
**Skynet: Speech-enabled Conversational Agent to Support Second
Language Acquisition**

by Seth Horrigan

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences,
University of California at Berkeley, in partial satisfaction of the requirements for the
degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Professor John Canny
Research Advisor

(Date)

* * * * *

Professor Coye Cheshire
Second Reader

(Date)

Contents

1	Introduction.....	1
2	Prior Work.....	2
2.1	Virtual Learning Environments.....	3
2.2	Automated Tutors.....	4
2.3	Intelligent Agents.....	7
2.4	Learning Games.....	8
3	Motivation.....	11
4	Design.....	19
4.1	Speech Recognition.....	20
4.2	Linguistic Analysis.....	26
4.3	Conversational Agent.....	27
4.4	Speech Synthesis.....	34
4.5	Graphical User Interface.....	37
5	Evaluation.....	39
6	Future Work.....	42
6.1	Embodied Agent.....	43
6.2	Plain English.....	43
6.3	Adaptive Vocabulary.....	44
6.4	Repair Dialog.....	45
7	Conclusion.....	46
	References.....	47

1 Introduction

Language learning has been a lively field of research for over a century. Much of the basic research in the field has been either in the realms of sociology or of education. In the former case, researchers like Piaget, Vygotsky, Leontev, and Chomsky have searched for the social and psychological significance of language and how it is formed [1,2,3,4,5]. By understanding speech and writings, we can understand the human mind, and perhaps by understanding how speech is formed we can understand how the mind develops. On the other end of the spectrum, education researchers have studied language learning as a practical necessity. They have both studied first language learning to understand how to convey the information needed both to language-impaired students and to those that follow a regular development pattern. Additionally they have studied second language acquisition (SLA) as a part of their field itself. Since students are increasingly compelled to learn second languages in order to function in a world economy, the need for second language (L2) education is keenly felt and educators have strong incentive to understand what does and does not work in educating language learners, be it their second, third, fourth or even later language. Differences in individual learners, the impact of the environment, effective exercises, common mistakes and proper positive reinforcement all fall under the purview of education researchers.

Computer science has long been harnessed in the pursuit of education, and language education is no exception. Journals such as Technology and Language Learning, Journal of Interactive Learning Environments, Teaching English with

Technology, Computer Assisted Language Instruction Consortium Journal (CALICO), Journal of Computer Assisted Learning, and Computer Assisted Language Learning: An International Journal as well as books such as Marina Dodigovic's "Artificial Intelligence in Second Language Learning" are entirely devoted to this intersection of education and computational resources, particularly in language learning. In this project, we explore new intersection of language learning theory and computation. By combining automatic speech recognition (ASR), natural language processing (NLP), artificially intelligent conversational agents, and computer speech synthesis (TTS), our system, Skynet, provides speech-enabled conversational agent that freely provides the resources of a skilled language partner to any learner with a capable computational device. This speech agent, described in section 4, recognizes the learner's speech using ASR, analyzes it for common linguistic errors, provide the text to a conversational agent that formulates a response, and finally synthesizes the response for the learner to hear. Using this system, learners can actually practice conversational skills with an intelligent computer agent and receive linguistic feedback tailored to their particular speech. Furthermore, the agent adapts its conversational discourse to the perceived skill of the user, expanding the available topics of conversation as the learner builds more experience.

2 Prior Work

This work has strong connections to a large number of diverse fields of study. Some previous commercial applications of simple artificial intelligence or dialogue to learning

systems are extremely similar; likewise, elements of this project can be found in fields ranging from research on general human cognitive abilities to machine translation.

Below, we discuss some related research that has influenced this project.

2.1 Virtual Learning Environments

Interactive virtual learning environments seek to engage learners in compelling activities that educate learners obliquely. Formerly, the term Edutainment was used to describe programs in this area. Recently, “serious games” or “edugames” have become more popular due to the negative association from earlier systems where the educational aspect was an interruption that needed to be endured before continuing the game [6].

Effective games in this space make the learning integral to the game tasks. Since popular games already require players to develop a good understanding of the game mechanics as well as a wide variety of cognitive and motor skills to excel, serious games can leverage this design for educational goals. Games like “Where in the World is Carmen Sandiego?” and “Timez Attack” succeed in this respect, but they lack a key component on which successful 3rd generation (3G) learning platforms rely: social interaction. While it is important to carefully construct the learning content, 3G platforms must recognize that the social context of learning needs to be central and not an afterthought [7]. The popularity of massively multiplayer online games (MMOGs) has already naturally directed serious games in this general direction. Systems like the Shrine Educational Experience, Neopets, and Whyville emphasize the social construction of learning by the collaborative design of the environment [8,9,10]. In Whyville, for example, children

engage in educational activities like collaborative virtual scavenger hunts and creating dances using vector arithmetic in order to earn experience-based currency which can be applied to purchasing items within the virtual world [10]. Learners explore the world with friends and fellow learners and create increasingly elaborate avatars to display to fellow Whyvillians. While Skynet does not currently provide the same sort of fellow-learner-influenced social construction of learning that 3G learning platforms aim for, it does allow the learners to socially construct their language skills through simulated conversation.

2.2 Automated Tutors

Among others, University of Memphis and Carnegie Mellon University (CMU) have extensively studied using automated tutors to assist in learning. The Autotutor system developed at the University of Memphis tutors primary and secondary students in math and physics. It emphasizes the constructivist school of educational thought – guiding students to build their own understanding rather than trying to explicitly convey methods or mental models [11]. Interestingly, although multiple studies of Autotutor have shown that it can be quite effective, the researchers at University of Memphis also concluded that automated tutors are not inherently well-suited to math or science education. Rather, dialog systems like those used by Autotutor are best suited to subject matter that is “verbal and qualitative” [12]. They also found that such systems work best when the shared knowledge between the tutor and the student is low to moderate. When the shared knowledge is too high, the students expect too “precise a degree of

mutual understanding” [12]. Fortunately, when using the tutor on a “verbal and qualitative” task with moderate shared knowledge as Skynet does, an automated tutor can work well. Graesser et al. point out that tutors do not need to fully understand the learner in order to advance the dialog; humans are very good at interpolating to impute meaning and significance [13].

CMU’s cognitive tutors have also been extensively deployed and tested, especially in the primary education system around Pittsburgh. Some of the cognitive tutors, such as the reading tutor [14,15,16] and the Fluency program for pronunciation practice [17,18] are very closely related to the goals and design of Skynet.

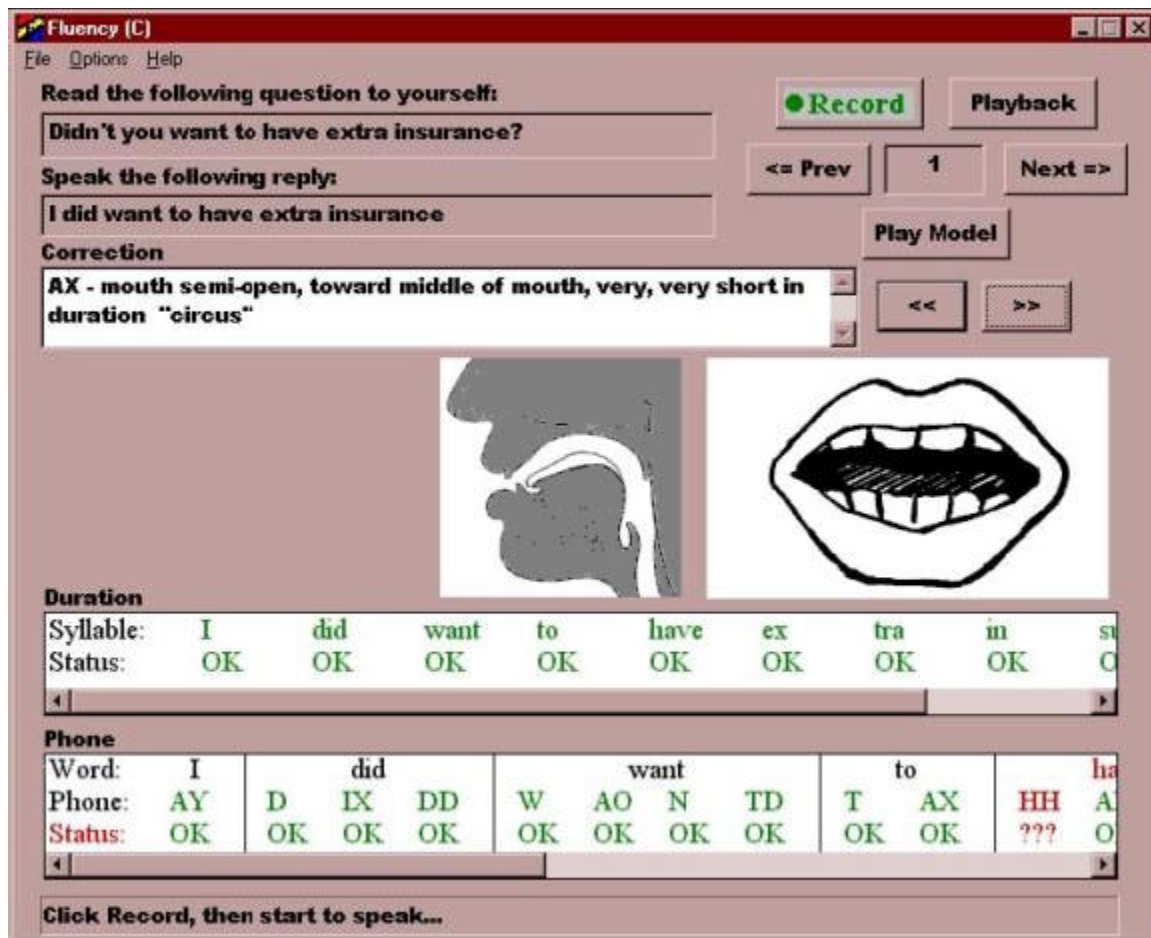


Figure 1. Fluency project at CMU

The empirical findings from the studies of deployed cognitive tutors are highly relevant to the design of Skynet. One key finding from the early tests of the reading tutor is that users will not ask for help, even when they clearly need it. Learners are not good at identifying when they need assistance, and even when they do realize they need assistance, they will usually waste time trying dead-end approaches instead of asking for help [18]. If the learners develop new skills or cognitive abilities through these repeated failures, they are useful, but in the empirical studies they did not develop new skills. Instead, they wasted their time and learned less than peers. Given this

information, Anderson et al. recommend that automated tutors provide assistance at proper times without any prompting [18]. Other studies by the same group also indicate that starting with vague hints and gradually increasing specificity does not work in practice; rather, the system should provide the clearest assistance possible initially [17,18]. This, of course, must be weighed against the potential for annoyance from overly-frequent interruptions to offer help. As such, Skynet only provides linguistic feedback when it is fairly certain that a specific class of common error is occurring, and it provides highly specific meta-linguistic information on how to correct the error.

