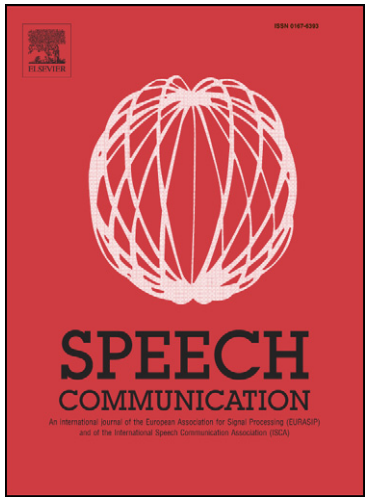
Embodied conversational agents in Computer Assisted Language Learning

Preben Wik, Anna Hjalmarsson



Embodied conversational agents in Computer Assisted Language Learning

Preben Wik, Anna Hjalmarsson

PII: S0167-6393(09)00085-5
DOI: 10.1016/j.specom.2009.05.006
Reference: SPECOM 1812

To appear in: *Speech Communication*

Received Date: 2 July 2008
Revised Date: 18 May 2009
Accepted Date: 19 May 2009

Please cite this article as: Wik, P., Hjalmarsson, A., Embodied conversational agents in Computer Assisted Language Learning, *Speech Communication* (2009), doi: 10.1016/j.specom.2009.05.006

Embodied conversational agents in Computer Assisted Language Learning

Corresponding author:
Preben Wik
Centre for Speech Technology
KTH, Lindstedtsvägen 24
SE-10044, Stockholm, Sweden
preben@speech.kth.se
Phone: +46-8-7906293
Fax: +46-8-7907854

Anna Hjalmarsson
Centre for Speech Technology
KTH, Lindstedtsvägen 24
SE-10044, Stockholm, Sweden
annah@speech.kth.se
Phone: +46-8-7906293
Fax: +46-8-7907854

Abstract
This paper describes two systems using embodied conversational agents (ECAs) for language learning. The first system, called Ville, is a virtual language teacher for vocabulary and pronunciation training. The second system, a dialogue system called DEAL, is a role-playing game for practicing conversational skills. Whereas DEAL acts as a conversational partner with the objective of creating and keeping an interesting dialogue, Ville takes the role of a teacher who guides, encourages and gives feedback to the students.

Keywords: Second langue learning, dialogue systems, embodied conversational agents, pronunciation training, CALL, CAPT

1. Introduction

An important aspect of research on computer assisted language learning (CALL) and computer assisted pronunciation training (CAPT) at the Centre for Speech Technology (CTT), KTH, is its focus on using embodied conversational agents (ECA) for language learning. Using the person metaphor rather than the desktop metaphor as an instructional interface could be beneficial in CAPT for several reasons:

- Users interacting with animated agents have been shown to spend more time with the system, think that it performs better, and enjoy the interaction more compared to interaction with a desktop interface (Walker et al., 1994; Koda & Maes, 1996; Lester & Stone, 1997; van Mulken & Andre, 1998).

- Speech is multimodal and we communicate more than just verbally through our facial expression. It is well established that visual information supports speech perception (Sumby & Pollack, 1954). Since acoustic and visual speech are complementary modalities, introducing an ECA could make the learning more robust and efficient.

- Subjects listening to a foreign language often make use of visual information to a greater extent than subjects listening to their own language (Burnham & Lau, 1999; Granström et al., 1999).

- The efficiency of ECAs for language training of deaf children has been demonstrated by Massaro & Light, 2004. Bosseler & Massaro, 2003 have also shown that using an ECA as an automatic tutor for vocabulary and language learning is advantageous for children with autism.

- ECAs are able to give feedback on articulations that a human tutor cannot easily demonstrate. Augmented reality display of the face that shows the position and movement of intra-oral articulators together with the speech signal may improve the learner's perception and production of new language sounds by internalizing the relationships between speech sounds and the gestures (Engwall, 2008).

We believe that using ECAs for language learning holds great promise for the future of CALL and CAPT. The challenge of making a virtual complement to a human tutor, or classroom teacher, that is infinitely patient, always available, and yet affordable, is an intriguing prospect.

This article describes two systems in which ECAs or computer-animated talking heads are used. Both systems are designed for language learning, but because of their very different roles and agendas, they behave differently, and have different design criteria.

The first system, called Ville, is a virtual teacher, guiding, encouraging and giving corrections and feedback on a student's pronunciation and language use. Ville and the underlying design criteria for the system are described in Section 3. A version of Ville without pronunciation analysis, but with logging and data collection abilities, has also been used by foreign students at KTH; this version is described in Section 4.

The feedback a learner of a new language (L2) receives when talking to a language teacher differs dramatically from the feedback one usually gets when talking to a native speaker. When a student makes a pronunciation error, a teacher or a proficient CAPT system should give explicit feedback on the L2 learner's utterance. In a communicative context however, when two people meet in a real dialogue, the pragmatic content of the exchange is what matters, and if there are pronunciation errors, they are usually not commented on.

The second system, called DEAL, is a role-play dialogue system for conversation training. The ECA in DEAL does not comment on a user's performance, but acts as a conversational partner, with the objective of creating and maintaining an interesting conversation. The feedback one can expect from a conversational exchange is back-channels and, when necessary, clarification questions. Mutual intelligibility is what is sought, and when it fails, communication breakdown is the result.

DEAL can also be seen as a meta-task, within the framework of Ville, of being a diagnostic gate-keeper to the next level of training. (A gate-keeper is a gaming term referring to entities protecting, or guarding gates to new levels in the game that the player must overcome in order to pass to the next level). If the student is able to successfully interact with the ECA in DEAL, it is a sign both to the student and to the system that the student is able to use the content of the previous lesson in a

communicative context. The DEAL system is described in the end of the article in sect. 6.

2. Expressive abilities of the ECAs

The ECAs, or talking heads, that are developed at KTH (Beskow, 2003), have the ability to link phonemes to visemes, thus synchronizing speech with lip movements. The architecture supports both synthetic speech from text (TTS) and pre-recorded utterances. TTS still has some limitations that may affect the CAPT program in a negative way, considering the fact that Ville is supposed to be a pronunciation model for the students. Ville's voice is therefore created from pre-recorded utterances. However, the ECA in DEAL has a TTS voice, since its utterances need to be generated in the course of the dialogue.
The ECAs can also move other parts of the head. Non-verbal signals such as head, eye, and eyebrow movements are used to signal stress, prominence, encouragement, or discourse changes such as turn-taking. The role-plays developed for DEAL should also include an aspect of drama, and the agent in DEAL should thus be able to display emotions such as surprise, anger, or joy.

