[*Translating and the Computer 26*, November 2004 [London: Aslib, 2004]

Development and Fielding of the Phraselator® Phrase Translation System

Ace Sarich, PE
VoxTec Division of Marine Acoustics, Inc.

Lieutenant Colonel Jim Bass, PhD
Unites States Army

Horacio Franco, PhD
SRI International

Introduction

Over the past five years, VoxTec, a division of Marine Acoustics, Inc., teaming with SRI International, has developed and refined and fielded the Phraselator®, a handheld one-way voice-to-voice phrase-based language translator. The Defense Advanced Research Projects Agency (DARPA) and a DARPA Small Business Innovative Research (SBIR) grant have funded the research and development and initial production. VoxTec has evolved the concept from a PC based system to a handheld device. After the 9/11 terrorist attack on the World Trade Center, Phraselator development was accelerated. Initial deployment of the prototype Phraselators to U. S. Military forces operating in Afghanistan in support of Operation Enduring Freedom began in March, 2002. Lessons learned from Afghanistan were folded into the redesigned Phraselator P2. To date, over 2,000 Phraselators P2s have been manufactured and delivered to users around the world.

The Phraselator is basically a user independent voice actuated phrase matcher. When a user speaks a known phrase into the microphone, the automatic speech recognizer (ASR) matches the input phrase with a known phrase in the phrase database. An output translated recording in the target language is then played through a speaker. With the Phraselator, the user is able to provide information, give directions or orders, or ask

questions that have an easily conveyed response. While simple in concept, developing a reliable, robust and flexible system for military users has been the heart of this effort.

The Phraselator is the hardware component of the phrase translation system (PTS) consists of the hardware, application and automatic speech recognition (ASR) software, and module management system (MMS). The current Phraselator P2 is a ruggedized, weather resistant device with superior audio input and output.

A typical phrase module consists of 100 to 1000 phrases translated and recorded into one or more target languages. Phrases are grouped into categories for purposes of improving accuracy for large modules and for navigating phrase databases. The phrases convey the translated meaning of the input phrase and are not necessarily an exact translation. In fact, a short input phrase may be matched with a long output translated phrase; e.g.:

> Input: This is a computer translator.
>
> Output (translated): This is a computer translator. It translates my words into your words. It does not understand your language.

The translated output is a recording of a human speaker. As such, the translation can be male or female, adult or child; be in the correct language and dialect for a particular region; have the appropriate intonation and emphasis; and accurately translate idiomatic or context appropriate phrases.

Phrases are initially recorded as WAV files and then converted into MP3 format for Phraselator application. MP3 files require about one tenth the size of a WAV file. In MP3 format, 1,000 recorded phrases require about two to three megabytes of computer storage space. The Phraselator stores the phrases on removable Secure Digital (SD) cards. These cards range in capacity from 64MB to 1+GB.

VoxTec's module management system (MMS) consists of: Module Builder™, a toolkit for rapidly building custom phrase modules; a content database of over 15,000 phrases in over 50 different languages; and Lingua Port which allows users to download modules from the database for auto installation onto their Phraselator.

This paper will provide an overview of the Phraselator PTS history and development; highlight the capabilities and limitations of phrase-based translation; present the concept of operations, development considerations and technology behind the Phraselator; highlight current operational use by military forces in support of Operation Enduring Freedom and Iraqi Freedom; and present future Phraselator development and fielding in both military and commercial applications.

Approaches to Speech Machine Translation

Ideally, we all would like a machine that you can speak anything you want into it, and then it automatically speaks translation in another language. While this approach may

not be practical for reasons discussed below, an alternative practical, albeit limited, approach is the phrase translation approach.

