# Contextual Inquiry, Task Analysis, Design Sketches

One World – Your Translation Companion

## Team Members

* Kenny Risk
* Alex Cheng
* Angapparaj Kalimuthu
* Daniel Mejía V.

## Problem and Solution Overview

Imagine you are making a conference call on your phone or computer with people from different parts of the world. Ideally, in those situations each person could speak in the language they feel more comfortable with. With the current advances in automated speech recognition (ASR) and machine translation (MT) technology, this is now becoming a possibility. However, there are huge challenges not only in terms of technologically improving the quality of speech recognition and translation, but also in terms of defining natural and effective user interfaces.

We want to tackle problem of designing a user interface that would enable effective and smooth communications across different languages. Having perfect speech recognition and translation systems would facilitate the task but in reality errors and imperfections in these technologies are expected and so far unavoidable. Therefore, the interface model needs to change in order to embrace the possibility errors and perhaps involve the user for resolution.

The focus of our project is to solve the problem of enabling natural cross-language communications, and for that our proposal is to create a flexible speech translation system where users can communicate naturally but also have the option provide feedback when needed in the case errors. Some of the key design questions that our interface must answer are the following:

* Text output or speech output
* Speech recognition correction model
* Video output
* Automatic or manual submissions of translated messages
* Display of content to aid assessment of recognition and translation quality.

## Contextual Inquiry Participants

For our first two contextual inquiry exercises, we chose to observe people who conduct business meetings in different languages through the help of an interpreter. More than ever before, businesses lean on global partners to achieve shared goals. Against this backdrop, we felt that interpreted dialogue in a business setting was a critical audience to capture. If we were unable to provide significant value to such users, it would be difficult to call our application a success.

As a person working at a foreign company, Kenny's workplace afforded us the opportunity to conduct contextual inquiry in a real business setting, so we took advantage of this. Kenny works at Nintendo of America, and interpreted meetings through video conferencing between NOA and its parent company in Kyoto, Japan are very commonplace. For the purposes of conducting contextual inquiry, Kenny attended two such meetings (the other project group members were unable to attend given the confidential nature of the meetings). One complication was the fact that it was impossible to ask questions during the meetings, as that would interrupt the flow of the conversation. Instead, valuable insights were obtained through discussion after the meetings had ended. Kenny's command of Japanese prevented him from being able to completely become a novice participant, but by the same token, he was in a unique position to be able to observe the interpreter in their interactions with both the American side and Japanese side.

### Participant #1

The first meeting that Kenny attended was a biweekly meeting between members of his team at work, and members from Japan discussing points of interest surrounding technical support that NOA provides to third-party licensee game developers. There were five participants from Japan, and five from Kenny's team, including the interpreter and Kenny himself.

### Participant #2

The second meeting that Kenny attended was a meeting discussing high-level issues surrounding Nintendo’s network infrastructure and service level agreements. There were six participants from Japan, four participants from NOA (not including Kenny), and the interpreter, who was also from NOA. Originally scheduled for two hours, the meeting lasted close to four. While nobody from Kenny's team other than himself was a participant, he was able to attend the meeting after asking a separate internal group that provides technical interpretation services throughout the company whether there were any meetings that he could attend as an observer. The interpreter for the meeting was from this group. There was another candidate meeting that was considered, but Kenny was unable to attend due to its confidential legal subject matter, so this network infrastructure-related meeting was the best opportunity to observe a topic about something different from the first meeting.

### Participant #3

For the third contextual inquiry, we wanted to try something different. Communication systems are useful in a wide variety of situations and a system that enables people from different languages to talk naturally to each other can open many possibilities. We did not want to tackle the problem too broadly, but we also did not want to restrict it only to people who currently use interpreters. Therefore, we decided to scope it and target at least one other type of audience, and for that we chose the people who speak only a little of some foreign language and then struggle trying to make themselves understood when needing to use it to speak to someone. The third participant was a great representative of this audience.

Mark is Daniel’s co-worker. He travels for pleasure very frequently and he likes to get involved and understand different cultures. On multiple occasions he has found the need to communicate directly with people in a different language to make travel arrangements or just to get by while traveling. In those occasions, Mark has tried to do his best especially when he knows the language a little (he mentioned specifically that he speaks some very basic Spanish and French). While he was not coincidentally planning a trip at this precise time, we knew he had done it before and this presented a great opportunity for contextually observing him while recreating the situation.

