Usability Issues in an Interactive Speech-to-Speech Translation System for Healthcare

Mark Seligman
Spoken Translation, Inc.
Berkeley, CA, USA 94705
mark.seligman@spokentranslation.com

Mike Dillinger
Spoken Translation, Inc.
Berkeley, CA, USA 94705
mike.dillinger@spokentranslation.com

Abstract

We describe a highly interactive system for bidirectional, broad-coverage spoken language communication in the healthcare area. The paper briefly reviews the system's interactive foundations, and then goes on to discuss in greater depth issues of practical usability. We present our Translation Shortcuts facility, which minimizes the need for interactive verification of sentences after they have been vetted once, considerably speeds throughput while maintaining accuracy, and allows use by minimally literate patients for whom any mode of text entry might be difficult. We also discuss facilities for multimodal input, in which handwriting, touch screen, and keyboard interfaces are offered as alternatives to speech input when appropriate. In order to deal with issues related to sheer physical awkwardness, we briefly mention facilities for hands-free or eyes-free operation of the system. Finally, we point toward several directions for future improvement of the system.

1 Introduction

Increasing globalization and immigration have led to growing demands on US institutions for healthcare and government services in languages other than English. These institutions are already overwhelmed: the State of Minnesota, for example, had no Somali-speaking physicians for some 12,000 Somali refugees and only six Hmong-speaking physicians to serve 50,000 Hmong residents (Minnesota Interpreter Standards Advisory Committee, 1998). San Francisco General Hospital, to cite another example, receives approximately 3,500 requests for interpretation per month, or 42,000 per year for 35 different languages. Moreover, requests for medical interpretation services are distributed among all the wards and clinics, adding a logistical challenge to the problem of a high and growing demand for interpretation services (Paras, et al., 2002). Similar situations are found throughout the United States.

It is natural to hope that automatic real-time translation in general, and spoken language translation (SLT) in particular, can help to meet this communicative need. From the viewpoint of research and development, the high demand in healthcare makes this area especially attractive for fielding early SLT systems and seeking early adopters.

With this goal in view, several speech translation systems have aimed at the healthcare area. (See www.sehda.com, DARPA's CAST program, www.phraselator.com, etc.) However, these efforts have encountered several issues or limitations.

First, they have been confined to narrow domains. In general, SLT applications have been able to achieve acceptable accuracy only by staying within restricted topics, in which fixed phrases could be used (e.g., www.phraselator.com), or in which grammars for automatic speech recognition (ASR) and machine translation (MT) could be optimized. For example, MedSLT (Bouillon et al, 2005) is limited to some 600 specific words per sub-domain. IBM's MASTOR system, with 30,000 words in each translation direction, has much broader coverage, but remains comparable in lexicon size to commercial MT systems of the early 1980s.

Proceedings of the Workshop on Medical Speech Translation at HLT-NAACL 2006, pages 1–8,
New York, New York, June 2006.

Granted, restriction to narrow domains may often be appropriate, given the large effort involved in compiling extensive lexical resources and the time required for deployment. A tightly focused approach permits relatively quick development of new systems and provides a degree of flexibility to experiment with different architectures and different languages.

Our emphasis, however, is on breaking out of narrow domains. We seek to maximize versatility by providing exceptional capacity to move from topic to topic while maintaining adequate accuracy.

To provide a firm foundation for such versatility, we "give our systems a liberal arts education" by incorporating very broad-coverage ASR and MT technology. Our MT lexicons, for example, contain roughly 300,000 words in each direction.

But of course, as coverage increases, perplexity and the ASR and MT errors due to it increase in proportion, especially in the absence of tight integration between these components. To compensate, we provide a set of facilities that enable users from both sides of the language barrier to interactively monitor and correct these errors. Putting users in the speech translation loop in this way does in fact permit conversations to range widely (Seligman, 2000). We believe that this highly interactive approach will prove applicable to the healthcare area.

We have described these interactive techniques in (Dillinger and Seligman, 2004; Zong and Seligman, forthcoming). We will review them only briefly here, in Section 2.