Machine Translation Approach

Over the years text-to-text machine translation (MT) has been developed and are available. While not perfect, theses programs have been found useful for both the casual user and as an aid for professional translators. MT performance tends to degrade with sentence complexity, technical jargon, and idiomatic expressions. For true speech-to-speech machine translation (SMT) obtaining an accurate translation is compounded. For SMT the general steps are:

- Human speech in the source language (SL) input via a microphone
- Analog to digital conversion of the audio input
- Acoustical model phonetically matches words with the sound of what you said.
- Language model guesses the most likely word based on context and frequency.
- Speech engine combines the acoustical model and language model information to guess the words and produce the SL text output.
- MT engine translates SL text to target language (TL) text.
- Text-to-speech (TTS) engine produces synthesized speech out.

This process is reversed for translation from the TL back to the SL. While this approach is a long-range, desirous approach, it has several limitations:

- Works for simple, non-complex phrases; the more complicated the phrases, the less effective the translation,
- Speech-to-text (STT) and TTS engines support a limited number of languages and extension to new languages is difficult.
- MT engines support a limited number of generally commercially viable languages.
- Requires high quality speech audio input.
- User needs to train the speech engine to his or her voice
- Requires powerful processor and large amount of memory.

While there are prototype systems using this approach, they are generally considered to not be reliable or robust enough for practical use at this time for true two-way voice machine translation and have the limitations stated above.

Phrase Translation Approach

A more reliable and robust approach to SMT is a phrase-based translation system. This is, in reality, not a translator; but a speech actuated phrase matcher. When a known phrase in the SL is spoken into the device it is matched with a recorded phrase in the TL and played through a speaker. For this approach, there two methods for matching the SL audio input with the output recording.

The first is a system using analog pattern matching. This approach requires the user to speak and record every phrase in the phrase database prior to using the device. When the

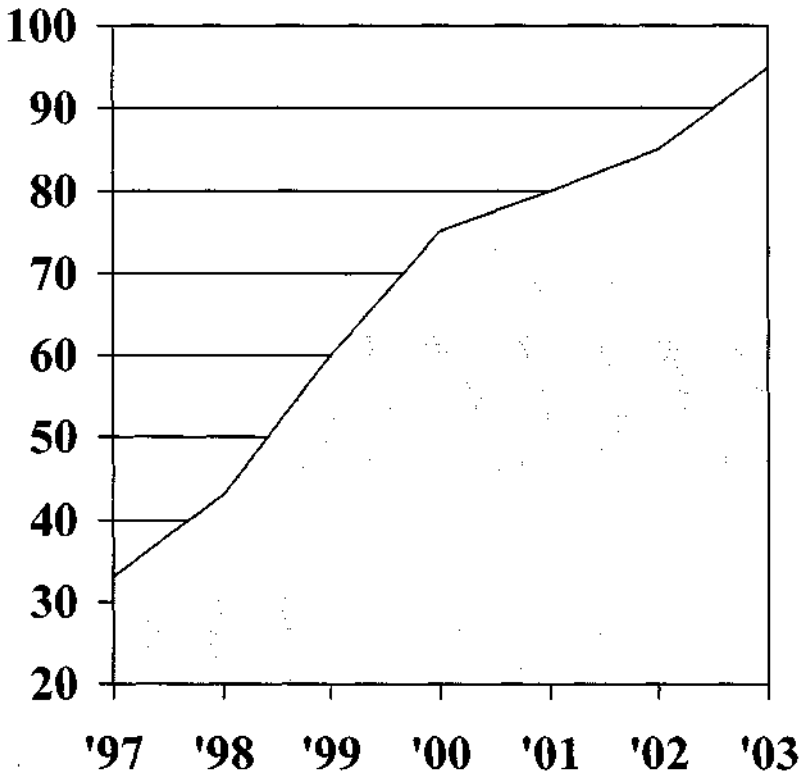
using the device, the user speaks the phrase and it matches the analog voice pattern with one of the recorded patterns and then plays the appropriate translation recording. This is simple and requires little processing power. It is, however user dependent and requires the user to say the phrase exactly as he recorded it.

The second approach, used by the Phraselator, is to use the acoustical model and language model from the automatic speech recognizer (ASR) engine as described above and match the phonetic content of the input with the phonetic content of one of the phrases in the phrase list. This approach is generally user independent, allows for inexact input, and can be modified easily. It does, however, require more processing power.

