Computer Speech Recognition and Language Learning: A Case Study

Don Ploger
Florida Atlantic University

Abstract
In 1968, Douglas Engelbart gave "The Mother of All Demos", showing a computer with a graphical interface controlled by a pointing device. Those experimental technologies of 1968 have become part of every day life, as have speech recognition and computer translation.

This case study follows the progress of a man learning a second language using readily available technology. When J spoke his native English, Google Translate achieved an accuracy rate greater than 95%. When J spoke Mandarin words that a native Chinese speaker could easily understand, Google Translate had an accuracy rate less than 40%. Using this feedback, J decided to learn French. In the beginning, there were many false negatives, words that a sympathetic native speaker understood, but the computer program did not. Despite these setbacks, there was long-term progress, and the methods are described.

These results provide some evidence to guide engineering practice. The human translator succeeded because she knew the extent of the speaker's limited vocabulary. The computer translation failed because it wrongly selected words that the speaker did not even know. This comparison suggests that translation could be improved by adjusting the vocabulary to include only words that the learn knows. The end result should be a computer that understands human speech and performs accurate translations with an interface that is easy to use.

Introduction

To err is human; to forgive, divine.
Alexander Pope [1]

To err is human, to really foul things up requires a computer.
William Vaughan [2]

"Engineering shaped a century and changed the world." (National Academy of Engineering, 2014). [3] It is important to communicate these great accomplishments to engineering students in engineering survey courses (Landis, 2013). [4] However, engineers are a very small percentage of the population, and it is important for these achievements to reach wider audiences.

This study examines one of those achievements: the modern computer. Engineers have made computers part of the public's lives. Fifty years ago, a computer was extremely expensive and filled an entire room. The average person could not possibly use it.

In December 1968, Apollo 8 orbited the moon: a remarkable achievement. That mission required, among many other engineering accomplishments, a computer that could withstand the pressures associated with the launch and re-entry of a spacecraft. The solution was the Apollo Guidance Computer (AGC), the first computer that used integrated circuits as essential components (NASA, 1987) [5].

That same month of 1968, Douglas Engelbart gave "The Mother of All Demos," showing a computer with a graphical interface controlled by a pointing device (Engelbart, 1970) [6]. The user could communicate with the new computer simply by typing, which was impossible with commercial computers of the day (Stanford University Libraries, 2005 [7]; Doug Engelbart Institute, 2014) [8]. Even more amazing, he or she could control the video screen with a pointing device. To achieve this demo, the combined power of several University computers was necessary (Palfreman & Swade, 1991, p. 98) [9]. As a result, many people believed it was impossible to have a commercially available computer that would be easy for people to use (WGBH Television, 1992)[10].

Despite all of the engineering accomplishments of the 1960s, computers remained difficult for non-engineers to use. For many people, Vaughan's joke, "to really foul things up requires a computer," expressed their frustration. That quote entered the lexicon, and has continued for half a century.

Engelbart's work has been recognized within the engineering community: "1968: The computer mouse makes its public debut during a demonstration at a computer conference in San Francisco. Its inventor, Douglas Engelbart of the Stanford Research Institute, also demonstrates other user-friendly technologies such as hypermedia with object linking and addressing. Engelbart receives a patent for the mouse 2 years later." [3]

Those experimental technologies of 1968 have become part of every day life. The computer that took the Apollo spacecraft to the moon had less power than a solar calculator that costs less than $20 today.

In addition to the wide range of consumer electronics products, there is also speech recognition and computer translation.

Proof of concept

This is a proof of concept case study. There was no attempt to test a theory of learning, and there was clear recognition that this experience may not apply to all second-language learners. The question of the study was very practical:

> Can speech recognition programs help someone learn a second language?

The goals are, in some ways, similar to those of Englebart. The primary goal is to create something that works in helping a person learn a foreign language. There are, however, also important differences. Englebart needed state-of-the-art technology (for 1968) in order for his demo to work. In contrast, this study used common existing technology: a smart phone and the Google Translate App that is available free of charge.

Although the primary emphasis is practical, there is a concern with communication to non-engineers. Communicating with the general public is a well-established issue in engineering education: “[E]ven though we are able to create complex objects that could save lives or millions of dollars, it is all worth nothing if we cannot communicate our ideas or work with others.” Dunsmore, Turns, & Yellin, 2011) [11]. This study is also guided by previous work on making engineering concepts accessible to a wider audience (Ploger, Shankar, Nemeth, & Hecht, 2013) [12].