With this in mind, we asked Mark to make a reservation over the phone, in Spanish, at a hotel in Cartagena, Colombia. The exercise was performed in his office, where he had his computer and any other tools that he could need for preparing and communicating. As part of the introduction, Daniel explained the mechanics of the exercise and what we were trying to accomplish. He was encouraged to do whatever was natural, and to speak through the process mentioning what was easy, what was problematic, and what had been useful in the past. As the apprentice, Daniel observed and listened to Mark, but also probed him to stir the exercise in a direction that would give us insight for the product in mind.

## Contextual Inquiry Results

### Results for Participants #1 & #2

Both meetings were conducted via videoconference. There was a single person who conducted the flow of the meetings, and this flow was determined by an agenda that was prepared and communicated to the attendees prior to the meeting. For both meetings that Kenny observed, the conducting person was on the Japanese side. There was also a single interpreter for both meetings, and in both cases, this person was from the American side. The interpretation was not concurrent with what was being said--in other words, one person would say something, and the interpreter would translate it after it was said. The conductor would go through the list of discussion items, and describe a question or discussion point. The interpreter would translate that question or statement, and from there, there would be discussion back and forth, with the interpreter translating in between each step.

In observing the meetings, one point of interest was the granularity of translations. In the first meeting, there was a certain individual on the Japanese side who would make a very long statement without waiting for the interpreter, making it difficult for the interpreter to recall every detail that was said. Notes were being taken by the interpreters constantly, and it was obvious why: to provide as accurate a translation as possible. (Side note: grammatically-speaking, there are many things that are in reverse order between English and Japanese. For example, objects of transitive verbs appear after the verb in English, but before the verb in Japanese. Taking good notes is thus crucial in order to reconstruct sentences to match the target language's word ordering.) For note-taking purposes the interpreter in the first meeting utilized his laptop, while in the second meeting, the interpreter used pen and paper. Certain people seemed to be more considerate of the interpreter (or perhaps deliberate in their effort to make sure what they were saying was communicated as correctly as possible), as they would say one to two sentences and pause for the interpreter's translation, and then continue with their thought after that.

One common feature in both meetings was the presence of stretches of time where there would be lengthy discussion between only members of a particular side. In these cases, the interpreter would not translate each part of what was being said, but instead a single member would prepare a summarizing remark of the discussion that just took place and the interpreter would just translate that. In the lengthy second meeting, some of these sub-discussions on the Japanese side, for example about automated versus manual testing, would last upwards of 15 minutes. Following the meeting, an IT Manager on NOA's side expressed his regret in not being able to understand what was individually being said and by whom, rather than hearing just a summarizing remark after consensus had been reached. Valuable information that would be strategically useful--namely, which person on the Japan team had which opinion and so on--was lost in this case, and was difficult for the interpreter to both follow and translate for the NOA members in real time.

Here are some other noteworthy findings that were found while observing the meetings:

* Interpreter as active facilitator - In the second meeting, the interpreter on several occasions presented pinpoint questions to a certain side based on what he heard from the other side and his understanding of the topic matter. This proved helpful in moving the discussion along. The first interpreter took a less active stance and simply provided interpretation of the remarks.
* Latency - While it is obvious that interpretation would increase the latency of the conversation by about a factor of two, for contentious issues, the efficiency seemed to reduce even further due to the language disconnect. This was particularly the case for the second meeting. However, when asked, the participants understood the inevitability of interpretation adding time to the meeting, and they didn't seem to feel disjointed by the delay, albeit if not tired from an extremely lengthy meeting.
* Summary vs. word-for-word - Often times, the interpreter would summarize what was said as opposed to providing a word-for-word translation. Kenny found this to be effective more often than not. While sometimes information was lost in summarizing, he felt that its positives in terms of efficiently communicating the message outweighed its negatives.
* Lengthy pauses - given that abstract thoughts were being formulated on the fly about contentious issues, there would often be lengthy pauses between statements or even words of a single statement.
* Interpreter feedback - In the second meeting, when the interpreter missed fully understanding a given statement, he would ask, "Could you run that by me again?" (The interpreter in the first meeting did this to a lesser degree and would instead do his best to translate given the provided information.) Also, on one occasion, the second interpreter said he was unsure of what the translation of "fund settlement" would be in Japanese prior to translating it. The NOA members reassured him that the Japan team would understand if he just said the term as an English word.
* Terminology - Many terms that would be unknown to outsiders (such as Kenny) were being used frequently. Separate from the two meetings that were observed, an interpreter of legal meetings at NOA expressed to Kenny how important and involved legal terminology is. Often times with such terms, literal translations would simply yield confusion on the other side. A direct mapping of terms (and a good memory if you are a human translator) is thus necessary with such important concepts.