A second limitation of current speech translation systems for healthcare is that bilingual (bidirectional) communication has been difficult to enable. While speech-to-speech translation has sometimes proven practical from the English side, translation from the non-English side has been more difficult to achieve. Partly, this limitation arises from human factors issues: while naïve observers might expect spoken input to be effortless for anyone who can talk, the reality is that users must learn to use most speech interfaces, and that this learning process can be difficult for users who are less literate or less computer literate. Further, many healthcare venues make speech input difficult: they may be noisy, microphones may be awkward to situate or to pass from speaker to speaker, and so on.

Our group's approach to training- or venue-related difficulties for speech input is to provide an array of alternative input modes. In addition to providing input through dictated speech, users of our system can freely alternate among three other input modes, using handwriting, a touch screen, and standard bilingual keyboards.

In this paper, we will focus on practical usability issues in the design of user interfaces for highly interactive approaches to SLT in healthcare applications. With respect to interactivity per se, we will discuss the following specific issues:

- In a highly interactive speech translation system, monitoring and correction of ASR and MT are vital for accuracy and confidence, but can be time consuming – in a field where time is always at a premium.
- Interactivity demands a minimum degree of computer and print literacy, which some patients may lack.

To address these issues, we have developed a facility called *Translation Shortcuts*™, to be explained throughout Section 3.

Section 4 will describe our approach to multimodal input. As background, however, Section 2 will quickly review our approach to highly interactive – and thus uniquely broad-coverage – spoken language translation. Before concluding, we will in Section 5 point out planned future developments.

2 Highly Interactive, Broad-coverage SLT

We now briefly summarize our group's approach to highly interactive, broad-coverage SLT.

The twin goals of accuracy and broad-coverage have generally been in opposition: speech translation systems have gained tolerable accuracy only by sharply restricting both the range of topics that can be discussed and the sets of vocabulary and structures that can be used to discuss them. The essential problem is that both speech recognition and translation technologies are still quite error-prone. While the error rates may be tolerable when each technology is used separately, the errors combine and even compound when they are used together. The resulting translation output is generally below the threshold of usability – unless restriction to a very narrow domain supplies sufficient constraints to significantly lower the error rates of both components.

As explained, our group's approach has been to concentrate on interactive monitoring and correction of both technologies.

First, users can monitor and correct the speaker-dependent speech recognition system to ensure that the text that will be passed to the machine translation component is completely correct. Voice commands (e.g. **Scratch That** or **Correct <incorrect text>**) can be used to repair speech recognition errors. Thus, users of our SLT enrich the interface between ASR and MT.

Next, during the MT stage, users can monitor, and if necessary correct, one especially important aspect of the translation – lexical disambiguation.

Our system's approach to lexical disambiguation is twofold: first, we supply a *Back-Translation*, or re-translation of the translation. Using this paraphrase of the initial input, even a monolingual user can make an initial judgment concerning the quality of the preliminary machine translation output. (Other systems, e.g. IBM's MASTOR, have also employed re-translation. Our implementations, however, exploit proprietary technologies to ensure that the lexical senses used during back translation accurately reflect those used in forward translation.)

In addition, if uncertainty remains about the correctness of a given word sense, we supply a proprietary set of Meaning Cues™ – synonyms, definitions, etc. – which have been drawn from various resources, collated in a database (called SELECT™), and aligned with the respective lexica of the relevant MT systems. With these cues as guides, the user can monitor the current, proposed meaning and select (when necessary) a different, preferred meaning from among those available. Automatic updates of translation and back translation then follow. Future versions of the system will allow personal word-sense preferences thus specified in the current session to be stored and reused in future sessions, thus enabling a gradual tuning of word-sense preferences to individual needs. Facilities will also be provided for sharing such preferences across a working group.

Given such interactive correction of both ASR and MT, wide-ranging, and even jocular, exchanges become possible (Seligman, 2000).

As we have said, such interactivity within a speech translation system can enable increased accuracy and confidence, even for wide-ranging conversations.