The Case Study

This case study focuses on the work of one native-born 62-year-old American. J had previously taken eight years of school French (from Grades 5 through 12). He took one semester of college French (which he completed with a C+ to satisfy the foreign language requirement). At the beginning of this case study, J understood approximately 1000 French words, but his pronunciation was only reasonably accurate for about 200 words.

J’s wife is Chinese and was born in China, and J had spent a total of ten weeks in China. As a result, he had some familiarity with Mandarin. J understood approximately 300 Chinese words, but his pronunciation was only reasonably accurate for about 40 words.

An Attempt to Learn Chinese

In one example, the spoken material consisted of the Chinese words for the numbers 1 through 10. J first counted to 9. In each case, the Google translate correctly transcribed those nine spoken Chinese words.

When J spoke the Chinese words for the numbers 1 through 9, the iPhone correctly transcribed:

“1 2 3 4 5 6 7 8 9”. In this case, it was correct for each spoken number.

When J spoke the Chinese words for the numbers 1 through 10, the Google Translate result was:

1 2 3 4 5 6 7 8 9 **is**

In this case, the translation was correct for each of the first 9 spoken numbers, but the computer program incorrectly transcribed the Chinese word for “ten”.

After hearing “1 2 3 4 5 6 7 8 9,” there is an expectation that the next word would be “10.” The human translator correctly recognized the word. This error could easily be corrected with a simple program that could recognize consecutive numbers. There is, however, an even simpler solution. If the vocabulary size were limited to the first ten numbers, then the Google translate would not have made any errors in that brief passage.

Although J was highly motivated to learn Chinese, he concluded that it would be too difficult for the following reasons:
He could not master the pronunciation.
He could not use speech recognition effectively because the computer only recognized a small percent of the words native speakers understood.
He could not read the language.

Initial Attempt to Learn French
Instead, J chose to master French, at a conversational level, since he and his wife will visit Paris in the next year.
He had taken eight years of school French and a semester at university.
He could read the language.
He could recognize approximately 1500 French words.
His pronunciation, although not good at the beginning, was adequate to begin using speech recognition.
In addition, many familiar movies had both French dialog and subtitles.

When J spoke familiar French sentences, the speech recognition program made many mistakes. But a native speaker, familiar with J’s accent, could have easily understood the words. In some cases, there are words common to both French and English (such as mademoiselle, Louvre, metropolitan): When spoken in the English mode, the program was over 95% accurate. In French mode, initially, the program was less than 50% accurate. This demonstrates that Google translate is making errors in J’s French because of his accent.

Creating a Simple Script
In response to these many errors, J decided to make a simple example. Working with a native speaker, J created a 15-word script in French. With practice, J pronounced those words so that Google Translate was accurate in more than 95% of the cases.

Bonjour Madame; Bonjour Monsieur.
Comment allez-vous?

S'il vous plait.
Est-ce que vous comprenez l'anglais?
Merci.

Hello Madam; Hello sir.
How are you?
Please.
Do you understand English?
Thank you.

Extending the Script

With this success on a very simple example, J used the basic script as the starting point for further learning. At this point, J had a dramatic improvement in his pronunciation as measured by both the speech recognition program and a native speaker. Therefore, progress was now based upon accurate pronunciation.

The first additional script contains 58 words and is shown in Appendix 1. This includes one sentence from the Basic Script: "Est-ce que vous comprenez l'anglais?" [Do you understand English?].

The second additional script contains 40 words and is shown in Appendix 2. The first three sentences were provided by the native speaker. The following four sentences were generated by J, but edited by the native speaker.

The third additional script contains 60 words and is shown in Appendix 3. J generated the original version of this script after watching the 1963 movie, *The Birds*, dubbed in French. The native speaker provided the essential corrections to the grammar.

Results Summary

There were four major milestones in using speech recognition to learn a foreign language.