Performance Progress

Over the past years, there have significant improvements in the overall performance of the Phraselator PTS as compared to its predecessors and can be evaluated in a number of areas. The below graph highlights the performance progress.

- Reliability and robustness: No hard crashes or lockups; fail soft and recover; absolute control of command and phrase modes; operation in noisy or vocal environment
- Phrase translation accuracy: 95% phrase translation accuracy; rapid translation; alternate phrasing capability; ability to translate inexact phrasing. Early versions of the PTS software has translation accuracies that were often less than 80%. This was particularly true in noisy environments.
- Ease of use and training: New user trained in less than one hour; minimal number of commands and functions. Early versions of the PTS were not particularly intuitive of easy to use by the untrained person.



- Speaker independence: capable of being used by different users with no or little registration. The PC based MLT ran Dragon Dictate and Dragon Naturally

Speaking for speech recognition. The user needed to train the system for his voice. The current Phraselator is user independent and can be used by male or female with minimal adaptation (two to three phrases).

- Rapid module build capability: module build in less than two weeks; automated phrase addition in the field. Early module were hand crafted. It could take up to two months to build a module. Words not in the dictionary needed to be built by an engineer. Using Module Builder, a module can now be built in a day.
- Voice/audio interface: Closed notebook operation; all commands or prompts via audio or voice. Using earlier PC versions of the PTS required the user to use the graphic user interface when using it. The Phraselator now permits eyes free and hands free operation.
- Ease of software installation and setup: automated installation of speech recognition and application software and modules in less than 15 minutes. Since early versions of the PTS ran the commercial Dragon speech-to-text software on a PC, loading the software and calibrating it for individual use took time. Phraselator software and modules load in less than a minute.
- Software size, hardware requirements and portability: Can software be ported to systems with limited processing power and storage? The initial application software, speech-to-text software, and phrase modules could require 500MB of storage on a PC. The same size module on the handheld Phraselator requires less than 30MB.

Phrase Translation System (PTS) Components

The Phraselator is basically a user independent voice actuated phrase matcher. When a user speaks a known phrase into the microphone, the automatic speech recognizer (ASR) matches the input phrase with a known phrase in the phrase database. An output translated recording in the target language is then played through a speaker. With the Phraselator, the user is able to provide information, give directions or orders, or ask questions that have an easily conveyed response. While simple in concept, the key is the development of a reliable, robust, integrated and flexible system. Components of the Phraselator are discussed below.

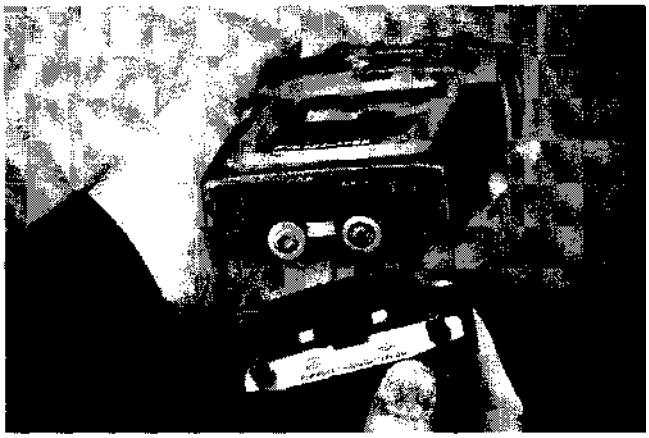
Phraselator P2 Hardware

Leveraging off of the success of the Phraselator model 1100, the P2 was developed. It is in essence, a ruggedized PDA design optimized for audio performance and field use. It consists of a base unit which houses the display and core

electronics, and an audio plug with microphone, speaker, and audio circuitry. The general specifications are:

- WinCE.NET 4.2 OS
- Intel® 400 MHz XScale PXA 255 Processor
- 64 MB Flash RAM on-board (256 MB max)
- SD/MMC slot
- Class D digital amplified two-watt speaker
- Daylight and night readable TFT color touch screen
- Battery charging/power management
- Long-lasting lithium-polymer battery
- AA battery capable
- Audio in and out jacks
- Standard mini USB connector

Design Considerations



Since the P2 was developed with military field use in mind there are a number of design considerations which may not be applicable to a device for commercial or consumer use.