In order to achieve varied and natural movements of the ECAs, a library of head and face movements has been developed. A sequence of movements (such as raising and lowering the eyebrows) is stored as a '*gesture*', and a sequence of *gestures* can be stored as an '*event*' or in a '*state*'. An *event* is something that happens during a specified time frame, with a start and an end. A nod of the head together with a smile, as a confirmation that the student has done something correct, is an example of an *event*. A *state* is a loosely connected chain of *gestures*, without a defined start or end. The ECA is always in a *state* and will stay in that *state* until some external event causes another *state* to begin. The *state* 'idle' for example, contains several types of blinking with the eyes, slight puckering of the mouth, tilting of the head, slightly turning the head left or right, where every such *gesture* has a weighted chance of occurring. Unless the student is actively interacting with the software, the agent is in the *state* 'idle'. In Ville, there is also a higher level collection of re-occurring 'conversational acts' (for example 'give praise', correct, incorrect) which are *gestures* and pre-recorded utterances with a common semantic meaning. A feedback expression like, for example, '*Correct*' contains several *gestures* where Ville nods his head in various ways, and several pre-recorded utterances like *'Correct', 'Ok', 'Good', 'Yes'*. Because *gestures* and utterances are selected independently of one another, it creates the impression of a larger and more natural variability in Ville's expressive repertoire through combinatorics. Should a version of Ville be created in a new target language, the library of *gestures*, *events*, *states*, and conversational acts can be kept intact, and only the recordings of actual words need to be replaced with equivalent recordings in the new target language.

3. Ville - The virtual language teacher

Ville is a virtual language teacher, who guides, encourages and gives feedback to students who wish to develop or improve their language skills. Ville takes on the teacher's role, selecting the words the students should say. This is a great advantage for the analysis stage in Ville, facilitating the task of correcting pronunciation errors, by having an implicit hypothesis of what kind of pronunciation errors a student is likely to make. The focus of Ville is on pronunciation and perception exercises designed to raise the awareness of particular aspects of the language that are known to

be difficult for many L2 learners. Ville also helps with vocabulary training, providing a model pronunciation of new words, and drilling students in memorization exercises.

The first implementation of our virtual language teacher teaches Swedish to foreign university students at KTH, but our aim is to create a more general language tutor, with placeholders for language specific modules and user specific applications (Engwall et al., 2004; Wik, 2004).

3.1.Contrastive analysis

Our vision is that Ville should be able to detect and give explicit feedback on all types of pronunciation errors that a language student is likely to make. Rather than giving a numerical score for how native-like a student's pronunciation is, which is common in current state-of-the-art CAPT, we want the system to be able to pin-point what type of pronunciation error the student makes in linguistic/phonetic terms. A numerical score will give a quantitative measure of the statistical similarity to the acoustic target, but such measures are difficult for the student to understand, and have been deemed insufficient to help students improve (Neri et al., 2002a). Knowledge-based rules that describe pronunciation difficulties will enable us to give feedback that is more similar to what teachers do, which will be more instructive and easier for the students to understand.

In order to help L2 learners' improve their pronunciation, insight into the types of error they make must first be obtained, and the implementation of detectors for specific pronunciation errors must be based on such insights. This strategy requires an interdisciplinary approach based on language-instruction pedagogy, phonetics, and speech technology. From the experience of language teachers we know that many difficulties L2 learners have are predictable, often based on the influence of their native language (L1). More specifically, difficulties are likely to occur for the learner of a new language if a distinction that carries meaning in the new language (L2) does not exist in the learner's native language. For example, L2 features not used to signal phonological contrast in L1 will be difficult to produce and perceive for the L2 learner (Flege, 1998; McAllister, 1997).

By comparing the inventory of languages and using some form of contrastive phonetic, phonotactic, or prosodic analysis, a list of potential difficulties can be obtained (Ellis, 1994; Meng et al., 2007). Not all features and contrasts in a language are, however, a source for pronunciation errors, and not all pronunciation errors are equally detrimental. A list of priorities is needed when deciding what detectors to include, and where to put the initial focus: Some errors are more common than others, and can be given priority based upon their frequency of occurrence. Some errors are easier for the student to correct than others. Some error-detectors are easier to build than others, and some pronunciation errors are perceived by native speakers as more serious than others, resulting in misunderstandings and communication breakdown.

3.1.1. Common pronunciation errors in Swedish as L2

Bannert, 2004 has investigated pronunciation difficulties in second language learners from 25 L1 languages, with Swedish as target language. Some of the most serious errors according to Bannert are:

- lexical stress
  (insufficient stress marking, or stress on the wrong syllable)
- consonant deletion in a consonant cluster before a stressed vowel

- vowel insertion (epenthesis) in, or before a consonant cluster
  (/spanien/ ("Spain") becomes /espanien/)
- vowel quantity
  (The duration of a stressed vowel is neither long nor short)
- vowel quality
  (difficulties with Swedish vowels not present in L1)
- liquids
  (no difference is made between /l/ and /r/ sounds)

Our initial work on creating detectors for anticipated pronunciation errors is based on Bannert's work. Our focus has been on designing detectors for segmental errors that are possible to analyse using forced alignment, rather than statistical methods that require large amounts of training data that is not yet available. Other types of detectors are anticipated when our L2-Swedish database described in Section 4.1 has been created.

3.2.Pronunciation error detectors in Ville

The sound processing in Ville is to a large extent done using Snack, an open source sound processing tool developed at KTH [1], in conjunction with N-align, the CTT aligner tool (Sjölander, 2003). Our current detectors are built on top of these components.

3.2.1. Vowel quantity detector

Students practice on carefully selected word-pairs, where the vowel duration of the stressed syllable changes the meaning of the word. Vowel quantity errors (failing to make a distinction between long and short vowels), are detected using forced-alignment, by identifying and time-marking phones, based on a transcription of what is being said and the speech wave of the student recording. The time segments are then normalized, and compared with a reference recording of the word. The Swedish language has complementary distribution, (i.e. a long vowel is followed by a short consonant and vice versa), so the relative duration of the following consonant must also be taken into consideration.

3.2.2. Lexical stress detector

Lexical stress is manifested by giving one of the syllables in a word more prominence. Commonly measured acoustic correlates to stress are pitch, intensity and duration. In Swedish, duration has been considered the most important correlate, which has also been noted in Dutch and English (Sluijter, 1995). Pedagogically, it therefore makes sense to encourage L2 learners to use duration as the predominant variable. Since different phones have different intrinsic durations, one syllable may be longer than another without being perceived as prominent. This intrinsic duration must be normalized away before syllable length can be used as a measure for lexical stress. The lexical stress detector uses average phone durations measured from a number of speakers (Carlson et al., 2002), a syllable divided transcription, and forced-alignment to estimate which syllable is the most prominent in a student recording.

[1] Snack: downloadable at http://www.speech.kth.se/snack/

This estimate is then compared with a reference in order to determine if the student has placed the stress on the correct syllable or not.

3.2.3. Insertion and deletion detectors

Insertion and deletion errors are predictable in the sense that a mispronunciation hypothesis can be created in conjunction with certain consonant clusters. For example, many native Spanish speakers will produce a consonant cluster with an initial /s/ in Swedish by inserting a vowel in front of the /s/: 'Stockholm' thus becomes 'Estockholm'. Both our insertion and our deletion detectors employ forced-alignment in parallel on both the original transcription and on a modified transcription, where the hypothesised insertion or deletion is included. The aligner returns a distance measure (using viterbi search). By comparing the distance of the original transcription from the hypothesis transcription, we can determine the likelihood that an insertion or deletion was made.

