KTH Computer Science
and Communication

The Virtual Language Teacher

Models and applications for language learning using embodied conversational agents

PREBEN WIK

Doctoral Thesis
Stockholm, Sweden 2011

TRITA-CSC-A 2011:09

ISSN-1653-5723

ISRN–KTH/CSC/A--11/09-SE

ISBN 978-91-7415-990-5

KTH School of Computer Science and Communication

SE-100 44 Stockholm

SWEDEN

Akademisk avhandling som med tillstånd av Kungliga Tekniska Högskolan framlägges till offentlig granskning för avläggande av teknologie doktors-examen i tal- och musikkommunikation med inriktning på talkommunikation torsdagen den 26 maj 2011 klockan 09.30 i F3, Kungliga Tekniska Högskolan, Lindstedtsvägen 26, Stockholm.

Abstract

This thesis presents a framework for computer assisted language learning using a virtual language teacher. It is an attempt at creating, not only a new type of language learning software, but also a server-based application that collects large amounts of speech material for future research purposes.

The motivation for the framework is to create a research platform for computer assisted language learning, and computer assisted pronunciation training.

Within the thesis, different feedback strategies and pronunciation error detectors are explored.

This is a broad, interdisciplinary approach, combining research from a number of scientific disciplines, such as speech-technology, game studies, cognitive science, phonetics, phonology, and second-language acquisition and teaching methodologies.

The thesis discusses the paradigm both from a top-down point of view, where a number of functionally separate but interacting units are presented as part of a proposed architecture, and bottom-up by demonstrating and testing an implementation of the framework.

Acknowledgements

I would like to start by thanking my supervisor Björn Granström, not only for giving me the opportunity to pursue this line of work, but also for having been a constant support during my period as a PhD student at KTH. He was the first person I contacted when applying for a job at KTH. We met with a mutual interest for combining technology and language learning, and I think our two visions coincided somehow. In particular during the final part or my thesis work when I at times showed signs of despair, he returned my frustrations with firm but friendly encouragement and guidance on how to progress. Perhaps he saw already then (before I did) the work I had ahead of me.

I am also grateful to my co-supervisor Olov Engwall for all his constructive criticism and insightful comments on my work, as well as for the opportunities I have gotten to cooperate with him.

Thanks also to Anders Askenfelt, our head of department, for running the lab so well, and the other professors at the lab, Rolf Carlson and David House for constant encouragement and good spirit in the lab, and to Gunnar Fant for creating the lab in the first place.

I would like to thank the language unit with Margaretha Andolf at the helm, who initiated the Ville for SWELL work, and put me in contact with Cecilia Melin Weissenborn, Lars Cederwall, Anita Kruckenberg, Phillip Shaw, y Elena Salazar Reyes, la guapa y amistoso Mexicana que siempre me habla en español, and many other good people.

I am grateful to the graduate school of language technology, GSLT to which I have been an associate, for being a provider of financial support, PhD courses, and great retreats. It has also been a great opportunity to meet and get to know many good people in the field of language technology.

The people at NTNU in Trondheim: Jacques Koreman, Åsta Øvregaard, Olaf Husby, Egil Albertsen, Sissel Nefzaoui, Eli Skarpnes: the CALST team that is putting together a Norwegian version of Ville.

I would also like to thank Kåre Sjölander the maker of 'Snack' and the CTT-aligner used extensively in Ville, Jonas Beskow for the heads (and the music) and some beautiful Tcl-hacks, Bosse Thorén, for good ideas and help with Ville - a kindred spirit in language learning pedagogy, Anne-Marie Öster for pointing

me in the direction of what has become essential literature for me, and Julia Hirschberg for being a great ambassador for Ville around the globe.

I also feel indebted to David House for reading and giving comments on the thesis, and to Rebecca Hincks for proofreading the thesis, for our work together, and for great professional discussions, pointing me in the right direction in the SLA literature.

I am grateful to Anna Hjalmarsson, my roommate during most of my time at KTH, and collaborator in several projects, thanks for both creative and fun times together, Giampiero Salvi, my roommate for part of the time, Jens Edlund, for beer and creative discussions at Östra Station, most of the time. Ananthakrishnan, अनंत, बहुत से दिलचस्प विचार-विमर्श और तुम्हारे सहयोग के लिए धन्यवाद!, and Samer Al Moubayed, -Habibi, it has been such a great pleasure working with you. أن يل فـرشل منإ .(شتا يت ياك) في يف لضفأ قيدص يل قيدص نوك ،كقيدص نوكتس انتقادص لمآ قليوط

I thank the dialogue team: Joakim Gustafson, Anna Hjalmarsson, Gabriel Skantze, Jens Edlund, Mattias Heldner, I'm almost in…, and Grötgänget – Kjell Elenius, Mats Blomberg, Inger Karlsson, Per-Anders Jande and the others who made the lunches become so much more than just food. Oh, this work would not have been the same if I had not spent all that time playing 'Innebandy' on Thursdays: Marco, Gael, Kjetil, Giampi, Roberto, Marius, Kjell, Jocke, Mats etc…It's been great fun!

And to all other colleagues at TMH who are not mentioned explicitly, you have all contributed to the team spirit that makes me feel that our lab is outstanding!

Kim Sørenssen thank you for insightful comments and discussions (no, they have no branches!) and Shie Ing-Ping 我上次欠妳的. Michiel Schotten – "je bent de volgende!"

And to all my friends in the 'non-academic world' Tocke Wingårdh, Johan Björkdahl, Olle Lindeberg, Stefan Bernards, to mention some of the dozen or more people who tried to distract me as much as they could from my work in order to make me see some other sides of things. Life, and consequently this thesis would not have been the same without that 'wow' factor.

Last but not least: my big and wonderful and complicated family in Norway, Sweden and Taiwan – sisters and brothers, half-brothers, parents, step-parents and extra-parents: Roar, Siri, Bente, Gunnar, Thildy, Frode, Felicia, Julia, Bim-

bim, Kristine, Ebba, David, Nina, Anders, Jonna Adam, Vera, 姐姐 ,姐夫 ,二哥 , 維廷 ,偉韓 ,龍龍 婉怡

And to my closest family: I am deeply indebted to both Li-Hui Chen my wife, friend and life-partner through a considerable part of my life: 給我親愛的老婆 , 願妳生命中的一切都美好, and to my children Ronya and Anton whom I am so proud of. It would not have been possible without you, and I know you are glad that it is done.

Thank you all!

Table of content

Publications and contributors

Some of the work presented in this thesis has already been published in journals and conference proceedings, and some of the work presented has been done in collaboration with others. The publication list below specifies the details of the collaborations.

Wik, P., & Granström, B. (2010). Simicry - A mimicry-feedback loop for second language learning. In *Proceedings of Second Language Studies: Acquisition, Learning, Education and Technology*. Waseda University, Tokyo, Japan.

Simicry presented in chapter 8 is partially based on discussions with Granström. All the work and writing of the paper was done by Wik.

Wik, P., & Escribano, D. (2009). Say 'Aaaaa' Interactive Vowel Practice for Second Language Learning. In Proc. of SLaTE Workshop on Speech and Language Technology in Education. Wroxall, England.

The vowel game described in chapter 5 was done in cooperation with Escribano. He was a master thesis student at KTH, and was supervised by Wik. Escribano wrote some of the code, conducted the user test, and implemented the first version of the vowel-game, based on a proposal and original idea of Wik. Wik wrote the paper.

Wik, P., Hincks, R., & Hirschberg, J. (2009). Responses to Ville: A virtual language teacher for Swedish. In Proc. of SLaTE Workshop on Speech and Language Technology in Education. Wroxall, England.

The user study described in chapter 10 was done in cooperation with Hincks and Hirschberg. Most of the work regarding the replies from the questionnaires was done by Hincks. Hirschberg did most of the statistics on the user tests, and all three participated during the user tests and in writing the paper. Hincks and Wik did the recruitment of students. Wik wrote all the code.

Wik, P., & Hjalmarsson, A. (2009). Embodied conversational agents in computer assisted language learning. Speech Communication, 51(10), 1024-1037.

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, Hjalmarsson, and Brusk all contributed to the ideas behind the DEAL system described in chapter 9. Implementation of the system described in the DEAL section was done in cooperation with Hjalmarsson. Hjalmarsson was responsible for implementing the dialogue modules set up in Higgins (Pickering, Galatea and Ovidius). Hjalmarsson and Wik co-wrote the dialogue manager, and Wik developed the graphical user interface. Wik wrote the Ville section and Hjalmarsson and Wik co-wrote the DEAL section.

Finally, there is some unpublished work done in cooperation with companies and institutions that I would like to accredit.

Ville for SWELL presented in chapter 7 was done in cooperation with the language unit at KTH. Cecilia Melin Weissenborn, Lars Cederwall and Matts Bengtzén were responsible for selection of words and pictures.

The Langofone project described in chapter 12 was done in cooperation with Tobial Meschke at Luli Media group and Sirocco mobile.

The CALST project presented in chapter 13 is an effort to create a Norwegian version of Ville, and is done in cooperation with NTNU, UiO, and EVO. The development is a teamwork done at NTNU by Preben Wik, Jacques Koreman, Åsta Øvregaard, Olaf Husby, Egil Albertsen, Sissel Nefzaoui, Eli Skarpnes and Öyvind Bech.

List of publications

Wik, P., & Granström, B. (2010). Simicry - A mimicry-feedback loop for second language learning. In *Proceedings of Second Language Studies: Acquisition, Learning, Education and Technology*. Waseda University, Tokyo, Japan.

Picard, S., Ananthakrishnan, G., Wik, P., Engwall, O., & Abdou, S. (2010). Detection of Specific Mispronunciations using Audiovisual Features. In *International Conference on Auditory-Visual Speech Processing(AVSP)*. Kanagawa, Japan.

Engwall, O., & Wik, P. (2009). Are real tongue movements easier to speech read than synthesized?. In *Proceedings of Interspeech*, Brighton, England.

Engwall, O., & Wik, P. (2009). Can you tell if tongue movements are real or synthetic?. In *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP), Norwich, England*.

Engwall, O., & Wik, P. (2009). Real vs. rule-generated tongue movements as an audio-visual speech perception support. In *Proceedings of Fonetik 2009*, Stockholm, Sweden.

Wik, P., & Escribano, D. (2009). Say 'Aaaaa' Interactive Vowel Practice for Second Language Learning. In *Proc. of SLaTE Workshop on Speech and Language Technology in Education*. Wroxall, England.

Wik, P., Hincks, R., & Hirschberg, J. (2009). Responses to Ville: A virtual language teacher for Swedish. In *Proc. of SLaTE Workshop on Speech and Language Technology in Education*. Wroxall, England.

Wik, P., & Hjalmarsson, A. (2009). Embodied conversational agents in computer assisted language learning. *Speech communication, 51*(10), 1024-1037.

Beskow, J., Engwall, O., Granström, B., Nordqvist, P., & Wik, P. (2008). Visualization of speech and audio for hearing-impaired persons. *Technology and Disability, 20*(2), 97-107.

Wik, P., & Engwall, O. (2008). Can visualization of internal articulators support speech perception?. In *Proceedings of Interspeech 2008* (pp. 2627-2630). Brisbane, Australia.

Wik, P., & Engwall, O. (2008). Looking at tongues – can it help in speech perception?. In *Proceedings of Fonetik 2008*, Gothenburg, Sweden.

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), Toronto, Canada.

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.

Wik, P., & Granström, B. (2007). Att lära sig språk med en virtuell lärare. In *Från Vision till praktik, språkutbildning och informationsteknik* (pp. 51-70). Nätuniversitetet.

Wik, P., Hjalmarsson, A., & Brusk, J. (2007). Computer Assisted Conversation Training for Second Language Learners. *Proceedings of Fonetik, TMH-QPSR, 50*(1), 57-60, Stockholm, Sweden.

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*, Farmington, USA.

Nordenberg, M., Svanfeldt, G., & Wik, P. (2005). Artificial Gaze - Perception experiment of eye gaze in synthetic faces. In *Proceedings from the Second Nordic Conference on Multimodal Communication*, Gothenburg, Sweden.

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.

Wik, P. (2004). Designing a virtual language tutor. In *Proc of The XVIIth Swedish Phonetics Conference, Fonetik 2004* (pp. 136-139). Stockholm, Sweden.

Wik, P., Nygaard, L., & Fjeld, R. V. (2004). Managing complex and multilingual lexical data with a simple editor. In *Proceedings of the Eleventh EURALEX International Congress*. Lorient, France.

List of abbreviations

ALM= Audio-Lingual Method

ASR = Automatic speech recognition

CAH = Contrastive Analysis Hypothesis

CALL = Computer-Assisted Language Learning

CAPT= Computer-Assisted Pronunciation Training

CPH= Critical period hypothesis

CLT = Communicative Language Teaching

CTT = Centrum för Talteknologi (Center for speech technology)

ECA = Embodied Conversational Agent

fMRI = functional Magnetic Resonance Imaging

IPA = International phonetic association

L1 = First language (mother tongue)

L2 = Second language (target language)

NS = Native speaker

NNS = Non native speaker

PED= Pronunciation Error Detection

SLA = Second Language Acquisition

VLT=Virtual Language Teacher/Tutor

–1–

Introduction

Research consistently show that people with foreign accents are judged to be less intelligent, less trustworthy, less competent, less educated, and more unpleasant to listen to (Cunningham-Andersson, 1996; Gluszek & Dovidio, 2010).