2.3 Intelligent Agents

The Computer Science and Artificial Intelligence Laboratory (CSAIL) at the Massachusetts Institute of Technology (MIT) has worked extensively with conversational interfaces and dialog systems. In fact, in 2003, Tien-Lok Jonathan Lau published his thesis on something he calls the Spoken Language Learning System (SLLS) [19]. SLLS, like Skynet, approaches the question of supporting second language learning using intelligent agents, but it does so in a very different manner. SLLS uses the Galaxy communicator systems developed at CSAIL [20] to try to provide automated phone conversations for language practice. The system uses a phrasebook system and follows a specific lesson structure, instructing the learner to “fill in” their parts of the conversation verbally. The system employs Galaxy’s voice control system to initiate, conduct, and complete the lesson over the telephone. SLLS was designed for English speakers learning Mandarin, but was never formally tested [19].

CSAIL's other experience with "conversational interfaces" has also been influential in the design of Skynet. Their extensive experience in building telephone-based informational systems has yielded design principles that can be applied to a variety of conversational interfaces. Victor Zue's overview of conversational interfaces identifies three types of systems, system initiative (e.g. "Please say the destination city"), user initiative (e.g. "I want to directions to visit my grandmother"), and mixed initiative where both the system and the user can drive the interaction as appropriate [21]. In general system initiative is easy to implement and easy for the computer to get right, but it is of limited usefulness because it only works for tasks that are essentially form completion. User initiative is very hard to implement because the user can state or request anything and correctly categorizing and responding to this can be very difficult. Skynet is a mixed initiative system. While it does allow the learner to say whatever he or she wishes to, it can only respond within its domain, and it will redirect the topic of conversation to something it can understand if necessary.

2.4 Learning Games

In 2007, CSAIL also published work on a web-based "learning game" for Mandarin. The system was web-based and used speech recognition to score how close the students were to the "proper pronunciation" of the words. Although the authors called the system a "game," the only ludic aspect to the system is that it scores the speaker and provides them more difficult words and sentences as they advance [22].

More in the actual domain of games, the Mobile and Immersive Learning for Literacy in Emerging Economies project (MILLEE) at Berkeley has investigated using mobile phones to provide simple electronic versions of physical games normally played by children in India. MILLEE harnesses the familiarity of the children with the traditional games to engage them in the tasks developing literacy in English [23].

Meanwhile, the University of Southern California (USC) and an independent game development company called Coccinella are developing three dimensional (3D) immersive game environments to instruct students in second languages. Coccinella's game, aptly named 3D Language, uses the Torque game engine to provide a 3D town environment in which the player can explore the surroundings, speak with computer characters, and learn about the culture of the country where the language is spoken [24]. The program includes tutorials on the language and culture, speech recognition software to give feedback on accuracy, scripted dialog to produce a realistic conversation, and speech synthesis so that the learner can actually hear the language as well as see it. USC's system, called Tactical Iraqi, was designed with funding from the Defense Advanced Research Projects Agency (DARPA) to train US military personnel in Iraqi language and culture. It is very similar in structure to 3D Language, but is built using the Unreal Tournament 3D game engine. Tactical Iraqi also includes simulated physical actions using the avatar to practice proper cultural activities [25]. When Tactical Iraqi was evaluated with actual US Marines, the Marines reported finding it very engaging and they performed statistically better on language tests than their

counterparts given traditional classroom instruction. These sorts of games are an obvious complement for Skynet. The 3D Language and Tactical Iraqi are clearly more immersive than Skynet, but they are severely constrained by the authoring time. They suffer from the branching problem in tightly scripted conversations – in order to exhaustively describe all plausible interaction patterns, very large amounts of human time are required to script the possible actions on each side. Skynet, on the other hand, allows the learner to say whatever he or she wishes and attempts to respond in a believable fashion.

2.4.1 Interactive Drama

While it has not yet been directly applied to language learning, interactive drama has some very interesting possibilities. The seminal work in the field to date is Michael Mateas's *Façade* [26]. *Façade* provides a 3D environment in which the character and move and interact. The drama simulates a full-scale verbal and physical interaction with two believable characters with robust personalities. The overall effect is accomplished using pre-recorded sentences for each of the actors, combined with a complex drama management system revolving around "beats." Each beat is a dramatic context advancing the plot, such as discussing the furniture, mixing drinks, or characters fighting about a past vacation. The player is given both local agency to interrupt, request something, pick up objects, hug the characters, and so on, and global agency to change the overall outcome of the dramatic interlude. In order to achieve high believability in a

mixed initiative system with open-ended player dialog and actions, the reactions in each beat require a staggering amount of scripting (about 2 persons years for 20 minutes of interaction), but it is still substantially less than with a true branching system since unaccountable actions will either evoke no response or move the player to a new beat. This sort of drama could be useful both to increase the robustness of the interaction with Skynet and in language games like Tactical Iraqi where cultural interaction could be modeled robustly and realistically in a variety of contexts.

3 Motivation

This project is grounded in and motivated by previous findings from research in education and language learning. We discuss much of the research that has directly motivated the design and creation of Skynet below.

One of the most influential of Lev Vygotsky's learning theories is the so-called Zone of Proximal Development (ZPD). The National Research Council's report on human learning strongly emphasizes this idea along with the complementary concept of scaffolded learning [27]. The idea of the ZPD is simple but also very powerful: students, be they young or old, have a set of skills with which they are comfortable. This forms the core of their skills. Outside of this core is a set of tasks or ideas that they can use with assistance from others who are competent in that area. This is the Zone of Proximal Development. Even further out are tasks that they cannot accomplish even with help, either because they have not yet developed the necessary skill or the necessary knowledge. Students who remain in their core will never develop further, but by

assisting them to move into their ZPD, teachers can help them expand their skills – to enlarge the core of skills and knowledge with which they are individually comfortable (see Figure 2).

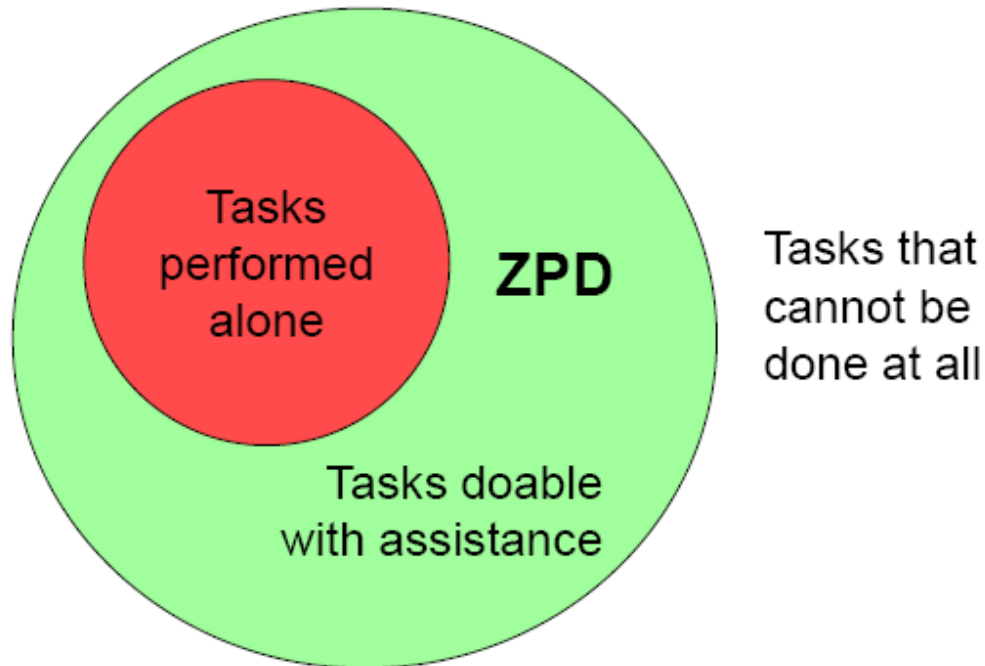


Figure 2. Zone of Proximal Development

This process of supporting the students to move them through the ZPD is often referred to as “scaffolding.” This follows from the idea of a physical scaffold, wherein the scaffold supports the structure as it is constructed until it is able to stand on its own. Once the structure is able to stand, the scaffold can be removed and the structure is whole in itself. In the same way, scaffolding in learning provides students with support to begin building their skills themselves. Once the skills are developed, the scaffold can be removed and a new one constructed in their new ZPD. This idea was first introduced

by Jerome Bruner in the 1950s and was developed over the next two decades, especially in [28]. *How People Learn* from the National Research Council devotes an entire chapter to giving practical examples of scaffolding in classroom settings [27]. Numerous discourses on Vygotskian theory expound on this concept and elucidate connections between the ZPD and the structure of learning systems [29, 30, 31, 32, 33]. This concept drives the overall design of Skynet, but the specific implementation is further influenced by a variety of findings in language learning research.

