Field Testing the Tongues Speech-to-Speech Machine Translation System

**Robert E. Frederking*, Alan W Black*, Ralf D. Brown*,
John Moody†, Eric Steinbrecher†**

*Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
{ref+, awb+, ralf+}@cs.cmu.edu

†Lockheed Martin Systems Integration
Owego, New York, USA
{john.moody, eric.steinbrecher}@lmco.com

Abstract

The Tongues portable, rapid-development, speech-to-speech machine translation system was developed specifically to allow a realistic field-test of a deployable prototype. In this paper we will describe the system, its field-testing using regular US Army officers and naive Croatians, and the evaluation of these tests. The evaluation includes analysis of answers to a questionnaire, analysis of system transcript logs, and the authors' qualitative observations. The overall result of the test was that while the system did successfully aid translation, it requires further development before it would be ready for regular field use.

1. The Tongues System

The Tongues system was funded by the US Army to support the mission of the US Army chaplains, who are increasingly called upon to deal with local populations, usually without the benefit of human translators. It is thus intended to be used by a trained US Army chaplain with a completely naive and untrained non-English speaker.

The architecture and user interface of the Tongues system were based in large measure on the Diplomat system (Frederking et al., 2000). The speech recognition system used was the open-source Sphinx II (Huang et al., 1992); the translation system was a EBMT/MEMT (Example-Based MT/Multi-Engine MT) system (Brown, 1996; Frederking and Nirenburg, 1994; Brown and Frederking, 1995) very similar to that in Diplomat; and the synthesis system was the open-source Festival (Black et al., 1998).

While the initial system was specifically to demonstrate translation in both directions between English and Croatian, the design was also required to allow rapid development for new languages. To ensure rapid development, the entire project was only allowed to take one calendar year, including contractual arrangements, hiring language experts, etc. The total development effort was similarly restricted: six senior research personnel (the authors of this paper) provided an estimated total of about two (2) full-time person-years of effort. In addition to the senior staff, there were also part-time Croatian informants, chaplains, and some student programmers. We should note that some of the translation data used to train the system was collected for the Diplomat project (Frederking et al., 2000).

In addition to rapid development, the system was not permitted to be restricted to a narrowly-limited domain, but had to be wide-coverage. (Both of these properties were important for the chaplains' envisioned activities.) Since we were to build a broad-coverage system in a short period of time on a small budget, data-driven approaches were the only reasonable choice. In order to provide in-domain conversational data, we arranged at the start of the project to record a number of chaplains in role-playing conversations of the type they expected the device to encounter. Fortunately, the chaplains were familiar with role-playing exercises, and all had relevant field experiences to re-enact. Both sides of the conversations were spoken in English. These were digitally recorded with head-mounted microphones at 16KHz in stereo (one speaker on each channel), as this was closest to the intended audio channel characteristics of the eventual system. In all, we recorded 46 conversations, ranging from a few minutes to 20 minutes length. This provided a total of 4.25 hours of actual speech.

The recorded conversations were hand-transcribed at the word level, and translated into Croatian by native Croatian speakers. The English recordings were used for training the English speech recognition models. The transcripts and their translations were added to the EBMT system's example base of parallel sentences. A subset of the Croatian translations were read by native Croatian speakers to create data for the Croatian speech recognizer, as described elsewhere (Black et al., 2002). This simple approach appears to be surprisingly adequate.

Simply stringing together a recognizer, translator, and synthesizer does not make a very useful speech-to-speech translation system. A good interface is necessary to make the parts work together in such a way that a user can actually derive benefit from it. Using our experience from the earlier Diplomat system, we designed the Tongues interface to be asymmetric, with the Croatian side being as simple as possible, and any necessary complexity handled on the English side, since the chaplain would be trained and practiced in using the system. Even the English side was not terribly complex (see Figure 1).

We included a back-translation capability, to allow a user with no knowledge of the target language to better assess the quality of the translation. (We could not use the approach of generating paraphrases from meaning representations, since the system does not use any meaning represen-

tations.) We also included several user-requested features, such as built-in pre-recorded instructions and explanations for the Croatian (since the Croatian speaker is completely naive regarding the device and the chaplain's intentions), emergency key phrases (such as "Don't move!"), and enhancements such as being able to modify the translation lexicon in the field, so that the system could be tuned to more specific tasks.