It is a socio-political tragedy that immigrants are alienated because of language barriers. Although an increased tolerance towards foreign accents would be desirable both from a commonsensical and a socioeconomic point of view, it seems like the mechanism of stigmatizing deviant behavior is perhaps an inevitable aspect of societies everywhere.

As Falk (2001) describes it: *we and all societies will always stigmatize some condition and some behavior because doing so provides for group solidarity by delineating 'outsiders' from 'insiders'.*

Face-to-face oral communication is our primary mode of communication, and for someone who does not have a firm command of this medium the result is often some form of alienation. This holds for native speakers with some form of speech disorder, and for second language (L2) learners.

If individuals have to deal with prejudice because of their accent, and this stigma is likely to prevail, L2 learners stand more to gain than just better communication skills from improving their pronunciation.

Although L2 learners apparently have so much to gain by achieving a native-like accent, the fact remains that a large majority of healthy, intelligent adult L2

learners have a persistent accent even after years of exposure to a new language.

Notions of a neurologically-based critical period for second language acquisition (SLA) have been with us in various forms and disguises for over fifty years now (Bongaerts et al., 1997). The critical period hypothesis (CPH) states that complete mastery of a language, first or second, is not possible for learners who start to learn after a certain age. The loss of neural plasticity has been suggested as the primary cause by those who believe that attainment of native-like proficiency after a certain age is, in principle, impossible. This notion is however, not uncontested.

How is it that 'neuroplasticity' - the ability of our brains to create new neural pathways - would be applicable for recovery from stroke and other traumas, but not for improving foreign accents? Growing evidence for our ability in adult life to train our minds to bring about lasting changes both in physical and psychological health and wellbeing (c.f. Begley, 2009), could perhaps have implications also on the notion of CPH?

As argued by Long (1990): even one single post-critical-period L2 learner with an underlying competence indistinguishable from that of native speakers would suffice to reject the CPH.

Several researchers do in fact report on late L2 learners who have acquired native or near-native pronunciation proficiency in a new language (c.f. Neufeld, 1978; Bongaerts et al., 2000; Piller, 2002; Abrahamsson & Hyltenstam, 2004). This does not say that everybody can acquire a native-like accent; it simply states that some people for some reasons have managed do so.

Bongaerts et al. (1997) suggest, after having reviewed the circumstances for a number of people who have acquired native-like levels of pronunciation, that a combination of input, motivational, and instructional factors may compensate for the neurological disadvantages of a late start.

According to Bongaerts et al. (1997), key factors for successful L2 speakers are:

- High motivation to achieve accent-free pronunciation
- Unlimited access to L2 speech
- Intensive training in L2 perception and L2 production

These characteristics are not very specific, but suggest that perhaps new training methods and technological innovations could indeed help people acquire near-native pronunciation, or at least help people improve their pronunciation regardless of age.

In a single generation, technological revolutions have transformed the way we live, work and do business. Thirty years ago, we could not have predicted that something called the 'internet' would lead to such dramatic changes in our lives.

New technological innovations are shaping our future in all aspects of life, and also in the educational domain, technology is likely to soon take on a larger and more important role. We may stumble in the dark as to exactly where the next breakthroughs will come, but rather than trying to predict what will come, we may propose in what direction we would like the technology to move. By taking on visions and pointing out some direction that we would like technology to bring us, we may in fact participate in shaping the future in that direction.

One such vision is the idea of a virtual language teacher. 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.

1.1 The ultimate language teacher

If we don't have to take today's technological limitations into consideration, and are allowed to dream up a machine (or a human) with any features and abilities that come to mind, what would the ultimate language teacher be like? To actually create such a machine is not feasible of course, but it might be an interesting starting point to see what kind of obstacles are present and what challenges lie ahead.

One of the rudimentary requirements that one would expect from such a teacher would be natural language understanding. This would however in many researchers' opinion require *strong-AI*, which means solving the central artificial intelligence problem and making computers as intelligent as people. A daunting prerequisite as it may seem, it is only one of many features one would want the ultimate language teacher to have.

In addition to general human intelligence he (or she) would have highly specialized skills in linguistics and phonetics, and understand which features from the learner's first language would affect the learner's aptitude in the target language. Not only transfer phenomena, but a full taxonomy of all possible errors a learner can make, would be part of the teacher's knowledge.

Coupled with this knowledge would be acutely tuned sensors, for detecting all kinds of fine details in the pronunciation of a learner, enabling specialized exercises to remedy all kinds of errors.

He would be a master of feedback, giving just the right amount of praise and criticism to entice optimal performance, and know when to be demanding and when to be forgiving. With perfect timing he would give encouragement when it is needed the most.

He would be well versed in all the important literature on learning theories of second language acquisition, cognitive science, and skill building, in addition to pedagogical skills that would make the learning compelling and effective.

He would be user configurable, not only in order to transform into a learner's preferred gender, ethnicity, or age, but also to cater to conscious or subconscious preferences of cognitive load, type of feedback, rate of introduction to new material, and rate of repetition on old material.

In addition, awareness of the various approaches and ways of individualized learning, and the ability to assess the learning style of every individual would be desirable in order to tailor every lesson to a learner's preferred learning style.

He would know when it is time to rest in order to achieve maximum retention from the lessons, and know when to switch to another exercise in order to keep the attention and motivation at a maximum.

He would always be available, day or night, in your PC or in your mobile device. Respectfully at a distance unless called for, but untiringly monitoring your efforts, committed to follow your progress.

He would in addition be funny, motivating, and engaging, patient, persuasive, observant, committed, respectful, inspiring, entertaining, stimulating, compelling, demanding, energetic, determined, intelligent, …

1.1.1 Thesis goals

With such a vision in mind, which of all these features are possible to implement with current technology and which will continue to be science-fiction for an unknown time to come? To try to implement the full vision of even the technologically feasible aspects of such a virtual language teacher could keep a team of researchers, developers, and pedagogues busy for years, and is for natural reasons well beyond the scope of this thesis.

The overall hypothesis of this thesis is still that it will be possible to create a fully functional virtual language teacher. If it is built using the cyclic software development process known as the iterative and incremental development process, it is possible to take on a subset of the tasks that the ultimate language teacher would be expected to handle and incrementally expand towards the target. The main research questions that will be in focus are:

Which parts of the ultimate language teacher should be addressed first, and which can, or must, wait? Will it be possible to reach the 'critical mass' of functionality and robustness needed to let real language learners use it?

What aspects of language should or can be taught using this paradigm?

Will it be possible to implement different teaching methods?

How can the inevitable errors that language learners make be addressed?

Will it be possible to use the first iteration as a knowledge source for the next generation VLT?

How can portability be addressed?

1.1.2 Outline of the thesis

This thesis is organized as follows:

Chapter 1 (this chapter) is an introduction presenting the overall vision and motivation that this work is based on.

Chapter 2 gives an introduction to skill-building. Initially skill-building on a neurological level is presented and the concept of automaticity is introduced. Then the relationship between motivation, games and learning is discussed, and finally an overview of general learning theories and language learning theory is presented.

Chapter 3 presents the linguistic, phonetic, and phonological aspects that the work is built around. Both exercises and feedback are based on explicit knowledge drawn from these disciplines of science.

Chapter 4 introduces the overall architecture of the VLT the way it looks in its present iteration. The Ville framework is a combination of a VLT and a data collection tool deemed necessary to further the development of the VLT.

Chapters 5 - 10 present a varied set of different types of exercises, learning strategies and feedback implementations, divided into a linguistic hierarchy of different levels.

Chapter 5 shows how some low-level associative learning can take place in the Ville framework. It gives two examples on the segmental level, one in the perception domain, taking advantage of the concept of minimal pairs, and one in the production domain. The production exercise is an example of how real-time transmodal feedback can be utilized to help language learners discover new and unknown configurations of their articulators.

Chapter 6 establishes the concept of pronunciation error detectors (PEDs), which are designed around specific phonetic or phonological issues on a supra-segmental level, which is hypothesized to aid language learners in raising the awareness of these particular problems.

Chapter 7 investigates whether the same concepts that are introduced in previous chapters can be useful also for acquisition of new vocabulary. Adhering to the dual-coding theory, a picture component is added to every word, using the same architecture to create vocabulary and writing exercises.

Chapter 8 focuses on the sentence level, and introduces a new concept called Simicry. This is a paradigm in which exposure to large amounts of meaningful input, in the form of formulaic language is facilitated, and where prosodic aspects of language can be practiced. A different type of feedback, similarity measures, is investigated, and two different ways to interact with the VLT within the Simicry paradigm are explored.

Chapter 9 opens with an introduction to spoken dialogue systems and how they may be utilized for CALL purposes. Then follows an introduction to DEAL, a game based spoken dialogue system for CALL, exploring the CLT teaching methodology, with focus on communicative skills rather than focus on form, or correct pronunciation.

Chapter 10 is a short-term user study, investigating how some of the functionality of the Ville system is being received by L2 learners.

Chapter 11 is a more elaborate user study, done over a longer time, to investigate the impression users of the system have after having used the system at home and over a longer time. In the study different groups also received different versions of the program to investigate the effectiveness of specific aspects of the system.

Chapter 12 looks at portability issues and investigates how the Ville framework can also be utilized in mobile devices. Langofone, a mobile phone application for language learning is presented.

Chapter 13 presents a Norwegian version of Ville. Portability from the point of view of localization are discussed and issues regarding how well the underlying framework is able to handle language specific differences..

Chapter 14 presents future work and conclusions.

–2–

Skill building

When learning a second language, what we really want to achieve is automaticity.

This chapter starts with a general description of what automaticity is, and some theories of learning are introduced. Then it moves on to describe how motivation to learn can be promoted and the relationship between games and learning, and finally how learning is supported through different teaching methods.

2.1 Automaticity

Automaticity is achieved when a task can be performed almost effortlessly. The learner does not have to think about individual steps and can carry out the task from start to end while thinking about other things. When someone speaks their first language it is to a large extent an automatic process. We often do not think of which words to use or which grammatical constructs to apply, much less how to shape the mouth or move the tongue in order to create the right sounds, where to place the stress or how to adjust the pitch for a sentence to sound right. An L2 learner on the other hand might struggle with all these aspects of a new language, and from this point of view, what the L2 learner is aiming at is to a large extent to develop automaticity.

One of the fundamental questions to ask is thus: how do we achieve this?

Dramatic changes in brain activity can be seen on fMRI scans as automaticity develops. Schneider & Chein (2003) demonstrate how performance is increased, but the cognitive load is reduced thus less attention is needed on the task at hand and attention can be given to other processes or tasks. Schneider & Chein state:

"Automaticity leads to fast, parallel, robust, low effort performance, but requires extended training, is difficult to control, and shows little memory modification. In contrast, controlled processing is slow, serial, effortful and brittle, but it allows rule-based processing to be rapidly acquired, can deal with variable bindings, can rapidly alter processing, can partially counter automatic processes, and speeds the development of automatic processing".

2.1.1 The declarative procedural model

There is today compelling evidence that language depends on brain systems that are also used for other functions (c.f. Ullman, 2001; Tomasello, 2005). Ullman's declarative/procedural model of language predicts common computational, processing, and anatomic neural substrates for language and non-language functions (Ullman, 2001). A consequence of this is that knowledge about skill building in other domains may be equally adequate for language learning. Regardless of the activity, skilled performances share many common elements, including goal oriented behavior, improvements in performance with practice and training, use of feedback for error correction, and conservation of cognitive resources with improved performance.

It has been proposed that skill acquisition proceeds through phases characterized by qualitative differences in performance. A framework for skill acquisition was proposed by Fitts (1964) based on observations that different cognitive processes are involved at different stages of learning. Fitts distinguishes three phases: cognitive, associative, and autonomous.

At the first stage (cognitive) the learner needs to use cognitive/intellectual processes to understand the nature of the task. The learner must attend to outside cues and feedback about his performance as a guide to the learning. This is the slowest and most error-prone stage. The learner then enters the associative phase, where inputs are linked more directly to appropriate actions and performance time and error rates are reduced. Finally, according to Fitts, a learner may enter the autonomous stage where the task no longer requires conscious control and may be performed concurrently with other tasks.

Perhaps the best known general theory of skill acquisition is Anderson's adaptive control of thought (ACT-R). Anderson (2002) makes a distinction between declarative and procedural knowledge. Declarative knowledge is basically the body of facts and information that a person knows (factual, know-what knowledge), whereas procedural knowledge is the set of skills a person knows how to perform (know-how). While in the declarative stage, instructions are encoded in the brain as a set of facts. These facts are retained in active working memory while performing the task, and used by some interpretive mechanism. As a person practices, procedures specific for the task develop and the need for the active maintenance of declarative knowledge about how to do the task are no longer required. Performance continues to improve through something Anderson calls tuning. Through processes like generalization, discrimination and strengthening of appropriate rules in the newly developed procedures performance improve gradually.