Audio Performance

The quality of the audio input will have a major impact on recognition accuracy and translation speed. Particular attention must be paid to the microphone type and quality, microphone directionality, microphone enclosure, audio circuitry, button noise, and electronic noise.

Size and Weight vs Ruggedness and Weather Resistance

For field use the device needs to withstand shock and vibration and environmental effects such as rain, dust, and heat. The size and weight of the device are directly coupled to these.

Battery Life

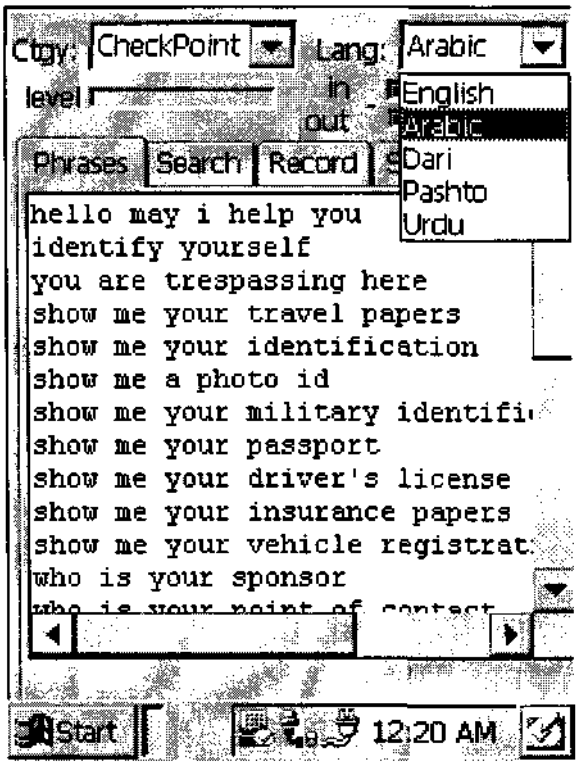
For military operations battery life is a major consideration. In the Phraselator design, particular attention was paid to maximizing the energy capacity of the battery, selecting a low power processor, and optimizing power management. Also, the Phraselator accepts AA batteries.

Modularity

As mentioned, the P2 has a rugged base unit (RBU) and an audio plug. The modular design was selected to allow for a future upgrade path for microphone input and speaker design; and to allow other plug in devices such as GPS devices, communications plugs, and card readers.

Application Software

When the user presses the push-to-talk (PTT) button and says one of the phrases in the phrase database, the client application receives voice commands and spoken phrases and plays back the appropriate translation after the user selects from a list of best-guesses. The application is configured by loading the language and phrase module(s) from a remote connection (either via serial port or modem) onto a removable SD card. The client application program itself can also be updated in the field via a telephone connection to the device. The application runs in less than 8MB.



When the application is loaded, the user may select from one of the modules available on the SD card. Each phrase module consists of: (1) the recognition grammars, vocabularies, and pronunciation dictionaries to configure the recognizer for a complete task, including the command-and-control phrases, (2) the output prompt waveforms for playback in any number of languages, and (3) any additional information or data to make the client application work properly in the field for a particular application.

For improved recognition accuracy or translation speed for large modules the user may select a category. Typically, modules of up to 500 phrases translate at near real time. As the number of phrases approaches 1,000 there is a noticeable delay in translation speed on the order of from one to two seconds, particularly if the audio input is poor or there is high vocal background noise. Also, as the phrase database increases in size, there is a more likely possibility that the Phraselator will have do discriminate between similar sounding phrases or words. When a category is selected, only the phrases in that category will be recognized, plus the phrases in the "basic phrases" category.