Accuracy of translation is, in many healthcare settings, critical to patient safety. When a doctor is taking a patient's history or instructing the patient in a course of treatment, even small errors can have clinically relevant effects. Even so, at present, healthcare workers often examine patients and instruct them in a course of treatment through gestures and sheer good will, with no translation at all, or use untrained human interpreters (friends, family, volunteers, or staff) in an error-prone attempt to solve the immediate problem (Flores, et al., 2003). As a result, low-English proficiency patients are often less healthy and receive less effective treatment than English speakers (Paras, et al., 2002). We hope to demonstrate that highly interactive real-time translation systems in general, and speech translation systems in particular, can help to bridge the language gap in healthcare when human interpreters are not available.

Accuracy in an automatic real-time translation system is necessary, but not sufficient. If healthcare workers have no means to independently assess the reliability of the translations obtained, practical use of the system will remain limited. Highly interactive speech translation systems can foster the confidence on both sides of the conversation, which is necessary to bring such systems into wide use. In fact, in this respect at least, they may sometimes prove superior to human interpreters, who normally do not provide clients with the means for judging translation accuracy.

The value of enabling breadth of coverage, as well as accuracy and confidence, should also be clear: for many purposes, the system must be able to translate a wide range of topics *outside of* the immediate healthcare domain – for example, when a patient tries to describe what was going on when an accident occurred. The ability to ask about interests, family matters, and other life concerns is vital for establishing rapport, managing expectations and emotions, etc.

3 Translation Shortcuts

Having summarized our approach to highly interactive speech translation, we now turn to examination of practical interface issues for this class of SLT system. This section concentrates on Translation Shortcuts™.

Shortcuts are designed to provide two main advantages:

First, re-verification of a given utterance is unnecessary. That is, once the translation of an utterance has been verified interactively, it can be saved for later reuse, simply by activating a **Save as Shortcut** button on the translation verification screen. The button gives access to a dialogue in which a convenient *Shortcut Category* for the Shortcut can be selected or created. At reuse time, no further verification will be required. (In addition to such dynamically created *Personal* Shortcuts, any number of prepackaged *Shared* Shortcuts can be included in the system.)

Second, access to stored Shortcuts is very quick, with little or no need for text entry. Several facilities contribute to meeting this design criterion.

- A *Shortcut Search* facility can retrieve a set of relevant Shortcuts given only keywords or the first few characters or words of a string. The desired Shortcut can then be executed with a single gesture (mouse click or stylus tap) or voice command.

NOTE: If no Shortcut is found, the system automatically allows users access to the full power of broad-coverage, interactive speech translation. Thus, a seamless transition is provided between the Shortcuts facility and full, broad-coverage translation.

- A *Translation Shortcuts Browser* is provided, so that users can find needed Shortcuts by traversing a tree of Shortcut categories. Using this interface, users can execute Shortcuts even if their ability to input text is quite limited, e.g. by tapping or clicking alone.

Figure 1 shows the Shortcut Search and Shortcuts Browser facilities in use. Points to notice:

- On the left, the Translation Shortcuts Panel has slid into view and been pinned open. It contains the Translation Shortcuts Browser, split into two main areas, Shortcuts Categories (above) and Shortcuts List (below).
- The Categories section of the Panel shows current selection of the **Conversation** category, containing everyday expressions, and its **Staff** subcategory, containing expressions most likely to be used by healthcare staff members. There is also a **Patients** subcategory, used for patient responses. Categories for **Administrative topics** and **Patient's Current Condition** are also visible; and new ones can be freely created.
- Below the Categories section is the Shortcuts List section, containing a scrollable list of alphabetized Shortcuts. (Various other sorting criteria will be available in the future, e.g. sorting by frequency of use, recency, etc.)
- Double clicking on any visible Shortcut in the List will execute it. Clicking once will select and highlight a Shortcut. Typing **Enter** will execute the currently highlighted Shortcut (here "Good morning"), if any.
- It is possible to automatically relate options for a patient's response to the previous staff member's utterance, e.g. by automatically going to the sibling **Patient** subcategory if the prompt was given from the **Staff** subcategory.