ACT-R maintains the same three stages as Fitts (1964) in the process of skill acquisition, but also provides an explanation of the phenomena associated with these three stages (Cognitive stage, associative stage and autonomous stage).

The cognitive stage: In the first stage the learner receives instructions about a skill in declarative form. Explicit information is given in order to provide clear and concise rules and sufficient examples, which the learner can interpret and rehearse, thereby raising awareness of and internalizing the skill. The processing in this stage is conscious, deliberate, slow, and requires full attention.

Associative stage: The major development of this stage is *knowledge compilation*. During the associative stage, a process of proceduralization takes place, converting declarative facts into production form. The learner should here be offered opportunities for abundant repetition within a narrow context (which is what drills are all about).

Autonomous stage: After a skill has been compiled into a task-specific procedure, the learning process involves an improvement in the search for the right production. In this stage, the procedure becomes more and more automated and rapid. The process underlying this stage is tuning. Three learning mechanisms serve as the basis of tuning: generalization, discrimination, and strengthening.

Studies on vocabulary and grammar acquisition have shown that these general models of skill acquisition also apply to development of automaticity in gram-

matical aspects of second language acquisition (DeKeyser, 1997; Robinson, 1997). Can we assume that the same general principles would also apply to the development of automaticity in pronunciation proficiency?

2.1.2 Motor skills and pronunciation

Acquiring proficient or native-like pronunciation is primarily a psychomotor skill, mediated by the sensory and motor cortex.

Pronunciation as a psychomotor skill differs from other aspects of language in a number of ways, but first and foremost because muscle movements are involved. Muscle movements in general are performed with either *open-loop* or *closed-loop* motor control. *Closed-loop* motor control uses perception to consciously adjust muscle movements. It is a stimulus-response loop, where each adjustment takes at least 200 ms. It is schematically divided into *sensation* (transmission from sensory receptors to brain)*, perception* (classify retrieve)*, response selection* (formulate course of action) *and response execution* (signal from brain to muscles) (Schmidt & Lee, 2005).

Speech is a very fast and complex skill, requiring precise coordination of hundreds of muscles, yet the average phoneme typically lasts less than 100 ms, (Fant, 2004), and is thus from a psychomotor point of view too fast to be done by *closed-loop* motor control. *Open-loop* motor control is much faster and is the execution of preprogrammed movements (motor programs) without perceptual feedback.

The learning of motor skills can be seen as the construction of "generalized motor programs" i.e., a sequence or class of automated actions that are triggered by associative stimuli, habit strengths, and re-enforcers, and can be executed without delay (Anderson, 2002), not too different from the proceduralization process described in the previous section.

Although normal speech uses open-motor control, schematic relations among movement parameters and outcomes can be built-up, modified, and strengthened in closed-loop motor control by perceptual feedback.

Pronunciation errors have to do with incorrect coordination and movement of muscles in the tongue, lips, and jaw etc., and hence require reprogramming of existing motor programs, or the creation of new ones. If an L2 learner encounters a novel sound in the L2 there are two possible (subconscious) outcomes.

Either a new phonological class (with corresponding new motor programs) is created, or the sound is deemed as the same as an existing phonological class and an existing motor program is strengthened. Proceduralized knowledge, once formed, is believed to be committed to a specific operation and cannot generalize to other uses.

Our neural apparatus is highly plastic in its initial state, but the initial state of second language acquisition (SLA) is no longer a plastic system; it is one that is already tuned and committed to the L1 (Ellis, 2006). A consequence of this is that sounds in a foreign language that are similar, but not the same, may by L2 learners be perceived as the same. As a result the learner would erroneously use and reinforce an existing motor program, and the more this erroneous association is strengthened the more rigorous it becomes and the more difficult it becomes to change it. This process is often referred to as phonological fossilization.

2.1.3 Hebbian learning

Hebb's theory states that when one neuron participates in firing another, the strength of the connection from the first to the second will be increased (Hebb, 1949). This principle is called Hebbian learning and often referred to as "Cells that fire together wire together". This simple algorithm attempts to explain "associative learning" in which simultaneous activation of cells leads to increases in synaptic strength. Hebbian learning principles are used in connectionist models and simulations of human learning, and some researchers have proposed that perceptual learning of speech categories depend on an unsupervised Hebbian learning process.

This model offers an explanation to why perceptual discriminations between sounds not contrasted in one's own native language are so difficult to acquire. McClelland et al. (2002) suggest that many failures of learning in adulthood may reflect a paradoxical tendency of the mechanisms of learning to reinforce inappropriate or undesirable responses. If an input elicits a pattern of neural activity, Hebbian learning will tend to strengthen the tendency to elicit the same pattern of activity on subsequent occasions. If the response is useful and constructive, the brain will learn to reinforce it. If the response is inappropriate or undesirable, Hebbian learning will still tend to reinforce it.

As stated by Ellis (2006) *"paradoxically perhaps, it is the very achievements of associative learning in first language acquisition that limit the input analysis of L2 and that result in the shortcomings of SLA."*

For example McClelland et al. (2002) investigated the well-known difficulties that Japanese listeners experience when attempting to discriminate between the sounds /l/ and /r/ in English. The range of sounds treated in English as /r/ and /l/ are all mapped to the same (apparently /l/-like) percept in Japanese. Once this is established, further presentations might elicit Hebbian learning, which will simply strengthen the tendency of each sound to elicit the same percept, which would be counterproductive to discrimination. McClelland et al. (2002) showed that this could be overcome by using sounds that strongly exaggerated the contrast between /r/ and /l/, thus forming new speech categories in an unsupervised fashion (without feedback). Interestingly, their findings also indicated that feedback may modify Hebbian-based learning or recruit additional learning systems, indicating that Hebbian learning is not fully sufficient to account for all aspects of learning.

2.1.4 The phonological loop

Working memory is a theoretical construct within cognitive psychology that refers to the structures and processes used for temporarily storing and manipulating information. Although there are a number of theories on the theoretical structure, the best known that has received wide acceptance is the model of working memory from Baddeley (1992).

This theory proposes that two "slave systems" are responsible for short-term maintenance of information, and a "central executive" is responsible for the supervision of information integration and for coordinating the slave systems. One slave system, called the visuospatial sketchpad, is assumed to hold information about what we see. The second is called the phonological loop (sometimes referred to as auditory loop or articulatory loop), and it deals with sound and phonological information.

The phonological loop consists of two parts: a short-term phonological store with auditory memory traces that are subject to rapid decay, and an articulatory rehearsal component that can revive the memory traces. The first component, the phonological memory store, can hold traces of acoustic or speech based material. Material in this short term store lasts about two seconds unless it is

maintained through the use of the second subcomponent, the articulatory rehearsal component.

Music is also processed in the phonological loop. When a song or tune gets latched onto the phonological loop, it is rehearsed in a constant loop to prevent decay (which explains why sometimes, you can't seem to get a song out of your head). Acoustic information is assumed to enter into the phonological store automatically. Also visually presented language can be transformed into phonological code by silent articulation. This transformation is facilitated by an articulatory control process. The phonological store acts as an 'inner ear', remembering speech sounds in their temporal order, whilst the articulatory process acts as an 'inner voice' and repeats the series of words (or other speech elements) on a loop to prevent them from decay.

The phonological loop is believed to play a key role in both first and second language acquisition, both in learning and rehearsing new vocabulary, and in learning the novel phonological forms of new words (Baddeley et al., 1998).

Another important function of the phonological loop is correction. Hearing your own speech may make you aware of errors. That is, auditory feedback is sent through one's own speech comprehension system, parsed, and consciously monitored and evaluated (e.g. "No that didn't sound right" as an internal reflection on an attempt to pronounce something).

Auditory monitoring of the acoustic environment can ensure intelligibility by adjustments of the sound level, speaking rate, and prosody, but is too slow to have a direct impact on speech commands. It is believed that auditory feedback is employed for tuning in the associative and autonomous phases of learning.

2.1.5 Proprioceptive feedback

Not only language learners make pronunciation errors. In fact, native speakers are also constantly making errors which may be corrected by means of self-repair.

Actions can be incorrect with respect to some external criterion (for example the linguistic rules of a given language) or actions may be judged as errors with regard to some internal standard, i.e. a person's intentions form the starting point from which correctness and incorrectness have to be decided.

Most of the self-monitoring corrections are not made by monitoring through the phonological loop, but by efferent, tactile and proprioceptive feedback (the sensing of where your limbs - and in this case of speech articulators - are and where they are moving to).

Self-repairing is typical for most human motor skills and refers to the correction of errors without external feedback or prompting. Voice modulation is thought to be primarily proprioceptive It has been estimated that one out of ten of all our utterances contains some sort of revision activity (cf. Nakatani & Hirschberg, 1994).

Errors can be made as response-selection errors, i.e. a wrong motor program is selected, which is, however, executed perfectly, or they can be made as response-execution errors, where the correct motor program is selected but something goes wrong in its execution.

Speakers can monitor their utterances for a multitude of distortions: at the conceptual level, in the lexical selection, syntactic construction or in the sound form encoding. It may also be directed towards suprasegmental characteristics, such as sound level or prosody.

2.2 Motivation

Successful language learning depends to a large extent on the individual learner. Motivated people were able to learn a foreign language just as successfully 500 years ago as one does today.

Learning a language requires a substantial effort, and the motivation for doing so varies both over time and between individuals. 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 a stronger motivator for others. Motivation can also come from the pleasure of learning (intrinsic motivation), or from the task itself (task motivation) to mention some sources.

Two broad classes of motivation are often mentioned in this context: Extrinsic motivation - external incentives (such as money, grades, or prizes) for a person to perform a given task. Intrinsic motivation - internal motivation to do some-

thing because it either brings pleasure or because learners think it is important, or they feel that what they are learning is significant.

Although there is research suggesting that extrinsic motivation such as rewards may reduce the intrinsic motivation for learning (c.f. Lepper et al., 1973), various types of motivation are to a certain extent additive, in the sense that the overall motivation for learning something may be increased if for example an instrumental motivation is added to an already existing integrative motivation.

Task motivation (the enjoyment of doing the task at hand) certainly seems to be additive - i.e. it is more motivating to do a task that is fun/enjoyable than a task that is not, all else being equal.

2.2.1 The involvement load hypothesis

In addition to motivation, processing activities also influence memory performance. Retention in long term memory depends on how deep information is processed during learning. Laufer & Hulstijn (2001) proposed the involvement load hypothesis, which states that the amount of involvement in the task that learners are engaged in will affect the retention of unfamiliar vocabulary. They mention three components of task-induced involvement: need, search, and evaluation. Need is a motivational construct while search and evaluation come from the cognitive dimension. Need is the motivation to learn target words. Search occurs when the learner has to find the meanings of target words or the word form for words indicated by target concepts. Evaluation involves comparison of a target word with other words. The involvement load hypothesis attempts to draw attention only to vocabulary learning in a second language, but other processing models will suggest the same thing. For example Baddeley et al. (1998) write:

"In general, information that is encoded in terms of a rich and detailed representation of the world is likely to be more accessible than material that is processed in terms of a simpler or more impoverished scheme"

Task-based, interactive exercises and the use of sound, pictures, agents, and games, will not only enrich the learning by making it a more worthwhile experience to learn. By presenting content to be learned in a rich multimodal environment, a more robust memory trace is also created and thus the retention will

be increased. Motivational and cognitive factors may hence fuse during learning activities and influence the outcome of the skill building.

2.2.2 Dual-coding theory

Similarly, the dual-coding theory of memory retrieval states that the human information-processing system consists of two separate independent channels: an auditory channel for processing auditory and verbal information, and a visual channel for processing visual input and pictorial representations. Memory for verbal information is enhanced if a relevant visual is also presented (Paivio & Clark, 1991), and words that are associated with objects or imagery techniques are more easily learned than those without (Chun & Plass, 1996).

2.2.3 Games

Games build exclusively on task motivation. The reason people play games is the desire for a worthwhile experience. Game developers thus focus on finding ways to give players enjoyment and have in their strive for success developed several effective design strategies both to get and to keep players engaged and motivated throughout a game.

Many game designers view games as cognitive learning environments. According to Koster (2004), learning is really the mechanism that allows for fun. He states:

"Fun from games arises out of mastery. It arises out of comprehension. It is the act of solving puzzles that makes games fun. In other words, with games, learning is the drug"

Koster indicates that fun essentially derives from the player's brain attempting to find patterns and succeeding in doing so. It is a feedback mechanism from the brain when successfully exercising survival tactics. Much of what humans perceive as fun stems from activities that aid in survival. Things that made cavemen better cavemen, such as stalking, running, and throwing, contain the same mechanics that many modern games have at their core.

Crawford (1984) also claims that the fundamental motivation for all game-playing is to learn, and that game-playing has a vital educational function for any creature capable of learning. He states:

"Games are thus the most ancient and time-honored vehicle for education. They are the original educational technology, the natural one, having received the seal of approval of natural selection."

