Evaluation and Revision of a Speech Translation System for Healthcare

Mark Seligman and Mike Dillinger

Spoken Translation, Inc.
mark.seligman@spokentranslation.com
mike.dillinger@gmail.com

Abstract

Earlier papers have reported on Converser for Healthcare, a highly-interactive English↔Spanish speech translation system for communication between patients and caregivers, and upon an extensive pilot project testing the system at a San Francisco medical center, part of a very large healthcare organization. This historical paper provides for the first time details of the resulting evaluation and fully describes the associated system revisions to date.

1. Introduction

Spoken language translation systems are now in operation at Google and Microsoft/Skype, and multiple applications for spoken language translation (SLT) or automatic interpreting are also available – SpeechTrans, Jibbigo, iTranslate, and others. However, widespread use remains in the future for serious use cases like healthcare, business, emergency relief, and law enforcement, despite demonstrably high demand.

In spite of dramatic advances during the last decade, both speech recognition and translation technologies are still error-prone. While the error rates may be tolerable when the technologies are used separately, the errors combine and even compound when they are used together. The resulting translation output is often below the threshold of usability when accuracy is essential. As a result, present use is still largely restricted to use cases – social networking, travel – in which no representation concerning accuracy is demanded or given.

The speech translation system discussed here, Converser for Healthcare, applies interactive verification and correction techniques to this essential problem of overall reliability.

Earlier papers ([1], [2], [4], [5], [6], [7], [8], [9]) have reported on this highly-interactive system for English↔Spanish communication between patients and caregivers, and upon an extensive pilot project in 2011 testing Version 3.0 of the system at a San Francisco medical center, part of a very large healthcare organization ([9]). This paper provides for the first time details of the resulting evaluation and fully describes the associated system revisions to date, yielding the current Version 4.0. The paper is partly of historical interest, since the pilot took place four years ago – a long time in computer years. However, most of the issues raised by the evaluation remain current, and will be discussed below.

For orientation, Section 2 of this paper will review Converser's basic interactive facilities, as common to both Versions 3.0 and 4.0. Section 3 gives the results of the pilot project, as seen in the independent evaluation commissioned by the healthcare organization. Section 4 then details the revisions which were made for Version 4.0 in response to this feedback and other lessons learned. Section 5 offers an extended example of the revised system in use. We conclude in a final section.

2. The Converser System

We now briefly describe Converser's approach to interactive automatic interpretation, restricting description to core elements common to Version 3.0 (as used in the pilot project discussed in Section 3) and to the revised Version 4.0 (to be described in Sections 4 and 5 below). We'll concentrate on the system's verification/correction and customization features.

First, users can monitor and correct the speech recognition system to ensure that the text which will be passed to the machine translation component is completely correct. Speech, typing, or handwriting can be used to repair speech recognition errors.

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

The 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 ([4]), have also employed re-translation. Converser, however, exploits 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, the system supplies 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 when necessary select a different, preferred meaning from among those available. Automatic updates of translation and back-translation then follow.

The initial purpose of these techniques is to increase reliability during real-time speech translation sessions. Equally important, however, they can also enable even monolingual users to supply feedback for off-line machine learning to improve the system. Until now, only users with some knowledge of the output language have been able to supply such feedback, e.g. in Google Translate.

Converser adopts rather than creates its speech and translation components, adding value through the interactive interface elements to be explained. Nuance, Inc. supplies speech recognition; rule-based English↔Spanish machine translation is supplied by Word Magic of Costa Rica; and text-to-speech is again provided by Nuance.

The Converser system includes Translation Shortcuts™ – pre-packaged translations, providing a kind of translation memory. When they're used, re-verification of a given utterance is unnecessary, since Shortcuts are pre-translated by professionals (or, in future versions of the system, verified using the system's feedback and correction tools). Access to stored Shortcuts is very quick, with little or no need for text entry. *Shortcut Search* can retrieve a set of relevant phrases given only keywords or the first few characters or words of a string. (If no Shortcut is found to match the input text, the system seamlessly gives access to broad-coverage, interactive speech translation.) A **Translation Shortcuts Browser** is provided (on the left in Figure 1), so that users can find needed phrases by traversing a tree of Shortcut categories, and then execute them by tapping or clicking. Shortcuts are fully discussed in [8].