Because the Shortcuts Browser can be used without text entry, simply by pointing and clicking, it enables responses by minimally literate users. In the future, we plan to enable use even by completely illiterate users, through two devices: we will enable automatic pronunciation of Shortcuts and categories in the Shortcuts Browser via text-to-speech, so that these elements can in effect be read aloud to illiterate users; and we will augment Shared Shortcuts with pictorial symbols, as clues to their meaning.

A final point concerning the Shortcuts Browser: it can be operated entirely by voice commands, although this mode is more likely to be useful to staff members than to patients.

We turn our attention now to the Input Window, which does double duty for Shortcut Search and arbitrary text entry for full translation. We will consider the search facility first, as shown in Figure 2.

- Shortcuts Search begins automatically as soon as text is entered by any means – voice, handwriting, touch screen, or standard keyboard – into the Input Window.
- The **Shortcuts Drop-down Menu** appears just below the Input Window, as soon as there are results to be shown. The user has entered "Good" and a space, so the search program has received its first input word. The drop-down menu shows the results of a keyword-based search.
- Here, the results are sorted alphabetically. Various other sorting possibilities may be useful: by frequency of use, proportion of matched words, etc.

- The highest priority Shortcut according to the specified sorting procedure can be highlighted for instant execution.
- Other shortcuts will be highlighted differently, and both kinds of highlighting are synchronized with that of the Shortcuts list in the Shortcuts Panel.
- Arrow keys or voice commands can be used to navigate the drop-down list.
- If the user goes on to enter the exact text of any Shortcut, e.g. "Good morning," a message will show that this is in fact a Shortcut, so that verification will not be necessary. However, final text not matching a Shortcut, e.g. "Good job," will be passed to the routines for full translation with verification.

4 Multimodal input

As mentioned, an unavoidable issue for speech translation systems in healthcare settings is that speech input is not appropriate for every situation.

Current speech-recognition systems are unfamiliar for many users. Our system attempts to overcome this training issue to some extent by incorporating standard commercial-grade dictation systems for broad-coverage and ergonomic speech recognition. These products already have established user bases in the healthcare community. Even so, some training may be required: optional generic Guest profiles are supplied by our system for male and female voices in both languages; but optional voice enrollment, requiring five minutes or so, is helpful to achieve best results. Such training time is practical for healthcare staff, but will be realistic for patients only when they are repeat visitors, hospital-stay patients, etc.

As mentioned, other practical usability issues for the use of speech input in healthcare settings include problems of ambient noise (e.g. in emergency rooms or ambulances) and problems of microphone and computer arrangement (e.g. to accommodate not only desktops but counters or service windows which may form a barrier between staff and patient).

To deal with these and other usability issues, we have found it necessary to provide a range of input modes: in addition to dictated speech, we enable handwritten input, the use of touch screen keyboards for text input, and the use of standard keyboards. All of these input modes must be completely bilingual, and language switching must be arranged automatically when there is a change of active participant. Further, it must be possible to change input modes seamlessly within a given utterance: for example, users must be able to dictate the input if they wish, but then be able to make corrections using handwriting or one of the remaining two modes. Figure 3 shows such seamless bilingual operation: the user has dictated the sentence "Tengo náuseas" in Spanish, but there was a speech-recognition error, which is being corrected by handwriting.

Of course, even this flexible range of input options does not solve all problems. As mentioned, illiterate patients pose special problems. Again, naïve users tend to suppose that speech is the ideal input mode for illiterates. Unfortunately, however, the careful and relatively concise style of speech that is required for automatic recognition is often difficult to elicit, so that recognition accuracy remains low; and the ability to read and correct the results is obviously absent. Just as obviously, the remaining three text input modes will be equally ineffectual for illiterates.

As explained, our current approach to low literacy is to supply Translation Shortcuts for the minimally literate, and – in the future – to augment Shortcuts with text-to-speech and iconic pictures.

Staff members will usually be at least minimally literate, but they present their own usability issues.