### Results for Participant #3

In the first two inquiries, the goal was to see an actual speech translation setting in action and get some insight on the mechanics of using a third person in between. For the third inquiry, on the other hand, the goal was to observe how people without interpreters manage to communicate, what tools they find useful, and what is particularly painful for them that could be solved with a speech translation system. Mark was very helpful in this regard.

The first thing Mark did was to prepare the questions to ask and make sure he knew how to say them in Spanish. He used Bing translator (for disclosure, Daniel and Mark both work in Bing, but nowhere close to Bing Translator), and he looked up the questions he didn’t know how to ask as well as specific words within the questions which he said are handy sometimes. Mark mentioned that this was a tedious process, but that he had no alternative. Even if obvious, this started to exhibit how much he would value being able to speak in English.

Mark then made the call. He explained that he knew very little Spanish and started asking questions about the hotel. In general he was able to articulate the questions well, but he had trouble understanding the answers, especially when there were numbers involved. He had to ask for repeats multiple times and be a little ashamed about it. Mark complained about the phone signal making it even harder to understand. Moreover, he suggested that it would be really helpful to have the actual text of what was being said to facilitate the understanding and also to see the spelling of words to be able to look them up. In general, he did not want to use the translator in the middle of the conversation because that would make the other person wait, but he also mentioned that people are often (but not always) understanding and allow him to take his time seeing that he is struggling. He did use the computer twice while talking on the phone. First he looked up a word he wanted to say, and second he tried to find a Spanish word he heard. For the second case, he actually did an image search with what he understood for the word “castillo” which means castle, and he instantly got an answer visually. Mark mentioned that many people might not do it, but that looking up images is a very efficient way to get the meaning of a word instantly.

As part of the wrap up, Daniel explained what he had observed, and they discussed more in detail some of the important issues. Specifically, Mark emphasized the importance of being able to see the actual text which is an important detail that has implications in our design choices for the interface. Mark also talked about it being useful to have some time to actually assimilate what was being said and prepare the response which has a connection to the finding in the first inquiries in which participants showed acceptance for certain latency increases. Finally, when talking specifically about ways to solve the issue with a speech translation system, he said that online translators are not perfect and that longer and more complicated sentences might result in errors. He also noted that sometimes protocols flow naturally to ensure correct understanding, and that one option was to answer a question in the form of “You asked me X, and my answer is Y."

### Anecdotal Observations

Alex and John recently traveled to Peru to visit a friend Erin. Alex does not speak any Spanish and John knows only enough to communicate very basic words. Erin, on the other hand is fluent. When we all went out with Erin’s friend (who only speaks Spanish with very little English), Erin essentially translated for them. Our interaction would involve many pauses and waiting for the translation and interaction. My friend John, on the other hand required less translation help from Erin, only when stumbling on certain words or phrases. The takeaway is that different levels of language skills require different levels of the involvement of the translator.

Raj recently had an opportunity to serve as a translator for a very interesting demographic - his mother. Raj's parents were visiting from India and while his father is fluent in English and Hindi, his mother is able to converse only in Tamil, which is a language native to the southern part of India. When Raj's parents paid a visit to a friend of Raj called Mayank, who is from northern India, a very peculiar situation arose. Mayank's parent primarily speak Hindi which is native to that part of the country. Parents being parents, there was a very animated discussion between both set of parents with one side speaking Tamil with a splattering of English thrown in and the other side speaking Hindi. This was facilitated by Raj functioning as a translator since he knew all three languages in the mix. Interestingly, several parallel conversations can pop up which can make the task of a single human translator difficult. The value of solutions like automated speech translation (e.g. One World) become more apparent at these times.

## Task Analysis Questions

### 1. Who is going to use the system?

Our target audience is anyone who needs to communicate in an unfamiliar language. We focused on two specific audiences and took them as our starting point through contextual inquiries. The first type of audience consists of people conducting business with someone from a foreign country. Usually such business meetings happen through video conferencing. An interpreter is usually present to facilitate the communication. Our system could help remove the need for a human translator. The second type of audience consists of people who travel and need to communicate in a foreign language they are not fully familiar with. Our system could be used to allow the travel to make travel arrangements naturally without having to struggle with different languages.