Another term often used by both developers and players is 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 essentially 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 motivating force to play.

2.2.4 Game-based learning

The same design principles that are used by game developers are finding their way into the educational field. 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).

A *serious game* is a game designed for a primary purpose other than pure entertainment, such as education, defence, city planning or scientific exploration (Iuppa & Borst, 2007).

If the original purpose for game-playing was educational, and was invented even before the advent of man (as stated by Crawford, 1984), it may seem like a paradox that we are now re-inventing game based learning. Some are however calling it a paradigm shift from conventional learning into harnessing the power of games for learning (Squire & Jenkins, 2003).

2.2.5 Flow

One such design principle that computer game designers try to integrate into the game design is the much publicized phenomena of 'flow' (Csikszentmihalyi, 1991). What has been described as the optimal experience of performing and learning is an experiential state of complete absorption or engagement in an activity, to the extent that people lose track of time and self-consciousness.

Flow is an experience *"so gratifying that people are willing to do it for its own sake, with little concern for what they will get out of it, even when it is difficult or dangerous"* (Csikszentmihalyi, 1991)

Flow experiences consist of eight elements, as follows:

- a task that can be completed
- the ability to concentrate on the task
- that concentration is possible because the task has clear goals
- that concentration is possible because the task provides immediate feedback
- the ability to exercise a sense of control over actions
- a deep but effortless involvement that removes awareness of the frustrations of everyday life
- concern for self disappears, but sense of self emerges stronger afterwards
- the sense of the duration of time is altered

In order to maintain a person's Flow experience, the activity needs to reach a balance between the challenges of the activity and the abilities of the participant. If the challenge is higher than the ability, the activity becomes overwhelming and generates anxiety. If the challenge is lower than the ability, it provokes boredom (see Figure 1).

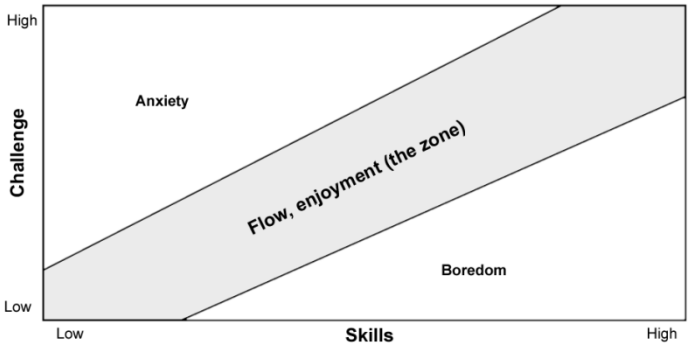

Figure 1 An activity needs to have a balance between the challenge of the activity and the skill level of the participant in order to maintain a person's flow experience. If the challenge is too high it generates anxiety. If the challenge is too low it provokes boredom.

To manage this, many games are developed using concepts such as dynamic difficulty adjustment (DDA) in order to automatically change parameters, scenarios and behaviors in the game in real-time.

The fusion of cognitive and motivational constituents found in flow has been shown to allow for improved performance and skill development, and the flow state has been shown to have positive impact on learning (Webster et al., 1993).

The concepts of flow and DDA are such that they could also be taken into account when designing CALL and CAPT systems.

As mentioned by Sweetser & Wyeth (2005) *"If a game meets all the core elements of Flow, any content could become rewarding, any premise might become engaging."*

2.3 Learning theories

Second language acquisition (SLA) theories have historically gone hand in hand with research in psychology, philosophy of mind, and learning theories. Three main theoretical schools of learning theory, Behaviorism, Cognitivism and Constructivism are often mentioned in the literature, and a short introduction is offered here as a background.

2.3.1 Behaviorism

Behaviorism is a theoretical framework developed in the early 20th century, which dominated psychological theory and research on learning for a large part of the twentieth century. Mental states and consciousness were at the time considered impossible to measure objectively, and only observable data in the behavior of a person (or animal) were of interest. Beliefs, thoughts, and other inner mental experiences were ignored, and the mind was treated like a black-box.

Associative learning is the main characteristic of behaviorist learning theory. The doctrine is associated with the work of Pavlov on classical conditioning (a dog's response to a bell by salivating after repeated exposure to simultaneous bell ringing and food), John Watson (considered the founder of behaviorism), and B.F. Skinner.

Skinner was interested in the learning process viewed as behavior modification, and tried to discover conditions that produce and control learned behavior. He developed or refined the methodology called operant conditioning, where learning can be equated with conditioning, or habit forming, and is the result of a three-stage process: Stimulus > Response > Reinforcement.

Skinner also designed a learning methodology called *Programmed instruction*, adhering to the principles of operant conditioning, where instructional content was broken down into small units, and correct responses were rewarded early and often.

Behaviorism has in its investigations of conditioning revealed some associative learning mechanisms by which information about our environment are detected and stored. Some low-order basic principles of learning, common to all animals have been encountered, and behaviorist-style learning ties in naturally with Hebbian learning mentioned in section 2.1.3, which is increasingly substantiated by neurological research (Goertzel, 2006). The behaviorists' basic mechanism of learning as a *Stimulus* > *Response* > *Reinforcement* feedback loop is a universal mechanisms that can be used to make a pigeon peck on a window or to teach a language learner aspects of a new language.

2.3.2 Cognitivism

Under the development of cognitive psychology, and what is known as the cognitive revolution in the 1960s and 1970s, behaviorism dropped out of favor. The cognitive revolution was not a refutation of behaviorism, but an expansion, since the behaviorist paradigm was somewhat restrictive in terms of what it allowed researchers to investigate. The existence of internal mental states and the inner workings of the human mind had been regarded as impossible to investigate objectively, and this changed as result of the cognitive revolution.

The advent of computers provided researchers with a first working model of human thought processes and enabled them to look at mental functions as information processing models, and to map human cognition in terms of mental representations. New theories of learning were developed where mental experiences such as beliefs, hopes, expectancies, emotions, and motivation were recognized as playing an important part in the learning process.

2.3.3 Constructivism

Whereas behaviorism and cognitivism view knowledge as external to the learner and the learning process as the act of internalizing knowledge, constructivism claims that individual learners construct mental models to understand the world around them, and that learning is a reconstruction rather than a transmission of knowledge.

The theory suggests that learners construct knowledge and meaning from an interaction between their experiences and their ideas, and promotes "learning by doing" rather than by structured instruction. The learner should play an active role in the learning process and the facilitator should adapt the learning experience based on what the learners want to do.

Constructivism is associated with the ideas of Piaget and Vygotsky, and Vygotsky's "zone of proximal development" where learners are challenged within close proximity to, yet slightly above, their current level of development. By experiencing the successful completion of challenging tasks, learners gain confidence and motivation to embark on more complex challenges (resembling the skill-challenge relationship of flow described in section 2.2.5)

Constructivists criticize previous learning theories for neglecting the unique personality of each student. According to the social constructivist approach, the instructor should have the role of facilitator and be more like a consultant and coach than a teacher in the traditional sense, emphasizing that each learner is a unique individual with unique needs and backgrounds. A facilitator needs to display a different set of skills than a teacher. A facilitator provides guidelines and creates the environment for the learner to arrive at his or her own conclusions, whereas a teacher gives answers according to a set curriculum.

2.4 Language teaching methods

A large number of language teaching methods have in the last century been presented as the best solution to language learning. The direct method, the humanistic approach, enlightened eclecticism, the natural approach, the silent way, suggestopedia, total physical response, etc. Larsen-Freeman & Long (1991) state that "at least forty 'theories' of SLA have been proposed".

A brief overview of some methods that have had a great impact in the history of language teaching is here offered as a historic background. These methods represent very different approaches to language teaching and will serve as examples in the following discussion on what is possible and appropriate to teach using a virtual language teacher (VLT).

2.4.1 The grammar-translation method

The grammar-translation method was developed for the study of 'dead' languages such as Latin and Ancient Greek. It involved little or no attention to pronunciation or communicative competence, but relied heavily on reading and translation, mastery of grammatical rules and accurate writing.

2.4.2 Audio-lingual method (ALM)

During and after World War II the need for foreign language proficiency in listening and speaking skills set the stage for a 'revolution' in language teaching methodology. What initially was called the "Army Method" later became what is known as the Audio-lingual Method (ALM).

The new method was firmly rooted in behaviorism, which was the dominant psychology and learning theory at the time. The operant and classical conditioning methodologies of behaviorism were developed into exercises. Language skills were seen as 'habit-formation' that was best taught through repetition, drills, imitation and memorization of language patterns and dialogues. The contrastive analysis hypothesis developed by Lado (1957) was also integrated in the method and exercises with minimal pairs were used extensively. Great importance was attached to pronunciation, and the development of automaticity. Errors should be removed as soon as possible as they would otherwise become 'bad habits' that would be difficult to remove later, creating phonological fossilization.

In the 60s, along with the cognitive revolution and criticism of behaviorism, came also criticism of ALM. As cognitive psychologists developed new views on learning in general, arguments were put forward that mimicry and rote learning (learning by repetition) was not sufficient and that language learning involved affective and interpersonal factors, and that thinking processes themselves led to the discovery of independent language rule formation (rather than

"habit formation"). ALM was criticized for its emphasis on rote learning, and mindless repetition drills, not focusing on building the students' intrinsic motivation to learn, but viewing language competence as mere 'habit forming'. Many new methodologies entered the arena of second language teaching during this period with heavy competition between rival methods. The biggest and most influential method is what has become known as communicative language teaching (CLT).

2.4.3 Communicative language teaching (CLT)

CLT is not one method but an "umbrella" term covering a variety of methods, but has since the mid 70s been considered the accepted norm among language teaching methods. The basic premise is that language learning is learning to communicate, not learning structures, sounds or words. CLT has been seen as a response to ALM with what its critics considered an over-emphasis on repetition and accuracy, which ultimately did not help students achieve communicative competence in the target language. CLT has placed an emphasis on learning to communicate through interaction in the target language. Unlike the ALM, its primary focus is on helping learners create meaning rather than helping them develop perfectly grammatical structures or acquire native-like pronunciation.

Since the late 90s and onward the strict adherence to "one method" changed and SLA has entered an era by many referred to as the "post method" era (Dörnyei, 2009).

2.5 Skill building summary

Some very different approaches to language teaching have at different times in SLA history been dominant methods. Grammar translation, emphasizing explicit knowledge, Audiolingualism, focusing on automatizing language skills through memorization and repetition drills, and CLT, with focus on implicit learning through exposure of meaningful communication have each been highly influential. In their pure forms all three approaches have been criticized and found lacking (Dörnyei, 2009). Contemporary SLA researchers no longer adhere to pure teaching methods, and therefore in the current "post-method" era

of language instruction the question is not so much "which method is best", but rather "which combination of ingredients is best".

To facilitate automatization, a system should involve explicit initial input components that are then 'proceduralised' through practice (Ellis, 2006; DeKeyser, 2007). According to Dörnyei (2009) the key to the effectiveness of the associative stage of proceduralizing is to design interesting drills that are not demotivating.

Games can not only make it more interesting, and thus increase time on task, but have also the potential to change the learning experience into something entertaining, and by the merits of being fun change the motivation for doing the task, possibly increase the uptake of the information, thus making every minute involved in the skill-building task more effective.

–3–

Phonetics, phonology and CAPT

What is there to learn?

A fundamental part of learning a new language is to get to grips with the salient features of the language. When viewed from a particular level of abstraction, all human languages are built up in strikingly similar ways. At the lowest level there are some arbitrary sound units that function as symbolic message units (phonemes) that in themselves lack meaning. Only through combining these units together into larger chunks of sound combinations do they carry meaning (words). These arbitrary sound chunks will through an agreement among the speakers of a language represent various semantic units in the world (symbolic reference), and by placing these chunks in a particular order (sentence) they carry a larger meaning, displaying agency among the units etc. All spoken languages display some levels of phonetic, morphological, syntactic, semantic, and pragmatic structure.

When we look closer however, we find a wide variation of alternative solutions on how to encode messages. As for the phonetic aspects, one would think that with the same physical apparatus (lungs, vocal tract, tongue, and lips etc.), human languages would end up using pretty much the same sounds, but that is not the case. According to The UCLA Phonological Segment Inventory Database (UPSID), who have made a phonetic inventory of 451 of the world's languages, the number of phonemes in each different language spans from 11 to141 (Maddieson, 1980).

All phonological systems are based on contrasts, just as the whole linguistic system is based on contrasts from a structural linguistic point of view. In the words of the father of structuralism, Swiss linguist Ferdinand De Saussure: *"A linguistic system is a series of differences of sound combined with a series of differences of ideas."* (De Saussure, 1986)

Counting the number of phonemes in a language means in effect counting the number of sounds that for a speaker of the language create a perceptual contrast that could change the meaning of a message. But the sound can sometimes vary a lot and still be considered the same phoneme. What is considered an allophone in one language may have a contrastive phonetic meaning in another. In addition, many languages have also evolved ways to represent change by contrasting for example the duration or the pitch of a sound.