1. J could not speak Mandarin at a satisfactory level: He could not count to "ten". The computer program did not understand J's pronunciation for "ten", even though it was obvious to a human translator. When J spoke familiar words in Mandarin, the accuracy rate was less than 40%. The human translator succeeded because she knew the extent of the speaker's limited vocabulary. The computer failed because it wrongly selected words that the speaker did not know. J concluded that Mandarin was too difficult.

2. Speaking familiar French sentences, J found that the speech recognition program still made many mistakes. A native speaker, familiar with J's accent, could understand his words. In some cases, there are words common to both French and English (such as Mademoiselle, Louvre, metropolitan): When spoken in the English mode, the program was over 95% accurate. In French mode, initially, the program was less than 50% accurate.

3. Working with a native speaker, J created a simple script in French. With practice, J pronounced those words so that Google Translate was consistently 98% accurate.

4. J demonstrated a step-by-step improvement in his pronunciation as measured by both the speech recognition program and a native speaker. There were a total of 202 words in the 4 scripts used (a complete transcript is shown in the proceedings paper). Initially, accuracy ranged from 43 to 62%. At the end of the initial phase of the study, the assessment showed the following results: Of the 202 words, the Google Translate App correctly typed 192: an accuracy rate of 95.0%. For the 10 errors, J repeated the passage: the Google Translate App correctly typed each word with (at most) two re-takes.

Discussion

The results did confirm that existing speech recognition programs provided definite benefit to a person learning a second language. Although the study began as a practical investigation, it is relevant to important theoretical concepts in both engineering and human cognition.

False negative errors

The first relevant concept is that of a false negative, when the learner says a word that is understood by a native listener, but not by the speech recognition program. The importance of false-negative errors is well known, and these false-negatives can decrease motivation (Cai et. al., 2013). [13]

Attribution theory

It is especially important to reduce false negatives because of attribution: users intuitively respond to the computer as if it were a person. If there are many false negatives, the computer is is considered to be unhelpful. In a study of native speakers using speech recognition with a computer game, false negative errors inhibited their enjoyment of the game. For example, one user wrote that false negatives “made me less engaged, because I felt like [the game] was counting off for something I knew” (Cai et. al., 2013). [13]

In a classic study, Heider & Simmel (1944) [14] showed a video in which two triangles and a circle move on a screen. Nearly every one of the students interpreted the action, not as the movement of geometric shapes, but as a story involving three people. Most reported that the two triangles were fighting for the affections of a young woman, with the smaller triangle defeating the bully (Heider and Simmel (1944) animation). [15]

To understand the effect of false-negative errors on motivation, it is useful to consider attribution theory. Humans naturally interpret the movement of objects as being the result of a conscious entity. In the classic study, Fritz Heider found that college students interpreted the movement of graphical objects on a video as being characters with thoughts and feelings.

People tend to see cause and effect relationships even where there is none. (McLeod, 2010) [16]. Attribution applies to speech recognition. People interpret the computer, not as a program that runs on a particular machine, but as another person. In particular, there is a strong tendency to evaluate that other entity as either helpful or unhelpful.

There are also well-established software engineering methods to reduce the number of false negative errors. It is well known that errors are relatively low when the working vocabulary is small. This suggests a definite intervention. Since the second language learner only knows a very small number of words, the program could limit its search space to only those words.

Learning a Second Language

Research on language development indicates that there is a critical period for learning a language (Lenneberg, 1967) [17]. Learning a language is easy during the first three years of life. After age 7, however, it is much more difficult (Kuhl, et. al. 2005). [18]

The study also provided a detailed example that is consistent with language development. It is well established that there is a critical period for the language learning, and this is consistent with common sense. Each of us learned our native-language without conscious effort by age 4. Most of us spent the same number of years struggling to learn a second language without success.

The existence of a critical period does not it all mean that it is impossible to learn a second language, but it does mean that powerful methods are needed (Pimsleur, 2013) [19]. This study suggests that existing instructional methods could be enhanced with speech recognition.

Vocabulary size

It is well established that the size of vocabulary increases the error rate (Waibel & Kai-Fu, 1990) [20]. However, over the past four decades, the maximum vocabulary size for large speech recognition has increased substantially (Huang, Baker, & Reddy, 2014) [21]. In fact, for real-time natural language dictation systems in the late 1990s the vocabulary size essentially became unlimited. The process was made to appear as seamless as possible to the interactive user. However, the problem remains a challenge because modeling new words is still far from seamless when seen from the point of view of the models, where the small-sample models are quite different from the large-data models (Huang, Baker, & Reddy, 2014). [21]

These finding apply to native speakers. For language learners, the situation is quite different. In the beginning, a learner has a very small vocabulary. In the very early stages, that vocabulary is less than 100 words. Despite this, the speech recognition program searches the entire dictionary. The result, as was seen in the early phases of the case study, was many false negative errors.