In their compilation of research in effective teaching strategies for young language learners, Lightbown and Spada have identified specific practical aspects of language learning that are highly effective in acquisition [34]. Particularly salient to this discussion, they investigated what sorts of classroom structure and feedback are effective in correcting mistakes and advancing students' speaking skills. To distill a wealth of information into a single piece of advice: in order to most effectively develop new language skills, students need an opportunity to practice speaking (not just repeating set phrases, but constructing the speech themselves) combined with structured feedback on what they do correctly and incorrectly. In their studies of corrective feedback, Lyster and Ranta found that there were 6 distinct types of linguistic feedback given in classrooms [35]. Based on empirical evidence they ranked the frequency of each type as show in Table 1. Many of these types of feedback occurred in concert. For example, repetition and meta-linguistic feedback would often occur together, as in the example in Table 1.

Recasts

S1 When you're phone partners, did you talk long time.
T When you were phone partners, did you talk for a long time.
S2 Yes, my first one I talked for 25 minutes.

Elicitation

S My father cleans the plate.
T Excuse me, he cleans the ?
S Plates?

Clarification requests

T How often do you wash the dishes?
S Fourteen.
T Excuse me. (Clarification request)
S Fourteen.
T Fourteen what? (Clarification request)
S Fourteen for a week...

Meta-linguistic feedback

S We is...
T We is? But it's two people, right? You see your mistake? You see the error? When it's plural it's we are.

Explicit correction

S The dog run fastly.
T 'Fastly' doesn't exist. 'Fast' does not take -ly. That's why I picked 'quickly.'

Repetition

S He's in the bathroom.
T Bathroom? Bedroom. He's in the bedroom.

Table 1. Lyster and Ranta's six types of linguistic feedback [35]

Lyster and Ranta also evaluated the effectiveness of each type of feedback. Interestingly, although recasts were the most common type of feedback, accounting for more than half of the classroom feedback, they were the least effective in correcting mistakes. Elicitations and meta-linguistic feedback were most likely to help students correct their mistakes [35,36].

In classroom settings, prompting, identifying, and correcting such errors is very difficult to achieve. The instructor is usually viewed as the oracle of knowledge with which the students interact in very short, limited bursts. Classroom time does not allow the students to individually practice their conversational skills with the instructor as needed. On the other hand, private tutors can provide the necessary apparatus and time to properly facilitate language learning; however, most students do not have the resources to contract a private tutor. Skynet does deliberately emulate human tutors in many respects. For example, effective human tutors regulate their speech to accommodate the perceived abilities of the learner, both in terms of grammar and in terms of vocabulary [34], just as Skynet does to attempt to match the learner's ZPD. Already deployed software such as Rosetta Stone and Pimsleur effectively incorporate research in language learning to support individual learners [37], but even so, they cannot provide the interactive language tutoring and the structured feedback suggested by Lightbown and Spada's research. Granted, the software should not be expected to be as effective and useful as a trained human, but it can provide a good approximation.

In order to design Skynet to be useful to an actual population of language learners, we chose a specific group around which we could mold the specifics of the system. In Menlo-Atherton High School, found in the Palo Alto School District in Northern California, a large proportion of the students - over 10% in recent years - are immigrants who begin their secondary education with little or no experience in English [S Kayton, personal communication]. These students are expected to simultaneously

learn English and engage in courses in English. Interestingly, they are expected to take courses in English literature at a high school level before they have even learned English at an elementary school level [S Kayton, personal communication]. Our current design has been largely influenced by the needs of this group of learners. For these students, Skynet can provide an initially very simple conversational partner allowing them to hear English speech and attempt to communicate in English. As they advance in skill, it can use more robust vocabulary and engage the students in more diverse domains; slowly moving the students through their ZPD and expanding their overall skill.

In Patricia Porter's studies, she found that "learners talked more with other learners than they did with native speakers" [38]. Not only did they talk more, they did not make more mistakes with an intermediate-level speaker than with an advanced or native speaker [39]. This indicates that L2 learners do not really need a native speaker or a master of the language in order to improve. While their fellow students cannot offer the level of accurate grammatical feedback, they can provide "genuine communicative practice which includes negotiation of meaning." Porter also found that the learners showed the greatest improvement when they were given what she called the "sender" role; that is, when the learner was the one directing the conversation and the partner, either fellow student or more advanced speaker, was mainly the respondent [40].

The type of speech used in practice also largely determines how well the students learn the language. Lightbown and Spada discuss the differences between "natural" settings and "instructional" settings. In the instructional setting, speech

production is mostly forced in order to illustrate a particular concept - be it a grammar rule, a colloquial usage, or a vocabulary term. The focus is on the speech itself rather than on the information conveyed. In natural speech, learners are actually trying to communicate to accomplish some task. These tasks may be as simple as trying to communicate interest in a television show to a fellow student or as complex as trying to work with others understand a physical mechanics problem. Lightbown and Spada point out that these types of natural speech most often occur at work or in social interaction where the majority of peers are native speakers of the second language and the learner must direct his or her speech towards these individuals to communicate. Recognizing these differences, some education theorists have attempted to design content-based and task-based instructional environments. Our research project attempts to capitalize on the positive aspects of both by providing practice where the focus is actually on the content and interaction, but there is feedback provided to correct mistakes. Lynne Cameron, in her book *Teaching Languages to Young Learners*, also discusses the parallel issue of what she calls “real language” [38]. In order for learners to foster conversational skills, students should be able to converse about topics in their realm of knowledge using terminology and grammar that is appropriate to their demographic [38]. Cameron admits though that “real language” is very difficult to define since it changes dramatically based on age and location: children aged 7 talk about different topics in very different ways from 9-year-olds. While our system is not perfect in this regard, by

initially targeting high-school students and by constructing a mixed initiative system we allow the learners to direct the conversation using what is to them “real language.”

Cameron also discusses how autonomy and self-regulation are very important in language learning and should be encouraged whenever possible. She furthermore provides evidence that overly high demands lead to formulaic responses and do not expand the ZPD. Instead, learners should be able to advance at their own pace, practicing as they feel comfortable. It is important for the system to make the students feel comfortable and successful at the beginning to allow them to explore and while simultaneously pushing the boundaries of their skill [38].

We must balance these principles with the actual capabilities of the modern technologies available and try to support language learning as best as possible using what is available or can be built. Given time, a human tutor is able to negotiate meaning and figure out what a student is attempting to say even if words are mispronounced or sentences are grammatically lacking. Although a computer cannot do this highly effectively yet, it can cheat. By looking for key terms and key constructions, conversational agents can move the discussion forward even if the entire sentence is not fully understood. Additionally, since even given well-formed L1 text current NLP techniques cannot always understand, conversational agents are already designed to effectively equivocate in a human manner until a common frame or context can be reestablished. In the case of SLA, the equivocation and reliance on the user initiative

allows the learner to be the “sender” and move the discussion forwards as he or she sees fit. It is not perfect, but we hope that it will be “good enough.”

4 Design

At a high level the conceptual design of Skynet is fairly straightforward. As Figure 3 shows, there are four main components.

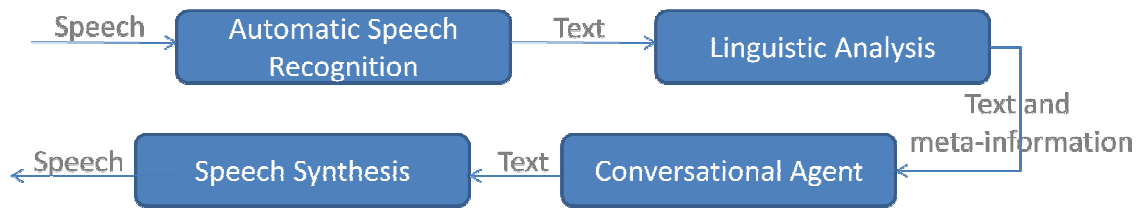


Figure 3. Anatomy of Skynet

First, the speech recognition engine attempts to determine what the speaker says. Second, the linguistic analysis uses very simple natural language processing to identify common errors and provide feedback. Third, the conversational agent parses the speech and determines the response based on the input. Fourth, the response from the conversational agent (or from the feedback engine) is synthesized for the learner to hear. The entire system is provided as a free, open-source software package designed to operate on both Microsoft Windows operating systems and distributions of the Linux operating system. We hope this will allow the system the easiest distribution and the largest impact possible.

4.1 Speech Recognition

Skynet uses the open-source Sphinx III-0.8 automatic speech recognition engine developed at Carnegie Mellon University [41]. Sphinx is a hidden Markov model (HMM) based speech recognition engine. While Sphinx-III is not the only open-source HMM-based ASR system, it is fairly widely known and is undergoing active development and improvement based on on-going research in ASR techniques. Sphinx III-0.8, the most recent version of Sphinx-III, was released in January of 2009. As the name implies, Sphinx-III is not the first in the series of Sphinx speech recognition engines from CMU. Sphinx is also not the only complete ASR system developed through research at CMU; in particular, the Janus speech recognition system at CMU is also fairly widely used, although it is not available to the public. Sphinx-III's immediate predecessor, Sphinx-II, is still fairly widely used. Sphinx-II is provided under the same permissive open-source license as Sphinx-III but relies on older speech recognition technology. While Sphinx-II is not being actively developed, it is still useful for many applications. Specifically, the semi-continuous density Gaussian mixture models (GMMs) on which Sphinx-II relies are particularly effective at modeling the acoustics of very similar speakers. Given a set of speakers, or given a single speaker, very similar or identical to the speaker to be recognized, the acoustic model constructed from the utterances provided by these speakers can provide very good recognition accuracy (e.g. ~90%) on a medium-vocabulary task (e.g. 1000 words) [42]. The other main advantage of the older Sphinx-II is that the simpler earlier techniques used in speech recognition are generally faster

than the current state-of-the-art systems. This means that for simple tasks, the older technology may suffice and may be able to hypothesize text from speech more quickly than the current advanced techniques.