Other features that enhance the overall usability and functionality include;

- Audio or text verification of phrase prior to playing the recording.
- Ability to switch languages via vocal commands or touchscreen.
- Ability to switch modules via vocal commands or touchscreen.
- Record capability that allows the user to record a response in the TL so that it can be played and translated at a later time.
- Search on word feature. The user speaks a word into the Phraselator and all phrases with that particular word are displayed. The user may then play one of the phrases.

ASR

SRI's DynaSpeak Speech Recognition Engine features run-time configurable grammars and vocabularies, and state-of-the-art speaker-independent continuous-speech performance. DynaSpeak was scaled down and ported to the Phraselator.

The DynaSpeak ASR engine uses up to 10 MB of DRAM to allow for higher-accuracy acoustic models and a large dictionary of over 10,000 words. Furthermore, DynaSpeak contains adaptation algorithms that improve recognition performance for speakers using the system for more than several minutes at a time. The engine recognizes American English input speech with grammars containing up to 5,000 unique phrases.

Content or Modules

A typical phrase module consists of 100 to 1000 phrases translated and recorded into one or more target languages. Phrases are grouped into categories for purposes of improving accuracy for large modules and for navigating phrase databases. The phrases convey the translated meaning of the input phrase and are not necessarily an exact translation. In fact, a short input phrase may be matched with a long output translated phrase.

The modules are stored on SD cards in MP3 compressed audio format. 1,000 recorded phrases requires about 5MB of SD card storage. Thus, a typical 128MB SD card can store 20,000 or more recordings.

Modules may be obtained in one of three ways: Existing modules may be obtained from VoxTec directly or via the web portal; VoxTec can build custom modules; or the user can build his own module using Module Builder software.

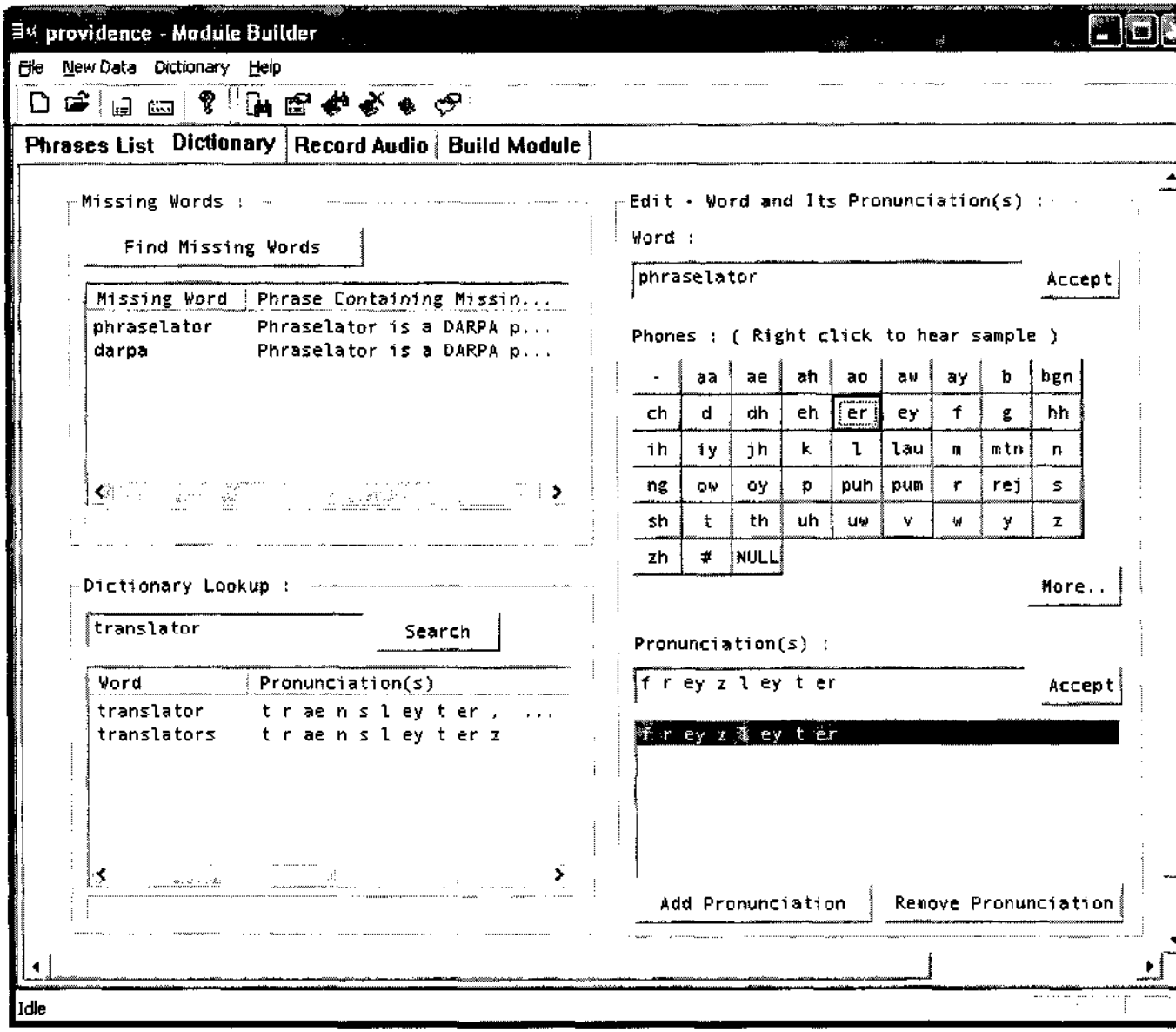
Modules may be developed by subject matter experts (SME) in conjunction with VoxTec personnel. VoxTec works with the SMEs to develop phrases and then builds the modules with translations in the desired languages. VoxTec uses contract language translation support to build the modules. Alternatively, using the Module Builder software, the users can build their own modules.