There appears to be a logical remedy: adjust the vocabulary size to fit that of the language learner. If the learner only can speak 100 words in the new language, then the speech recognition program would only search those 100 words. There is evidence from this case study that this would improve performance: in almost every false negative error, the computer typed a word that the learner did not know.

Under favorable conditions, a vocabulary size of 2000 to 3000 words provides a very good basis for language use. Nation & Waring (2011) [22]

There is also evidence that adjusting the vocabulary size improves speech recognition performance. Working with native speakers who were speaking at a high rate, Cai, Miller, & Seneff (2013) [13] observed high rates of false negative errors, which lead to a decrease in motivation. They modified the program to include only commonly used words, and performance improved. With that improved performance, participants became more interested in the task.

The Metaphor Of The Target

It is possible to have a speech recognition accuracy scale similar to that of marksmanship. The goal of a speech recognition system is to type the words that a person says. For a second language learner, the system often fails to type a word that a native listener can easily understand. This is a false negative.

To understand a false negative, it is helpful to consider firing a rifle at a target. It is not simply a question of a hit or a miss. Did it hit the bull's-eye? Did it miss the target completely? Did it miss the broad side of the barn? In marksmanship, a shot is not simply given a pass-fail grade: it is given a score. It is likewise possible to give a score to any word spoken.

In marksmanship, the U.S. Army defines four levels of accuracy (U.S. Army Board Preparation Guide) [23]. For Rifle record fire, these standards (out of a possible score of 40) are shown as follows:

Rating	**Standard**
Expert	36-40
Sharpshooter	30-35
Marksman	23-29
Unqualified	22-Below

One can consider a speech recognition scale similar to that of marksmanship. The center of the bull's-eye is a pronunciation at the level of a native speaker. The broad side of the barn corresponds to the computer's ability to match spoken words to a specific script. On the one hand, achieving native level competency in pronunciation is an unrealistic goal. On the other hand, failure to match words in a given script indicates that the learner needs to spend more time listening before attempting to speak a foreign language.

In the proposed four levels, the highest level is the full vocabulary list on Google Translate and the lowest level is completely incomprehensible. Between these two extremes, there are three clearly defined intermediate levels. Even if a person is minimally capable, if the computer has the script of what he or she is saying, it will be able to understand. On the other hand, the pronunciation level that would be understood with current speech recognition software would be given a high score, but not the maximum. Even in the early stages, then, a learner would not be encouraged to continue practice. And even in relatively advanced stages, the learner would get feedback to the effect of, "Very good, but you can do even better."

In between these two extremes, there are several important levels of efficiency that are relatively easy to define. One can consider the minimally acceptable level, at which the speech recognition system can correctly identify the words spoken from a list of words at the speaker can pronounce. The next higher-level would also include the 1000 most common words in the language, a level that is generally considered acceptable for travelers to succeed in the country.

Figure 1 shows a framework that defines four levels of proficiency in terms of vocabulary size.

<u>Blue</u>: The sentence can be understood by a sympathetic native listener, considering the full vocabulary list on Google Translate.

<u>Green</u>: The sentence can be understood when the native listener has a list of the **learner's vocabulary**.

<u>Yellow</u>: The sentence can be understood when the native listener has the **script** that the learner is reading.

<u>Red</u>: The sentence cannot be understood, even when the native listener has the script that the learner is reading.

Figure 1. Four Levels of Proficiency in Pronunciation

It is not be realistic for a non-native speaker to attempt to achieve the highest level. On the other hand, if the performance is at the lowest level, a learner might do better to focus on listening before attempting to speak.

It is the three middle levels that are important for the second language learner. It would be helpful if the software could understand a non-native speaker whenever a sympathetic native listener could understand. It would also be helpful for the system to provide feedback on the level of pronunciation.