Another project from CMU capitalized on this particular aspect of Sphinx-II in order to develop a real-time mobile speech recognizer: PocketSphinx. PocketSphinx is essentially Sphinx-II using only fixed-point calculations, and with the Viterbi search code tightened up [42]. SphinxTiny, a recent system based on Sphinx-III, applies the same ideas as PocketSphinx to Sphinx-III, reducing execution time on mobile devices substantially [43]. There is also a version of Sphinx entitled Sphinx-4, but despite the naming, it is not in fact a successor to Sphinx-III; rather, Sphinx-4 is a slightly earlier version of Sphinx-III rewritten in Java. Sphinx-4 is also being actively developed in parallel with Sphinx-III [44].

Aside from the work at CMU, there are numerous other speech recognition engines available under a variety of licenses. The Hidden Markov Model Toolkit (HTK) is available under a slightly less-permissive open-source license than Sphinx [45], but is also widely used. Proprietary systems such as Microsoft's Speech Recognition and Nuance's Dragon systems are often seen in speech-enabled applications.

A variety of factors went into the decision to use Sphinx-III over any other speech recognition system. Sphinx-III is available under an open-source license that allows it to be distributed (with the source if desired) free of charge for most purposes; this is a major consideration as we wish Skynet to be available as a free teaching tool to anyone

who wishes to use it. Sphinx-III provides high accuracy on speaker-independent large vocabulary tasks. This is vital since the language learners cannot train the system to their own speech patterns in the target language due simply to the fact that they do not yet speak the target language. The fact that Sphinx-III is designed for large vocabulary tasks enables Skynet to work with a cornucopia of conversational domains simultaneously, thereby supporting general speech learning in place of limited vocabulary or constrained sentence practice.

The general structure of Sphinx-III (and most ASR systems) is shown in Figure 4.

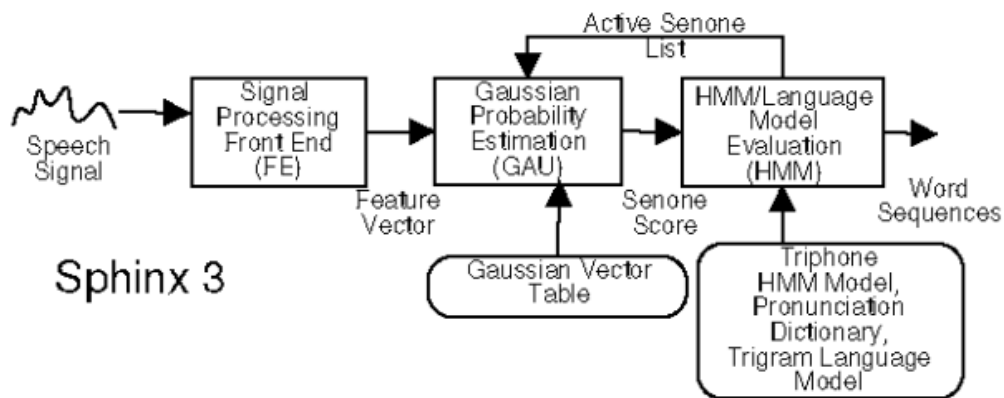


Figure 4. Anatomy of Sphinx-III, adapted from [41]

The front end is responsible for processing the raw audio signal captured from the microphone input device. It then converts from the time domain into the frequency domain and extracting the actual speech signal after scrubbing the signal to clean as much background noise or microphone noise possible. While Sphinx-III conforms to the European Telecommunications Standards Institute (ETSI) guidelines, and thus can work with any ETSI front end, Skynet uses the CMU Sphinxbase front end specifically designed

for use Sphinx-II and Sphinx-III. Sphinxbase provides good accuracy and very low latency. Using Sphinxbase we provide the Mel-frequency cepstral coefficients (MFCCs) using cepstral mean normalization (CMN). Older ASR systems may also use pure Discrete Fourier transforms (DFT), linear prediction coefficients (LP) or perceptual linear prediction coefficients (PLP). PLP is especially useful if the speakers using the system are very different from the speakers used in training the system and was specifically designed to allow ASR systems to be trained on adult speech then used to recognize children [46]; however, since we have a corpus of speakers very similar to our target audience both in age and accent, this is no advantage to us. Also, a recent comparison among the various methods for signal extraction using a variety of ASR front ends indicated that MFCCs paired with relative spectra filtering (RASTA) where a high-pass filter is applied to the MFCCs provides the highest accuracy [46,47,48]. This is especially interesting since RASTA was initially developed as a complement for PLP. Unfortunately, Sphinxbase does not currently support RASTA filtering, so we must rely on the slightly-less-accurate but more common CMN where the filtering is achieved by subtracting the short-term average of cepstral vectors [48]. In order to derive the original MFCCs from the raw audio, Sphinxbase takes the Fourier transform of the signal, maps the powers of the spectrum obtained onto the Mel scale using triangular overlapping windows, takes the logs of the powers at each of the frequencies, then takes the discrete cosine transform of the log powers. The resulting amplitudes of the cosine transform are the MFCCs given to the acoustic modeler to identify phonemes [49].

In the acoustic modeling stage, vectors of forty consecutive MFCCs representing a single frame of speech (with usually 100 frames per second) are matched against a context-dependent sub-phonetic unit called a senone in Sphinx's parlance. Each senone against which the feature vector is matched is a mixture of Gaussian ellipsoidal distributions within the feature space. The number of possible senones is determined when the acoustic model is initially trained. The number of Gaussian mixture parts used to represent the senone is also determined in the training process. In the model Skynet currently uses there are 1000 senones with 16 Gaussian mixture parts in each. The senones are hypothesized and combined to form phonemes that are then combined into words. In general, vowel sounds are identified and the modifications in the vowels created by the consonants preceding and following are used to identify the consonants used. The actual sounds are then matched against a dictionary of words matching the phonetic units comprising a word to the word itself.

Once likely phonemes are identified and the possible words that the phonemes could form are identified, the speech recognition system uses an HMM to determine from the language model the actual words of the utterance. As would be expected, a Viterbi search is used to find the best possible path. This language model is also constructed based on a corpus of sentences that represent likely speech. This allows it to identify key features such as the fact that articles like "a" or "the" will almost always precede a noun, with the exception of cases in natural speech where the speaker may say something like "the...the...the blue boat, I think."

In the prototype, achieving high recognition accuracy is paramount. In order for the conversation to be at all helpful and in order for Skynet to be able to actually provide any feedback on the speech that the learner is using, the prototype must be able to get at least an approximate idea of what the learner is trying to say. Of course, speech recognition is a difficult task even under optimal conditions when the speaker can speak clearly in a noise-free environment with a good quality microphone and can train the engine to their own speech patterns. Unfortunately, we have an even more difficult task because we need the system to be able to understand a diverse set of speakers whose speech patterns are changing rapidly as they develop their language skills. The evaluation will show if it is sufficient, but at present the prototype acoustic and language models are trained on a corpus of sentences as close to the speakers as we could find.

In 2007, the Center for Spoken Language Understanding (CSLU) at the Oregon Health and Science University released a corpus of speech they collected using their CSLU Speech Toolkit [50,51]. The corpus is a collection of spontaneous and prompted speech from 1,100 between Kindergarten and Grade 10 in the Forest Grove School District in Oregon [51]. There were approximately 100 children at each grade level, and all read a set phonetically-balanced simple words, sentences, or digit strings. In addition, the children recorded spontaneous natural speech; each beginning with a recitation of the alphabet and containing a monologue around one minute long. The corpus also provides the human transcriptions of the speech necessary to train the acoustic and

language model. From this corpus, we are able to select speakers of similar age and with similar accents to our target audience and train our acoustic model to their speech. The transcriptions of the spontaneous speech is also useful as the core of the language model, as the speech that the participants produced speaking with the CSLU Toolkit should be similar to the speech that learners will use practicing with Skynet.

Nevertheless, we must be sensitive to the fact that some of the learners will have much lower levels of speech production skills in the target language – English in this case – and thus will produce much more “broken English.” Also, we can supplement the language model training data with additional conversational data obtained from telephone conversations gathered in other corpora, specifically from the 2003 NIST Language Recognition Evaluation project which is a combination of elements from the CALLFRIEND, CALLHOME, and Switchboard-2 corpora [52].

4.2 Linguistic Analysis

Once the speech recognition has provided a hypothesis of the speech, the text is given to the linguistic analyzer and the conversational agent to analyze and formulate a response. The linguistic analysis is currently only rudimentary. It is being built from the ground up to match the goals of the project. Currently, the text that is produced by the speech recognizer is run through a part-of-speech tagger, and the result is matched against common word ordering errors among native Spanish speakers learning English as a second language. In the future, this will be expanded to include many other classes of common errors. Since the speech recognition is not perfect, errors may not actually

have occurred. If a specific class of error is seen more than a threshold number of times, a preset meta-linguistic feedback message is given to the agent in addition to the text to which it should respond. Following Lyster and Ranta's findings, the feedback dialog provides both the meta-linguistic feedback describing the class of error and elicits the correct form of the speech. The speaker will be asked to repeat the sentence in the grammatically correct form (recognized using Sphinx in forced alignment mode) before continuing in order to reinforce the correct form. In the majority of cases though, the text from the speech recognition engine is simply passed on verbatim to the conversational agent.