### 2. What tasks do they now perform?

In business settings, there are wide variety of tasks that need communication with foreign parties. Cross-team communication within multinational companies, and partner meetings with international partners are among the many tasks that that require people with different native languages to talk to each other. In these situations, translators are often present at the meeting (virtually or not). As detailed in the contextual inquiry, the business parties usually understand that there is a translator present and pause once in a while to allow the translator to translate the sentence (or set of sentences). It is not as natural as normal conversation in the same language, but people understand and adjust accordingly.

In a travel setting, people might need to make arrangements, book a hotel, rent a car, find local rentals, etc. People who travel frequently often rely on a phrase book, or a translator app on their smart phone to help them maintain simple conversations in the other language. In some cases, both parties share some (though often limited) knowledge of the common language, and they communicate by using very simple expressions.

### 3. What tasks are desired?

The desired tasks are all those already performed by people in an unnatural and uncomfortable manner. Whenever there is a need to speak to someone in a different language (e.g business settings or travel settings), we would like people to be able to do it naturally in their own language and let the computers perform the translation for you. Additionally, by making it easy to communicate we'd like to enable people to become comfortable and open to tasks they currently do not do because of language barriers.

### 4. How are the tasks learned?

The task of communicating with a foreign language speaker through an interpreter, whether online or offline, or in a business setting or otherwise, simply comes from experience. The key point is the recognition that you are not simply communicating through an interpreter, but communicating with the interpreter as well. You need to ensure the interpreter understood your message and any hidden implications for your message to be delivered properly.

In terms of using our system, with the current technology, there will be errors on both the recognition side and the translation side. And as a result, the users will have to learn how to recognize that there is an error in the system as well as ways to correct it.

### 5. Where are the tasks performed?

As mentioned above, business meetings and travel settings are common scenarios for the use of a translator, and so conference rooms and offices are common locations for performing the tasks. With the system we propose, a translator would not need to be present at the site of communication. In the travel scenario, one might imagine being on the phone making a reservation to a hotel in a foreign country. He/she can use our system to simply speak in the native tongue and the system will speak in the foreign language on the other end. When the hotel receptionist responds in the foreign language, our system will translate that back to user.

One area to point out related to location is that background noise in the environment (and the quality of the microphone) has a dramatic effect on the accuracy of the speech recognition, but this is a technical challenge that can be overcome.

### 6. What’s the relationship between customer & data?

If you are speaking through an interpreter and have no access to a computing device or recording mechanism, there may be no data to speak of at all. In an interpreted business meeting, notes may be taken as the conversation proceeds, but it is difficult to have a word-for-word log of what was said, outside of recording the video or audio. Also, confidentiality in business settings is always a concern, and interpreters may need to get proper clearance. With an automated system capable of delivering a quality interpreting experience, this would no longer be a problem.

In terms of an automated system, data could be collected during the correction process. The idea is that if user corrects a recognition and/or translation, we can “save” those customizations in their system so the next time the same sentence is uttered, the translation could automatically set to the correct one.

### 7. What other tools does the customer have?

Once again, there are a wide variety of situations where interpretation may become necessary, and the tools available would depend on the circumstances. Video conferencing software is often used in the business setting for interaction between business partners. With video, the users have access to the gestures, expressions and motions of the other party. These are often crucial to the communication to provide intent. The parties often will have access to their laptop and/or notebook; when they encountered a word that they don’t understand, they can write it down and ask for clarification later. A professional translator also often writes down the conversation to ease translation.

In the travel setting, a customer may very well have access to a computer or smart device, which connected to the web, can leverage online translation services. The traveler can also use a phrase book and guide book which often have all the common phrases used. The use of photos / images to communicate was found to be useful for travelers as well.

### 8. How do customers communicate with each other?

The goal of our system is to enable communication with one another in their native language. The business setting video conferencing, instant message communication, email and face to face meeting can all use translator service (in person or virtual) to facilitate the communication. In the travel scenario, the user can communicate to a local travel agent, hotel receptionist, airport agent, other travelers, resident of the foreign country, and so on in a number of ways: phone, email, instant message chat and in person. Different means of communication will require the use of a different tool. For example, in a dinner setting in the travel scenario, taking out a laptop to help with the communication might not be the most appropriate – a smart phone app might perform better.

Another note of interest in the role of the translator in the communication. In some ways, the translator has to interpret the meaning of the phrase and utterances before he/she can translate the sentences. If the meaning is not clear, it is often the case that the translator asks clarification before performing the translation. In a machine translation setting, this is often not possible for the system to be aware of the ambiguity in the communication.