The final system ran on a Windows-based Toshiba Libretto, running at 200MHz with 192MB of memory. At the time of the project (2000) this was the best combination of speed and size that was readily available. The system was equipped with a custom touchscreen, so that the Croatian-speaker would not need to type or use a mouse at all. Aware that the system might be used in situations where the non-English participant would be unfamiliar with computer technology, we included a microphone/speaker handset that looks like a conventional telephone handset. This has the advantage of provided a close-talking microphone, thus making speech recognition easier, while coming in a form factor that will be familiar to most people. We have provided a more detailed description of the development of the Tongues system elsewhere (Black et al., 2002).

Our design provides abundant opportunities for user error correction, in an effort to enable cooperative users to communicate well enough to accomplish significant tasks that they could not accomplish without the system (or a bilingual human interpreter), despite the error-prone nature of current speech recognition, broad-coverage rapid-development machine translation, and speech synthesis. Determining whether we have met such a goal requires task-based evaluation; while error rates of components are useful information, the real system-level issue is whether communication is achieved, and at what level of effort.

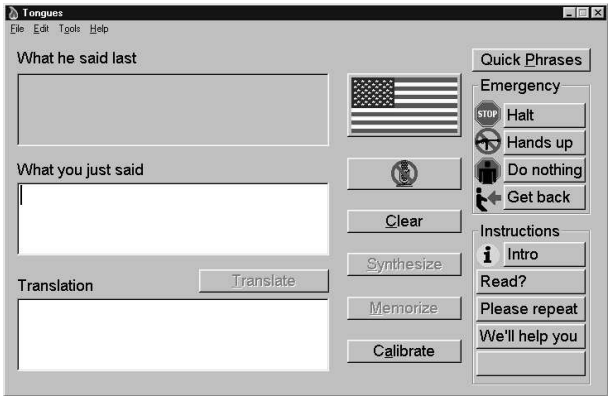

Figure 1: Tongues User Interface.

2. The Field Test

The US Army ACT-II program under which Tongues was funded is designed to result in field tests of deployable prototypes. Accordingly, in April 2001, representatives of the development team traveled to Zagreb, Croatia, with representatives of the US Army chaplains. We had arranged in advance to have native-Croatian speakers available as conversation partners. This was done by contacting someone at the University of Zagreb, and hiring them as a local organizer. They were instructed to recruit a large number of potential test subjects varying in gender and age, with as little English knowledge as possible.

Since the principal domain of the translation system was interaction with refugees, we prepared a number of refugee scenarios for the Croatian subjects and American chaplains to act out using the translation device. The scenarios were in the intended domain, involving refugees, medical supplies and getting general directions. The refugee side of each scenario was translated into Croatian. We also prepared a questionnaire for each participant, produced translated Croatian questionnaires, and after the test had the Croatian responses translated in to English.

We then travelled to Croatia. Over a three-day period, at the University of Zagreb, naive Croatians were brought into the room knowing only that they were supposed to enact the scenario that they had just been given with a US Army officer, who would be using a translation device (see Figure 2). The Croatian only knew the refugee side of the scenario, while the US officer only knew the Army side of the scenario. The actual Croatian subjects consisted of 21 speakers, male and female, of various ages ranging from young teenagers through adults. Each dialog was logged by the system to allow further analysis.

Figure 2: Tongues in use in Croatia.

3. Analysis of Results

As mentioned above, we generated questionnaire responses and system transcript logs in the course of our tests. We also directly observed the conversations and took notes. Our subjective impression of the results was that the conversations went reasonably well about one half of the time. In addition to cases where the parties failed to complete their tasks, the system was often frustrating to use, due to the large amount of user error correction often required, and the corresponding slowness of the dialogue.

Difficulties described by the participants range over all the components; but our subjective impression was that the speech components performed quite acceptably; the translation component was the weakest link. (This was especially surprising to us given that the speech components were not trained on a large amount of data.) In particular, as

our rapid-development translation system contains no internal representation of the meaning of the utterance, the only method for feedback of the translation results to the (monolingual) user is (independent) back-translation. This risks doubling the error rate, and a bilingual team member in fact observed that often an English-to-Croatian translation that was basically correct would be rejected by an English-speaker because the back-translation was seriously garbled.

Before presenting more detailed qualitative analysis, we will present the results of the questionnaire responses and an analysis of the system transcript logs.