4.3 Conversational Agent

Skynet relies on RebeccaAIML for the artificial intelligence. RebeccaAIML is an open-source implementation of the AIML standard [53]. AIML, or the Artificial Intelligence Markup Language, was created by Dr. Richard Wallace between 1995 and 2002 [54]. The AIML standard is a specification for parsing text and responding to it. Using AIML, Dr. Wallace created A.L.I.C.E., a conversational agent that has won the Leobner Prize Contest for Most Human Computer three times [55]. RebeccaAIML is just one implementation of the standard. It is designed with an interface for C, C++, Java, .Net/C#, and Python with an Eclipse plugin for editing AIML categories, and it is operational both under Microsoft Windows and under Linux operating systems. There are other implementations of AIML available such as Program R in Ruby, Program P in Perl, Program D in Java, and Programs O in PHP, but RebeccaAIML offers the most

portability and the most functionality. There are other modern well-known Artificial Conversation Entities (ACEs) such as Jurgen Primer's Jabberwock, Rollo Carpenter's Jabberwacky bots George and Joan, Robert Medeksza's Ultra Hal, and the most recent Leobner Prize winner Fred Robert's Elbot. These agents, all designed specifically for the purpose of carrying on a conversation with humans, are variously known as chat bots, chatterbots, chatterboxes, or talk bots. In addition, there are tools such as the website Personality Forge devoted entirely to the scripting of new chat bots. While artificial intelligence is a major area of research in computer science, at present the work being done on ACEs is confined almost exclusively to the commercial and private sector.

In 1966, Joseph Weizenbaum published a paper about his program ELIZA. ELIZA represented very simple natural language processing to simulate a Rogerian therapist [56]. ELIZA was one of the first chat bots and is still one of the most famous. The program matched the text that "patients" entered against a set of around 200 patterns and reformulated statements as questions or prompts for more information. For example, given the statement "My mother hates me," ELIZA would respond with "Who else in your family hates you?" While ELIZA managed to convince many patients that it was in fact a human to whom they were talking, research in this sort of artificial intelligence (AI) did not progress very far. Instead, AI research turned to unequivocal tasks like object recognition, image and video search, and part-of-speech tagging. These techniques could be tested, verified, and directly applied to needs elsewhere in computer science.

Meanwhile, the concept of ELIZA was expanded and applied elsewhere. Most modern chat bots are built on the same principles, and AIML bots are no exception. Some of the ACEs, such as Rollo Carpenter’s Jabberwacky bots, attempt to learn patterns from interactions with humans. AIML bots, also known as ALICE clonebots or ALICEbots, rely instead entirely on human scripting of patterns. For Skynet, we chose to use AIML for much the same reason that ALICE clonebots are very prevalent across the Internet. AIML is a completely open standard, the implementations of the standard, like RebeccaAIML, are released as free, open-source software under the GNU Public License (GPL), and the core of the “ALICE brain” is also freely distributed as the Annotated A.L.I.C.E. AIML (AAA). Additionally, other versions of the AAA are freely available in French, Spanish, German, Italian, and Portuguese, allowing the same technology to be used for other languages. While Weizenbaum’s ELIZA matched around 200 patterns, the core of ALICE matches over 40,000, giving it a much more robust interaction spaces [54].

The AAA brain is composed of various groupings of AIML categories by topic. Each category is designed to match a specific pattern and provide a response to it. For example, if the language learner asks “what is your name,” it will match a category like in Table 2.

```
<category>
  <pattern>WHAT IS YOUR NAME</pattern>
  <template>My name is John.</template>
</category>
```

Table 2. Category in AIML

In order to actually match a specific pattern it is often necessary to first simplify the input it is given. This process is accomplished recursively as shown in Table 3.

(1) Symbolic Reduction: Reduce complex grammar forms to simpler ones.
(2) Divide and Conquer: Split an input into two or more subparts, and combine the responses to each.
(3). Synonyms: Map different ways of saying the same thing to the same reply.
(4) Spelling or grammar corrections.
(5) Detecting keywords anywhere in the input.
(6) Conditionals: Certain forms of branching may be implemented with <srai>.
(7) Any combination of (1)-(6).

Table 3 - from [The Anatomy of ALICE \[54\]](#)

The categories are stored together in individual AIML files which are merely logical groupings by topic or theme of the XML making up the categories (see Table 4).

Patterns	File	Description	Last Modified
1314	Adverbs.aiml	Omits adverbs w/o changing logical semantics	Sep 10 2005
230	AI.aiml	Knowledge about A. I. and robots	Sep 10 2005
195	ALICE.aiml	Specifically mentions ALICE by name	Sep 10 2005
13	Astrology.aiml	Star signs	Sep 9 2005
3565	Atomic.aiml	Categories with patterns without wild cards	Oct 10 2005
38	Badanswer.aiml	Client can teach bot new replies.	Jun 30 2005
559	Biography.aiml	Famous personalities	Sep 9 2005
36	Blackjack.aiml	The Card Game	May 11

			2006
3041	Bot.aiml	Knowledge about the bot's personality	Sep 9 2005
35	Botmaster.aiml	Knowledge about the botmaster	Sep 9 2005
1213	Client.aiml	Getting to know the client	Oct 10 2005
138	Computers.aiml	Knowledge about computers and software	Sep 9 2005
45	Date.aiml	Date and time using Pandorabots formatted AIML date tag	Sep 9 2005
4987	Default.aiml	Non-committal replies to imprecisely matched inputs	Oct 10 2005
10	Dialog.aiml	Display the recent conversation history	Sep 7 2005
57	Drugs.aiml	Politically Incorrect Opinions	Nov 21 2009
123	Emotion.aiml	Emotional responses depend on personality type	Sep 9 2005
23	Food.aiml	Culinary knowledge	Sep 9 2005
843	Geography.aiml	Places and locations	Sep 10 2005
1	Gossip.aiml	Spreading rumors.	Sep 10 2005
20	Happy.aiml	Knowledge about the past	Jan 2 2009
10	History.aiml	Knowledge about the past	Sep 10 2005
230	Human.aiml	Replaces AI.aiml for a more "human" bot	Oct 10 2005
9	Humor.aiml	Take your chances	Sep 10 2005
25	Inquiry.aiml	Gather Information about the client without repeating questions	Sep 10 2005
110	Integer.aiml	Simple Integer Addition in AIML	Jul 6 2005
99	Interjection.aiml	Yes and No	Sep 10 2005
2	IU.aiml	Default categories for inputs starting with I and YOU.	Sep 10 2005
1459	Knowledge.aiml	General knowledge	Sep 10 2005
8	Literature.aiml	Books and fiction	Sep 10 2005
31	Luckyslots.aiml	Slot machine game	Jun 30

			2005
33	Money.aiml	Economics	Sep 7 2005
61	Movies.aiml	Film	Sep 10 2005
8	Multiple.aiml	Multiple Choice Test	Sep 10 2005
17	Music.aiml	Musical tastes and trends	Sep 7 2005
8980	Parts.aiml	Simplifies some past participle expressions	Sep 7 2005
67	Personality.aiml	Simple personality test	Sep 10 2005
16	Philosophy.aiml	Epistemology and Metaphysics	Sep 7 2005
3	Pickup.aiml	Pickup Lines	Sep 10 2005
24	Politics.aiml	More Politically Incorrect Opinions	Nov 21 2008
5	Predicates.aiml	Manage client predicates	Sep 10 2005
714	Psychology.aiml	NLP and Assertiveness Training	Sep 9 2005
5366	Reduce.aiml	General Purpose Symbolic Reductions	Oct 10 2005
880	Reducer.aiml	Symbolic reductions using <srail><star/></srail>	Oct 10 2005
9189	Reductions.aiml	Atomic Reductions	Oct 10 2005
288	Religion.aiml	Leave this out unless you want a Protestant Christian robot	Oct 10 2005
141	Salutations.aiml	Hello and Goodbye	Sep 10 2005
25	Science.aiml	Scientific Knowledge	Sep 9 2005
148	Sex.aiml	ALICE has been called a prude.	Sep 9 2005
252	Spam.aiml	Filters out some spam found in log files	Sep 7 2005
24	Sports.aiml	Sports Talk	Sep 9 2005
60	Stories.aiml	Telling Stories	Sep 7 2005
6	Stack.aiml	Manipulate a stack of topics	Sep 10 2005
1355	That.aiml	Categories with <that>	May 11 2006
4	Utilities.aiml	Useful debugging AIML	Sep 10

		categories	2005
440	Wallace.aiml	Information about the Archbotmaster.	Sep 10 2005
58	Wordplay.aiml	Anagram word game.	Feb 26 2006
5	Xfind.aiml	Search for external knowledge on the web.	Sep 10 2005
567	update.aiml	Recent AIML additions.	Oct 10 2005
47205	Categories (including duplicates)		

Table 4. Elements of Annotated ALICE AIML (<http://www.alicebot.org/aiml/aaa/> accessed 5/1/2009)

In order to actually simulate a conversation partner, it is necessary to maintain context from the conversation. AIML does this by providing variables which are set initially and can be modified throughout the conversation, and by tracking the most recent statements by both parties. The <that> tag refers to the agent's last statement allowing it to take previous statements into account (see Table 5). Table 5 also shows a variable, called a "property" in AIML parlance, being retrieved. In this case, the property is the name of the person to whom the agent is speaking (as defined by the speaker).

<pre> <category> <pattern>KNOCK KNOCK</pattern> <template>Who is there?</template> </category> <category> <pattern>*</pattern> <that>WHO IS THERE</that> <template><person/> who?</template> </category> <category> <pattern>*</pattern> <that>* WHO</that> <template>Ha ha very funny, <get name="name"/>.</template> </category> </pre>

Table 5. Context in AIML

AIML also maintains context by tracking the current topic. Topic tags wrap groups of categories and define the topic of that group. The AIML category search will attempt to match patterns within that topic before looking elsewhere. The topic is maintained as an AIML property and can be set at any time within a category; thus, the agent can adjust the topic in response to conversational details from the learner.

As the categories are usually inherently independent of each other, especially when considered at the level of an AIML file or general topic, it is straightforward to add or remove parts of the agent's brain as needed. This allows us to begin with a very simple agent and expand the topics available and the complexity of the speech as the speaker shows more comfort with a given level of speech. This can also decrease the initial loading time for the entire system. Using this capability we can try to match the speech produced by Skynet to the current level of the learner. By storing the previous state and restoring it with the agent we can allow the agent to evolve with the skill of the learner.