VoxTec's module management system (MMS) consists of: Module Builder™, a toolkit for rapidly building custom phrase modules; a content database of over 15,000 phrases in over 50 different languages; and Lingua Port which allows users to download modules from the database to a PC for auto installation onto their Phraselator.

Module Builder runs on a Win/Intel computer and greatly simplifies the process of building a custom module. There are two versions of Module Builder: a version for professional users, and a version for use in the field. Using Module Builder, the basic steps are:

- Develop phrase list
- Type or import list into Module Builder
- Build words not in the ASR dictionary using a user friendly tool (see figure below).

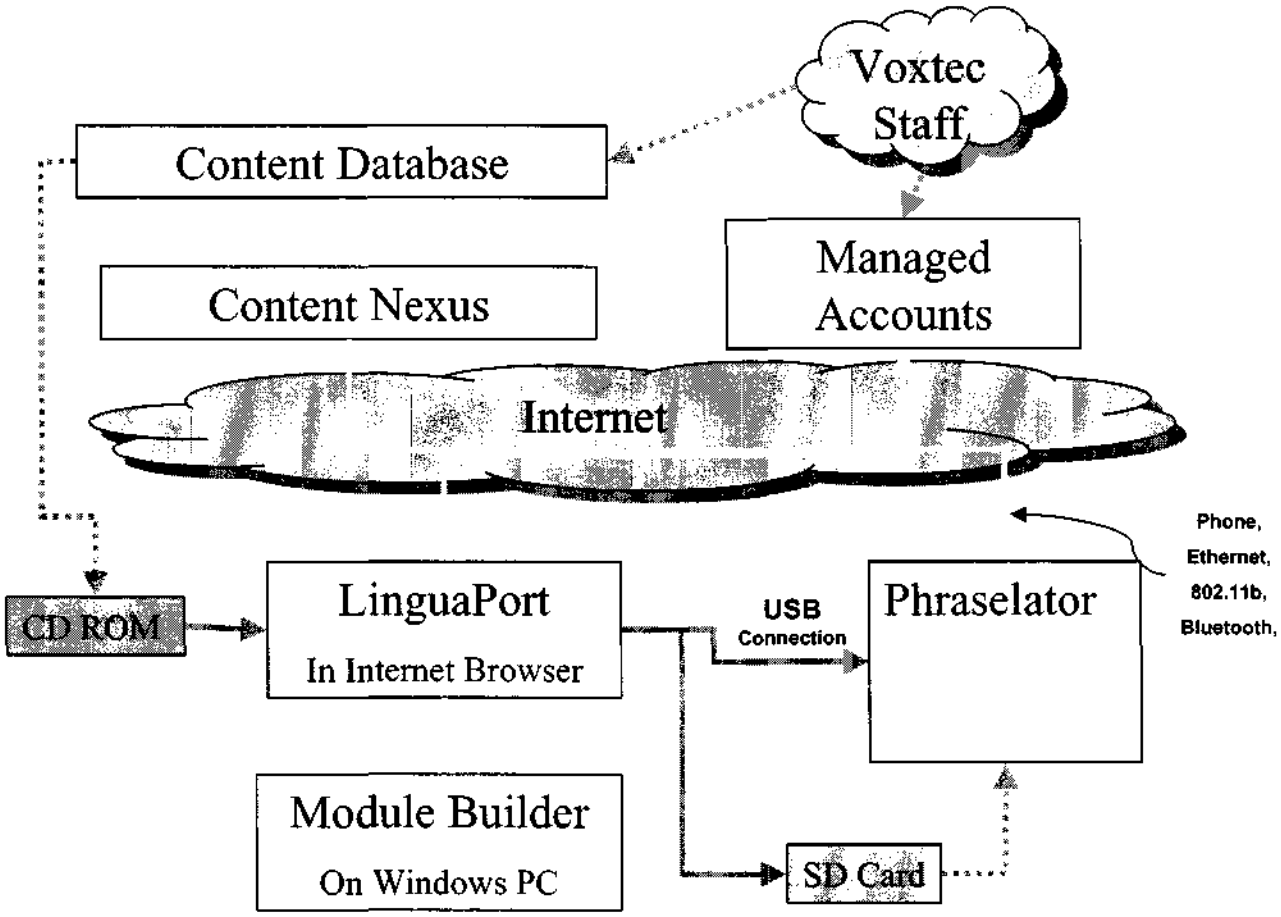
- Do voice recordings for English and target languages.
- Compile module.
- Transfer module to SD card.



When VoxTec builds a module it is very important that the translation conveys the correct meaning in the particular domain with the appropriate voice inflections. After the phrase list is developed, it is provided to a language translation service for text translation into the desired languages. Often it is necessary to specify the nature of the module, specific dialects or domains such as medical, maritime law enforcement; male to male, male to female, elder to junior translations; aggressive or passive translations.

When doing the voice over recordings of the translations, the speaker has to act out the phrases as he or she records them. After the recordings are completed, they may be

played back to another speaker of that language to verify the meaning of the translation in context.



Training and Support

As the Phraselator PTS transitions from a research and development program to a fielded system, necessary training and support must be provided. VoxTec has developed training materials and provided training around the world.

Development and Fielding

There is a large potential market for the Phraselator: business and tourist travelers, U. S. Military forces overseas, flight attendants, law enforcement, humanitarian and medical assistance, and others.

Users

Phraselator users may be broadly classified as *casual, professional,* or *dedicated.*

Casual

A tourist or traveler would use the Phraselator much like a traveler's phrasebook available from Fodors or Berlitz. For this user the Phraselator needs to be affordable, small, reliable, and easy to operate. The envisioned Phraselator would be packaged as a consumer item with about 2000 phrases translated into one to four languages. The user would be able add a limited number of additional phrases on the fly via audio prompts. Additional modules or translations would be available on flash cards or downloads.

