible to determine the semantic error rate in terms of user acceptance of sentences.

Although SER is around 40% for physicians, the percentage of rejected sentences is lower (around 32%) because of the nature of the recognition strategy adopted here, which makes the loss of information (e.g. no articles) trivial and ensures a binary behavior: it either produces a well constructed sentence or no primary recognition at all (instead of producing arbitrary or nonsense results).

Most of the sentences that were not accepted due to bad recognition were either out of coverage (e.g. full sentences when using V.1) or very short sentences such as “un poco”.

	V.1	V.2
SemER overall	317/1006 (31.5%)	339/1015 (33.4%)
SemER Doctors	179/516 (34%)	203/533 (38.0%)
SemER Patients	138/490 (28.2%)	136/482 (28.2%)

Table 6. Percentage of rejected sentences because of bad recognition.

The overall performance according to SemER is quite similar for V.1 and V.2. Not surprisingly, the greatest difference appears on the physicians' side when using V.2 (38%), while the number of sentences rejected by patients remains constant across systems.

The resulting ranking is correlated with previous scores of WER and SER, showing that the number of rejections is primarily due to recognition problems.

5 Translation quality

In order to measure the quality of the translations generated by both systems, a human and an automatic evaluation were performed on the utterances that make up the dialogs between patients and doctors. In this section the results of these evaluations are summarized and discussed.

5.1 Human metrics

Given the intended context of use and purpose of the system, that is, communication between a doctor and his patient for a successful diagnosis, the evaluation of the translation quality focused on the end usage of the produced translations. Therefore, this evaluation tried to leave purely linguistic aspects on the side, i.e. instead of judging the syntactic or linguistic aspects of the translations, the evaluator's task consisted on indicating whether the message from a patient was correctly sent to the doctor. For this purpose, the 4-point scale chosen relates the meaning of a sentence to its potential to create misunderstandings or false communication between a doctor and his patient. The following scale was given to the evaluators:

- 4: The translation is completely correct. All the meaning from the source is present in the target sentence.
- 3: The translation is not completely correct. The meaning is slightly different but it represents no danger of miscommunication between doctor and patient.
- 2: This translation doesn't make any sense, it is gibberish. This translation is not correct in the target language.
- 1: This translation is incorrect and the meaning in the target and source are very different. It is a false sense, dangerous for communication between doctor and patient.

The Spanish language Interpreters of the Dallas Childrens Hospital kindly provided three evaluations for each sentence. As shown in Table 7, the judges favored the restricted version of the system (V.1), which only outperforms V.2 by around +0.03.

	Phy V.1	Phy V.2	Pat V.1	Pat V.2
Human	3.71	3.55	3.58	3.56

Table 7. Result of the human evaluation.

Table 8 gives a breakdown into individual categories.

Cat	Phy V.1	Phy V.2	Pat V.1	Pat. V.2
4	70.7%	68.0%	80.9%	73.1%
3	24.0%	25.8%	14.8%	17.4%
2	2.2%	2.2%	0.9%	2.0%
1	2.3%	3.5%	3.2%	7.1%

Table 8. Translation quality by category.

As can be seen in Table 8, the less restricted version V.1 gets the highest percentage of totally correct sentences, while V.2 scores the highest ratio of translations deemed as dangerous (7.1%). If we analyze why this figure is so high, we can observe that again, most of the dangerous translations are due to the false treatment of negation, second comes the tense problems (**Q**: are you allergic to antibiotics? **A**: sí a la penicilina **T**: I **was** allergic to penicillin), but in many examples, the evaluators choice does not quite seem justifiable.

We can conclude that for the human metrics, V.1 is better than V.2. We will now see if the automatic metrics follow the same direction.

5.2 Automatic metrics

For the automatic evaluation a set of standard metrics were chosen, including WER, Position Independent Error Rate (PER), Precision/Recall/F-

measure and BLEU/NIST. For the application of these metrics 3 reference translations were created.