### 9. How often are the tasks performed?

The tasks are performed when there is a need for communication with people who speak different languages. You can think of millions of scenarios within travel, business, immigration, education, language learning, etc., and they will have specific recurrences. In the two scenarios that we focused on in our contextual inquiry, business meeting between external teams usually occur in a weekly basis. For travel, this may happen a few times a year depending on the person.

### 10. What are the time constraints on the tasks?

Efficiency is often important for communication and interpretation inherently adds latency. Interpretation concurrent to ongoing speech by trained professionals is an often expensive option and quality may suffer for two languages that are significantly different. If an automated system would make too many errors and it becomes too time consuming to recognize the error and correct them, then people will cease to use the system and resort to the old way of communication with their non-native tongue. In a business setting, with international partners, the addition of a translator increases the duration of the meeting and reduces productivity. In a travel setting, say making a reservation over the phone, the time dramatically increases whether speaking the foreign language, using machine assisted translation, or using human translator assistance. The delay could cause annoyances to the other party in both the business and travel scenarios.

### 11. What happens when things go wrong?

To state the obvious, if a message is miscommunicated, this could have negative effects, so as a participant it is crucial to understand the possibility of misinterpretation and reconfirm any sensitive material. If a translation system fails to consistently perform a “good enough” translation in a timely manner (see # 10), then the user will most likely give up on the system. In a business setting, this can be extremely frustrating and counterproductive to the business despite the cost saving of not hiring an interpreter. In a travel scenario, miscommunication causes uncomfortable situations (asking for repeat many times) and shame. It also forces the person to sometimes agree without fully understanding and without having a way to verify what was being said. People could get discouraged by this and stop making arrangement that require a foreign language. In the best case, users will begin to adapt to the system by speaking more clearly and using shorter sentences and simpler vocabulary.

## Task description and analysis

As mentioned before, communication systems have a wide variety of uses and they enable people to perform very diverse tasks. Nevertheless, we decided to focus on two specific but different types of audiences that could particularly benefit from a translation system and would enable us to learn and understand different perspectives on how the system could be used to improve their current tasks and enable new possibilities.

The first type of audience consists of business people needing to conduct work-related meetings with co-workers or partners who do not share the same native language. The context of the meetings can vary significantly, however. Local teams of a multinational company can have regular status meetings with a team outside the main location, executives can have high priority meetings with important partners, and entrepreneurs might need to present an idea to some foreign investors. In all cases, both sides of the conversation will need to express some concepts and ideas, and understand what the other side has to say. The first task we designed is specific to this type of audience and it consists on effectively having a meeting with a person or group of people whose native language you are not fluent with. More specifically the first task is the following:

***"Your team is close to signing a big contract with a manufacturing company in Taiwan. Today is a big day and you're very excited because this might mean a big promotion for you. You are expecting a call for a video-conference tonight with the VP of Communications on the Taiwanese company and he is part of the board in charge of making the decision. All you need to do is to politely salute, explain the main terms of the contract which you are very familiar with, and then answer his questions."***

The difficulty of this first task is easy because it consists of mostly going over a list of well-known terms. The person will be the one doing most of the talking and in a very granular manner by stating a set of sentences instead of doing a long speech or having a two-sided discussion.

The second type of audience we decided to target consists of regular people who try to communicate in a foreign language but do so very uncomfortably and are not effective. People not part of business settings will likely not have access to human interpreters unless they are very wealthy. In many situations those people know some basics of the foreign language and try to use that to communicate but are not able to do it effectively. This might lead simply to uncomfortable moments or even to frustration and in some cases it might result in avoiding contact with foreign people and cultures who do not share a common language. One specific but representative example of this audience is people who travel. While making travel arrangements or while actually staying somewhere outside your home country, there is often the need to communicate with people who cannot speak your language. The second task is the following:

***"You are planning to go to the world cup in Brazil but everything is very expensive. In Rio hotels are mostly run out or extremely pricey and so you want to get in touch with someone local to rent a room. One of your friends gives you the number of a Brazilian girl who might have a room available and so you want to call her and ask her for details. Specifically, you want to know the size of the room, the location, whether there are landmarks nearby, and of course the price. Also you would like to be friendly and get to know a little about her in case you end up choosing her location."***

The difficulty of the second task is moderate because now there is more focus on the conversation and there are two sides to it. The person needs to both effectively state the questions and also understand the answers clearly. There is more room for errors that in the first task. Nonetheless, the set of questions and possible answers are still granular which reduces the complexity.