Future Directions

The future of engineering will build on the successes of past work. Englebart's demo combined a crucial engineering invention (the X-Y position locator) with a vision of the future. Success depended upon a revolution in both hardware and software and because both of these revolutions occurred, every child can now use a mouse.

In the case of this proof of concept, everything was done using only currently available materials. Creating an actual product for second language learning would not require any changes in computer hardware, nor does it require a revolution in software engineering. It does require good programming and following established principles for improving the accuracy of speech recognition.

The general principal, accuracy increases by reducing vocabulary size, is true in all cases. For native speakers, current speech recognition systems are so efficient that it is possible to consider all of the words in the dictionary and still have a remarkably high accuracy rate. For language learners, however, the tradeoffs are different. The learner does not need the full dictionary. In this case, reducing vocabulary size could lead to a substantial increase in the accuracy of the computer's performance.

Central to this study was the high frequency of false negative errors. These errors occur because both human and computer are fallible. The error rate is very low for native speakers, but for an adult learning a new language, a certain level of error is inevitable. Human listeners can forgive these errors. Speech recognition systems potentially have that same capacity for forgiveness.

Humans use context -- the meaning of a sentence -- to make predictions about words that are not pronounced clearly. Current speech recognition programs cannot use context nearly as well as a person can. In the near future, speech recognition programs will not be able to use meaning, but they can modify vocabulary size. The effect of decreasing vocabulary size for the language learner would greatly increase the accuracy. The result would be that the learner would feel that "the computer understands me." Of course this is an illusion, but it is a useful illusion which can help someone with the considerable challenge of learning a second language.

Acknowledgements: The author thanks Stephanie Kujawa for many helpful suggestions during the development of the manuscript; Penelope Fritzer for helpful comments; and Maikane Deroo for many helpful suggestions during the development of this study and for ensuring that the French is accurate and *au courant*.

Bibliography

[1] Pope, A. An Essay on Criticism. Retrieved from http://poetry.eserver.org/essay-on-criticism.html

[2] Quote Investigator (2010). To Err is Human; To Really Foul Things Up Requires a Computer. Retrieved from http://quoteinvestigator.com/2010/12/07/foul-computer/

[3] National Academy of Engineering (2014). Greatest engineering achievements of the 20th century. Retrieved from http://www.greatachievements.org

[4] Landis, R. B. (2013). Studying Engineering: A Road Map to a Rewarding Career, 4th Edition. Los Angeles, CA: Discovery Press.

[5] NASA. (1987). Computers in Spaceflight: The NASA Experience. Computers On Board The Apollo Spacecraft. Retrieved from http://history.nasa.gov/computers/Ch2-1.html

[6] Engelbart, D. (1970). X-Y position indicator for a display system. U. S. Patent US 3541541 A. Retrieved from http://www.google.com/patents/US3541541

[7] Stanford University Libraries. (2005). MouseSite: The Demo. Retrieved from http://web.stanford.edu/dept/SUL/library/extra4/sloan/mousesite/1968Demo.html

[8] Doug Engelbart Institute (2014). Doug's Call to Action. Retrieved from http://www.dougengelbart.org/about/vision-highlights.html

[9] Palfreman, J. & Swade, D. (1991). The dream machine: Exploring the computer age. London: BBC Books.

[10] WGBH Television (1992). The machine that changed the world. Produced by WGBH Television in Boston MA, in cooperation with the British Broadcasting Corp., with support from ACM, NSF and UNISYS. Retrieved from https://www.youtube.com/watch?v=iwEpKy_7mYM

[11] Dunsmore, K., Turns, J., & Yellin, J. M. (2011). Looking toward the real world: Student conceptions of engineering. Journal of Engineering Education, 100(2), 329-348.

[12] Ploger, D., Shankar, R., Nemeth, A. & Hecht, S. (2013). Engineers and Other People: Communicating Engineering Results to a Larger Audience. Proceedings of the Conference for Industry and Education Collaboration (CIEC), Phoenix, Arizona, February 6-8, 2013.

[13] Cai, C. J., Miller, R., & Seneff, S. (2013). Enhancing Speech Recognition in Fast-Paced Educational Games using Contextual Cues. Speech and Language Technology in Education Proceedings. Grenoble, France - August, 2013. Retrieved from http://www.slate2013.org/images/slate2013_proc_light_v4.pdf

[14] Heider, F., & Simmel, M. (1944). An experimental study in apparent behavior. The American Journal of Psychology, 57, 243-259.