4.4 Speech Synthesis

Festival is a multi-lingual speech synthesis engine under active development at the Centre for Speech Technology Research (CSTR) at the University of Edinburgh and the Language Technologies Institute at Carnegie Mellon University [57]. Like Sphinx-III and RebeccaAIML, Festival is open-source, free software designed for use both on Microsoft Windows operating systems and under the Linux operating system. Festival is actually a collection of a number of speech synthesis algorithms designed to work with the

Festvox tools. Festvox is a toolset designed to construct a variety of types of speech synthesis models. At a high-level speech synthesis works as shown in Figure 5.

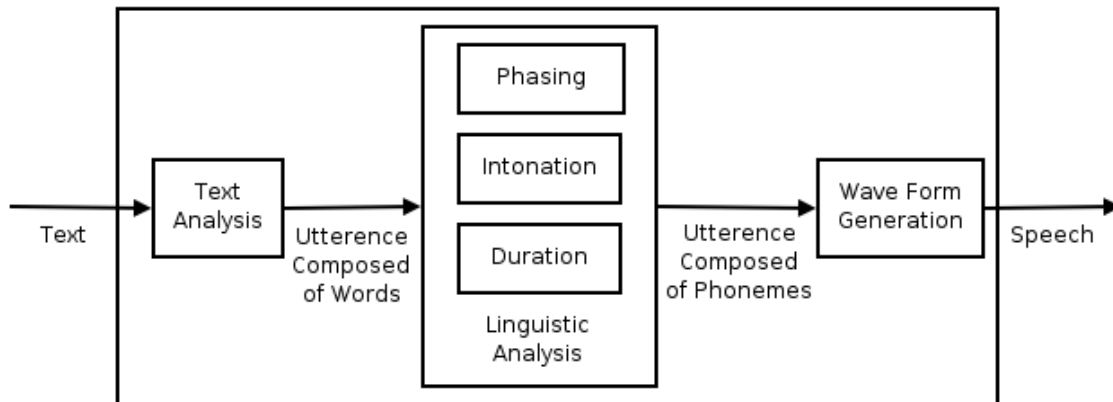


Figure 5. Anatomy of a speech synthesis system

Festival offers configurable, language-independent tokenization and text-to-phoneme (also known as grapheme-to-phoneme) conversion. This includes the phoneset, lexicon, letter-to-sound rules, tokenizing, part-of-speech tagging, and intonation and duration modules: everything up to the waveform synthesizer. The engine has been used to synthesize English, Spanish, French, German and Welsh speech by its developers. Initially, at the waveform generation stage, Festival only offered a variety of diphone synthesis models, including Linear Predictive Coding (LPC), Pitch Synchronous Overlap Add Method (PSOLA) and the more recent MBROLA based on Multi-Band Re-synthesis Pitch-Synchronous OverLap-Add (MBR-PSOLA) [58,59]. Diphone synthesis concatenates sound-to-sound transitions, called diphones, from a set of all possible diphones in a language in order to generate the waveform of the words. Since there are a relatively small number of possible diphones in most languages – only about

800 in Spanish and 2500 in German – this requires minimal memory and processing to generate the waveform [57]. The actual method, such as MBROLA, determines the prosody which is superimposed on the concatenated diphones. Although techniques like sinewave synthesis and formant synthesis are optimal for generating intelligible speech quickly, in our system, naturalness of the voice is as important if not more important than clarity. Understanding prosody and rhythm in the second language is vital to communication.

More recently, Festival has added hidden Markov model based waveform synthesis and unit selection method synthesis. In the HMM method, the frequency spectrum (vocal tract), fundamental frequency (vocal source), and prosody are modeled using HMMs and waveforms are generated based on the maximum likelihood from the HMM. In the unit selection method, input speech is segmented into phones, diphones, half-phones, syllables, morphemes, and words. These units are reassembled according to a weighted decision tree to generate the waveform. This can provide very good naturalness since it uses minimal digital signal processing (DSP), relying instead on merely assembling the component parts correctly. Unfortunately, these methods can require very large amounts of space to store all the various units of a robust model.

At present Skynet uses a free female diphone residual LPC voice created by the CSLU Speech Synthesis Research group. This will be replaced by a robust unit selection model as soon as an appropriate free one is found.

4.5 Graphical User Interface

The discussion thus far has covered the unseen architecture of Skynet, but Skynet also relies on a graphical user interface to assist in learning in addition to the audible speech interaction. The visual interface is designed to facilitate learning and to incorporate foundational language learning principles as much as possible.

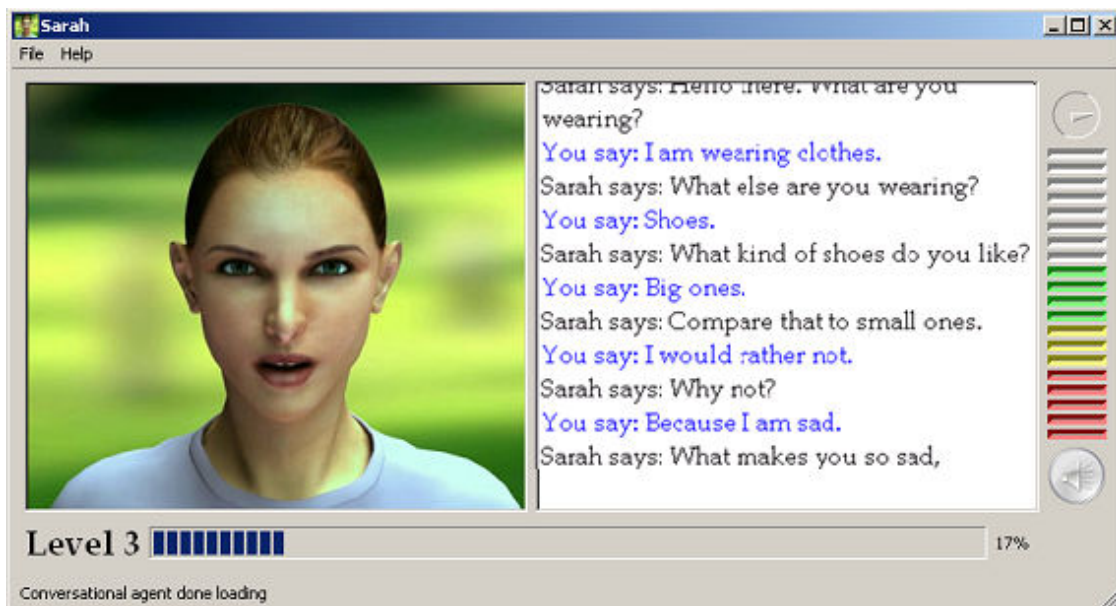


Figure 6. Skynet graphical user interface

The dominant feature in the interface is the static image of the conversational partner (see Figure 6). Research on the effects of media richness suggest that when initially attempting to form a social connection, presenting an image of the partner promotes “affection and social attraction” [60, 61]. Additionally, richer media, such as video, “facilitate social perceptions...and perceived ability to evaluate others’ deception and expertise” [62]. We try to anthropomorphize the conversational agent and foster a social connection to the computer character by giving it a face. Providing a face or an

avatar can be detrimental when it is not actually an accurate visual representation of a person to whom the learner is actually speaking [63, 64]. In particular, making the avatar “too human” when it is not actually human will disturb people as the subtle differences make the avatar “creepy” [65]. We avoid this by providing only a static image, seeking to foster the initial social connection and to give the learners a focal point for their interaction with the character while not distracting them with extraneous or misleading movement. We furthermore give the agent a human name, Sarah, to allow the learners to perceive it as person and interact in conversation with it more naturally.

While young children learn primarily through aural and verbal interaction, once an individual is literate, withholding textual reinforcement can severely disadvantage the learner [38]. Since our expected audience is already literate, we provide a textual transcript both of what the speech recognition engine hypothesizes and the response that the conversational agent generates. The hypothesized speech is displayed both so that students can check how well they are actually pronouncing the words and so that responses to misrecognized speech do not confuse students.

Lastly, we attempt to incorporate achievement motives as suggested by persuasion theory. Computer games have long incorporated achievements and competition in order to engage players and to persuade them to complete tasks [66]. We attempt to apply this same idea in a very simple manner by expanding the functional conversation ability of the agent as they interact more. This both scaffolds

the learner and encourages them to continue their interaction. Also, we provide an experience progress bar towards the next level based on the quantity of speech generated by the learner. As the learners actually speak with the agent, they can see their experience build up, achieving certain levels as they advance (see Figure 6). As this experience accumulates and the learner moves on to new levels, additional conversation topics are added. For example, at level 6, among others, sports and common literature are added, and at level 8, philosophy, politics, and religion become available for discussion. While the system can be cheated by simply making non-sense noises to generate hypotheses and accumulate experience, we are more concerned with providing learners with a sense of accomplishment and efficacy as they practice than with strictly controlling advancement [67].

5 Evaluation

The system has not yet been evaluated, but we have formulated a plan to evaluate it in two stages. The first stage is entirely qualitative and will be used to gather data used to improve the system later. In this stage, Skynet will be deployed to 3 L2 course instructors and 15 L2 students for evaluation. They will be asked to interact according to a semi-guided template over the course of a week. They will be given a set of basic tasks to try out, such as finding the best volume to speak, talking about themselves, and gaining enough experience to reach level 6. This will provide a basis for the later discussions with participants about the positive and negative aspects of Skynet. They will also be asked to experiment as they want and keep track of what they do so that we

can get an idea of what learners will likely naturally do with the system. The participants will be interviewed in semi-structured interviews to collect information about 5 dimensions of the software:

1. What they like about it
2. How well they think it understands their speech
3. What topics they discussed with the agent
4. What topics they would like to discuss with the agent
5. How they would like to improve it

In addition to the quantitative interviews, the actual audio and text from the participants' interactions will be logged for use in improving the acoustic and language models and identifying empirically what topics were most useful and where the system failed.