Identical facilities are available for Spanish as for English speakers: when the Spanish flag is clicked, all interface elements – buttons and menus, onscreen messages, Translation Shortcuts, handwriting recognition, etc. – change to Spanish.

3. Pilot Project and Evaluation

We now turn to a pilot project which tested Converser for Healthcare, Version 3.0, in three departments (Pharmacy, Inpatient Nursing, and Eye Care) of a large hospital complex belonging to a major US healthcare organization.

The hardware and software used in the project have been described and assessed in [6]. Accordingly, our focus here will be on the user experience. We rely on the healthcare organization's internal report, based on a commissioned survey by an independent third party, an experienced medical interpreter from an accredited local agency. While the report itself is proprietary, we'll reproduce its findings in essence.

First, however, several preliminary points are in order concerning stumbling blocks for the pilot project. As we will see below, all of these impediments have by now been removed as a result of the striking infrastructure advances over the four years since the pilot concluded.

Converser Version 3.0 was designed to cooperate with the then-current Dragon NaturallySpeaking, to be installed separately, and thus required *speaker-dependent* speech recognition: each speaker had to register his or her voice. This process took two or three minutes, including a 30-second speech sample; and, while this interruption was no great burden for English-speaking staff members, it usually made speech recognition from the Spanish patients' side impractical.

Microsoft's handwriting recognition was integrated into the system for both languages; but correction of errors was tricky at the time, so that this addition, too, incurred a training cost.

One more speed bump resulted from a software feature intended for customization: patients and staff could be registered in Converser, so that their names could appear in transcripts, and so that various personalization features could be added later. However, registration of the login user was required rather than optional; and this process necessitated still more training time.

Taken together, these obstacles necessitated 45-minute training sessions for participating staff members.

Further, because the experiments predated the era of modern tablets, portability was inferior to that available now, while physical set up was much less convenient ([9]). On the first-generation tablets used, for instance, it was necessary to manually configure the physical buttons which turned the microphone on and off.

With these initial obstacles in mind, we can now review the results of the organization's evaluation.

Project goals. The organization's goals for the project were stated in terms of the problem to be solved, as follows: (Throughout, we closely paraphrase the original language of the report.)

- Members' [i.e. patients'] language needs remain unmet in many situations throughout the … organization. Since the needs vary from situation to situation, no single solution can be expected.
- Different interpretative solutions need to be tested and analyzed to determine their best fit on multiple variables such as setting, situation, type of patient, etc.
- Accuracy of translation and member acceptance of technology-based interpretive services vs. in-person interpretation need to be assessed.

The independent interviewer observed 61 real-time translation interactions – some involving spoken input, some with typed or handwritten input – and solicited reactions from most of the staff and patients involved. (A few patients declined to answer the questions.) Interviews included both open-ended requests for reactions and prepared questions.

Patients' reactions. Positive comments from patients included the following:

- "cool"
- "useful" **– 5 mentions**
- "looks good" "well done"
- "would help"
- "good tool" **– 2-3 mentions**
- "I would recommend it"
- Even if translation was not 100%, it was always understood
- "Perfect and clear" **– 2 mentions**
- Saving time – don't have to wait for an interpreter
- "I like it"
- "I like the idea of it"
- Good for emergencies **– 2 mentions**

Less positive or negative comments included these:

- GUI too complicated (need larger buttons, crowded screen, …) **– 6 mentions**.
- Literacy issues: some immigrants can't read or write **– 6 mentions**
- Font size too small **– 3 mentions**
- "Too technical for me" "I don't like computers": family say elderly can't use **– 8 mentions**
- Quality of Sound/Volume issues **– 6 mentions**
- Handwriting didn't work **– 6 mentions** (Note: usage was limited)
- Worries about quality of translation **– 2 mentions**
- Keyboard issues (hard to use, pen is faster …) **– 5 mentions**
- Problems with English voice **– 2 mentions**
- System slow or froze **– 6 mentions**
- Hard to use tablet in hospital **– 1-2 mentions**

Some general patient comments:

- Training (for users) would be needed **– 4 mentions**
- Product would be "ideal" with voice recognition **– 4 mentions**
- A lot of mixed comments – They like the system but worry others (elderly, less literate) will struggle with it. (These comments came largely from partial or full English speaking members.)
- Would rather have an in person interpreter **– 4-5 mentions**

Staff reactions. Positive staff comments:

- Good for short interactions
- Writing was easier than talking
- Typing was easier than talking
- You can verify translations better vs. Language Line **– 2-3 mentions**
- I would use it if no other options
- Portability is good

Less positive or negative staff comments:

- Occasionally missed a sentence
- Computer literacy of members is a real issue. **– 3 mentions** (Also, elderly can't double-click fast enough.)
- User Interface – buttons crowded
- Translations were a bit odd
- Slow
- Hard for patients to write on the tablet in bed **– 2 mentions**
- Takes valuable time for the system to process

General staff comments:

- Training of patient's voice for DragonNaturallySpeaking would be needed.
- But time is limited already (i.e. no time in visit to train patients) **– 4 mentions**
- Training for staff and providers needed **– 3 mentions**
- This product is really more needed for Cantonese/Mandarin here in San Francisco.
- The system needs a formal introduction (so that the system can describe itself: for English providers to use with Spanish members).

Summary. Overview of patient and staff evaluations:

- High praise for the "idea." Higher than the actual experience of it
- Translation quality definitely "good enough" as rated by Members/Patients
- Limited English speakers (who can get along) would still use to verify the conversation and ensure completeness.
- Issues of literacy and computer literacy impact applicability
- Even though the system had issues (low to fair GUI, slow processing, lack of recognition of voice etc.), members partial or full English speakers thought it was "cool."
- Most people, and especially those who lacked English skills, preferred an in- person interpreter, although one person noted it saves time waiting for an interpreter, and a provider commented it saved the wait for Language Line.
- Good for emergencies
- Hard for members to use tablet in the hospital
- A number of patients declined to use in hospital but we lack data as to why.

Patient responses to six significant questions are tabulated in Table 1. The rightmost column shows the percentage of respondents who replied to each question with Completely or Mostly.

Most significantly, when asked whether the system met their needs, of the 79% of interviewed patients who answered the question, 94% responded either Completely or Mostly.

Table 1: **Patient responses to six questions**

Patient Evaluation	**% Answered Question**	**Completely or Mostly**
Did this meet your needs?	79%	94%
Was it accurate?	79%	90%
Was it easy to use?	72%	57%
Prefer handwriting question	67%	68%
Prefer using keyboard	67%	17%
Prefer to use handwriting and keyboard	67%	12%

4. System Revision

Having conveyed the organization's own assessment of the Converser for Healthcare 3.0 pilot project, we go on to describe the revisions prompted by it.

First and foremost, there was a glaring need to facilitate speech input from the Spanish side. This goal implied implementation of *speaker-independent* speech recognition; and this has been carried out by exploiting advances in Dragon NaturallySpeaking. Auxiliary third-party software has also been required to enable adaptation of Dragon software for use on desktop and tablet computers.

The need was also obvious for reduction in setup and training time. The following improvements reduce total warm-up to a few minutes for both staff and patients.

- The requirement for registration of the login user has been relaxed: registration is now optional, so that users can begin using the system immediately at startup time.
- An on-screen microphone button has now been substituted for the physical buttons previously used, so button configuration is no longer needed.
- Microsoft handwriting recognition has improved to the point that its correction facilities can be learned independently. Likewise, the company's on-screen keyboard now supports larger keys, so that on-screen typing has become more practical.
- Delivery of Converser via the Web will be enabled, so that only installation of the client software, providing access to a virtual desktop, will be required.

Another clear need has been to speed the interactions. While numerous staff members (and, separately, their managers) praised the ability to verify translations, others also stressed that verification consumed limited time. To balance these competing wishes, we have implemented a new set of icons allowing quick switching between Pre-Check and Post-Check modes. In the latter mode, useful when speed is more important than accuracy, speech recognition and translation are not checked in advance of transmission; but *post*-verification is still enabled, since back-translations are still generated and now appear in the bilingual transcripts (see Section 5). A **Rewind Button** has been supplied as well, so that erroneous or unsatisfactory translations can be quickly repaired and retransmitted. These new controls operate separately for English and Spanish speakers, so that, for instance, a doctor can pre-check when appropriate while allowing the patient to respond without distractions.