3.3.Perception exercises

The first step in learning a new sound contrast is to be able to perceive it. If someone is unable to perceive a linguistic contrast, they are not likely to be able to produce it correctly either. The perception component of Ville does not contain speech technology components, but is still a very important part of the system from a pedagogical point of view. The exercises are designed to raise the awareness of the specific sound contrasts that are known to be difficult for many L2 learners. Minimal pairs are very useful for intuitively exposing learners to contrasts that exist in the target language (L2) but not in their native language (L1). In minimal pair exercises for vowel quality for example, a pair like /bita/-/byta/ ('bite' vs. 'swap') is presented on the screen, and Ville will randomly say one of the words. The student's task is to identify which word was uttered, and click on it. Ville will then give verbal feedback on the student's choice. Lexical stress exercises are performed in a similar fashion, by Ville saying a word and the student selecting the word type that was uttered. The classes are in this case not binary pairs, but a three by three grid with symbols representing different stress patterns and number of syllables, c.f. Figure 1.

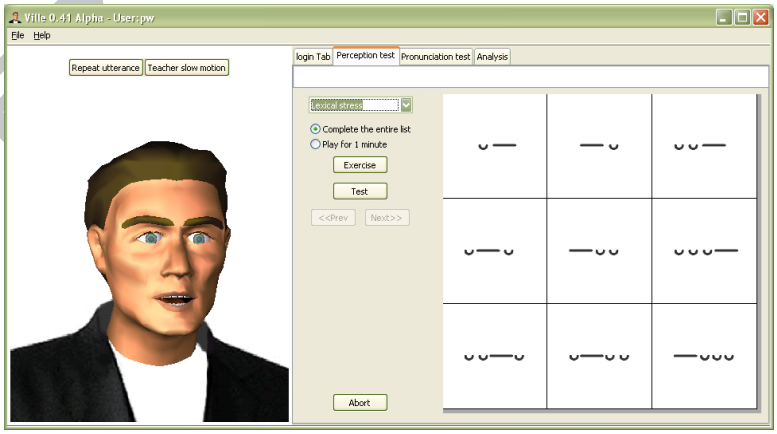


Figure 1. A grid for lexical stress perception exercise: u is an unstressed syllable, — is a stressed.

3.4.Vocabulary

The central importance of vocabulary knowledge for language competence is clear. Exercises for vocabulary training are therefore implemented as part of the virtual language teacher. 'Flashcards' are a popular and efficient vocabulary acquisition tool,

where a deck of physical cards is normally used. A word is written in the L1 on one side of the card and in the L2 on the other. The card can be used two ways: Look at the L1 word and translate it to L2, or look at the L2 word and translate it to L1. The cards the student has learned can be discarded, so that they focus on the words they find difficult. The same paradigm is utilized in Ville, but with voice and pictures added to every word. Students can click on a card and hear Ville pronounce the word, make their own recordings of the words, and get feedback on their pronunciation.

Figure 2. Picture of the flashcards and recording section of the program

3.5.Feedback and when to intervene

One of the most important aspects of being a teacher, virtual or real, is to give correct and relevant feedback. It is not easy for a virtual language teacher to know when it is appropriate to give feedback, how verbose it should be, and when it is better to refrain from talking. Having the opportunity to give verbal, multimodal feedback, does not in itself mean that it is always the best thing to do.

In an early version of Ville, pronunciation exercises included verbal feedback, in which Ville commented on the results from the detectors. In a vowel length exercise for example, Ville could say "Good, but your /a/ was a bit too long - try again, say: badda". Reactions from beta testers of the system revealed that such interventions were at first perceived as good, but that they soon became irritating and tiresome. This is similar to findings of Eskenazi, (1999), who stated that “Interventions can appear to users as being either timely or irritating. Bothersome interventions tend to be caused by either recognition errors or by a system that intervenes too frequently and is too verbose.”

We have hence implemented this part of the system so that, rather than using verbal feedback, the result from the pronunciation detectors is shown as iconic ‘traffic lights’. Red or green circles will light up for the active detectors after a student recording Figure 3

.

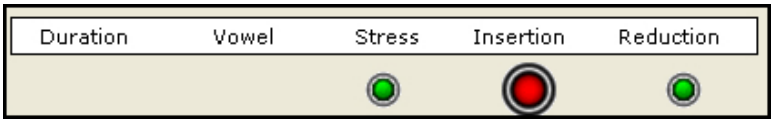


Figure 3. Feedback on pronunciation is displayed with red (large) and green (small) lights.

The advantage with this type of visual feedback is that many lights can be shown in parallel, and the student will quickly be able to get an overview of how his performance was rated. If the student wishes to know why one of the circles was red, he can click on the circle, and a new page will appear with more detailed information such as graphs, or spectrograms, accompanied by the verbal feedback (Figure 4). If, on the other hand, this is a recurring error, and the student feels that he has already understood the information, he can simply note that the visual feedback indicates that the error is still occurring, and move on.

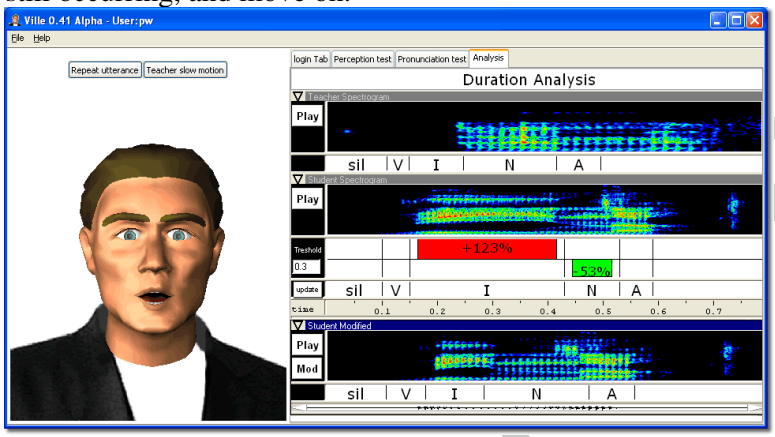


Figure 4. Detailed feedback on duration analysis appears only if the student actively chooses to see it.

Giving correct and relevant feedback is, however, a skill that many real teachers will spend years to develop, and that requires a great deal of intelligence, creativity, and sensitivity to master. Some teachers are clearly more successful than others at it and much more research is needed on how to successfully apply successful feedback didactics in CALL and CAPT (Engwall & Bälter, 2007; Neri et al., 2002b).

4. Ville 1.0

Foreign students at KTH can study Swedish as a second language at the Unit for Language and Communication. The growing demand for classes, due to the large increase in foreign Master students in recent years, have inspired the language unit to seek new methods in their language training strategies. As a complementary self-study resource the unit has created a web-based beginner course in Swedish called SWELL (Swedish for Elementary Learners). For pronunciation and vocabulary practice, a version of Ville has also been offered since May 1st 2008, in a first attempt to test Ville on real students, with real needs to learn a new language.

The Unit for Language and Communication wanted initial focus to be on vocabulary and provided pedagogical expertise to provide a training vocabulary of approximately 750 words and pictures divided into 27 lessons. This version of Ville contains no speech analysis, and no feedback is given on pronunciation (since analysis and identification of errors is a prerequisite for giving that type of feedback). Students are offered three types of exercises in this version.

Perception exercises: Ville says a word and the student must identify the corresponding picture (as described in Section 3.3)., Listen-and-click vocabulary exercises divided into perception tests, perception exercises, and perception games are offered (see Figure 5).