Metric	Phy V.2	Phy V.1	Pat V.2	Pat V.1
BLEU	0.31	0.55	0.42	0.54
NIST	4.58	5.74	5.39	5.91
WER	0.52	0.38	0.41	0.36
PER	0.47	0.34	0.36	0.31
Precision	0.65	0.73	0.76	0.77
Recall	0.68	0.79	0.69	0.73
Fmeasure	0.67	0.76	0.71	0.75

Table 9. Result of automatic evaluation

As can be noted from the figures in Table 9, the restricted version of the system achieves significantly higher scores than the unrestricted version. This seems to be a direct consequence of the type of utterances generated by each version of the system, the restricted one favoring short sentences that produce more matches against a set of references, as these do not present a wide variation. Additionally, this result is coherent with the results found in the previous section dealing with the recognition quality but disagrees with the system's usability results.

To study the correlation between the different metrics applied, the Pearson coefficient was calculated as in (Estrella et al., 2007), where the bootstrapping technique is employed to artificially generate a large number of sample points on which the correlations are calculated. The following table presents the correlation between the different metrics applied.

Metrics	Phy V.1	Phy V.2	Pat V.1	Pat V.2
H vs. B	0.25	0.23	0.40	0.29
H vs. N	0.21	0.23	0.35	0.32
H vs. Pr	0.25	0.32	0.44	0.36
H vs. Re	0.09	0.15	0.49	0.50
H vs. F	0.19	0.24	0.48	0.47
H vs. Per	-0.26	-0.31	-0.49	-0.41
H vs. Wer	-0.24	-0.29	-0.50	-0.42

Table 10. Correlation between the different metrics used to evaluate the quality of the translations.

Despite the weak correlation coefficients obtained for all the versions of the system, the restricted version on the patients' side achieves the highest scores, suggesting that the selected metrics could be used as preliminary indicators of the overall quality of the system, given that there are many short answers that are easily evaluated as correct by both human-based and automatic metrics. However, for both versions on the physicians' side the correlation is negligible, suggesting that these metrics should be carefully used.

The results obtained confirm that this type of evaluation provides the same ranking as the evaluations of the recognition quality and translation quality by human judges but do not reflect users' view of quality.

6 Conclusion and perspective

In this paper we assessed different aspects of the quality of an SLT system for the medical domain. The results obtained with the various methods applied show a discrepancy between usability measures (Section 3) and automatic measures of the translation/recognition quality. Comparing the results of Section 4.1 to those of Section 4.2 perfectly illustrate how the linguistic recognition is only slightly affected by some word errors, such as the omission of single words or the mishandling of negations. The latter has lead to certain translation problems (see Section 5) that could be solved thanks to the back- translation, thus not affecting the overall performance of the system but possibly explaining diagnosis errors. The human-based MT metric rank the more restricted version (V.1) first, contradicting the questionnaire answers, which ranked V.2 first and almost to the unanimity.

These results suggest that the quality measures applied here are somehow artificial since they do not correlate to the usability and subjective evaluations. It turns out that users give their preference to the (at that time) buggy system despite objective lower results. Users of such systems apparently do not expect a perfect translation but want a system that works well enough to avoid dangerous interactions. Furthermore, what is more important to users is a system to which they can adapt more easily and the freedom given by the less restricted version is more appreciated than the higher accuracy provided by the restricted one. The help tool has widely contributed to this capacity of adaptation.

Based on the present study future work could aim at combining the usability and standard metrics to achieve an enhanced, more meaningful quality measure. These results encourage to choose a global approach to evaluate SLT systems as described in the shared task (Rayner et al., 2008). Future work would consist in applying the scale described in the above cited paper, where all the

types of interaction problems (like in Section 3) are taken into account at the same level as the translation quality measure, to the same data to compare the final ranking obtained with this method.