A number of interviewees called for various improvements in the user interface. In response, we have supplied large fonts for all on-screen elements (the exact size can be selected); added prominent icons for easier switching between English and Spanish speakers; enabled adjustment of the text-to-speech volume and speed, for easier comprehension; and added a quick way for staff to introduce Converser to patients, making use of our Translation Shortcuts. (We've also added more new Shortcut categories – including food, physical therapy, and mental health – since these browsable and searchable fixed phrases proved popular with staff members.)

5. Extended example

This section provides an example of the revised system in use. New elements introduced in the previous section are highlighted in *italics*.

Depending on the platform, the system can offer up to four input modes: speech, typing, handwriting, and touchscreen. To illustrate the use of interactive correction for speech recognition as well as machine translation, we assume that the user has clicked on the round red ***Mic Button*** to activate the microphone (Figure 1).

Still in Figure 1, notice the ***Traffic Light Icon*** and two ***Earring Icons***. These are used to switch between *Pre-check and Post-Check Modes* for translation and speech recognition, respectively. Both icons are currently green, indicating "Full speed ahead!" That is, verification has been temporarily switched off: the user has indicated that it is unnecessary to pre-check either ASR or MT before transmitting the next utterance, preferring speed to accuracy.

Just prior to the figure's snapshot, the user said, "San Jose is a pleasant city." Since verification had been switched off for both ASR and MT, these functioned without interruption. The speech recognition result appeared briefly (and in this case correctly) in the **Input Window**. Immediately thereafter the Spanish translation result (also correct in this case) appeared in the right-hand section of the **Transcript Window**, and was immediately pronounced via text-to-speech. Meanwhile, the

original English input was recorded in the left-hand section of the transcript.

Also on the English side of the transcript and just below the original English input is a specially prepared back-translation:[1] the original input was translated into Spanish, and then retranslated back into English. Proprietary techniques ensure that the back-translation means the same as the Spanish. Thus, even though *pre*-verification was bypassed for this utterance in the interest of speed, *post*-verification via the transcript was still enabled. (The **Transcript Window**, containing inputs from both English and Spanish sides and the associated back-translations, can be saved for record-keeping. *Inclusion of back-translation is new to Version 4.0.* Participant identities can optionally be masked for confidentiality.)

Using this back-translation, the user might conclude that the translation just transmitted was inadequate. In that case, or if the user simply wants to rephrase this or some previous utterance, she can click the ***Rewind Button*** (round, with chevrons). A menu of previous inputs then appears (not shown). Once a previous input is selected, it will be brought back into the **Input Window**, where it can be modified using any available input mode – voice, typing, or handwriting. In our example sentence, for instance, *pleasant* could be changed to *boring*; clicking the **Translate Button** would then trigger translation of the modified input, accompanied by a new back-translation.

In Figure 2, the user has selected the *yellow* ***Earring Icon***, specifying that the speech recognition should "proceed with caution." As a result, spoken input remains in the **Input Window** until the user explicitly orders translation. Thus there's an opportunity to make any necessary or desired corrections of the ASR results. In this case, the user has said "This morning, I received an email from my colleague Igor Boguslavsky." The name, however, has been misrecognized as "Igor bogus Lovsky." Typed or handwritten correction can fix the mistake, and the **Translate Button** can then be clicked to proceed.

Just prior to Figure 3, the ***Traffic Light Icon*** was also switched to yellow, indicating that translation (as opposed to speech recognition) should also "proceed with caution": it should be pre-checked before transmission and pronunciation. This time the user said "This is a cool program." Since the ***Earring Icon*** is still yellow, ASR results were pre-checked and approved. Then the **Translation Verification Panel** appeared, as shown in the figure. At the bottom, we see the preliminary Spanish translation, "Éste es un programa frío." Despite the best efforts of the translation program to determine the intended meaning in context, "cool" has been mistranslated – as shown by the back-translation, "This is a cold program."