Finally, we think that having an effective translation-based communication system could remove language barriers enable people to become more comfortable speaking and bonding with foreigners. If language is no longer an obstacle, one could think of people making and maintaining friendships with people far from their hometown who belong to different cultures and do not speak your language. One could also think of starting and keeping romantic relationships with foreign people while becoming familiar with your significant other's language. We decided, thus, to design a task that is not as common today but would be facilitated with the existence of our proposed system. The third task is the following:

***"You are getting tired of being single and you meet a girl through an online chat website. So far, the girl is great. She is nice and funny and she is from Japan which makes her much more interesting. The only problem is that she's not very fluent in English, and thus so far, you've been mostly talking to her through instant messaging with the help of both an online translator and your Japanese school friend. Nevertheless, it's now time to talk to her and know her voice, and you prefer to speak about serious topics using your voice (and without your friend involved). You prepared a speech and now you're ready to ask her how she feels about you and suggest that it's perhaps time to take the relationship to the next level and go visit her in Kyoto. "***

The difficulty of the last task is hard because there is no longer a limited set of granular and well-defined statements and answers. In this case, the person needs to effectively give a speech with multiple concepts and ideas, and then understand and respond to what the other side might say which is not necessarily predictable.

## Storyboarding

### Login Screens (shared between all designs)



Figure : Login Screens



Figure : Language settings page



Figure : Call connecting screen

### Design A



Figure : Main screen (Design A)



Figure : Error correction screen

### Design B



Figure : Main screen (Design B)

### Design C



Figure : Main screen (Design C)



Figure



Figure 9: Long speeches broken into sentences

## Selected Interface Design

Our main goal is to provide users with a user interface that excelled at providing the ability to communicate naturally and visualize the translation process to enable quick feedback and corrections. We quickly realized that this main translation screen was what the major design to focus upon, with the other supporting screens being more mainstream features for which good designs are already available in existing applications. The application we are designing will function similar to present day video conferencing apps with the option to translate between the parties on a call.

### Design Reasoning

Our final design is a based on a combination of the major features we explored while brainstorming. Several key points were recognized in early discussions as factors that could influence the final design and are discussed below.

#### Devices and input methods

The level of control and feedback that we wanted to provide called for large screen sizes. We observed that the interactions that we aim to improve normally happen at work places or homes where a PC is usually available. For our initial version we targeted desktop in order to add to existing video conferencing software because business settings were one of our primary targets. As a consequence, our main focus has been designing for a desktop application where the user interactions would be via mouse and keyboard. We believe this design can be improved with some slight adjustments for touch friendliness and enabling a mobile companion.

#### Correcting speech recognition

One central focus of the interface is the ability to quickly and effectively correct the recognized speech before it is translated and sent to the other party.

The first part of this problem is to detect if the speech recognition worked correctly. It becomes necessary to obtain feedback from the user to make this judgment. There are three options that were considered:

* Ask the user always for confirmation (may not allow a fluid conversation)
* Ask the user for confirmation only when the confidence is low (may be inconsistent, and possibly send incorrect messages sometimes)
* Always assume the translation is correct unless the user indicates otherwise

The decision we arrived at was to provide a timeout period before which the user can flag a translation as containing an error and would then need to correct the error and click send to send it out. If the user does not provide any indication before the counter expires then the message is sent out automatically. To speed up the process the user can hit send to force a message to be sent even before the countdown expires. This design has the best balance between being non-intrusive and flexible.

Once errors are identified then it becomes necessary to allow the user to make corrections. We brainstormed the following different approaches.

1. Use a graph representation with a default selection, and allow the user to propose an alternate traversal.
2. Display the best match and also show other possible matches for words which could have been misinterpreted. The user can drag and drop words to form the final sentence.
3. Show the best translation and allow the user to select and make corrections before the communication is sent out.
4. Have the user speak again.

The final decision was to go with option 3. Options 1 and 2 had the tendency to put too much burden on the user while the conversation is in progress. We rejected those options since we wanted the main interaction to be free of distractions as much as possible. Option 4 was briefly considered as a better way of allowing the user to correct mistakes by re phrasing the input. But persistent speech recognition errors can make repeated speech input frustrating and is not ideal in business situations where the terminology may be very specific and cannot be changed. For these reasons, option 3 was picked as it allowed the best balance between fine-tuned control and smoother user experience.