Since the encoding of contrast can take place in different layers, it follows that errors also take place in different layers. Wherever a phonetic contrast can be made - a failure to make that contrast (an error) can also be made.

3.1 Transfer and contrastive analysis

When someone is speaking a language other than their L1, we are often able to guess the person's L1 due to what is known as language transfer, or L1 interference. The importance of transfer in language learning should not be underestimated. As Ellis (1994) puts it:

"No theory of L2 acquisition is complete without an account of L1 transfer".

One way of looking at the transfer phenomena is by using some form of contrastive analysis. A contrastive analysis describes the structural differences and similarities of two or more languages. It has been used as a tool in historical linguistics to establish language genealogies, in comparative linguistics to create language taxonomies, in translation theory to investigate problems of equivalence, and in language learning.

3.1.1 The contrastive analysis hypothesis

The use of contrastive analysis in language learning was initiated with the Contrastive Analysis Hypothesis (CAH), developed by Lado (1957). Lado shared

the mainstream behaviorist view of language learning at the time that language is a set of habits, and learning is the establishment of new habits. Transfer of L1 articulatory habits was the root of the problems L2 learners had in acquiring a new language.

CAH claims that difficulties in language learning derive from the differences between the new language and the learner's first language, and that that errors in these areas of difference derive from first language interference and that these errors can be predicted and remedied by the use of contrastive analysis. Lado's perspective was that "those sounds that are similar to the learner's L1 will be easy to transfer, and those sounds that are different will be difficult."

The CAH was widely influential in the 1950s and 1960s, and as described in section 2.4.2 it was used extensively as part of the audio-lingual teaching method. However, from the 1970s its influence dramatically declined, due to the general decline of the audio-lingual method, structuralist linguistics, and behaviorism, with which it was closely associated.

3.1.2 Problems with CAH

The decline was partly due to the political/philosophical shifts at the time, but several problems with CAH were also pointed out.

One of the problems is concerned with the model of description of a language. The descriptions of individual languages have changed over time, in accordance with developments in linguistic theory, and there have been disagreements as to what such a description should include. CAH claimed to be applicable to all aspects of language, but even when looking only at the phonological aspects of it we can see that the granularity of the description could become a problem for the method.

A phone can for example exist in a language, but only in specific positions. If the phonetic description of languages only list which phones are part of the phonetic inventory, but leaves out at which position, a resulting contrastive analysis would conclude that an element that exists in both languages would cause no difficulties, whereas in fact it does if there is a mismatch in position.

For example: In Swedish /f/ exists in initial, medial and final position whereas in Vietnamese /f/ is part of the phonetic inventory, but exists only in initial position. /f/ in final position does cause a serious problem for many learners

with Vietnamese as L1 who wish to learn Swedish. This would be predicted from CAH if position was part of the description, and otherwise the difficulty would have gone by unnoticed.

The same problem could also occur if regional variations are not part of the description. By creating a description that is too general, what is sometimes referred to as "idealization", a contrastive analysis loses much of its power. Odlin (1989) explains that idealization of linguistic data is unavoidable since there are many minute variations in the speech of individuals who consider themselves to be speakers of the same language. He states that, *"The more idiosyncratic variations in a language, the less accommodating contrastive descriptions become."*

CAH claimed not only to be explanatory, but to also be able to predict what difficulties a language learner would have, based on their L1. As the claims of CAH came to be empirically tested, researchers found that there were many kinds of errors besides those due to interlingual interference that could neither be predicted nor explained by CAH (Odlin, 1989).

3.1.3 Alternative models

To come to terms with some of the criticism and difficulties found by empirical testing of the CAH, Eckman (1977) proposed to also include a dimension of linguistic universals in the CA model, in particular typological markedness. (A phenomenon A in some language is more marked than B if the presence of A in a language implies the presence of B; but the presence of B does not imply the presence of A.)

The Markedness Differential Hypothesis according to Abrahamsson (2004) states that:

- Those parts of the L2 that are different from L1 *and are more marked* than in the L1 will result in difficulties
- The level of difficulty a learner will have corresponds with the level of markedness
- Those parts of the L2 that are different from L1 but are *not more marked* than in the L1 will not result in difficulties.

Major (2001) has made an attempt to assimilate transfer phenomena with language universals such as a hierarchy of markedness, and sonority hierarchy, and

also includes an additional dimension of developmental learning in the ontogeny phylogeny model. Major's point is that pronunciation errors are not only dependent on L1 transfer but also changes over time as a learner develop proficiency in the L2.

Contrary to what CAH states, the Speech Learning Model (SLM) proposed by Flege (1995), claims that ***equivalent*** or ***similar*** sounds are difficult to acquire. This is because a speaker perceives and classifies similar sounds as equivalent to those in the L1 and no new phonetic category is established, whereas 'new' (dissimilar or different) sounds are easier to learn because the speaker perceives these differences and therefore establishes new phonetic categories. Beginners will perceptually assimilate most L2 categories to native ones and only if an L2 segment is sufficiently dissimilar will a new L2 perceptual category be established over time, (see section 2.1.3 on Hebbian learning).

It is problematic also within this framework to determine what easier or harder mean, and definitions of similar and dissimilar are not always clear-cut, but Flege's insight that *a learner's ability to perceive a sound contrast determines the difficulty of acquisition*, has important implications for what and how to teach an L2 learner.

3.1.4 Perceptual foreign accent

It is often common to place focus mostly on the L2 learner's production, since a learner's perception and interpretation of the phonetics, phonotactics, and prosody of the L2 are not directly observable. It is however clear from research by for example McAllister (1997) that a spoken foreign accent is accompanied by a perceptual foreign accent.

If the language learner when perceiving L2 speech subconsciously makes use of a template reflected by the phoneme categories of the L1, it will have adverse effects on how the L2 sounds are interpreted. Category formation for an L2 sound may be hindered by the mechanism of equivalence classification, i.e. an adult L2 learner may not be able to create two unique categories for sounds that are similar in the L1 and L2 and will therefore classify the L2 sound using the L1 category. Equivalence classification leads to the conclusion that if one cannot form a new category for an L2 sound, one cannot produce that sound correctly either (Flege, 1987), and learning to perceive L2 features correctly becomes a prerequisite of learning to produce them. The non-native perceptual

disadvantage has been shown to be stronger in background noise, and audiovisual perceptual training has been shown to improve both perception and pronunciation (Hazan et al., 2005)

Odlin (1989) states: *"individuals differ in their perceptual acuity, and it may be that only individuals with especially high phonetic sensitivity will be able to overcome most of the inhibiting influence of phonological patterns in the native language"*

3.1.5 Transfer and CAPT

Well designed CAPT programs may be able to offer learners an environment where differences are highlighted in such a way that learners become aware of them, and hence enhance the learner's phonetic sensitivity. Pragmatically, to be able to predict and rank which features are going to be the most difficult in advance is perhaps not necessary.

What is important is to be able to determine if a feature of the L2 is problematic for the learner or not. If this can be done by some diagnostic means so that relevant exercises for each individual can be compiled, learners could work with these exercises in an interactive way and individual differences can be catered for.

Errors do occur in L2 learners due to other factors than L1 interference, but there is little or no doubt that a learner's L1 will influence the way in which they approach and learn a second language. What has been debated is not if such a phenomenon exists, but whether the proposed theories are able to predict all errors, and degrees of difficulty, and in that sense whether they are useful as tools in SLA.

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 (cf. Ellis, 1994; Meng et al., 2007). One must however keep in mind that not all features and contrasts in a language are a source for pronunciation errors, and that not all pronunciation errors are equally detrimental.

As part of learning a new language it is essential to at least be presented with all the new sounds and other contrastive features of the language. A VLT could give a crash-course in the L2 phonology by means of presenting a contrastive analysis between the L1 and L2, which could give the learner an overview of

what elements exists in the L2, which elements are the same as in their L1 and which are new, and thus something to pay special attention to.

In the CALST project described in chapter 13, a novel way of using contrastive analysis is being implemented as a part of a Norwegian CAPT project using the Ville framework.

If contrastive analysis can be considered a 'top-down' approach to the explanation and prediction of potential difficulties, the bottom-up equivalent would be to empirically collect and analyze data from L2 learners, and ask experienced language teachers, and phoneticians with a background in foreign pronunciation and accent research about their findings.

3.2 Pronunciation error categorization

While transfer and contrastive analysis focus on difficulties in acquiring features of the new language from the learner's point of view, the perspective of how a particular pronunciation error or accent is perceived from the receiver's side is not covered. Some errors that may be noticeable will not cause any difficulties in understanding from a native listener's point of view, whereas other types of errors will cause serious problems for the intelligibility of an utterance.

Although it will perhaps be desirable for many people to reach native-like pronunciation, and learn to speak with no accent at all, it is as discussed elsewhere not necessarily the primary target. Regardless of what the final aim is, any learner will benefit from realizing the impact of various errors.

According to Bannert (2004) pronunciation training for adult immigrants in Sweden should primarily focus on teaching them to pronounce Swedish in such a way that people find it easy to understand what he or she is saying, which implies that it is ok to have an accent, but not any kind of accent.

Bannert (2004) investigated pronunciation difficulties in second language learners from 25 L1 languages, with Swedish as target language. The main motivation for this work was to create guidelines for teachers of Swedish as a foreign language. In order to get recordings that were representative for each language group, and that covered all aspects of pronunciation without making

the material, and hence the analysis part of the investigation, too large, the strategy was to keep the number of subjects low, and the size of the recorded material for each subject high. Subjects/informants with various L1 backgrounds were invited and a screening process was conducted, so that subjects that had what was considered a 'typical accent' for that L1 by a group of judges (teachers) were selected to participate.

The informants made recordings that were both read speech, and free speech guided by pictures and sequences of pictures. Both sentences and isolated words were recorded in each of the two categories. The read texts were designed to cover various aspects of the Swedish language and highlight pronunciation difficulties.

A comprehensive table listing the difficulties for each of the 25 languages was made, and in addition attention was given to re-occurring difficulties across L1 groups, as well as a categorization of errors based on the seriousness from an intelligibility point of view. Based on this analysis Bannert sorted errors on an intelligibility scale as a guideline for what aspects of pronunciation should be prioritized in pronunciation teaching. The most serious errors in ascending order according to Bannert are shown in Table 1.

The initial work on creating pronunciation error detectors for the Ville framework is inspired by Bannert's work as will be described in section 6.2

1	Lexical stress	Insufficient stress marking, or stress on the wrong syllable
2	Quantity	Wrong duration of a stressed vowel or postvocalic consonant (often neither long nor short)
3	Syllable structure	Incorrect number of syllables in a word.
4	Consonant clusters	Vowel insertion (epenthesis) in, or before a consonant cluster, or consonant deletion in a consonant cluster before a stressed vowel.
5	Rhythm	The relationship between stressed and unstressed syllables in a sentence is wrong
6	Vowel quality	Difficulties with Swedish vowels not present in L1

Table 1 The most serious errors for learners' of Swedish with respect to intelligibility according to Bannert (2004).

3.3 A brief introduction to Swedish phonology

Since the target language in this thesis is mainly Swedish, the characteristics and difficulties of Swedish in particular are described.

3.3.1 The Swedish vowel system

Swedish is notable for having a large vowel inventory, with 17-22 different monophthongs, depending on how one counts. There are nine vowels (/ʊ/ /ɔ/ /a/ /ɪ/ /e/ /ɛ/ /ʏ/ /ɵ/ /œ/), that occur in pairs of long and short, with a substantial quality difference apart from the length, thus 18 vowels in all. Because of the small difference in vowel quality between short /ɛ/ and /e/ in standard Swedish, it is sometimes counted as the same, thus 17 vowels as shown in Figure 2. Changes in vowel quality in many dialects (including standard Swedish) due to the vowels' position in a word (pre /r/ allophones) raises the count to 22 (Elert, 1995).

Many L2 learners of Swedish find the vowel system very complex and difficult to master. A CAPT system allowing language learners to practice this in a self paced manner, on their own computer at home, is therefore an attractive and potentially valuable asset (see chapter 5).

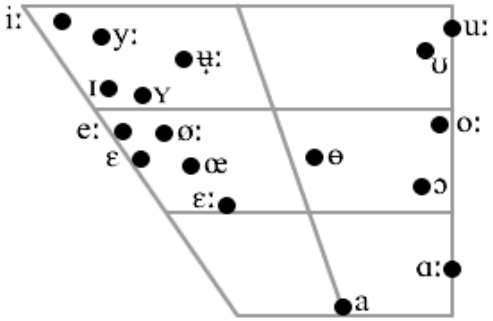

Figure 2 Vowel chart of the Swedish vowel system, with 17 monophthongs (Engstrand, 1999)

3.3.2 Consonants

There are 18-23 consonant phonemes in Swedish, of which two (/ɧ/ and /r/) show considerable dialectal variation. There are ten voiced /b d g n m ŋ v j l r/ and eight unvoiced consonants /p t k f s ç ɧ h/ in all dialects.