The second stage will quantitatively measure the effectiveness of Skynet in improving language skills. Participants will be given a short initial written and spoken language pre-test to determine their starting language skills. They will be grouped according to the language test as relatively low, medium, or high experience. Each group's members will be randomly assigned to the control group or one of the two intervention groups. Groups will be of the same size, hopefully between 6 and 8 participants each. The students in Menlo-Atherton High School, with whom we will be evaluating the system, are provided personal computers with a uniform set of software

running Microsoft Windows. We will rely on these computers to run the software for the two intervention groups.

The evaluation will be structured as a repeated measure design. The control group will not receive any treatment but will continue with their normal L2 education and continue interacting with native English speakers as usual. The first treatment group will be asked to follow the prescribed program for the Pimsleur language system for 3 hours per week. The Pimsleur's method comprises 30 minute lessons where the learners are instructed to listen to and repeat phrases along with a recording. It has previously been evaluated and proven to assist in language learning [37]. We will provide the Pimsleur English for Spanish Speakers course on CD and direct the participants to practice using the computer during the 3 hours each week in order to reduce possible confounds.

The second treatment group will be asked to practice conversation with Skynet for 3 hours per week. We will administer short standard tests for vocabulary knowledge, grammar knowledge, and conversational competence once every two weeks over the course of the six weeks of the study, then one more two weeks after the end of the study. The repeated measure design and the Pimsleur group should allow us to correct for intervention effects as well as to identify how much Skynet assists in SLA compared with a validated language learning system. The pretest and the control group relying on traditional language instruction should indicate how much Skynet assists in language development, as well as how that compares to a system like Pimsleur. The quantitative

tests will allow us to see if the expected statistically significant improvement due to practicing conversation with Skynet actually occurs.

6 Future Work

Roger Brown's and J.G. De Villiers's studies of the acquisition of "grammatical morphemes" shows that language children learning English a first language acquire the morphemes in roughly the same order regardless of specific circumstances (see Table 6) [68,69].

- Present progressive -ing (mommy *running*)
- Plural -s (two books)
- Irregular past forms (baby *went*)
- Possessive 's (daddy's hat)
- Copula (annie *is* a nice girl)
- Regular past -ed (she *walked*)
- Third person singular simple present -s (she runs)
- Auxiliary 'be' (He *is* coming)

Table 6. Grammatical morpheme acquisition in English L1 speakers

Interestingly, decades later researchers found that L2 learners from a given L1 also reliably follow specific morpheme acquisition patterns [70,71]. The order of morpheme acquisition is often similar to native learners of the language, but is also clearly influenced by the morphemes in the learner's first language. In particular, similar morphemes are often learned earlier than dissimilar morphemes [72]. At present, Skynet does not incorporate these findings in the design, although it is probably sensible to structure the responses generated by the agent so that it gradually introduces morphemes as the speech of the learner indicates that they may be in the his or her

ZPD. Along those same lines, we discuss below a few aspects of the project that have been designed or considered but have not yet been constructed.

6.1 Embodied Agent

Skynet is currently provided as a free, open-source software program designed to operate under both Windows and Linux. This affords it easy distribution to learners who might find it useful. Although it will have a much lower distribution potential, we would also like to investigate the effect of embodying the agent. As discussed above, media richness may suggest that having a tangible character with whom to talk would provide a better social connection to the agent. We would like to investigate if learners feel more at ease with a physical embodiment of the agent. We are specifically considering a humanoid representation similar to Karrie Karahalios's Ginger [73,74] in place of our current single-image computer interface.

6.2 Plain English

At present, Skynet uses the unmodified AAA as the core of the conversational agent. Certain of the terminology and grammatical structure of the responses may be awkward for new learners of English. Plain English is a standard subset of the English language advocated for clarity when conversing with those who speak English as a second language [75,76]. It is a more lax version of the Simplified English developed for communication in the aerospace industry. We intend to prune the AAA to either remove or simplify responses that do not fit the guidelines of Plain English. While it is true that native speakers of English will not constrain themselves to Plain English (especially since

most do not even know of its existence), for beginning learners of English it is probably more useful to provide this sort of simplified sentence structure and vocabulary. This is akin to the reduced speech used by successful instructors when conversing with beginning and intermediate L2 learners [34].

6.3 Adaptive Vocabulary

In order to further support the user in their zone of proximal development, we intend to filter the vocabulary adaptively. We plan to maintain a list of vocabulary that the learner has used thus far (as well as a core of language that every learner should recognize) and attempt to structure responses to rely on the vocabulary with which they are familiar as much as possible. This adaptation will be done in two ways. First, the response normally generated by the agent will be matched against the dictionary of previously seen terms. If possible, words will be replaced with synonyms already seen from the speaker. A minimal number of new vocabulary terms will be used. This is not to say that the agent will constrain itself to only vocabulary previously seen; in fact, doing so would jeopardize the ability of the agent to challenge the learner. Rather, it will try to keep from overwhelming the learner with too many new vocabulary terms at any given time. Second, when multiple possible categories match a given pattern from the learner, Skynet will choose the response that has the fewest out-of-vocabulary terms possible. Again, this is to prevent the learners from becoming overwhelmed by new language use.

6.4 Repair Dialog

The linguistic analysis is still fairly basic and can currently only recognize transposition and misplacement errors. There are many classes of other well-known errors associated with learners at a particular stage [68,69,70]. One great advantage that Skynet has is that these errors are common to almost all learners, so the feedback does not need to be tailored to the individual; instead, it can be designed for all learners with good confidence that it will recognize the same classes of errors when any individual makes them. The actual correction dialog can probably also be improved further. Currently, it just describes the class of error and elicits the correct form of the sentence. Based on research in intelligent tutors and given more time, we can design a more robust interaction for instruction.

Along those same lines, using the structure of AIML, we should be able to construct a more robust repair dialog. As mentioned earlier, the pattern matching is determined initially by topic and context. At present, if the input does not match a specific pattern, it will fall into a default case where the agent will either equivocate by producing something like “Hmm” or “Interesting,” or it will specifically ask for more information after indicating that it does not understand. This works fairly well, but we could just as well attempt to negotiate understanding in this same interaction. When the agent does not understand, it can immediately admit that it does not understand and ask the learner to repeat or rephrase the statement or question. If after multiple tries, the agent still cannot “understand,” it can equivocate to move the conversation

forward. This sort of interaction is fairly simple to implement – requiring only authoring of AIML categories, but making it highly effective will require observing how learners actually interact with it, then tuning the responses to match the most common action chains.

7 Conclusion

In this work we presented the motivation and design of Skynet, a speech enabled conversation agent created to support second language learning. We presented prior work that has informed the design of the system and enumerated sources of language learning research that have guided the ideals of the design. We also discussed the plan to evaluate the system with a specific set of language learners and described improvements we hope to make to Skynet in the near future. We hope that this will prove a useful tool for language learners and possibly become incorporated into other programs to support second language learning.

References

- [1] JV Wertsch. Vygotsky and the social formation of mind. Harvard University Press, 1985.
- [2] J Piaget, B Inhelder. The Psychology of the Child. Basic Books (Harper-Collins), 1969
- [3] H Ginsburg, S Opper. Piaget's Theory of Intellectual Development. Prentice Hall, 1969
- [4] AA Leont'ev and CV James. Psychology and the language learning process. Pergamon, 1981
- [5] L White. Universal Grammar and Second Language Acquisition. John Benjamins, 1989
- [6] C Conati, X Zhao. "Building and evaluating an intelligent pedagogical agent to improve the effectiveness of an educational game." Intelligent User Interfaces 2004.
- [7] J Laister, A Koubek . "3rd Generation Learning Platforms Requirements and Motivation for Collaborative Learning." Technikum Joanneum 2001.
<http://www.eurodl.org/materials/contrib/2001/icl01/laister.htm>. Accessed 5/1/09
- [8] N Di Blas, P Paolini, C Poggi. "3D Worlds for Edutainment: Educational, Relational and Organizational Principles." IEEE Pervasive Computing and Communications Workshop, 2005
- [9] SM Grimes, LR Shade. "Neopian Economics of Play: Children's Cyberpets and Online Communities as Immersive Advertising in NeoPets.com," International Journal of Media and Cultural Politics, 1(2): 181-198. 2005
- [10] C Galas, J Sun. "Why Whyville?" Learning and Leading with Technology. 2006
- [11] AC Graesser, N Person, Z Lu, MG Jeon, B McDaniel. "Learning while holding a conversation with a computer," Technology-based education: Bringing researchers and educators together, 2005
- [12] AC Graesser, S Lu, GT Jackson, HH Mitchell. "AutoTutor: A tutor with dialogue in natural language," Behavior Research Methods Instruments and Computers, 2004
- [13] AC Graesser, GT Jackson, EC Mathews, HH Mitchell. "Why/AutoTutor: A test of learning gains from a physics tutor with natural language dialog," Proceedings Cognitive Science, 2003
- [14] S Witt, SJ Young. "Language learning based on non-native speech recognition," Eurospeech, 1997
- [15] JR Anderson, CF Boyle, AT Corbett, MW Lewis. "Cognitive modeling and intelligent tutoring," Artificial Intelligence and Learning Environments, 1990
- [16] JR Anderson, AT Corbett, KR Koedinger, R Pelletier. "Cognitive tutors: Lessons learned," Journal of the Learning Sciences, 1995
- [17] M Eskenazi, S Hansma. "The Fluency pronunciation trainer," Proceedings STiLL-Speech Technology in Language Learning, 1998
- [18] M Eskenazi, Y Ke, J Albornoz, K Probst. "The Fluency Pronunciation Trainer: Update and user issues," Proceedings INSTiL2000, 2000