Their typing skills may be low or absent. Handling the computer and/or microphone may be awkward in many situations, e.g. when examining a patient or taking notes. (Speech translation systems are expected to function in a wide range of physical settings: in admissions or financial aid offices, at massage tables for physical therapy with patients lying face down, in personal living rooms for home therapy or interviews, and in many other locations.)

To help deal with the awkwardness issues, our system provides voice commands, which enable hands-free operation. Both full interactive translation and the Translation Shortcut facility (using either the Browser or Search elements) can be run hands-free. To a limited degree, the system can be used *eyes*-free as well: text-to-speech can be used to pronounce the back-translation so that preliminary judgments of translation quality can be made without looking at the computer screen.

5 Future developments

We have already mentioned plans to augment the Translation Shortcuts facility with text-to-speech and iconic pictures, thus moving closer to a system suitable for communication with completely illiterate or incapacitated patients.

Additional future directions follow.

- **Server-based architectures:** We plan to move toward completely or partially server-based arrangements, in which only a very thin client software application – for example, a web interface – will run on the client device. Such architectures will permit delivery of our system on smart phones in the Blackberry or Treo class. Delivery on handhelds will considerably diminish the issues of physical awkwardness discussed above, and anytime/anywhere/any-device access to the system will considerably enlarge its range of uses.
- **Pooling Translation Shortcuts:** As explained above, the current system now supports both Personal (do-it-yourself) and Shared (prepackaged) Translation Shortcuts. As yet, however, there are no facilities to facilitate pooling of Personal Shortcuts among users, e.g. those in a working group. In the future, we will add facilities for exporting and importing shortcuts.
- **Translation memory:** Translation Shortcuts can be seen as a variant of Translation Memory, a facility that remembers past successful translations so as to circumvent error-prone reprocessing. However, at present, we save Shortcuts only when explicitly ordered. If all other successful translations were saved, there would soon be far too many to navigate effectively in the Translation Shortcuts Browser. In the future, however, we could in fact record these translations in the background, so that there would be no need to re-verify new input that matched against them. Messages would advise the user that verification was being bypassed in case of a match.
- **Additional languages:** The full SLT system described here is presently operational only for bidirectional translation between English and Spanish. We expect to expand the system to Mandarin Chinese next. Limited working prototypes now exist for Japanese and German, though we expect these languages to be most useful in application fields other than healthcare.
- **Testing**: Systematic usability testing of the full system is under way. We look forward to presenting the results at a future workshop.

6 Conclusion

We have described a highly interactive system for bidirectional, broad-coverage spoken language communication in the healthcare area. The paper has briefly reviewed the system's interactive foundations, and then gone on to discuss in greater depth issues of practical usability.

We have presented our Translation Shortcuts facility, which minimizes the need for interactive verification of sentences after they have been vetted once, considerably speeds throughput while maintaining accuracy, and allows use by minimally literate patients for whom any mode of text entry might be difficult.

We have also discussed facilities for multimodal input, in which handwriting, touch screen, and keyboard interfaces are offered as alternatives to speech input when appropriate. In order to deal with issues related to sheer physical awkwardness, we have briefly mentioned facilities for hands-free or eyes-free operation of the system.

Finally, we have pointed toward several directions for future improvement of the system.

References

Pierrette Bouillon, Manny Rayner, et al. 2005. A Generic Multi-Lingual Open Source Platform for Limited-Domain Medical Speech Translation. Presented at *EAMT 2005*, Budapest, Hungary.

Mike Dillinger and Mark Seligman. 2004. A highly interactive speech-to-speech translation system. *Proceedings of the VI Conference of the Association of Machine Translation in the Americas*. E. Stroudsburg, PA: American Association for Machine Translation.

Glenn Flores, M. Laws, S. Mays, et al. 2003. Errors in medical interpretation and their potential clinical consequences in pediatric encounters. *Pediatrics*, 111: 6-14.

Minnesota Interpreter Standards Advisory Committee. 1998. *Bridging the Language Gap: How to meet the need for interpreters in Minnesota*. Available at: http://www.cce.umn.edu/creditcourses/pti/downloads.html.