Pronunciation exercises: Students are offered flashcards (as described in Section 3.4) and can record and play back their own recordings for self-evaluation.

Writing exercises: Spelling exercises are also offered, where Ville says a word and the students should write down what they hear.

Figure 5. A vocabulary perception exercise in Ville for SWELL

4.1.The need for a data collection

In order to develop new detectors for the pronunciation analysis part of the program, (as described in Section 3.1) we need to obtain speech data which contains typical pronunciation errors. For future development, we also see the need for a student profile, that is, a way to structure and store a student's performance and improvements, in order to create a lesson manager that can tailor lessons to a student's individual performance. These program enhancements are, however, difficult to develop without first having data on which to model the analysis, and such data is difficult to obtain without having the environment in which it occurs. Offering Ville to students within SWELL gives us the opportunity to collect real data from real students, and a way to resolve this dilemma.

Our approach is inspired by Lois Von Ahn's "Human Computation", and "Games with a purpose". von Ahn (2006) builds games that, when played by humans, produce useful computation as a side-effect. Through online games such as, for example, the ESP game[2], Verbosity[3], and Tag a tune[4], people are collectively solving large-scale computational problems in diverse areas such as image labelling, security, computer vision, Internet accessibility, adult content filtering, and Internet search. His players are offered entertainment, and provide researchers with brain power in return. Similarly, our users are offered education (hopefully with some entertainment value), and provide recordings and perception data in return. In both cases we are able to obtain user generated data for free, once the system has been built, in exchange for entertainment or education.

[2] http://www.gwap.com/gwap/gamesPreview/espgame
[3] http://www.gwap.com/gwap/gamesPreview/verbosity
[4] http://www.gwap.com/gwap/gamesPreview/tagatune

4.2.The Swedish L2 database

A data collection tool has been implemented in the version of Ville being offered to KTH students. All users were informed that they were taking part in a research project, and that, in order to use the program, they needed to give their consent that all data collected could be freely used for research purposes at KTH.

Every time a user makes a recording using the flashcards described in Section 3.4, the recording is saved locally on his computer, but it is also sent to our server. In addition to the actual speech files, an XML-file is created in which the text of the word spoken, the transcription, recording-style, and other information. are also stored. Since the user-generated data is stored in the same format as the content of the program, the student's own recordings can be retrieved and played back within the program, either by the student himself, or later by a researcher or a teacher.

The database currently being created by L2 learners of Swedish will eventually contain many instances of the same words spoken by different users. Many of the words are selected to highlight particular difficulties, and will be a good testbed for future pronunciation-error detectors. Recordings of the same word spoken by the same user at different times will further permit comparative studies of the user's improvement over time. Since all recordings are automatically annotated and saved, together with biometric data such as L1, sex, and age, this speech corpus could also in the future become a useful resource for other types of research. It could for example be used for comparative analysis between users of different groups (sex, age, country of origin, etc), or for ASR robustness or adaptation to non-native speech.

Perception exercises are also logged, and will be a valuable resource for determining which aspects of language learning the students are having difficulty with, and where to put the focus in future tool development. We can also determine which words are most difficult with regard to spelling or listening comprehension. We can also look at the distribution of exercises and recordings over time, per student, or per chapter. A synergy effect can be achieved when developing tools to analyze student performance on this level, since they will be similar to what is required for the lesson manager described in Section 4.1.

As of January 1st 2009, 402 students from 66 different countries have used Ville. The distribution of students grouped by continents is shown in Table 1. The users include both foreign students at KTH, (who have tried the software as part of the SWELL course) and external interested web users, who have downloaded the program without following a Swedish course at KTH.

Continent	Students	Countries	Perception Exercises	Writing Exercises	Recordings
Asia	170	20	523	79	1277
Europe	156	28	1345	574	4897
Africa	18	7	189	17	61
North America	7	1	77	8	92
South America	10	6	128	24	203
Central America	3	2	7	0	0
Oceania	4	2	8	0	0
Unknown	34	-	130	8	29
Total	402	66	2407	710	6559

Table 1: The number of exercises and recordings made by the students (grouped by continent)

Table1 shows that, although a substantial amount of users have tried the program, the number of exercises per user is in many cases very low. (170 students from Asia have for example only completed 79 writing exercises, which amounts to less than ½ exercise per student) This may in part be explained by users for whom the aim was not to learn Swedish, but who were merely curious to see what types of exercises and interaction that Ville offered. It may however also indicate that some users who did intend to use Ville to learn Swedish abandoned the program after a short time.

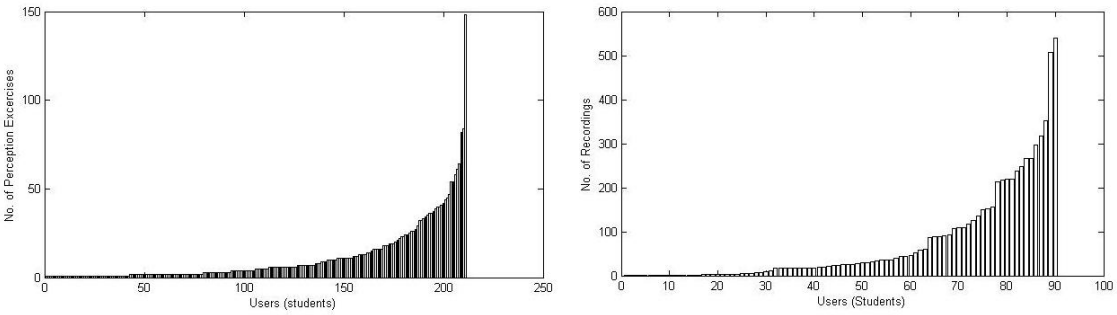


Figure 6. The distribution of perception exercises (left) and recordings (right) done per student.

The graphs in Figure 6 show that the total number of perception exercises (left) and recordings (right) are not evenly distributed among the students. Quite the contrary, it is reminiscent of a Zipfian distribution, where some students have been very active, while the contribution of the majority of users is quite small. The writing exercises follow a similar pattern. Only the 'active' students, i.e. those who have completed at least one exercise or recording are shown in the graphs. Note that, out of the 402 users, not more than 210 users have done any perception exercise, and only 90 users have made recordings.

4.3.Questionnaire

In January 2009 we also sent out a request to all the students at KTH who had used Ville, asking them to fill out a web-questionnaire in order to get some qualitative feedback. The questionnaire was anonymous, so we are not able to link student opinions to their performance. It focused instead on general questions that could give us indications of how students perceived the system, in order to help us improve it in the future.

Table 2 gives a summary of the questionnaire. All questions were multiple choice questions with the range 1–5 where 1 means "I do not agree at all" and 5 means "I totally agree". 59 students filled out the questionnaire.