- [19] TLJ Lau. "SLLS: An online conversational spoken language learning system," MIT Thesis Publishing, 2003
- [20] S Seneff, R Lau, J Polifroni. "Organization, communication, and control in the GALAXY-II conversational system," Eurospeech, 1999
- [21] VW Zue, JR Glass "Conversational interfaces: Advances and challenges." Proceedings of the IEEE. 2000
- [22] W Chao, S Seneff. "A Spoken Translation Game for Second Language Learning," Proceedings Artificial Intelligence in Education. 2008
- [23] M Kam, A Agarwal, A Kumar, S Lal, A Mathur, A Tewari, and J Canny. "Designing E-Learning Games for Rural Children in India: A Format for Balancing Learning with Fun," Proceedings DIS '08, 2008.
- [24] <http://3dlanguage.net/cms/>
- [25] WL Johnson. "Serious use of a serious game for language learning," Artificial Intelligence in Education: Building Technology, 2007
- [26] M Mateas, A Stern. "Facade: An experiment in building a fully-realized interactive drama," Game Developer's Conference: Game Design Track, 2003
- [27] JD Bransford, AL Brown, RR Cocking. How people learn: Brain, mind, experience, and school. National Academy Press, Washington D.C. 2004
- [28] DJ Wood, JS Bruner, G Ross. "The role of tutoring in problem solving." Journal of Child Psychiatry and Psychology, 17(2), 89-100. 1976
- [29] CB Cazden. "Adult assistance to language development: Scaffolds, models, and direct instruction." In R. P. Parker & F. A. Davis (Eds.), Developing literacy: Young children's use of language, 3-17. Newark, DE: International Reading Association. 1983
- [30] L Dorn. "A Vygotskian perspective on literacy acquisition: Talk and action in the child's construction of literate awareness." Literacy Teaching and Learning: An International Journal of Early Reading and Writing, 2(2), 15-40. 1996
- [31] AR Luria. "The development of writing in the child." In M. Martlew (Ed.), The psychology of written language: Developmental and educational perspectives, 237-277. New York: Wiley. 1983
- [32] JV Wertsch. Vygotsky and the social formation of mind. Cambridge, MA: Harvard University Press. 1985
- [33] E Bodrova, DJ Leong. "Scaffolding emergent writing in the zone of proximal development." Literacy Teaching and Learning, 3(2), 1-18. 1998
- [34] PM Lightbown, N Spada. How languages are learned. Oxford University Press. 1999
- [35] R Lyster. "Recasts, repetition, and ambiguity in L2 classroom discourse" Studies in Second Language Acquisition, 20/1: 51-81. 1998

- [36] R Lyster, L Ranta. "Corrective feedback and learner uptake: Negotiation of form in communicative classrooms." *Studies in Second Language Acquisition*, 19/1: 37-61. 1997
- [37] ISP Nation. *Learning Vocabulary in Another Language* (Cambridge Applied Linguistics). Cambridge University Press. 2001
- [38] L Cameron. *Teaching Languages to Young Learners* (Cambridge Language Teaching Library). Cambridge University Press. 2001
- [39] P Porter "How learners talk to each other: Input and interaction in task-centered discussions" R. Day (ed.): *Talking to Learn: Conversation in Second Language Acquisition*. Rowley Mass: Newbury House 200-222. 1997
- [40] MH Long, P Porter. "Group work, interlanguage talk, and second language acquisition" *TESOL Quarterly*. 19/2: 207-228. 1985
- [41] B Mathew, Z Fang. "A low-power accelerator for the SPHINX 3 speech recognition system," *Proceedings of the 2003 international conference on Compilers, architecture and synthesis for embedded systems*, San Jose, CA. 2003
- [42] D Huggins-Daines, M Kumar, A Chan, AW Black, M Ravishankar, A Rudnicky. "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," *Proceedings Interspeech*. 2006
- [43] S Horrigan, J Canny. "SphinxTiny: SphinxTiny: Mobile LVCSR Using Continuous GMMs," unpublished.
- [44] P Lamere, P Kwok, W Walker, E Gouvêa, R Singh, B Raj, P Wolf. "Design of the CMU Sphinx-4 decoder," *Proceedings Interspeech-Eurospeech*. 2003
- [45] PC Woodland, CJ Leggetter, JJ Odell, V Valtchev, SJ Young. "The 1994 HTK large vocabulary speech recognition system," *ICASSP-95*. 1995
- [46] H Hermansky, N Morgan, A Bayya, P Kohn. "Compensation for the effects of the communication channel in auditory-like analysis of speech.", *Eurospeech 91*, 1367--1370. 1991
- [47] PP Marsal, SP Font, A Hagen, H Bourlard, C Nadeu. "Comparison and Combination of RASTA-PLP and FF Features in a Hybrid HMM/MLP Speech Recognition System," *Proceedings ICSLP 02*. 2002
- [48] B Milner. "A comparison of front-end configurations for robust speech recognition," *Proceedings ICASSP 02*. 2002
- [49] M Xu, LY Duan, J Cai, LT Chia, CS Xu, Q Tian. "HMM-based audio keyword generation", in Kiyoharu Aizawa, Yuichi Nakamura, Shin'ichi Satoh. *Advances in Multimedia Information Processing - PCM 2004 5th Pacific Rim Conference on Multimedia*.
- [50] S Sutton, RA Cole, J de Villiers, J Schalkwyk, P Vermeulen, M Macon, Y Yan, E Kaiser, B Rundle, K Shobaki, P Hosom, A Kain, J Wouters, D Massaro, M Cohen. "Universal speech tools: The CSLU toolkit," *Proceedings ICSLP98*. 1998
- [51] K Shobaki, JP Hosom, RA Cole. "The OGI kids' speech corpus and recognizers," *Proceedings ICSLP 00*. 2000

- [52] "2003 NIST Language Recognition Evaluation," Linguistic Data Consortium, 1996, <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006S31>, Accessed 5/12/2009
- [53] F Hassanabad. RebeccaAIML <http://rebecca-aiml.sourceforge.net/>, Accessed 5/12/2009
- [54] RS Wallace. "The anatomy of ALICE," in Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer. Springer, 2008.
- [55] The Loebner Prize in Artificial Intelligence <http://www.loebner.net/Prize/loebner-prize.html>, Accessed 5/12/2009
- [56] J Weizenbaum. "ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine", Communications of the ACM 9 (1): 36-45, 1966
- [57] P Taylor, AW Black, R Caley. "The architecture of the Festival speech synthesis system," The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis. 1998
- [58] T Dutoit, H Leich. "MBR-PSOLA: Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database", Speech Communication, 13, 3-4. 1993
- [59] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, O. van der Vrecken. "The MBROLA Project: Towards a set of high quality speech synthesizers of use for non commercial purposes." Proceedings ICSLP. 1996
- [60] RB Rosenberg-Kima, AL Baylor, EA Plant, CE Doerr. "The Importance of Interface Agent Visual Presence: Voice Alone Is Less Effective in Impacting Young Women's Attitudes Toward Engineering," Persuasive07, 2007
- [61] JB Walther, C Slovacek, LC Tidwell. "Is a picture worth a thousand words? Photographic images in long term and short term virtual teams." Communication Research 28 (1), 105-134, 2001
- [62] SS Kahai, RB Cooper. "Exploring the Core Concepts of Media Richness Theory: The Impact of Cue Multiplicity and Feedback Immediacy on Decision Quality." Journal of Management Information Systems 20 (1). 2003
- [63] J Donath. "Mediated Faces." In M. Beynon, C.L. Nehaniv, K. Dautenhahn (Eds.). Cognitive Technology: Instruments of Mind: 4th International Conference. 2001
- [64] SA Golder, J Donath. "Hiding and Revealing in Online Poker Games." Proceedings ACM Conference on Computer Supported Cooperative Work. 2004
- [65] C-C Ho, KF MacDorman, ZAD Pramono. "Human emotion and the uncanny valley: A GLM, MDS, and ISOMAP analysis of robot video ratings." Proceedings of the Third ACM/IEEE International Conference on Human-Robot Interaction, Amsterdam. 2008
- [66] FM Harper, SX Li, Y Chen, JA Konstan. "Social Comparisons to Motivate Contributions to an Online Community" 2nd Int. Conf. on Persuasive Technology Persuasive07. 2007
- [67] F Pajares. "Self-Efficacy Beliefs in Academic Contexts: An Outline" Emory University, <http://www.des.emory.edu/mfp/efftalk.html>. Accessed 5/11/2009
- [68] R Brown. A First Language: The Early Stages. Cambridge, Mass: Harvard University Press. 1973

- [69] JG De Villiers, PA De Villiers. "A cross-sectional study of acquisition of grammatical morphemes" *Journal of Psycholinguistic Research*, 2/3: 267-78. 1973
- [70] JM Meisel, H Clahsen, M Pienemann. "On determining developmental stages in natural second language acquisition" in C. Pfaff (ed.): *First and Second Language Acquisition Processes*. Cambridge Mass. Newbury House. 1981
- [71] J Schumann. "The acquisition of English negation by speakers of Spanish: a review of the literature" in R.W. Andersen (ed.): *The Acquisition and Use of Spanish and English as First and Second Languages*. Washington, D.C.: TESOL 3-32. 1979
- [72] H Zobl. "A direction for contrastive analysis: the comparative study of developmental sequences." *TESOL Quarterly* 16/2: 169-83. 1982
- [73] K Karahalios, K Dobson. "Chit Chat Club: Bridging Virtual and Physical Space for Social Interaction." *ACM Conference on Human Factors in Computing Systems Extended Abstracts*. 2005
- [74] K Karahalios, T Bergstrom, M Yapchaian. "The ISEA Chit Chat Club. Ginger - An Installation." *ISEA San Jose, CA*. 2006
- [75] F Rook. *Slaying the English Jargon*. Society for Technical Communication, ISBN 0-914548-71-9. 1992
- [76] JM Williams. *Style, Toward Clarity and Grace*. University Of Chicago Press, ISBN 0-226-89915-2. 1995