3.1. Analysis of Questionnaires

Of 19 completed post-experiment questionaires, 3 stated that they felt the communication failed, 5 stated that it went well, with the remaining 11 thinking it was acceptable

In all, 21 dialogs took place, between different Croatian speakers and one of 5 chaplains. After the test, the Croatian participants were given a questionnaire to fill out. Their overall impression was as follows:

Overall	
Good	5
OK	11
Bad	3

That is, only 16% admitted failure. (We feel that this was clearly overly generous on the participants' parts.)

On asking the participants to identify the most difficult problems, they replied as follows:

User difficulties	
grammar/case	5
loudspeakers	4
translation	3
recognition	2
synthesis	2
speed	1

There was an obvious problem during the test with the small, built-in loudspeakers not being loud enough; this is clearly an easily solved problem (clearly easy for software developers, at least). As noted above, speech recognition and synthesis were both observed to be relatively acceptable.

We also asked the participants what they found easy:

What works?	
short sentences	10
nothing	4

Most participants quickly discovered that the system did not translate long, rambling sentences well. The second response belies at least one of the user's claims that the system was "okay".

3.2. Analysis of Transcript Logs

As mentioned above, each dialog was logged by the system to allow further analysis. Although the users did not always restart the system with each scenario, we can still easily identify 28 different dialogs from the logs. Of these, 10 are clearly just tests or single unconnected examples; this leaves a total of 18 dialogs that contain 3 or more Croatian turns. As we will see from the figures below, the English participant plays a much large role in the conversations than the Croatian participant, but useful information still passes between them even with a small number of Croatian turns.

Each dialog took between 14 and 68 minutes, with an average of 24 minutes. The 68 minute dialog took place on the first day and was the first real dialog to be carried out. The next longest one was 34 minutes, which confirms that this first conversation was somewhat of an anomaly.

The dialogs almost always started with a set of prerecorded instructions to the Croatian user. These explained who the US personnel were, their purpose, and some instructions on using the translation device. Additional prerecorded utterances were only rarely used elsewhere in the dialogs ("please repeat"). Although these utterances are technically English to Croatian turns, we have counted these as a separate type below ("Pre-rec") and not included these in our analysis of turn times, utterance sizes, etc.

Dialog	Duration	Pre-rec	English	Croatian
d6	68	4	14	8
d7	24	7	5	3
d9	19	12	10	9
d10	19	0	12	6
d11	28	6	15	4
d12	21	0	14	5
d13	19	6	8	3
d14	18	6	9	6
d15	17	6	9	4
d17	32	0	20	7
d18	14	6	11	4
d19	34	6	27	8
d21	23	0	17	3
d22	22	10	13	3
d23	24	6	12	4
d25	22	7	11	5
d26	16	7	14	8
d28	17	6	16	6

It is clear that the US Chaplain side controlled the conversation, as expected. The ratio of English to Croatian turns (excluding pre-recorded utterances) ranges from 4.3 to 1.1, with a mean of 2.67. Thus the English participant spoke more than twice as much as the Croatian. There were in fact only two cases where the Croatian side took two consecutive turns.

We also counted the number of words in each turn. This count includes all dialogs, though still excluding the prerecorded utterances. For the Croatian turns, words were counted in the English translation, not the original Croatian. This was done to normalize any difference in expressiveness in the two languages.

	words	turns	w per t
English	1019	218	4.67
Croatian	355	101	3.51

Turns thus tended to be short, direct sentences; as noted in the responses to our user questionnaires, longer utterances were more likely to contain errors, due both to the

fact that there were more words, and also because longer utterances are more likely to have more complex structure. The pre-recorded spoken instructions also explicitly inform the users that they should use short sentences.

3.3. Qualitative Analysis

It is important to note, and immediately obvious when participating in such a conversation, that communication through a translation device is not fast. Each person must speak, check the recognized form and possibly correct it, translate the utterance (possibly checking with back-translation), and then synthesize the result. Such devices thus do not enable truly spontaneous communication, as they deliberately allow the participants to review the translations and decide when they are adequate. It is possible for the component technologies (recognition, translation and synthesis) to become more streamlined, but it would be very difficult to achieve truly spontaneous, simultaneous translation.

In looking over the conversations, it is clear that the translations are often far from ideal, though usually understandable. For example in answer to the question "where are they?" the device produces "twenty minutes of village." The quality in the English to Croatian translations is similar, in our judgment.