Question	Mean	St.Dev.
1. Ville is an effective tool for practicing Swedish at home.	3,91	1,09
2. I feel more at ease practicing pronunciation with Ville than with a real teacher.	2,91	1,13
3. I felt at ease doing recordings, even though I knew that my recordings were logged for research purposes.	3,36	1,14
4. The talking head is a valuable feature of the program.	3,70	0,96
5. The talking head makes it easier to understand how new words are pronounced.	3,53	1,07
6. The possibility to zoom in on the mouth is a valuable feature of the program.	3,44	1,05
7. Ville's voice is clear and easy to comprehend.	3,76	1,00
8. The movements of Ville's lips correspond well with his speech (lip-sync).	3,43	0,89
9. The user interface (GUI) is intuitive and easy to use.	3,45	1,03
10. It is easy to see what the pictures in the program represent.	3,65	1,08
11. The vocabulary practiced in the program is well chosen.	3,76	0,86
12. User statistics (login page see above) is a valuable feature of the program	3,70	0,98
13 The perception exercises are useful.	3,93	0,90
14 I think the program has helped me improve my Swedish listening comprehension.	3,38	1,10
15 The pronunciation exercises are useful.	3,65	1,03
16 I think the program has helped me improve my Swedish pronunciation.	3,43	1,06
17 The listen and write exercises are useful.	3,60	1,03
18 I think the program has helped me improve my Swedish writing skills.	3,29	1,10

Table 2. Questionnaire and mean opinion, in terms of mean score and standard deviation.

There was also a free comment field where students could include their comments: Apart from some informative 'bug-reports' on misspellings and so on, we also got some great reviews. Here is a selection of comments:

- *"Generally i see Ville is very useful tool in learning Swedish words writing and pronunciation , but the only thing needs attention is the user interface , it is not so user friendly when choosing between tests and perception and other options"*
- *"It is really great, and helps a lot in learning the Swedish language."*
- *"I have mostly used Ville to learn vocabulary and it was really useful. Maybe there should be something like saying the word in English and we have to write it correctly after that."*
- *"This is an extremely helpful program for beginners to get a know-how of Swedish language, especially pronunciation, vocabulary etc."*

We can conclude that the system has been well received, at least by the students who filled out the questionnaire. It is interesting to note that the lowest score was on question 2 "I feel more at ease practicing pronunciation with Ville than with a real teacher" (reminding us of the programs role as a complement, not a replacement for a human teacher). Also note that people thought the program had helped them improve their pronunciation skills more than their listening and writing skills, even if many students made no, or very few recordings.

The number of Ville users and the amount of data is continuously increasing, but still limited. In order to get more data, we need both to increase the number of users, and to look at new ways to motivate the users we have to use the program more frequently.

5. Motivations for language learning

Learning a language requires a substantial effort, and the motivation for doing so varies both over time and between individuals. People learn a language for different reasons: A wish to be like the speakers of the language (integrative motivation) is often a strong motivating factor for younger learners, whereas the utility of what is learnt (instrumental motivation) is often a stronger motivator for adults. Motivation can also come from the pleasure of learning (intrinsic motivation), or from the task itself (task motivation) to mention some sources. Since the time spent in training is one of the key factors for language learning, increasing motivation and the time students spend practicing are crucial for a successful CALL application.

Game designers focus on finding ways to keep the players engaged and have in their pursuit of success, developed several effective design strategies both to get and to keep players engaged and motivated throughout a game. This is known as good gameplay. According to Prensky (2002) "Gameplay is all the doing, thinking and decision making that makes a game either fun or not". Good gameplay is what makes games addictive, and what makes millions of people spend a significant amount of their time and money on playing games. The pleasure of engagement is the motivation to play.

6. DEAL - A role-playing dialogue system for L2 learners

The same design principles that are used by game developers are starting to find their way into other fields as well. The notion of 'serious games' is an initiative focusing on using game design principles for purposes other than solely to entertain, for example training, advertising, simulation, or education (Iuppa & Borst, 2007). Good gameplay adds to any existing motivation to learn if there is one, and may otherwise create motivation by itself. The idea of transforming education and creating more engaging educational material by looking at the games industry has been suggested and described by several authors, for example Gee (2003) and Prensky (2001). To make our CALL and CAPT applications more effective we aim to look at new ways to motivate our users.

The DEAL dialogue system has been developed to serve as a multidisciplinary research platform in the areas of human-like utterance generation (Hjalmarsson et al., 2007), game dialogue (Brusk et al., 2007), and language learning (Wik et al., 2007). Our aim is to allow language learners to practice conversational skills in a fun and challenging context. Our objective is similar to that of the Tactical Language Training System (TLTS) (Johnson et al., 2004), in the sense that both systems are simulation games for the acquisition of language and cultural skills. However, where TLTS places focus on realism (teaching US military appropriate manners and phrases to be used on foreign ground), we wish to focus more on entertainment. In that respect our objective is closer to Façade, a one act interactive drama where the player's interaction affects the outcome of the drama, and where the goal of the interaction is to create a good story (Mateas & Stern, 2003). Our objective is also similar to that of the Nice project (Gustafson et al., 2004), in that we wish to create a game in which spoken

dialogue is not just an add-on, but is the primary means for game progression. Our ultimate goal is however on language learning.

6.1.Reasons for talking affects the design

Dialogue systems have traditionally been used mostly for information-seeking applications, such as train-schedule or weather-information systems. DEAL however, is a different type of system that is aimed at tutoring and entertainment. While the information-seeking type of system needs to be polite, unobtrusive and efficient, DEAL needs to be none of these things. Since the reason for talking to a dialogue system such as DEAL differs from an information-seeking application, the design criteria, and evaluation criteria will also be dramatically different. As an analogy we might consider some different reasons for walking.

1. One reason for walking might be to get somewhere, that is, the walking in itself is a hindrance for the real act, which is to arrive at the destination. Taking a shortcut or taking the car could be strategies to make it more efficient.
2. One could also choose to walk for health reasons, i.e. as a way to exercise. Then the walking in itself would be the reason for being on the road. Whether one enjoyed the walking or not, taking a shortcut would not be an option, since it would defeat the purpose of the walk.
3. Finally, one might wish to take a walk just for the pleasure of walking. A stroll in the park or a walk in the forest simply because it is an enjoyable and fun thing to do. Taking the car would be out of the question, since that would make one miss all the fun.

The first example is the walking equivalent of an information-seeking dialogue system. In such a system the user is not interested in the talking part, but wants to get an answer as quickly as possible. The talking equivalent of the second example could be to practice talking for language learning reasons. In a truly successful language learning application, the interaction would be so entertaining that the user would consider it to be a combination of 2 and 3: A game where the reason for playing would be just because it is fun (and at the same time getting useful exercise).

6.2.Evaluation of dialogue systems

As for evaluation, an information-seeking spoken dialogue system will be judged by factors such as efficiency in reaching task completion. A good system will thus try to minimize the number of turns needed (shorten the path). For DEAL the aim would be the opposite. The longer the conversation takes, and the more turns between the user and the system, the better. The interaction between the agent and the user – if successful, will take the form of a role-play, and user satisfaction will depend on things other than efficiency in task completion. Apart from creating a good story, the social competence and personality of the character may be considered important factors, as well as the response time of the system, and how well the system handles errors.