#### Audio vs. text translation

All our initial designs were in agreement that our user interface will allow real-time audio and video feeds to both people on a call. The point of contention was how to fit in the translation. One option was to play a digitized version of the translation once the speaker had finished speaking. The other option was to not transmit the original voice at all but to rather just play the digitized translation. We rejected the first option since it could introduce delays in the conversation, and the second option was also rejected since we understood during the contextual inquiry that the tone of the speaker’s voice conveys some nuances that cannot be replicated by a digitized translation. The final design we settled upon was to just play the original voice along with displaying the translated text and our third contextual inquiry was very helpful in making this decision. For the text display, our original exploration centered on a closed-captions like system where the text translation is superimposed over the speaker’s video image. We settled against this because of it looking unnatural due to sync issues. Instead we're adding a transcript/log of the conversation at the bottom.

#### Back translations

Back translation is the procedure in which a translator interprets a document previously translated into another language back into the original language. Back translation is an effective tool used to identify sensitive issues across cultures and is used by businesses that wish to ensure absolute accuracy of the translation. Some of our early prototypes were designed to show the back translation to the user in order to make it easy to identify speech translation errors. This idea was dropped in subsequent iterations because it had the potential to burden the user with too much text to read and parse before being able to proceed with conversation. Additionally, the usefulness of such back translation is questionable due to the inability to identify accurately whether the error was in the original translation or in the back translation.

### Interface description and Functionality Summary

The main functionality resides across three screens. The following section describes each screen in detail and provides an overview of the functionality exposed by each of them.

#### Home screen

The home screen is what the user sees as soon as they log into the application. This screen permits the user to manage their profile, language settings and contacts list. Figure 1 shows the design for this screen. This is also the place where new conversations can be initiated by selecting a contact from the list and initiating a call using the relevant buttons. Each contact on the contacts list is also annotated with the language that they have configured on their application. This setting is configured by each user on their home screen and it is shared with people that they add to their contacts list. This way the app can detect if translation is required for a call based on where the call initiator and recipient have different primary languages configured.

#### Language settings screen

The user can also reach the detailed language settings page shown in Figure 2 by clicking on the language shown in his profile. This dialog will let the user configure the languages that he is conversant in and also pick a default language that will be reported to his contacts and used for calls he makes. Here the user can setup advanced translation options, namely “My Pronunciations” and “My Dictionaries”. "My Pronunciations" is a feature where a user could create custom mappings between spoken words and how they are recognized as text. This is a feature that will be must-have for business clients based on our observations during the contextual inquire

Also "My Dictionaries" would be for defining custom dictionaries of mappings of words in different languages.  When a given dictionary is enabled, any speech-recognized word that is in the dictionary would be translated using the defined mapping if an entry for the target language is in the dictionary (Here we assume that the translation engine could support such a scheme).  This could be useful for domain specific terminology which could be a challenge to translate as we observed during our contextual inquiry at the Nintendo business meeting. This concept can be expanded into a scenario where parties of an interpreted conversation would share their terminology dictionaries with each other prior to a meeting using an import/export feature that would make this trivial to share such dictionaries with your contacts.

#### Connecting screen

This is the screen that is shown to the user while a call is in the process of being connected. This screen in illustrated in Figure 3. There are two versions of this, one shown to the person initiating the call and the other shown to the recipient of the call. The call initiator will be shown a simple screen that identifies the recipient as well as the language that the recipient has configured in their profile. The recipient sees a similar screen with the details of the caller and their language but also making it clear that it is a translated call because it imposes additional requirements. The recipient will have the option to change the language they want to use. If the recipient picks the same language that the caller has picked then the translation is no longer required and the caller will be notified of this.

#### Conversation screen

This is the main user interface that is shown to both parties in a conversation. The final design that we selected is shown in Figure 4. The top half consists of a status bar that lets the user tweak call settings and a translation window where the speech recognized, text version of the caller voice input is shown in the caller’s language along with the translated text in the language being used by the recipient. This section is interactive and allows the user to click and edit the text in case of speech recognition errors. In some cases clicking on an ambiguous word will result in the user being presented with a drop down list of alternate choice to speed up editing. In order to facilitate error detection and resolution, long speeches are broken into multiple granular sentences.

The bottom half of the screen shows the video feeds and the conversation log. The log serves as a record of the conversation thus far. The log can also be used by both parties to call attention to sections that need to be clarified or revisited, perhaps due to questionable translation or speech recognition errors that slipped though. When a piece of text is clicked in the conversation log, a question mark will pop up next to it. Clicking the question mark will highlight the text on both users’ screens thereby drawing their attention to the fact that maybe they need to rephrase of manually correct the specified section.