Acknowledgments

We would like to acknowledge Glenn Flores and the entire team at the Dallas Children's Hospital that made these test possible by welcoming us and participating to this study in February 2008.

References

Aberdeen, John, Oshika, Beatrice, Condon, Sherri, Harper, Lisa and Philips, Jon. 2005. MITRE.

Bouillon, Pierrette, Flores, Glenn, Starlander, Marianne, Chatzichrisafis, Nikos, Santaholma, Marianne, Tsourakis, Nikos, Rayner, Manny and Hockey, Beth Ann. 2007a. A Bidirectional Grammar-Based Medical Speech Translator. In *Proceedings of Workshop on Grammar-based approaches to spoken language processing within the 45th Annual Meeting of the Association for Computational Linguistics*. Prague, Czech Republic, June 29, pp. 41-48.

Bouillon, Pierrette, Rayner, Manny, Novellas Vall, Bruna, Starlander, Marianne, Santaholma, Marianne, Nakao, Yukie and Chatzichrisafis, Nikos. 2007b. Une grammaire partagée multi-tâche pour le traitement de la parole : application aux langues romanes. *Traitement Automatique des Langues (TAL)*, 47(3/2006).

Bouillon, Pierrette, Rayner, Manny, Starlander, Marianne and Santaholma, Marianne. 2007c. Les ellipses dans un système de Traduction Automatique de la Parole. In *Proceedings of Traitement Automatique des Langues Naturelles.* Toulouse, France, June 5-8, pp. 53-62.

Estrella, Paula, King, Maghi and Popescu-Belis, A. 2007. A New Method for the Study of Correlations between MT Evaluation Metrics. In *Proceedings of 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07).* Skövde, Sweden.

Jurafsky, D. Dan and Martin, James H. 2009. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition.*

Rayner, Manny, Bouillon, Pierrette, Flores, Glenn, Ehsani, Farzad, Starlander, Marianne, Hockey, Beth Ann, Brotanek, Jane and Biewald, Lukas. 2008. A Small-Vocabulary Shared Task for Medical Speech Translation. In *Proceedings of Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications (Coling 2008).* Manchester, 23 August, pp. 60-63.

Rayner, Manny, Estrella, Paula, Bouillon, Pierrette, Hockey, Beth Ann and Nakao, Yukie. 2009. Using Artificially Generated Data to Evaluate Statistical Machine Translation. In *Proceedings of GEAF Workshop at ACL-IJNLP.* Singapore, pp. 13-21.

Stallard, David. 2000. Evaluation Results for the Talk'n'travel System. In *Proceedings of Applied Natural Language Processing Conference.* Seattle, Washington.

Starlander, Marianne. 2007. Using a Wizard of Oz as a baseline to determine which system architecture is the best for a spoken language translation system. In *Proceedings of 16th Nordic Conference of Computational Linguistics (NODALIDA).* Tartu, Estonia, 24-26 may, pp. 161-164.

Starlander, Marianne, Bouillon, Pierrette, Chatzichrisafis, Nikos, Santaholma, Marianne, Rayner, Manny, Hockey, Beth Ann, Isahara, Hitoshi, Kanzaki, Kyoko and Nakao, Yukie. 2005. Practising Controlled Language through a Help System integrated into the Medical Speech Translation System (MedSLT). In *Proceedings of MT Summit X.* Phuket, Thailand, 12-16 Septembe, pp. 188-194.

Starlander, Marianne, Bouillon, Pierrette, Flores, Glenn, Rayner, Manny and Tsourakis, Nikos. 2008. Comparing two different bidirectional versions of the limited domain medical spoken language translator MedSLT. In *Proceedings of 12th annual conference of the European Association for Machine Translation.* Hamburg, 22-23 september, pp. 176-181.

Wang, Ye-Yi, Acero, Alex and Chelba, Ciprian. 2003. Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy. In *Proceedings of Workshop on Automatic Speech Recognition and Understanding.* St Thomas, US Virgin Islands.