The criteria a non-native speaker (NNS) has for judging a dialogue system are different compared to a native speaker (NS). When a misunderstanding between a user and a spoken dialogue system occurs, a NS knows he has done nothing wrong, and will ascribe the misunderstanding to a weakness in the system. A NNS on the other hand will often be critical of his own ability in the new language, and might instead ascribe the misunderstanding to his own pronunciation, or incorrect use of

vocabulary and grammar. A NNS will in a similar way be able to reason that if he is able to communicate with the system without communication breakdown, it can be seen as a confirmation of his abilities to communicate in the new language.
This is a situation where the difficulties of ASR in combination with foreign accent could be re-interpreted as a difficulty within the gameplay. Rather than trying to adapt the ASR to be able to handle the strong accent, it becomes a part of the challenge of passing the gate-keeper. This way of motivating students to adapt their pronunciation is reminiscent of what happens in the real world, and may serve as an implicit learning factor. The limitations of the ASR can then be read as a measure of the student's communicative skills, where the challenge is to be able to negotiate with an ECA. If the student fails, he will not be allowed to move to the next level, and must go back to Ville - the pronunciation teacher, who will provide the student with the exercises and feedback she needs in order to try again, and hopefully pass the gate-keeper at a later time. In DEAL the agent can also choose to avoid acknowledging non-understandings or possible misunderstandings by tactical maneuvers, such as counter-questions or changing the subject.

6.3.Domain

DEAL serves as an important complement to Ville; whereas Ville provides exercises on isolated speech segments, i.e. phone, syllable, word, and sentence level, DEAL adds the possibility of practicing these segments in the context of a conversation. Ville has the role of a teacher who gives you feedback and help when you encounter problems. DEAL on the other hand has the role of a native speaker, for example, a person with a service occupation, whom you need to communicate with using your new language. There are many potential everyday situations in which one may want to use a new language. Good choices of domains include situations that follow familiar *schemas* or *scripts*, i.e. episodic knowledge structures that guide us in our daily interactions. The choice of domain for the first implementation of DEAL is a trading domain. Our first trading scenario is in a simulated flea market. This domain was chosen for several reasons:

- A trading situation is a fairly restricted and universally well-known domain. It is something everyone is conceptually familiar with, regardless of cultural and linguistic background.
- It is a very useful domain to master in the new language
- The flea market allows for, and almost invites characters who are eccentric or otherwise out-of-the-ordinary in an interesting way.
- A flea market is a place where it is common to negotiate about price and to trade items. This type of negotiation is a complex process which includes both rational and emotional elements.
- The shop can include almost any type of item. (In a larger framework vocabulary just learned in Ville can easily become items in the shop).
- Second-hand items may have rich interesting characteristics or be defective and thus invite another type of conversation.

In summary it is a domain in which the user can engage in a dialogue that is well known but still includes elements of surprise, social commitment and competition (i.e. getting a good price).

6.4.Scenario

The basic teaching plan is for the language student is to use Ville in conjunction with DEAL. First Ville will teach the rudimentary vocabulary that is associated with the trade domain - that is: The numbers, some colors, a few objects like a clock and a teddy-bear, and a few phrases like "Do you have..." "How much does that cost" and so on. The student is then given a task to go to the nearby flea-market and use his newly acquired vocabulary in order to buy a given set of items from the shop-keeper in DEAL. The student is given a certain amount of money, but the money will not be enough to buy all the items on the student's list, unless he is creative. The stingy shopkeeper in the flea-market will try to get as much as possible for his goods. This scene can then unfold in different ways depending on what the student says. The willingness of the ECA to reduce the price of an item for example, may be affected by how the user gives praise or criticizes an item of interest, as, for example, in the dialogue below.

```
U1: I'm interested in buying a toy.
S1: Oh, let me see. Here is a doll.(a doll is displayed)
U2: Do you have a teddy-bear?
S2: Oh, yeah. Here is a teddy-bear. (a teddy-bear is
displayed, see Figure 2)
U3: How much is it?
S3: You can have it for 180 SEK
U4: I give you 10 SEK.
S4: No way! That is less than what I paid for it.
U5: Ok how about 100?
S5: Can't you see how nice it is?
U6: But one ear is missing.
S6: Ok, how about 150?
U7: 130?
S7: Ok, it is a deal!
```

Dialogue example 1

6.5. DEAL architecture

The dialogue system component in DEAL is based on Higgins, a spoken dialogue system developed at KTH (Skantze (2005)). Higgins includes modules for semantic interpretation and analysis. Pickering, a modified chart parser, supports continuous and incremental input from a probabilistic speech recognizer. Speech is unpredictable and chunking a string of words into utterances is difficult since pauses and hesitations will probably be incorrectly interpreted as end of utterance markers. This will be even more evident for second language learners whose conversation skills are not yet good and whose language contains disfluencies such as hesitations and false starts. Pickering uses context-free grammars and builds deep semantic tree structures. Grammar rules are automatically relaxed to handle unexpected, ungrammatical and misrecognized input robustly. The discourse modeler (DM), Galatea, interprets utterances in context and keeps a list of communicative acts (CA) in chronological order. Galatea resolves ellipses, anaphora, and has a representation of grounding status which includes information about who added a concept, in which turn a concept was introduced, and the concept's ASR confidence score. The system also contains an action manager (AM) described in 6.7, a communicative manager (CM), modules for text and text-to-speech generation (all described in 6.8) and a user interface (described in 6.6).

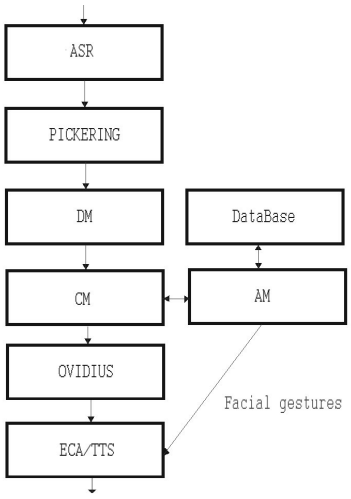


Figure 7. The DEAL architecture

6.6.The DEAL user interface

The user interface in DEAL is divided into six parts (Figure 8). The top part contains the shopkeeper, an ECA with the same expressive abilities as Ville, (described in Section 2) but with other attributes and characteristics and a very different agenda. The middle part of the user interface portrays the shop-counter, where any objects discussed between the user and the shopkeeper are shown, and where the financial transaction takes place if the negotiation results in an agreement. The pictures also give clues about the scope of the domain, that is, what can be talked about. Below the counter is a "notebook" with four tabs. The info-tab contains hints about things the user might try to say if the conversation has stalled. The wallet-tab contains the money the user has at his disposal. The things-tab holds a picture of all the items the user has managed to acquire. Finally, the text input-tab offers a text input field as an alternative to the automatic speech recognition.

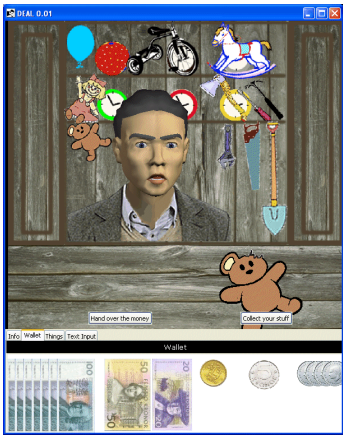


Figure 8. The DEAL interface with a stingy shopkeeper trying to sell a teddy-bear with a missing ear.