## Scenarios

### First scenario

The first scenario consists of using our system to perform the first task of conducting a business meeting for explaining the terms of a contract to a VP in a Taiwanese company.

As mentioned in the task, the VP in Taiwan in charge of making the call and so the first step in this scenario is to simply accept the call. Once there is an incoming call, you will see a screen like the one in Figure 3, and you will simply press the "Start call" button. The text of the screen implies that translation will be involved in the communication, and you are already aware of that because of the nature of the call.

After accepting the call you see a screen like the one in Figure 4 which is divided in two sections for each side of the conversation. You see the VP on the screen and you say "*Hello*." After saying that the text "Hello" shows up on the top of the screen. This top section shows the detected text and its translated counterpart in mandarin. You see a timer counting down and given that the text is correct you just wait for a couple of seconds and then the message is sent. You then see and hear the VP saying something, and in two seconds the text "Good evening" shows up on the bottom section in his side of the screen. This bottom section contains the log of the conversation in your language.

After saying hello, you then say "*I will go now go through the terms of the contract,*" you again see the detected text and the message being sent, and after that you see the VP on the screen simply nodding.

You then go through each of the terms of the contract using the system in the same manner and since you see the detections are always correct, you start hitting the "Send" button instead of waiting for a few seconds. Finally, you ask if there are any questions, and again you see the VP talking and after a few seconds, the text "No, everything looks perfect" shows up in your screen.

### Second scenario

The second scenario tries to accomplish the task of finding details about a place to rent in Brazil for the world cup.

In this case you are initiating the call. You received the contact from one of your friends and in her profile you see that she speaks Portuguese as expected. You click the "Call with translation" button, you see the intermediate "Calling.." screen (Figure 3) and after a few seconds, the girl responds. You see her on the screen talking, and after a couple of seconds the text "Hi, how can I help you?" shows up on the screen. You then say "*Hey, I'm the friend of Susan,*" but on the screen you see the text "Hey, I'm the fan of sushi." Before the message is sent, you click on the word "fan" and a drop-down menu with the words "fiend" and "friend" shows up so you select "friend." You do the same for "sushi," and since there are no suggestions you type the word "Susan." The timer starts counting again, and since the sentence looks right, you let it go. Susan's friend then says something starting with "*Olá ..*" and you soon see in the screen "Hello, how are you?"

After done saying hi, you start asking questions, inspecting them before they go and then seeing the answers from her. She is very friendly you get answer for your questions and everything goes smoothly until the end. You finally say "*goodbye*," and let message go. She then says something like "*bom comec.*." and you see the message "Good to start you." You don't really know what this means so you select the question mark next to the sentence to indicate it did not make sense. She then speaks again, and this time it sounds like "*bom conhec..*" and you see in the screen "Good to meet you." You respond the same and hang up the phone but having just found a great place to stay in Brazil.

### Third scenario

The third scenario consists of fulfilling the task of talking with a Japanese girl whom you intend to become serious with and perhaps visit her personally in Kyoto.

In this situation you again need to initiate the call. You have your speech on paper so you bring the sheet with you before making the call. Also you have used the system before and you have noticed that it sometimes has trouble understanding your sentence "How ya doing?" so you go to "Language settings" click on "Manage my pronunciations" and then say your sentence and type it manually. After you are done, you select your Japanese friend among your contacts and then click on "Call with translation." She then appears on the screen and very enthusiastically says "*Konnichiwa*" and then the text "Hello" appears on the screen. You say "*Hello, how ya doing. It's good to finally talk to you.*" The text was detected correctly and so you let it go. After she says that she is doing well, you decide to start with your speech.

"*It's been a pleasure talking to you for the past couple of months. You're interesting, you're funny, you're sweet and I truly enjoy every conversation we have. For these reasons, and despite our distance and different cultures...*"

As you speak you start seeing a list of separate sentences show up in the screen. You see "It's been a pleasure talking to you for the past couple of months", and since it's correct you let it go, right below you see the rest of the sentences you have said so far and they start coming up after the top one is sent. You just let them be sent automatically because you don't notice any errors. You then see that the sentence "For these reasons, and despite our distance and difference cultures.." is highlighted indicating that your friend did not understand what you meant. You therefore stop what you were saying and start again from there. This time you rephrase it as "Even though we're far and belong to different cultures..." and this time she understand and smiles at you.

You finish going through the rest of your speech letting every sentence go without finding any more problems. She then responds saying something and then the text on the screen says "I'm very touched, and I really like you too. I'd love to have you visit me over the summer."