Melinda Paras, O. Leyva, T. Berthold, and R. Otake. 2002. *Videoconferencing Medical Interpretation: The results of clinical tri*als. Oakland, CA: Heath Access Foundation.

PHRASELATOR (2006). http://www.phraselator.com. As of April 3, 2006.

S-MINDS (2006). http://www.sehda.com/solutions.htm. As of April 3, 2006.

Mark Seligman. 2000. Nine Issues in Speech Translation. *Machine Translation*, 15, 149-185.

Chengqing Zong and Mark Seligman. Forthcoming. Toward Practical Spoken Language Translation. *Machine Translation*.

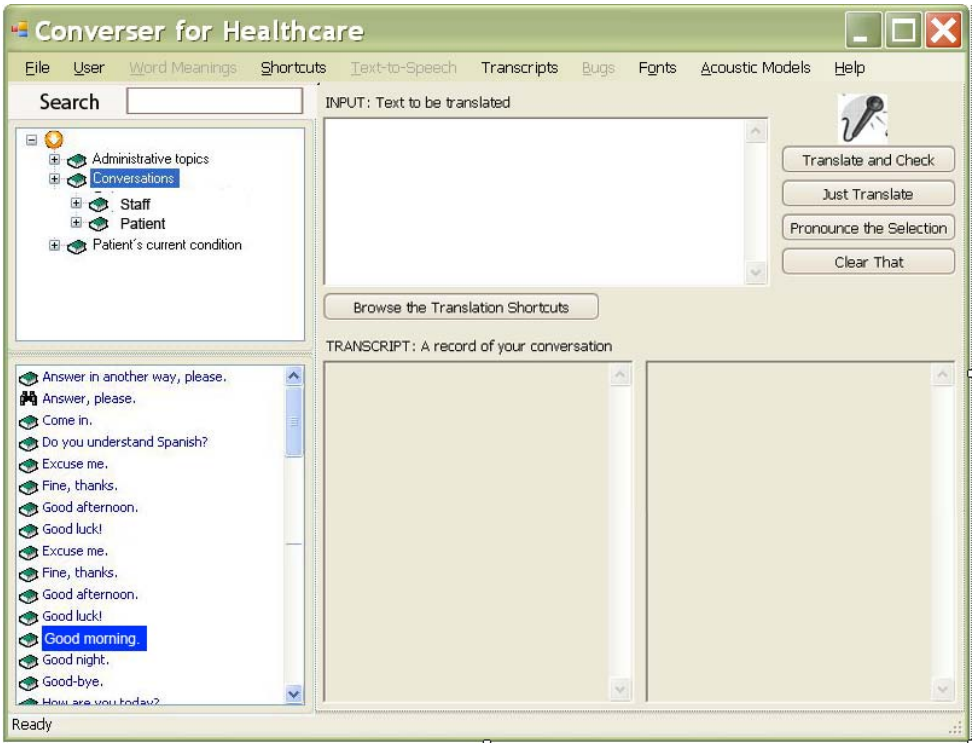

Figure 1: The Input Screen, showing the Translation Shortcuts Browser and Search facilities.