6.7.Action management in DEAL

Action management in DEAL, i.e. deciding what to say next, is currently done according to a set of simple rules based on the *script,* or episodic knowledge structures that guide us when we interact with a shopkeeper in a shop in order to buy a product. Communicative acts used in DEAL include OBJECT-REQUEST, PROPERTY-REQUEST, PRICE-REQUEST, SUGGEST-PRICE, DEAL and so on. Consequently, the user can request objects, ask about object properties, give price offers, and make deals. The haggling algorithm is a set of simple heuristics. These are based on the amount of the offer the user is suggesting in relation to the "retail price" which is stored in the system database. All objects in the database have obvious visual defects (e.g. the missing ear in Figure 8) and if detected and pointed out by the student, the agent reduces the price as shown in Dialogue example 1. The goal in DEAL is to build an emotional agent who is able to take initiative in the dialogue, if the student should fail to do so. As a first step in this development the agent looks angry or happy based on how the dialogue progresses. The agent responds with a smile to greetings and closings of deals. However, after long sequences of haggling or price offers from the student that are too low (less than 10% of the agent's initial price suggestion), the agent looks angry. The agent also takes initiative if no user input is provided, trying to bring the dialogue to a close. The action taken is based on the dialogue state; for example, if an object is in focus (on the table), the agent suggests a price for that object, and if no such object exists a new object is presented. An important characteristic of the system is that the goals of the agent and the student partly differ. Both have the goal to complete a successful interaction; however the agent wants to sell objects for as much as possible while the student wants to buy them for the lowest possible price. In terms of gameplay, buying an object for a certain price must be challenging. To make the bargaining trickier, the agent easily gets "fed up". After a fixed set of speaker turns haggling about the price of a certain object the agent claims to be bored and refuses to discuss that object anymore. Instead, he suggests a new one.

6.8.Human-like language generation

DEAL is still under development and yet to be evaluated. Since the agent's behavior is crucial for how the system is perceived, the main effort so far has been to try to build a system that can generate utterances in a human-like manner. To encourage the user to talk to the system as if talking to another human being the agent needs to be responsive and flexible. Long response times and simple non-flexible utterances using templates or pre-recorded speech are not acceptable, if we are aiming for a system with a dialogue that is diverse and engaging.

Humans produce speech incrementally and on-line as the dialogue progresses using information from several different sources in parallel (Brennan, 2000; Aist et al., 2006). We anticipate what the other person is about to say in advance and start planning our next move while this person is still speaking. When starting to speak, we typically do not have a complete plan of how to say something, or even what to say. Yet, we manage to rapidly integrate information from different sources in parallel and simultaneously plan and realize new dialogue contributions. To keep the response times constant and without unnaturally long delays, the dialogue system needs to be capable of grabbing the turn, hold it while the system is producing the rest of the message, and release it after completion. A corpus of human-human dialogues in the

DEAL domain was collected in order to study different human strategies on how to rapidly grab and maintain the floor while not knowing exactly what to say (Hjalmarsson, 2008). The data collection revealed a frequent use of linguistic devices often referred to as *cue phrases* or *discourse markers*. The function of these devices are to signal how new segments of speech relate to previous segments both within and between speaker turns. 81% of all speaker turns in the corpus contained at least one cue phrase and 21% of all words were labelled as cue phrases. Strategies that were frequently used to buy time while processing input was to generate filled pauses (e.g. eh, ehm), repetitions, or prosodic phrase-final lengthening. Incremental language production also requires that the system know how to alter, repair or refine previous utterances, such as when the system generates a response that was committed to too early and needs to revise it.

The generation task in DEAL is distributed over different modules. The communicative manager (CM) (see Figure 7) is responsible for detecting when it is appropriate for the system to speak and immediately initiates a new turn based on an early hypothesis from the input, or, if no such hypothesis is available, a grounding fragment or a filled pause. The CM also acts as an error-handling filter if the system is uncertain of some part of the incoming message. If the ASR confidence score for a particular entity is below a certain threshold, the CM generates a clarification CA without passing the message on to the action manager (AM), asking the user to clarify the entity before proceeding with the CA analysis any further. The system also utilizes the time it takes to make a complete analysis of input to ground the user's previous utterance (e.g. "ok a green watch"). This is done regardless of the fact that the availability of this object is unknown (i.e. the object could already be sold or not exist in the database). If the object for some reason turns out to be unavailable, the system revises its previous grounding segment and suggests another object (see dialogue example 2, *s1a* and *b*).

```
U1:     I want to buy a green watch.
S1a:    Ok, a green watch…
S1b:    … I'm sorry there is no green watch but I do have a
        red one.
U2:     Do you have a yellow one?
S3a:    mm a yellow watch…
S3b:    … here is one
```

Dialogue example 2

If the user input contains no reference to a particular entity, the CM generates neutral feedback such as "yes" or "ok". Since the DM not only keeps track of the user's CAs but also its own previous CAs the CM can modify new responses from the AM based the type of feedback used in the first turn segment. An object that has already been grounded with a full noun phrase is referred to with a pronoun in the second part of the system response (see dialogue example 2, *S3a* and *b*).

The AM is responsible for deciding which action to take based on new input from the user, or, if no input is detected, it initiates an action based on the previous dialogue state. When the AM has generated a response it is passed on to the CM, which is responsible for modifying the response based on the previous dialogue context. For example, the CM decides how entities should be referred to, e.g. determines whether to use referring expressions or full noun phrases, as well as turning full propositions

into elliptical constructions. The decisions are based on how well the entities are *grounded* in the dialogue, based on the confidence scores from the ASR and on whether these entities have been previously mentioned.

The CM forwards its message to Ovidius, the module responsible for realising the textual representation. Ovidius takes a system CA as input and generates a text that is subsequently realised acoustically by a speech synthesiser. Ovidius uses a set of template rules, working much like inverted Pickering grammar rules – they match the semantic tree structures and produce text strings. The acoustic realisation in the current version of DEAL is a combined set of pre-synthesized prompts and on-line text-to-speech generation. Feedback and other cue phrases as well as filled pauses are pre-synthesized prompts while the rest of the dialogue is synthesized speech generated online. The pre-synthesised elements have prosodic features, including F0 contour, speaker rate and energy, automatically extracted from the DEAL corpus. Since these elements are so frequently used, variation is essential or else the agent will sound too monotonous. An instance is randomly selected from a library of pre-synthesized prompts with the corresponding functional labelling (e.g. neutral-feedback, filled pause and so on). (For a more detailed description of the different functional classes see Hjalmarsson, 2008). Whether the functions of these elements are interpreted by users as intended is, however, still to be evaluated.

7. Conclusions

Both the Ville and the Deal systems described in this article are ongoing projects, under constant revision and development. We intend them to serve as platforms for research in narrower sub-disciplines such as utterance generation, error handling and turn-taking in dialogue systems, and as testbeds for research on feedback strategies, pronunciation detectors, and other CALL-related research areas.

The release of Ville to real students has been quite successful. Although it has officially only been tried within the KTH campus so far, more than 400 students from 66 different countries have downloaded and tried the program, and most of the students who filled out the questionnaire were satisfied with the program.
The method of collecting data (by offering education in exchange for student data) has proven very effective. A substantial amount of data has already been gathered, and new data is being generated continuously.

We have so far only scratched the surface of a large inter-disciplinary research area which has great potential to converge into something with a big impact on a large population group.

For those interested to read more about Ville and DEAL, the web sites www.speech.kth.se/ville and www.speech.kth.se/deal provide additional information and links to other publications.