Other specific observations we noted were that the users could not easily identify where the problems lay with the system. For example, if speech recognition produced and displayed a correct transcript, and then translation produced an unacceptable result, they would usually *respeak* the same utterance using the same words! Similarly, mistakes in the synthesizer were often erroneously attributed to the translator (and vice versa, despite the output text being visible in the user interface. Thus even if we provided separate user methods to add words to the recognizer, language model, and translation engine, it is clear that the user would not be able to identify which part (or parts) need to be updated. As there is strong user demand for such systems to provide methods of adaptation in the field, it is clear that the interface presented to the user to offer that adaptation needs more work.

A second observation was that the participants continued to use speech and did not resort to the alternative typing interface (although they were clearly aware of it), and only resorted to typing as a last resort. This may have been due to the fact the participants were asked to use the speech-to-speech translation device rather than being given the more abstract goal of achieving successful communication by the best means. The very small keyboard on the (required) small device may also have been a significant factor, in addition to the well-known preference many naive users have for speaking over typing.

We also note an interesting phenomenon with a limitation in the system in dealing with unknown words. Often such out of vocabulary words have direct cognates in the other language, and hence are directly understandable. We could see that some conjugations of the Croatian word for "kilometer" were not recognized by the Croatian speech recognition system, and hence failed to translate. When a word fails to translate, the system presents the word in its original language, but capitalized, in the translation output. For example, the recognized phrase "pet gje ometa" is translated as "five GJE OMETA"; given the context, it was clear to the English speaker that the Croatian speaker had said "five kilometers" (in Croatian). A similar example happened with the word "helicopter".

This point is important. We have two people cooperating and actively trying to communicate. Thus where cognates exist, the listener will understand and accommodate mis-recognitions.

We also noted that, as a consequence of the slowness of communication, the participants took more time to think about about they were going to say. Thus their utterances were on the whole more complete sentences than the fragments that one typically encounters in normal conversational speech. This factor almost certainly compensated for the fact that our Croatian speech recognizer was trained on read speech. Conversely, it probably slightly hindered English recognition, as that was trained on more spontaneous conversations.

The conversations took place in a quiet classroom situation, with little external noise. This helped both the speech recognition and the user understanding of the speech output. However, it is also worth noting that synthetic speech is much easier to understand when the written form of what is being spoken also appears on the screen in front of the (literate) listener.

Finally, we also noted that some English questions were answered with simple yes/no answers without using the device to translate them. The effort of translating simple one-word utterances (such as "da"), which can often easily be understood without knowing the language, was unnecessary.

4. Conclusions

We feel that this field test of the Tongues system was unusually rigorous and realistic, in that we tested the system using regular US Army officers speaking with naive Croatians who did not live in an English-speaking country. This was important, since if our system performed well in the field test, it would conceivably have gone into actual use.

The official report by the US Army participants was that the system is worth further development, since it is approaching the quality necessary for real use, but still requires further development before actual field use. We believe that this is actually quite a good result, given the current state of speech and MT technology, and especially the time, cost, and broad-coverage constraints of this project.

5. References

A. Black, P. Taylor, and R. Caley. 1998. The Festival Speech Synthesis System. http://festvox.org/festival.

A. Black, R. Brown, R. Frederking, R. Singh, J. Moody, and E. Steinbrecher. 2002. TONGUES: Rapid Development of a Speech-to-Speech Translation System. In *Proceedings of HLT-2002*, San Diego, CA, USA.

R. Brown and R. Frederking. 1995. Applying Statistical English Language Modeling to Symbolic Machine Translation. In *Proceedings of the Sixth International*

Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95), pages 221–239.

R. Brown. 1996. Example-based Machine Translation in the Pangloss System. In *Proceedings of COLING-96*, pages 169–174, Copenhagen, Denmark.

R. Frederking and S. Nirenburg. 1994. Three Heads are Better than One. In *Proceedings of the fourth Conference on Applied Natural Language Processing (ANLP-94)*, Stuttgart, Germany.

R. Frederking, A. Rudnicky, C. Hogan, and K. Lenzo. 2000. Interactive Speech Translation in the Diplomat Project. *Machine Translation Journal*, 15(1-2):27–42. Special Issue on Spoken Language Translation.

X. Huang, F. Alleva, H.-W. Hon, K.-F. Hwang, M.-Y. Lee, and R. Rosenfeld. 1992. The SPHINX-II Speech Recognition System: an overview. *Computer Speech and Language*, 7(2):137–148.