Professional

Law enforcement, military, medical, travel industry, humanitarian and other professionals need a customized PTS. These users would need mission or job related customized modules in multiple languages. Also they would need the flexibility to change or add to the modules and add additional languages. For the professional user, the Phraselator hardware would be basically the same and VoxTec would market a custom module build and translation service. Additionally, VoxTec would provide on-site training and module development services.

Dedicated

Professional users who want total control of the module development and translations would use the basic Phraselator hardware and software. For this user the PTS module build and database management software toolkit would be marketed separately.

Market Segments

The market can be divided into the following segments.

Travel Industry

The Phraselator provides a means of asking questions at check-in, providing general information, and providing emergency to non-English speaking travelers. This is a medium size specialized market. Potential travel industry users include:

- Aviation flight crews can use the Phraselator for providing safety and emergency information to passengers in a large number of languages.
- Airline counter agents can use the Phraselator for asking check-in questions, for aiding passengers, and for providing general information.

Homeland Security

This is a medium sized market consisting of both professional and dedicated users. The initial potential market is in excess of 30,000 users that include:

- Police and fire departments
- Fire departments
- Customs inspectors and border patrols
- Coast Guard inspections and safety

Humanitarian Assistance

This is a small, specialized market. Users include humanitarian assistance teams, and medical teams.

Medical and Dental offices

This market is potentially large. With populations shifts in the U. S. medical and dental professionals have to diagnose and treat non-English speaking patients.

Military

Military users make up another small, specialized market. Typical applications include peacekeeping operations, foreign military training, and ship boardings and inspections.

Tourist

By far the largest potential market is the international tourist and business traveler going. Global spending on travel and tourism has more than doubled over the last decade as the standard of living for most people in the world has risen and more countries have become accessible to tourists.

Global spending on tourism exceeds $500 billion. Each year over 20 million Americans travel abroad spending in excess of $50 billion; and over 50 million visitors come to the U. S. spending over $100 million.

Market Size and Trends

While the potential market size is large it is difficult to estimate. Preliminary surveys indicate broad and considerable interest in such a product such as the Phraselator. Below is a preliminary estimate of the market size.

Estimated PotentialMarket Size

	Size	% users	Total users
Business and Professional			
Flight Attendants	132,000	5%	6,600
Airline Check-in counters	166,000	5%	8,300
Government and Law Enforcement			
Police departments	704,000	5%	35,200
Fire departments	293,000	2%	5,860
Customs inspectors, Immigration	5,000	20%	1,000
Coast Guard inspections and safety	2,000	50%	1,000
Humanitarian Assistance and Medical			
Disaster relief	5,000	20%	1,000
Medical diagnostics	2,000	50%	1,000
Medical and Dental			
Physicians, PAs and nurses	2,500,000	1%	25,000
Dentists and hygienists	300,000	1%	3,000
Military			
Peacekeeping operations	5,000	10%	500
Training	2,000	25%	500
Ship boardings and inspections	100	50%	50
Consumer			
Tourist	20,000,000	1%	200,000
Language practice and training	2,000,000	1%	20,000
TOTAL ESTIMATED MARKET:	26,116,100		309,010

The Future

Everyone wants the true two-way voice machine translation system. This may be a long time in coming. If not two-way, then what? A more pragmatic approach that will yield incremental near term results is to take small steps:

- 1 ¼ way: User asks a question with an expected constrained exact response:
  - Source Language (SL): How old are you? Say the number.
  - Target Language (TL): twenty five
- 1 ½ way: User asks a question with an expected constrained response:
  - SL: How old are you?
  - TL: I will be twenty five in November.
- 1 ¾ way: Domain constrained two-way translation:
  - SL: How old are you?
  - TL: I was born November 19, 1979.
- 2 way: Unconstrained two-way:
  - SL: How old are you?
  - TL: Whatcha want to know for? Did you know you have a bug in your hair? I'll be a quarter century old Friday.

True two-way voice machine translation may be "too hard." What is certain is that it will take money and time to make it happen. DARPA has been a lead agency in this research. VoxTec and SRI have been privileged to be part of this development effort.