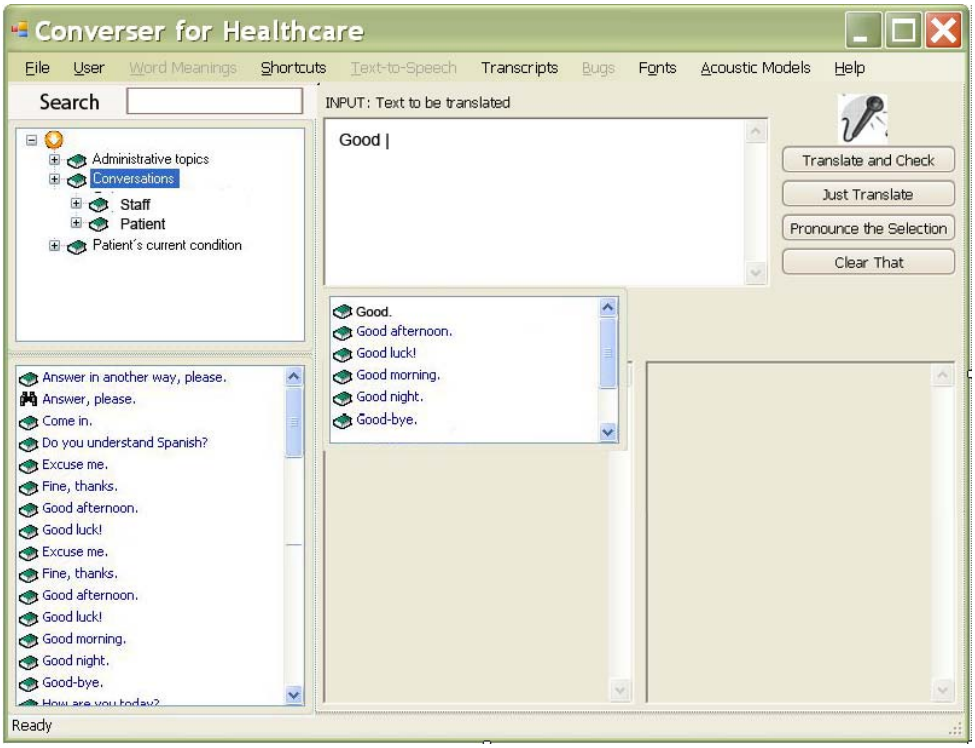

Figure 2: The Input Screen, showing automatic keyword search of the Translation Shortcuts.

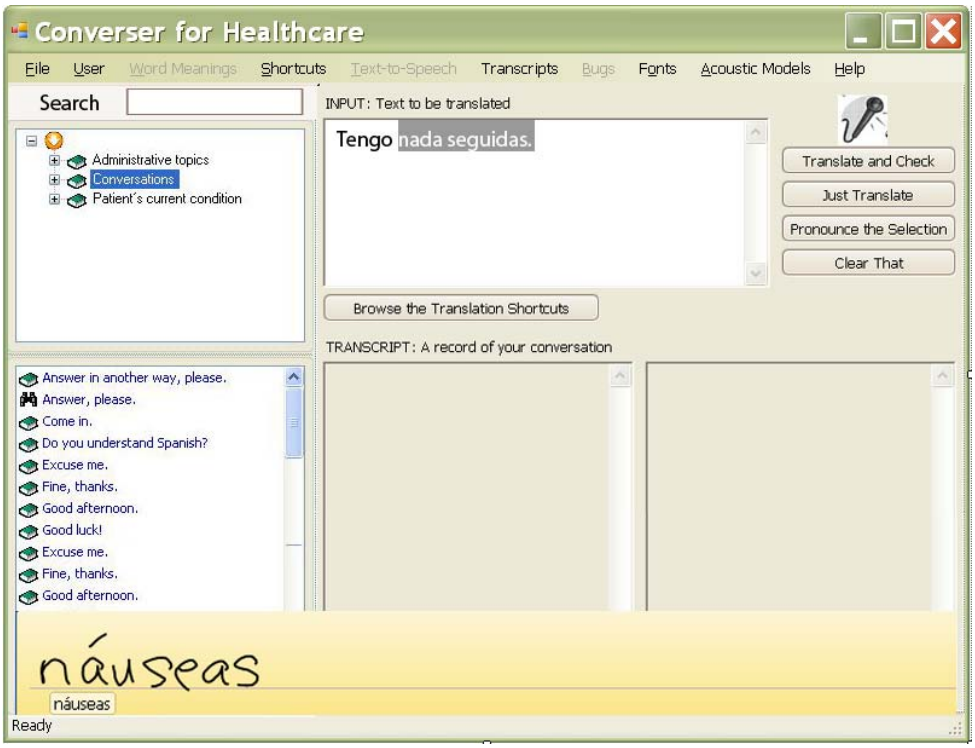

Figure 3: The Input Screen, showing correction of dictation with handwritten input.