In several dialects, including central standard Swedish, the combination of /r/ with dental consonants (/t, d, n, l, s/) produces retroflex consonant realizations, i.e. /t/ as /ʈ/, /d/ as /ɖ/, / n/ as /ɳ/, /l/ as /ɭ/, and /s/ as /ʂ/. Thus, /kɑːrta/ ("map") is realized as [kʰɑːʈa], /nuːrd/ ("north") as [nuːɖ], /vɛːnern/ ('Vänern') as [vɛːnəɳ], and /fɛrsk/ ('fresh') as [fæʂːk].

A set of three sibilants exists /ç ʂ ɧ/ (in addition to /s/), with phonemic contrast between /ç/ and /ɧ/ and allophonic variation between /ɧ/ and /ʂ/, although /ʂ/ and /ç/ are acoustically more similar than /ɧ/ and /ʂ/ (kärna vs. stjärna /çɛɳɑ/ vs. /ɧɛɳɑ/ alternatively /ʂɛɳɑ/).
Although by many considered to be the most difficult aspects of Swedish pronunciation for foreign students, these sounds are listed among the less important pronunciation goals in Bannert (2004), where confusion of /ç/ and /ɧ/do not cause any communicative problems as long as /ɧ/ is realized as /ʂ/ or /ʃ/.
The difficulties that learners of Swedish experience depend, as mentioned above, on their language background, but the Swedish consonant inventory is similar to that of Norwegian and so the table of minimal pair exercises developed for learners of Norwegian in the CALST project (Table 18 p.167) would also be applicable for learners of Swedish.

3.3.3 Phonotactics

Phonotactics defines the language-specific restrictions on what combinations of phonemes are permissible. Phonotactics deals with syllable structure, consonant clusters, and vowel sequences by means of phonotactical constraints.

Though not as complex as that of most Slavic languages, Swedish has what is considered a complex syllable structure similar to that of English. The onset has a limit of three consonants, though there are only six possible three-consonant combinations. The coda is less restricted due to additional suffix endings, but although there are theoretical constructions like for example "Ernstskts" with 8 consonants in the coda position (Abrahamsson, 2004), words with more than five consonants in coda are rare and would be a tongue-twister also for a native Swede.

Words with three initial consonants (for example: spricka, skriva, stryka, skvätta) and three to five consonants in the coda (e.g. falskt, skälmskt) are however common, and these consonant clusters cause considerable difficulties for many L2 learners of Swedish with less complex syllable structures, such as for example Spanish, Japanese, or Chinese.

3.3.4 Tones:

Swedish is actually a tonal language where word accents are differentiated, with two tones called acute and grave (sometimes called accent 1 and accent 2). Swedish intonation does not refer to high or low tones, but rather rising or falling pitch. The actual realizations of these two tones vary from dialect to dialect, and in Finland Swedish the tonal word accent is not used at all. It is hence possible to do without and still be understood (although with a notable accent).

3.3.5 Overview

Table 2 below gives an overview of Swedish phonology. Two main sections can be identified in the phonology of Swedish, (as with other languages): prosody and segments, which are illustrated in Table 2, upper and lower part respectively. With respect to prosody, Swedish has three phonologic contrasts: stress, quantity and tonal word accents. The rightmost lower column does not stand for a category of contrasts, but unites diverse phonological processes, where different segments influence the realization of others. The top two leftmost columns correspond to the two most important pronunciation goals, according to Thorén (2008), as well as Bannert (2004). Although there are some differences in opinion regarding the ranking internally, there is large agreement among Swedish L2 researchers that the temporal organization is the most important aspect of acquiring good Swedish pronunciation (see for example Kjellin, 2002; Bannert, 2004; Abrahamsson, 2004; Thorén, 2008).

Stress is in many ways considered 'the key' to Swedish pronunciation, and many segmental, temporal and tonal deviations by L2 learners are connected with a lack of control of the stress patterns in Swedish. Since for example the quantity distinction in Swedish is only realized in stressed syllables, someone who has difficulties mastering the lexical stress aspect will also make errors with respect to quantity.

1. Stress	2. Quantity	3. Tone /Intonation
'kanon – ka'non ['kɑ:nɔn]–[kɐ'nu:n] 'canon' – 'cannon' *'racket – ra'ket* ['rak:ət]-[rɐ'ke:t] 'racket' – 'rocket' *'Japan – ja'pan* ['jɑ:pɐn] – [jɐ'pɑ:n] 'Japan' – 'japanese' *armen – armén* ['ar:mən] – [ɐr'me:n] 'the arm' – 'the army' *'planet – pla'net* ['plɑ:nət] – [plɐne:t] 'planet' – 'the plane' Stress can be distinctive also on sentence level, e.g: ***'hälsa** på någon* – 'Greet someone' *hälsa **'på** någon* 'Visite someone'	*gl<u>a</u>s – gl<u>a</u>ss* [glɑ:s] – [glas:] 'glass' – 'ice cream' *v<u>i</u>t – v<u>i</u>tt* [vi:t] – [vɪt:] 'white' – 'white' *vila – villa* [vi:lɐ] – [vɪl:ɐ] 'rest' – 'villa' *s<u>ö</u>t – s<u>ö</u>tt* [sø:t] – [sœt:] 'sweet' – 'sweet' *b<u>u</u>s – b<u>u</u>ss* [bʉ:s] – [bɔs:] 'mischief' - 'bus' *r<u>o</u>t – r<u>o</u>tt* [ru:t] – [rʊt:] 'root' – 'rown' *vit – vitt* and *söt – sött* differ in Swedish grammatical gender.	Terminal tone ↘ Continuation tone → Accent 1 (acute) *ánden* 'the duck' *tómten* 'the building site' (Receives its tonal properties from the sentence intonation) Accent 2 (grave) *ànden* 'the spirit' *tòmten* 'Santa Claus' ↘
4. Vowels	**5. Consonants**	**6. Assimilations and the like**
"Hard"/Back vowels /a u ʉ* o/ "Soft"/Front vowels /e i y ɛ ø/ Qualitative/ Spectral difference between long and short allophone Substantial for /ɑ / and / ʉ / Negligible for /ɛ/ and /ø/ */ʉ/ is not a back vowel in a strict sense, but is grouped in the back vowel class because of similar impact on certain preceeding consonants.	Voiced */b d g j l m n r v ŋ/* Voiceless */ f h k p s t ɕʃ/* Phonotax: Swedish allows initial clusters of 3, e. g. *skriva* 'write' and finale up to 5, e.g. *skälmskt*, 'roguish' but in reality, a maximum of 3-4 finale consonants are realized.	/r/ + dental consonant = supradentals (retroflex sounds) *bord* [bu:ɖ] *svart* [svaţ:] *kurs* [køʂ:] 'table' 'black' 'course' Voicing assimilation *Onsdag* > ['ʊn:stɐ] Nasal assimilation *En bil* > [əm'bi:l] Word boundaries: Liaison *Han hete rAnders* Word-finale r-deletion before consonant: *Vad äte(r)Bengt?* No doubling of identical (or semi-identical) segments: *Vad läse(r)Rickard* *Vet du* > *Vetu* etc.

Table 2 Outline of main areas and contrasts in Swedish phonology. from Thorén (2008) with permission.

3.4 Summary

As mentioned in section 2.1.4, self-monitoring and ample time to practice will enable learners to acquire many aspects of the L2 by themselves. Many pronunciation errors are for example a result of confusion about the connection between sound and spelling, and then the cognitive basis of the problem is not perceptual. In such cases the most important elements of the CAPT program is to have relevant content containing all the necessary elements, and encourage time on task by making the acquisition process interesting and entertaining.

In some cases however, the learner will not be able to perceive a sound contrast in the L2, and will erroneously perceive and classify similar sounds as equivalent to those in the L1 due to transfer phenomena. Repeating what was noted by Flege (1995): a learner's ability to perceive a sound contrast determines the difficulty of acquisition.

Consequently, a CAPT program should have two roles: one role for exercises where learners are able to rely on their own perception as a feedback mechanism and regulator of success, and another role for the aspects of the L2 where they are not able to perceive a sound contrast.

Since the problematic aspects of the L2 are not the same for all learners, the CAPT system should be able to assist learners (and teachers) in identifying the problematic aspects for each individual, and work on these contrasts with special exercises.

Perception and production exercises designed to highlight such contrasts can then help, and relevant feedback that the learner is able to understand is essential in this process.

A CAPT system informed by L1 specific filtering should include perception elements in which learners identify and discriminate among problematic sounds.

It should also include production elements in which learners have to produce potentially problematic elements, and some mechanism to judge the learners' pronunciation must be part of the system for this approach to be viable.

–4–

The Ville framework

Ville is intended as a framework to support a multidisciplinary collaboration between language teachers, linguists/phoneticians, speech-technology researchers, and other technical experts.

The long-term vision has been to create a virtual language teacher (VLT) that can serve as a valuable addition to traditional classroom teaching, in that it is available when the learner has time, rather than when the teacher has time. This would hence allow for 'one-on-one' practice, and taking advantage of the computer's processing power, audiovisual capabilities, and 'infinite patience'.

It should also serve as a research tool with the ability to collect and store data from user interactions, and as a platform for research questions such as how to best detect pronunciation errors, how to give feedback in the most effective way, how to most effectively present new information to the learner, and how to make the learning experience an efficient and enjoyable activity.

In other words, the aim for this thesis has been to lay the necessary foundation for a much larger investigation and thus further research in the field of second language acquisition.

The design process of such a system needs to be both top-down and bottom-up. Top-down in the sense that functionality should be conceptually modularized into separate units and bottom-up by making an example application for one particular user group, with a specific linguistic background and with a specific target language. Building a framework that allows for growth, and starting with something basic that can be iteratively built upon.

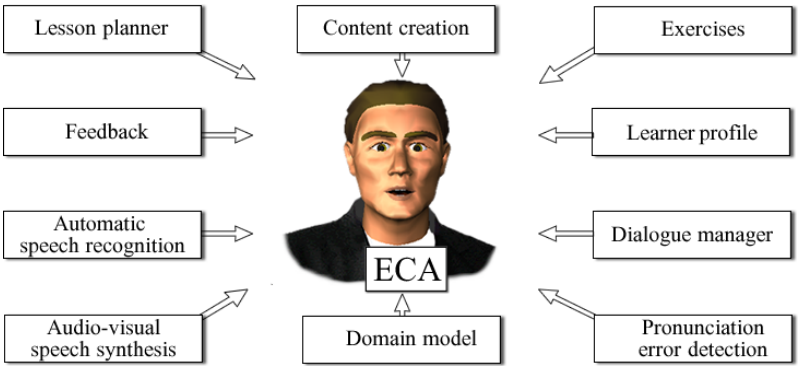

Figure 3 The Ville framework can be divided into functionally separate, but interacting units.

The vision has been to create a universal language tutor, with place-holders for language specific modules and user specific applications (Wik, 2004), as shown in Figure 3.

Separating general tools from user specific tools becomes an important issue, so that adaptation to a new user group may ultimately become a matter of changing some user-specific modules while all else can remain unchanged. Similarly, by separating linguistically universal tools from language-specific ones, adaptation to a new target language will be facilitated.

The long term goal is to make a tutor that would get to know people better the more they use the system, keep track of their improvements and tailor lessons based on their previous history and interaction with the system. The tutor should allow learners to practice dialogues as well as low-level phonetic details. The tutor should thus be able to correct learners' pronunciation and pay special attention to the particular weaknesses/needs the individual learner may have.

A hypothesis used during the development of Ville is that a system able to pinpoint what type of pronunciation error the language learner makes in linguistic/phonetic terms, rather than just a numerical score, will be more instructive and easier for learners to understand. In order to pursue this strategy, insight into the types of errors people make must first be obtained, and the implementation of detectors for specific pronunciation errors must be able to capture phonetic or phonological details in order to give appropriate feedback. This strategy requires an interdisciplinary approach based on language-instruction pedagogy, phonetics, and speech technology.

In order to build a system that is able to detect and give feedback on various kinds of pronunciation errors, we need data on what these errors are and what they sound like. There is, at least in many smaller languages such as Swedish and Norwegian, a lack of such data, so the motivation for the construction of this system is twofold.

A good way to get mispronunciation data would be to give language learners some piece of software where they can practice speaking their new language, and as a side effect collect their data.

The approach is inspired by the "Human Computation", and "Games with a purpose" research of von Ahn (2006). Capitalizing on the fact that there are still things humans do better than computers, von Ahn builds games that, when played by humans, help computers learn. Through online games such as, for example, the ESP game, Verbosity, and Tag a tune, people are collectively solving large-scale computational problems in diverse areas such as image labeling, security, computer vision, Internet accessibility, adult content filtering, and Internet search, by producing useful computation as a side-effect.

von Ahn's players are offered entertainment, and provide researchers with brain power in return. Similarly, although on a much smaller scale, the users targeted in Ville 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.

4.1 Domain model

Recognizing the fact that teachers and linguists, who are the domain experts, are not necessarily skilled computer users/programmers, it is imperative for a system such as Ville to clearly separate structure and content in order to allow for easy creation of content.