[15] Heider and Simmel (1944) animation. Retrieved from https://www.youtube.com/watch?v=VTNmLt7QX8E

[16] McLeod, S. A. (2010). Attribution Theory. Retrieved from http://www.simplypsychology.org/attribution-theory.html

[17] Lenneberg, E. (1967). Biological foundations of language. New York: John Wiley & Sons.

[18] Kuhl, P. K., Conboy, B. T., Padden, D., Nelson, T., & Pruitt, J. (2005). Early speech perception and later language development: implications for the" Critical Period". *Language Learning and Development*, *1*(3-4), 237-264.

[19] Pimsleur, P. (2013). How to learn a foreign language. New York: Simon & Shuster.

[20] Waibel, A. & Kai-Fu, L (Eds.). (1990). Readings in speech recognition. San Mateo, CA: Morgan Kauffman Publishers, Inc.

[21] Huang, X., Baker, J., & Reddy, R. (2014). A historical perspective of speech recognition. Communications of the ACM, 57(1), 94-103

[22] Nation, P. & Waring, R. (2011). Vocabulary Size, Text Coverage and Word Lists. Retrieved from http://www.robwaring.org/papers/CUP/cup.html

[23] U.S. Army Board Preparation Guide (2014). High, R. (Ed.) Trainers guide to successful marksmanship training. Retrieved from http://www.armystudyguide.com/content/army_board_study_guide_topics/m16a2/trainers-guide-to-success.shtml

Don Ploger, Ph.D., is an Associate Professor of Education at Florida Atlantic University. As a mathematics educator he is interested in how people solve novel problems and how essential mathematical knowledge is communicated from engineers to non-engineers. He is also involved in researching how people learn foreign languages using computer speech recognition and translation tools.

Appendices

Appendix 1: Expanding the Basic Script

Vous comprenez.
Est-ce que vous comprenez?
Oui, je comprends.
Non, je ne comprends pas.
Est-ce que vous comprenez l'anglais?
Bien sûr, je comprends l'anglais.
Est-ce que vous comprenez le français?
Je peux comprendre le français.
Si vous parlez lentement, je comprends le français □.
Parlez-vous français?
Oui je peux parler le français.
J'aime bien parler français, chaque jour.

You understand.
Do you understand?
Yes, I understand.
No, I do not understand.
Do you understand English?
Of course, I understand English.
Do you understand French?
I can understand the French.
If you speak slowly, I understand French.
Do you speak French?
Yes I can speak French.
I like to speak French every day.

Appendix 2: Movies

Tu es libre ce soir pour aller au cinéma ?
Le dernier film vient de sortir. Je veux le voir.
Je regarde la télévision .
Je regarde un film.
Je regarde un film à la télévision.
Je regarde un film en français.
J'aime bien le film, "Le Magicien d'Oz".

Are you free tonight to go to the movies?
The latest film is out. I want to see it.
I watch TV.
I watch a movie.
I watch a movie on TV.
I watch a movie in French.
I like the movie, "The Wizard of Oz."

Appendix 3: The Birds
1963: un film sort au cinéma, “Les Oiseaux.”
La star du film s'appelle Mélanie.
Les oiseaux attaquent les gens.
Mélanie va au restaurant.
Elle dit: “Il y a une guerre des oiseaux.”
Elle rencontre une dame, qui se trouve être une ornithologue.
“Une guerre contre les oiseaux? Je ne pense pas,” dit l'ornithologue.
“Mais, c’est vrai!” dit Mélanie.
“C’est dommage: Les oiseaux gagneraient.” dit l'ornithologue.
Il y a un homme dans le restaurant.
Il dit, “C'est la fin du monde.”

There is a famous movie, "The Birds."
The star of the movie was named Melanie.
Birds are attacking people.
Melanie goes to a restaurant.
She said: "There is a war against the birds."
She meets a lady who happens to be an ornithologist.
The ornithologist: “A war against the birds? I don't think so,"
Melanie: “But it’s true!”
The ornithologist: "That’s too bad: The birds will win."
A man in the restaurant says:
"It's the end of the world!"