Another indication of the error appears in the **Meaning Cues Window** (third from the top), which indicates the meaning of each input word or expression as currently understood by the MT engine. Converser 4.0 employs synonyms as Meaning Cues. (In the future, pictures, definitions, and examples may also be used.) In the present case, we see that the word "cool" as been wrongly translated as "cold, fresh, chilly, …".

[1] Proprietary, and branded as Reliable Retranslation™.

To rectify the problem, the user double clicks on the offending word or expression. The **Change Meaning Window** then appears (Figure 4), with a list of all available meanings for the relevant expression. Here the third meaning for "cool" is "great, fun, tremendous, …". When this meaning has been selected, the entire input is retranslated. This time the Spanish translation will be "Es un programa estupendo" and the translation back into English is "Is an awesome program." The user may accept this rendering, despite the minor grammatical error, or may decide to try again.

The new ***Traffic Light*** and ***Earring Icons*** help to balance a conversation's reliability with its speed. Reliability is indispensible for serious applications like healthcare, but some time is required to interactively enhance it. The icons let users proceed carefully when accuracy is paramount or a misunderstanding must be resolved, but more quickly when throughput is judged more important. This flexibility, we anticipate, will be useful in future applications featuring automatic detection of start-of-speech: in Green Light Mode, ASR and translation will proceed automatically without start or end signals and thus without demanding the user's attention, but can be interrupted for interactive verification or correction as appropriate. Currently, in the same mode, for inputs of typical length (ten words or less), the time from end of input speech to start of translation pronunciation is normally less than five seconds on a 2.30 GHz Windows 7 desktop with 4.00 GB RAM, and faster in a pending cloud-based version.

6. Conclusions

Following on earlier descriptions of Converser for Healthcare, Version 3.0, and a substantial pilot project which tested it at a leading San Francisco hospital, this historical paper has conveyed hitherto unpublished details of the resulting evaluation, as presented in the healthcare organization's internal reports, based in part upon interviews carried out by an independent third-party. We have also given an account of the system revisions in Version 4.0 which resulted from this feedback and from lessons learned independently.

We expect to release Version 4.0 in early 2016, and look forward to reporting the results.

7. Acknowledgements

The authors thank the many participants in the development of Converser for Healthcare and look forward to thanking by name the organization which sponsored the pilot project for Converser discussed herein.

8. References

[1] Mike Dillinger and Mark Seligman. 2004a. "System Description: A Highly Interactive Speech-to-speech Translation System." Association for Machine Translation in the Americas (AMTA-04). Washington, DC, September 28 – October 2, 2004.

[2] Mike Dillinger and Mark Seligman. 2004b. "A highly interactive speech-to-speech translation system." In *Proceedings of the VI Conference of the Association for Machine Translation in the Americas*. Washington, D.C., September-October, 2004.

[3] Yuqing Gao, Gu Liang, Bowen Zhou, Ruhi Sarikaya, Mohamed Afify, Hong-Kwang Kuo, Wei-zhong Zhu, Yonggang Deng, Charles Prosser, Wei Zhang, and Laurent Besacier. 2006. "IBM MASTOR system: multilingual automatic speech-to-speech translator." In *HLT-NAACL 2006: Proceedings of the Workshop on Medical Speech Translation*. New York, NY, June, 2006.

[4] Mark Seligman and Mike Dillinger. 2013. "Automatic Speech Translation for Healthcare:: Some Internet and Interface Aspects." TIA (Terminology and Artificial Intelligence) 2013: Proceedings of the Workshop on Optimizing Understanding in Multilingual Hospital Encounters. Paris, France, October 30, 2013.

[5] Mark Seligman and Mike Dillinger. 2012. "Spoken Language Translation: Three Business Opportunities." Association for Machine Translation in the Americas (AMTA-12). San Diego, CA, October 28 – November 1, 2012.

[6] Mark Seligman and Mike Dillinger. 2011. "Real-time Multi-media Translation for Healthcare: a Usability Study." Proceedings of the 13th Machine Translation Summit. Xiamen, China, September 19-23, 2011.