The domain model is where the language-specific information should reside. The words, pictures, and recordings are part of the domain model. In future versions of Ville with a wider scope, it can also include other linguistic characteristics of the target language such as the phonetic inventory (see chapter 13), the morphology, and syntax.

In order to have a creative collaboration between technical and domain experts and to iteratively develop the software, the domain model also needs to be open ended, that is, making room for a developer of a new PED for example, to add to the data structure without the need to restructure everything.

The core of the domain model is an XML-structure that is utilized by many of the modules in Ville. It is based on *WordObjects* and *SentenceObjects* with a number of attributes as shown in Figure 4. *WordObjects* can be embedded in *SentenceObjects*. *SyllableObjects* and *PhonemeObjects* to be embedded into *WordObjects*, are considered for CALST (See chapter 13), but not yet implemented. The underlying rationale for this structure is the view that different types of pronunciation errors and exercises belong to different levels or tiers in a hierarchical structure.

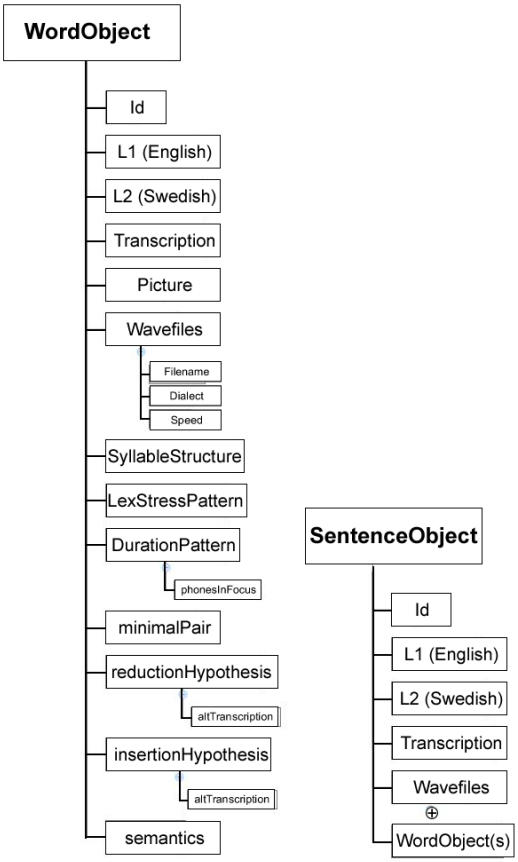

Figure 4 Outline of the elements contained in the XML-structure of *WordObjects* and *SentenceObjects* that constitutes the domain model in Ville

4.2 The embodied conversational agent

A central part of the Ville framework is the embodied conversational agent (ECA) that personifies the whole system. The name Ville is interchangeably used to describe the system, and as the name of the humanoid representation of the system.

4.2.1 The person metaphor

In the early computer days when the command line interface was still prevailing, the desktop metaphor was introduced as a way to make computers more user-friendly by making them resemble the common workplace at the time. The desktop metaphor was a very successful interface and is still ubiquitous in all operating systems today. By taking advantage of the knowledge people already have from other domains, using icons representing files, folders, file cabinets, trashcans and so on, an interface metaphor enabled users to immediately know how to interact with the user interface. The desktop metaphor works very well for office tasks, but other interface metaphors might be better suited for other tasks.

Using the person metaphor rather than the desktop metaphor as an instructional interface for computer assisted language learning (CALL) and computer assisted pronunciation training (CAPT) could be beneficial for several reasons:

- We talk to people, not to papers, folders or trash-cans. The reason for learning a language is ultimately in order to communicate with other people and a program that should mainly focus on speech in and speech out for language learning reasons is a very good candidate for the person metaphor.
- 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; Moreno et al., 2001).
- Speech is multimodal and we communicate more than just verbally through our facial expression. It is well established that visual informa-

tion 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 hard-of-hearing 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 articulation in ways 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 & Bälter, 2007).

Using ECAs for language learning holds a 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.

4.2.2 ECAs at CTT

The development and use of ECAs has been an important aspect of research at the Centre for Speech Technology (CTT), KTH for several years. The ECAs used in this project are created by Beskow (2003).

The ECAs have been used in a wide range of applications. For example, The Waxholm project, giving information on boat traffic in the Stockholm archipelago (Carlson & Granström, 1996); August, a synthetic August Strindberg, who offered tourist information on the centre of Stockholm (Gustafson et al., 1999); AdApt, a multimodal spoken dialogue system for browsing apartments on sale in Stockholm (Gustafson et al., 2000); SynFace, a talking head telephone support for the hearing-impaired (Beskow et al., 2004); and MonAmi, providing services for elderly and disabled persons (Beskow et al., 2009).

4.2.3 Expressive abilities of Ville

The ECAs developed at KTH have the ability to link phonemes to visemes, thus synchronizing acoustic speech with lip movements and other visible articulators. The architecture supports both synthetic speech from text (TTS) and pre-recorded utterances.

TTS still have some shortcomings with respect to, for example, prosody, which gives the technology certain disadvantages when used in a CAPT program. Considering the fact that Ville is supposed to act as a pronunciation model for the students, TTS is not yet mature enough for the task. Ville's voice has therefore been created using pre-recorded utterances.

The ECA used in a conversational setting, as will be presented in chapter 9, on the other hand, has a TTS voice. The utterances for the ECA in the DEAL domain need to be generated in the course of the dialogue, and because his role is that of a conversational partner, not a model of pronunciation, TTS is a suitable solution.

The ECAs can also move other parts of the head than the lips. Non-verbal signals such as head, eye, and eyebrow movements are used to signal e.g. prominence, encouragement, or discourse changes such as turn-taking. The agent is also able to display emotions such as surprise, anger, or joy.

In order to achieve rich, varied and natural movements in Ville, a library of head and face movements has been developed.

Gestures and events

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*.

States

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 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'.

Conversational acts

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.

Scenes

Gestures, ***events***, and ***conversational acts*** can finally be combined into ***scenes***, and a "bag of scenes" is currently the top level in the library of movements for Ville.

```
VilleSay:hello
SwitchToPane 2
$info(perceptionTab,combobox) current 0
LookAtPane
After 1000 (ms)
VilleSay "look at the top square"
makeHighlight perceptionTabCanvas 290 290 370 370
VilleSmile
move2g3 15.0
…
```

Figure 5 Excerpt of a short scene. Each line can be manifested in a variety of ways depending on how the underlying gestures and events are scripted.

The excerpt of a scene depicted in Figure 5 contains nine lines of code exemplifying how scenes can be built up. The scene describes for a learner what to do in a particular exercise.

- `VilleSay:` is a higher order procedure that selects a random recording and its corresponding .dat file (describing the synchronized lip movements) from the document object model (DOM) with the text attribute "hello", and lets Ville say the word (or sentence).
- `SwitchToPane 2:` changes the pane in the GUI displaying the perception pane.
- `$info(perceptionTab,combobox) current 0:` selects the first item in the comboBox on that pane, thus switching to the lexical stress perception exercise

- `LookAtPane:` Is an event, selecting one of several pre-programmed movements where Ville looks at the canvas to his left, indicating to the learner to pay attention to that part of the GUI.
- `After 1000 (ms):` simply pauses for 1 second.
- `VilleSay` "look at the top square": same as above.
- `makeHighlight $info(perceptionTab,canvas) 290 290 370 370:` Draws a red rectangle on a specific part of the canvas, thus highlighting an item on the canvas.
- `VilleSmile`: Selects one of several gestures of the type smile
- `move2g3 15.0:` Ville moves his head to a predefined position at a certain speed (15.0)

If a version of Ville is created in a new target language, the library of ***gestures***, ***events***, ***states***, ***conversational acts*** and ***scenes*** can often be kept intact, and only the recordings of actual words need to be replaced in the new target language with recordings of equivalent semantic meaning. Some scenes may of course be language specific, if they are for example explaining a specific phonological contrast between the L1 and the L2. Cultural differences could also affect the semantics of scenes and must be judged from case to case.

4.2.4 Utterance types

Based on their pragmatic function, there are several different types of utterances to consider in the current version of Ville, all using some or all of the expressions described above:

- ***Content:*** Words or sentences that are part of the material that the language learner should acquire. In for example lexical stress exercises, Ville will utilize a stress marker in the transcription of a word (see architecture section 4.1), and accompany every word with a head-nod ***gesture*** on the stressed syllable. Content is expressed in the L2 language.
- ***Explicit explanations*** of linguistic/phonetic aspects of the L2, such as for example the Swedish duration/quantity phenomena are also scripted as ***scenes***. Instructions and explanations are, in the current Swedish version of Ville, made in English (as the target language learn-

ers are international students in a university setting all with communicative competence in English), but see chapter 13 for a discussion of how this is different in the CALST project, where some of the learners do not speak English, and some are even illiterate.

- ***Instructions*** on how to do certain exercises are scripted as ***scenes***. They will be executed in a context sensitive fashion in a training scenario when the language learner presses the help button, or in a linear fashion. In for example a diagnostic test, a set of instructions and a set of exercises are queued in a list. Learners must follow a fixed set of exercises and instructions are preceding each exercise. Explaining for example the grid of symbols in the lexical stress perception exercise is useful for the learner first time, but not necessary to hear every time such an exercise should be done. Instructions are expressed in English.
- ***Feedback (short-term)*** Utterances such as 'correct', 'incorrect', 'good', etc. in the target (L2) language are scripted as ***conversational acts***. Although some of these expressions may be unknown to the learner initially, they are accompanied by visual gestures such as head-nods that in the context of the exercise will give the learner cues about their pragmatic meaning. There may of course be arguments that such visual expressions are not universal (a head-shake could mean yes instead of no) and cause confusion, but for the large majority of users this will hold true.
- ***Easter eggs*** is a miscellaneous group of ***scenes*** that do not have any specific pedagogical or instructional value, but are there solely to enhance the entertainment value of the program. These scenes are inserted sparsely and at random, to keep the user "*expecting the unexpected*". For example, when the program starts, Ville occasionally has a different hair color. Other Easter eggs could be a sneeze or a burp, followed by Ville blushing and saying 'sorry', or a short whistle, a laugh, or some other trivial but unexpected thing. If verbal, they are expressed in the target (L2) language.

  Some visual examples of the expressive power of the current version of Ville are shown in Figure 6.

Figure 6 Some examples of the expressive power of the current version of Ville

Surely, there is a lot more than this needed in order for an ECA to become 'alive', but it is interesting to note how even this rudimentary implementation has a surprisingly big effect on the user's experience while interacting with the program.

In a preliminary experiment it became clear that people's tendency to anthropomorphize would potentially be a very powerful asset. At the time Ville was only displaying a few idle moves (randomly blinking with the eyes, raising or lowering of the eyebrows etc.) and had no analysis of the learner's pronunciation in his repertoire. It was basically a 'tape recorder' with a head next to it. In the feedback we received from the learners, some said they had observed that Ville had frowned when they had pronounced something wrong, and one user described it as *"It felt like there was someone there helping me while studying"*.

4.3 Automatic speech recognition

Many CALL systems rely on automatic speech recognition to receive and/or analyze input from the learners. Automatic speech recognition (ASR) and Pronunciation error detection (PED) are functionally two different things. ASR determines *what* was said but not *how* it was said, which is the focus of PED.

The difference between these two aspects of speech also reflect the difference between the two opposing second language acquisition theories: Communica-

tive Language Teaching (CLT) and the Audio Lingual Method (ALM). CLT is focusing on the 'what aspect' (communicative abilities) whereas ALM is more concerned with the 'how aspect' (correct pronunciation).

ASR can be used as an aid in reading tutors (c.f. Mostow & Duong, 2009). By presenting text on the computer screen for L2 learners to read aloud, and following their progress by matching expected utterances to what the learner is saying, it is possible to give the learners positive feedback when they read correctly and show them that the ASR did not understand them when they make mistakes.

In a similar way a standard ASR can also be used as an assessment of speaker fluency, by calculating the rate of speech, since the rate of speech has been shown to correlate with speaker proficiency (Cucchiarini et al., 2000). Note that this still says nothing about the quality of the pronunciation.

A language learner can be quite fluent in both reading and talking, and still have poor pronunciation and a strong accent. This fact has also been one of the criticisms for the communicative teaching method, where focus on pronunciation is low, if it at all exists, and the reason why some have called for a return to a 'focus on form' (Doughty & Williams, 1998) as a reaction to the fact that graduates of CLT classrooms may produce language fluently but not very accurately.

Current ASR can also be used in CALL applications based on the CLT paradigm where communicative language skills are practiced. In CLT, correcting pronunciation errors is not what is sought, but rather to speak well enough to be understood. Under such conditions it is possible to use a standard ASR, as an indicator for learner's communicative abilities, or as the backbone in a dialogue system for language learning.

For example in DEAL, a role-play dialogue system for conversation training (described in section 9.2 and in Wik et al., 2007), the challenge of being understood by the ASR is part of the gameplay. In DEAL the ECA does not comment on a learner's performance, but acts as a conversational partner, negotiating meaning, with the objective of creating and maintaining an interesting conversation.