8. Acknowledgements

This research was carried out at the Centre for Speech Technology, KTH. The research is also supported by the Graduate School for Language Technology (GSLT). Many thanks to Jenny Brusk for her work in the DEAL project. We would also like to

thank Björn Granström, Rolf Carlson, Olov Engwall, Jens Edlund, and Julia Hirschberg for their valuable comments.

References:

Aist, G., Allen, J. F., Campana, E., Galescu, L., Gómez Gallo, C. A., Stoness, S. C., Swift, M., & Tanenhaus, M. (2006). Software Architectures for Incremental Understanding of Human Speech. In *Proceedings of Interspeech* (pp. 1922-1925). Pittsburgh PA, USA.

Bannert, R. (2004). *På väg mot svenskt uttal.* Studentlitteratur AB.

Beskow, J. (2003). *Talking heads - Models and applications for multimodal speech synthesis*. Doctoral dissertation, KTH, Department of Speech, Music and Hearing, KTH, Stockholm.

Bosseler, A., & Massaro, D. W. (2003). Development and Evaluation of a Computer-Animated Tutor for Vocabulary and Language Learning for Children with Autism. *Journal of Autism and Developmental Disorders, 33*, 653-672.

Brennan, S. (2000). Processes that shape conversation and their implications for computational. In *Proceedings of the 38th Annual Meeting of the Association for Computational Lingusitics*.

Brusk, J., Lager, T., Hjalmarsson, A., & Wik, P. (2007). DEAL – Dialogue Management in SCXML for Believable Game Characters. In *Proceedings of ACM Future Play 2007* (pp. 137-144).

Burnham, D., & Lau, S. (1999). The integration of auditory and visual speech information with foreign speakers: The role of expectancy. In *AVSP* (pp. 80–85).

Carlson, R., Granström, B., Heldner, M., House, D., Megyesi, B., Strangert, E., & Swerts, M. (2002). Boundaries and groupings - the structuring of speech in different communicative situations: a description of the GROG project. In *Proc of Fonetik 2002* (pp. 65-68). Stockholm.

Ellis, R. (1994). *The Study of Second Language Acquisition.* Oxford University Press.

Engwall, O., & Bälter, O. (2007). Pronunciation feedback from real and virtual language teachers. *Journal of Computer Assisted Language Learning, 20*(3), 235-262.

Engwall, O., Wik, P., Beskow, J., & Granström, G. (2004). Design strategies for a virtual language tutor. In Kim, S. H., & Young, D. H. (Eds.), *Proc ICSLP 2004* (pp. 1693-1696). Jeju Island, Korea.

Engwall, O. (2008). Can audio-visual instructions help learners improve their articulation? - an ultrasound study of short term changes. In *Proceedings of Interspeech 2008* (pp. 2631-2634). Brisbane, Australia.

Eskenazi, M. (1999). Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. *Language Learning \\& Technology, 2*(2), 62-76.

Flege, J. E. (1998). Second-language Learning: The Role of Subject and Phonetic Variables. In *STiLL-Speech Technology in Language Learning* (pp. 1-8).

Gee, J. P. (2003). *What video games have to teach us about literacy and learning.* New York: Palgrave Macmillan.

Granström, B., House, D., & Lundeberg, M. (1999). Prosodic cues in multi-modal speech perception.. In *Proc of ICPhS-99* (pp. 655-658).

Gustafson, J., Bell, L., Boye, J., Lindström, A., & Wirén, M. (2004). The NICE Fairy-tale Game System. In *Proceedings of SIGdial*.

Hjalmarsson, A., Wik, P., & Brusk, J. (2007). Dealing with DEAL: a dialogue system for conversation training. In *Proceedings of SigDial* (pp. 132-135). Antwerp, Belgium.

Hjalmarsson, A. (2008). Speaking without knowing what to say... or when to end. In *Proceedings of SIGDial 2008*. Columbus, Ohio, USA.

Iuppa, N., & Borst, T. (2007). *Story and simulations for serious games : tales from the trenches.* Focal Press.

Johnson, W., Marsella, S., & Vilhjalmsson, H. (2004). The DARWARS Tactical Language Training System. In *Proceedings of I/ITSEC*.

Koda, T., & Maes, P. (1996). Agents with Faces: The Effects of Personification of Agents. In *Proceedings of HCI* (pp. 98-103).

Lester, J., & Stone, B. (1997). Increasing believability in animated pedagogical agents. In Johnson, W. L., & Hayes–Roth, B. (Eds.), *Proc. of the First International Conference on Autonomous Agents* (pp. 16-21). Marina del Rey, CA, US.

Massaro, D. W., & Light, J. (2004). Using Visible Speech to Train Perception and Production of Speech for Individuals With Hearing Loss. *Journal of Speech, Language and Hearing Research, 47*(2), 304-320.

Mateas, M., & Stern, A. (2003). Façade: An experiment in building a fully-realized interactive drama. In *Game Developer's Conference: Game Design Track*. San Jose, California, US.

McAllister, R. (1997). Perceptual foreign accent: L2 users' comprehension ability. *Second-language speech: Structure and process*, 119-132.

Meng, H., Lo, Y. Y., Wang, L., & Lau, W. Y. (2007). Deriving salient learners' mispronunciations from cross-language phonological comparisons. In *Automatic Speech Recognition \\& Understanding, 2007. ASRU. IEEE Workshop on* (pp. 437-442).

Neri, A., Cucchiarini, C., & Strik, H. (2002b). Feedback in Computer Assisted Pronunciation Training: Technology Push or Demand Pull?. In *ICSLP* (pp. 1209-1212).
Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002a). The Pedagogy-Technology Interface in Computer Assisted Pronunciation Training. *Computer Assisted Language Learning, 15*(5), 441-467.
Prensky, M. (2001). *Digital game-based learning*. McGraw Hill.
Prensky, M. (2002). The motivation of gameplay. *On the Horizon, 10*(1), 5 - 11.
Sjölander, K. (2003). An HMM-based system for automatic segmentation and alignment of speech. In *Proc of Fonetik 2003, Umeå University, Dept of Philosophy and Linguistics PHONUM 9* (pp. 93-96).
Skantze, G. (2005). Galatea: a discourse modeller supporting concept-level error handling in spoken dialogue systems. In *Proceedings of SigDial* (pp. 178-189). Lisbon, Portugal.
Sluijter, A. M. C. (1995). *Phonetic correlates of stress and accent HIL dissertations 15*. Doctoral dissertation.
Sumby, W. .., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America, 26*, 212-215.
Walker, J. H., Sproull, L., & Subramani, R. (1994). Using a human face in an interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence* (pp. 85-91).
Wik, P., Hjalmarsson, A., & Brusk, J. (2007). DEAL A Serious Game For CALL Practicing Conversational Skills In The Trade Domain. In *Proceedings of SLATE 2007*.
Wik, P. (2004). Designing a virtual language tutor. In *Proc of The XVIIth Swedish Phonetics Conference, Fonetik 2004* (pp. 136-139). Stockholm University.
van Mulken, S. A., & Andre, E. (1998). The Persona effect: How substantial is it?. In *Proc. of HCI98* (pp. 53-66).
von Ahn, L. (2006). Games with a Purpose. *COMPUTER*, 92-94.