[7] Mark Seligman and Mike Dillinger. 2008. "Rapid Portability among Domains in an Interactive Spoken Language Translation System." *COLING 2008: Proceedings of the Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications*. Manchester, UK, August 23, 2008, pages 40-47.

[8] Mark Seligman and Mike Dillinger. 2006a. "Usability Issues in an Interactive Speech-to-Speech Translation System for Healthcare." HLT/NAACL-06: Proceedings of the Workshop on Medical Speech Translation. NYC, NY, June 9, 2006.

[9] Mark Seligman and Mike Dillinger. 2006b. "Converser: Highly Interactive Speech-to-speech Translation for Healthcare." HLT/NAACL-06: Proceedings of the Workshop on Medical Speech Translation. NYC, NY, June 9, 2006.

below.

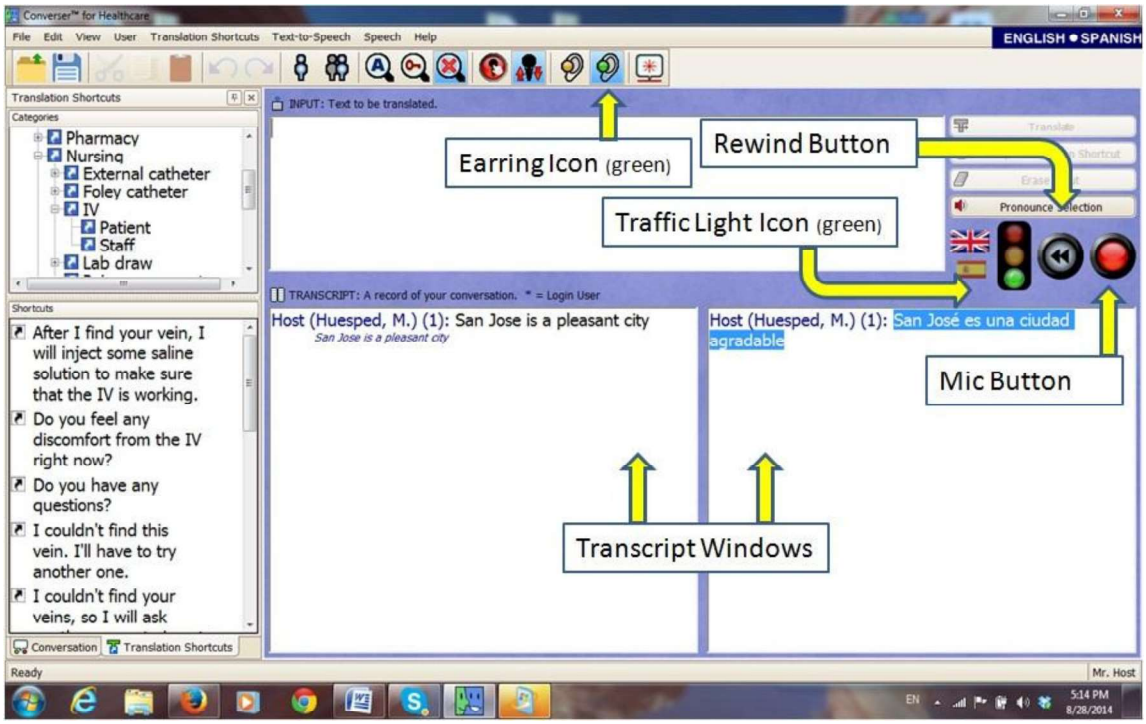

Figure 1: Earring and Traffic Light Icons are green: "Full speed ahead!"

Figure 2: Earring Icon is yellow: "Proceed with caution!"

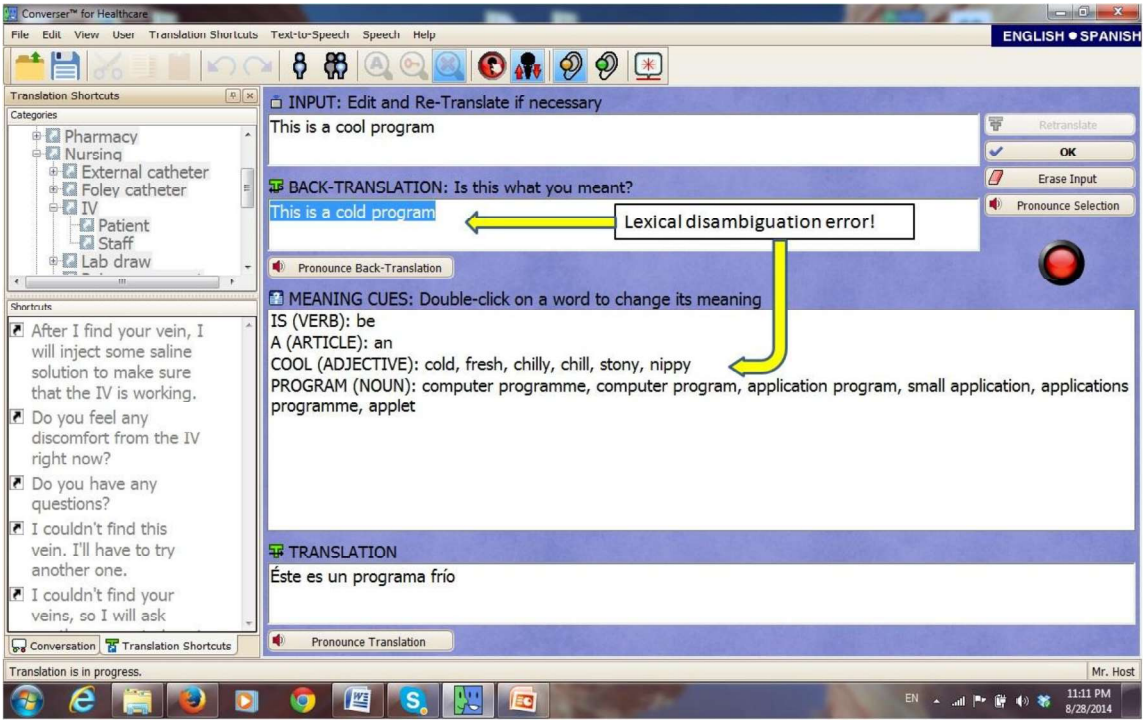

Figure 3: Verification Panel, with a lexical disambiguation error in *This is a cool program*.

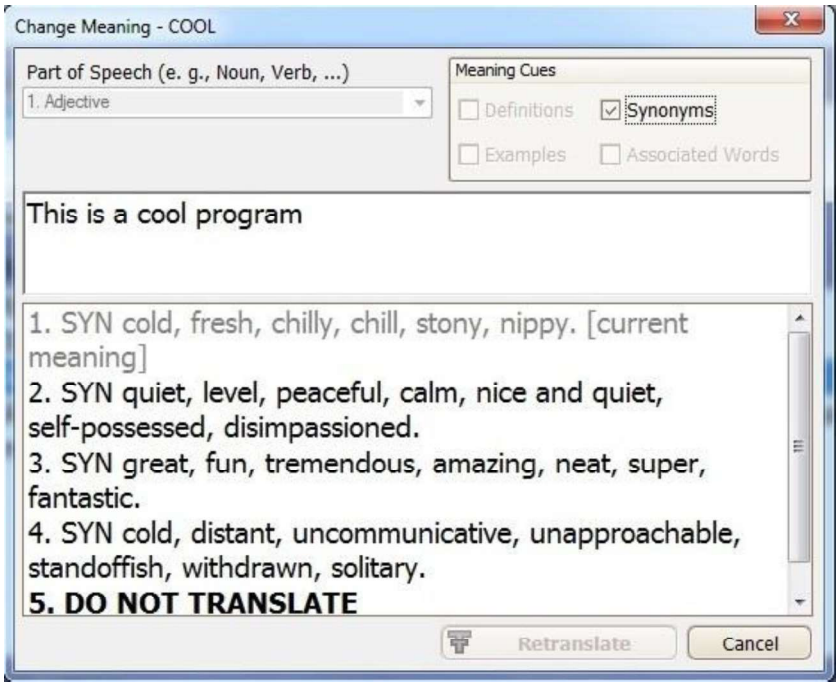

Figure 4: The Change Meaning Window, with four meanings of *cool*