The Tactical Language and Culture Training System (TLCTS) is a commercial CALL system with its roots in a DARPA-funded research project that was originally developed for teaching US military appropriate manners and phrases to

be used on foreign ground, but has since then evolved to also include non-military versions of the system (Johnson & Valente (2008)

TLCTS aim is to let people acquire functional skills in a foreign language. It contains a skill-building part in which learners practice saying words and phrases, and a simulated game world, where learners carry out missions, interacting with non-player characters. It combines game design principles and game development tools with learner modeling, pedagogical agents, and pedagogical dramas.

Earlier versions of TLCTS attempted to detect pronunciation errors on a continual basis in the skill building part (Mote et al., 2004). However, evaluations identified problems with this approach: it was difficult to detect pronunciation errors reliably in continuous speech (leading to user frustration), and the continual feedback tended to cause learners to focus on pronunciation to the exclusion of other language skills.

Developers of TLCTS have since adopted a different approach where the system does not report specific pronunciation errors in most situations, but instead provides a number of focused exercises in which learners practice particular speech sounds they have difficulty with. Johnson & Valente (2008) concludes that pronunciation error detection must be handled as a special case in special 'skill-building' exercises, and not be used in conjunction with ASR.

For evaluation of L2 pronunciation, an ASR must either be modified in some ways or one must use alternatives to ASR.

As described by Lee (2004a)

Efforts in integrating detailed knowledge, from acoustics, speech, language and their interactions, are hampered by the current ASR formulation as a "blackbox" of models trained to "remember" the training data, because it is not straightforward to integrate all available knowledge sources into the current top-down, knowledge-ignorant modeling framework.

4.3.1 Conventional ASR outline

In conventional ASR, typically the continuous speech signal is divided up into 20-25 ms windows spaced by 10ms and analyzed. For each such window, a feature extraction is done by means of digital signal processing techniques based on spectral analysis. The most popular speech frame representation is mel-frequency cepstral parameters (MFCC).

Speech frames (MFCCs) from the input signal are compared with and matched up against an acoustic model in order to be classified as one of the set of units in the acoustic model. The acoustic model in a conventional ASR is typically phone oriented. The most common model is a Hidden Markov model (HMM). Context dependent phone models are statistical models of a phone in a given context. In order to take coarticulation into account, each phone model is split into context dependent clones, called triphones. A link between the phone level description and the word level is usually given by a pronunciation lexicon, which typically provides a canonic, broad phonetic description of the pronunciation of each vocabulary word. A language model (typically n-grams) is used to restrict the possible sequence of words into the most likely during recognition. For a more detailed description, see Jurafsky & Martin (2000).

4.3.2 Alternative ASR models

An alternative to the traditional, spectrally based features (which aim at describing the speech signal as a set of phones), are features that are designed to describe the underlying speech production process. ASR based on features that are describing the interaction between the articulators involved in speech production, have been proposed by for example Tang et al. (2003), Lee (2004a), Siniscalchi et al. (2008). Feature sets used in these examples are based on place and manner of articulation or on Chomsky and Halles distinctive feature theory (Chomsky & Halle, 1968), including features such as sonority, voicing (voiced/unvoiced), manner, place etc.

For example, the ASR described in Siniscalchi et al. (2008) consists of a bank of speech event detectors and an event merger. The goal of each detector is to analyze the speech signal and produce a confidence score that pertains to some acoustic-phonetic attribute. The event merger then combines the event detectors' outputs and delivers evidences at a phone level. Although the system is not designed with CAPT in mind, results from such a detector-based system would give a CAPT system information on exactly the type of features that would allow it to give explicit feedback on a phonetic level.

Another alternative tried in for example the Demosthenes project (Deroo et al., 2000) and the ISLE project (Menzel et al., 2000), is to train the ASR on non-native speech (in addition to native speech). Such systems are able to recognize

non-native, deviant speech based on a given L1-L2 pair, and are also trained to recognize typical errors due to interference from a specific L1.

The Demosthenes database was made for French learners of Dutch, and the ISLE system for German and Italian learners of English. Such an approach can however only be adopted for specific L1-L2 pairs, and requires large amounts of training data by different speakers from the same L1. For smaller languages such as for example Swedish, with immigrants from a large number of countries with different L1, it will be prohibitively expensive to collect the training data needed to create such systems, unless it is done as part of some other activity (as for example, the data collection described in Ville, section 7.3).

An alternative that has been explored by Meng et al. (2007) is to extend a conventional ASR with set of context-sensitive phonological rules describing common mispronunciations in language learners. For example, all plosives and fricatives in Cantonese are unvoiced, and a native Cantonese speaker often substitutes the voiced fricative /v/ with an unvoiced fricative /f/ when speaking English. Hence, one may design the phonological rule (/v/ → /f/) to capture this particular phenomenon. This and other phenomena specific to the Cantonese-English language pair were collected to a set of phonological rules. Given the canonical pronunciation of a word and the phonological rules, a list of possible mispronunciations was obtained, and the lexicon of an ASR was then extended with these mispronunciation prediction rules, as alternative variants.

The results from such an ASR will also provide the CAPT system with important linguistic information that can be interpreted and formulated as corrective feedback to the learner.

4.3.3 Forced alignment

Forced alignment is a very useful tool in CAPT, since it allows to automatically segment an utterance where the expected input is known. Forced alignment can be described as an ASR with a language model restricted to the expected utterance. The acoustic model is the same as in a conventional ASR but the output text is already given. It is hence not used to find out what was said, but is providing information on *how* something was said with respect to the duration of the individual phones. By specifying a transcription together with the acoustic signal, an HMM and a Viterbi search can align segments of the acoustic signal with individual phones. Since the search space is dramatically reduced com-

pared to an ASR where every phone is given a probability at every frame the result is more efficient.

Since what is said is given, the system can more robustly calculate phone borders and durations without the danger of misrecognitions. A forced alignment is, however, more vulnerable than an ASR to cases where the user says something else than what is expected. This can be done either to 'test the system' or because the learner is not able to say what he or she is supposed to say. In such cases an incorrect transcription is forced onto the aligner with very unpredictable results.

Since features in many languages, as for example quantity and lexical stress in Swedish and Norwegian, are acoustically manifested as duration, a forced alignment is potentially a very useful tool in CAPT (Wik, 2004), and it is being exploited in PEDs described in chapter 6.

4.3.4 Pronunciation error detectors

Ideally, Ville should be able to detect and give explicit feedback on all types of pronunciation errors that a language learner is likely to make (although the system may decide not to for pedagogical reasons).

As described earlier, in order to provide learners with corrective feedback, not only a numerical score indicating how native-like their pronunciation is, the aim has been to design pronunciation error detectors (PEDs) that as much as possible are based on phonetic/phonological features.

The implementation of PEDs is an incremental task, and a list of priorities for deciding what detectors to build and in what order is needed.

- 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, and would thus be given priority because they would be a "high-yield investment".
- Some pronunciation errors are perceived by native speakers as more serious than others, resulting in misunderstandings and communication breakdown, and are thus more important to remove.

- Some PEDs are possible to build without large amounts of L2 data, and could thus be given priority for pragmatic reasons, in cases where such data is lacking.

The detectors developed so far are based on Bannert's work (see section 3.2), where the intelligibility from a native speaker's point of view has been given the highest priority. It makes sense pedagogically to start at that end, since those errors would be considered the most urgent to remove. Any learner, regardless of their level of ambition, is likely to start with the aim of reaching intelligibility and optionally moving on towards near-native or native pronunciation. They are also possible to build without data-driven techniques, as described in 6.2.

The PED architecture in Ville is based on the presence (or absence) of specific attributes in the XML-structure as specified in the domain model. Each PED should specify what type of input it requires (for example formants F1-F3, an acoustic model, aligner data, pitch, etc.), and each PED should specify what kind of output one can expect (a binary decision, a numerical score, a percentage etc), and whether it has any possibilities to adjust for 'overshoot' etc. (section 4.4.2).

The lexical stress detector for example, requires the element <lexicalStress> to be present in a *WordObject*, for it to be able to evaluate which syllable in a word is stressed.

Once a pronunciation error has been identified, the VLT must decide if and how to signal this error to the learner, through one of several different types of feedback.

4.4 Feedback

Giving correct and relevant feedback is one of the most important aspects of being a teacher, virtual or real. Yet many aspects of feedback require a great deal of intelligence, creativity, and sensitivity to master and is a skill that many real teachers will spend years to develop.

Here, as well as in many other design aspects of a virtual language tutor (VLT), human teachers and their virtual counterparts have some strong and some weak points respectively.

The strengths of human tutors compared to a VLT are associated with their insights into the psychology of human nature. By the merit of being human, they have emotions, empathy, morality, and other human traits that will give them guidance in when to give feedback, and how much feedback is appropriate. Some learners will be more sensitive than others and different learners will react differently on the same feedback. These are all qualities that are very difficult to apply to a computer program.

Computers, on the other hand, have the ability to make millions of precise and consistent measurements per second, which gives them certain advantages over the human, and will give the VLT complementary qualities to what a human tutor has. Also a computer's ability to sort, compare, and keep track of large amounts of data are qualities that give computers an advantage compared to humans.

4.4.1 Types of feedback in a VLT

Let us look at some different types of feedback that are associated with SLA and see how the VLT will measure up when compared to a human teacher.

Pronunciation error detection feedback:

This type of feedback has received the most focus and attention in CAPT research, and is perhaps the first type that comes to mind when feedback is mentioned in connection with CAPT.

Whether a human or a VLT is best equipped in this category is debatable.

Spoken feedback strategies from human teachers include: explicit correction, recasts, repetition, clarification requests, metalinguistic feedback and elicitation (Lyster & Ranta, 1997). Human teachers will use their intuition to choose which type of feedback to give at each moment, and also consider when not to give any feedback at all.

A VLT also has the option to use spoken feedback in the form of recasts (if what the learner is supposed to say is known), or explicit spoken feedback like for example "your tongue should press against the velum", but deciding which type of feedback to give at any given time is difficult to express algorithmically, and thus difficult to encode in a VLT.

For example Murphy (1991) stressed that teachers must be tactful when deciding on how and when to give feedback about student errors and that students may lose self-confidence if being corrected all the time.

Such affective factors are very individual and it is impossible to calculate or make any a priori assumptions about how a particular learner will react to the feedback given. For a CAPT system this should be a major concern since the necessary sensory apparatus and cognitive understanding to be able to evaluate a learner's emotional outcome of the system's actions are missing.

Technologically, CAPT systems often suffer from an inability to provide accurate and automatic diagnosis of pronunciation errors (Levis, 2007), whereas a human will intuitively be able to recognize if an utterance deviates sufficiently from an acceptable normative standard to be considered an error or not. Acoustically however, the difference between two 'acceptable' versions of an utterance (due to dialect, speed, emotion, etc.) can often deviate more than the difference between an acceptable and unacceptable utterance.

A distinction must here be made between the detector part of the system and the feedback presentation. The learner's production is first analysed by some form of classifier and the result of this will be given some feedback presentation. The technologically difficult part is to correctly assess whether an error has been made, the nature of the error, and possibly some remedial information. The feedback part is more of a pedagogical concern, regarding tactfulness and choice of modality (visual feedback, auditory feedback etc.)

However, from a feedback presentation point of view, given that the error detection was correct, the human instructor has a different palette of feedback options to draw from compared to a VLT.

Apart from oral feedback, a VLT has additional visual feedback options at its disposal. Not all visual feedback is necessarily good, given that the learner must be able to understand and correctly interpret the feedback for it to be meaningful. According to Neri et al. (2003) there are several examples of CAPT programs that have failed pedagogically by expecting learners to be able to interpret (or even imitate) waveforms or spectrograms. Neri states:

If those displays are available in a program, it is simply because of a choice made by the developers (possibly guided by marketing experts who consider technological innovations paramount to pedagogical requirements).

Several aspects of pronunciation still lend themselves well to be represented graphically. For example pitch contours and durational aspects of speech are easy to interpret. A VLT can also show articulatory movements that are not normally visible by making parts of the face transparent. Instructing learners how to change their pronunciation by showing a computer animated model of internal parts of the mouth has been demonstrated by Engwall & Bälter (2007).

Progression feedback:

This category relates to long term feedback of the learners' progression. It is a strong point for a VLT who has the ability to keep track of scores and completed exercises, and display such information in numbers or graphs. A human who has established a personal relation with the learner is able to remember general indications regarding the learner's progress, i.e. whether the learner has improved 'a lot' or 'just a little'. For a more comprehensive report the human is likely to resort to written records or a computer for assistance.

Encouragement feedback:

This type of feedback, aiming at increasing motivation and boosting the learner's self-confidence, is an area where a human (on a good day) will have an advantage. Psychological insights and good timing are human traits that will likely enhance the effect. It is however not an impossible task for a VLT, and to make the system positive, and generous in terms of praise to the learner is not likely to have any negative side effects. Additionally a VLT will not have a 'bad day', and its infinite patience will ensure a consistent positive attitude towards all learners.

Care must however be taken in the design so that the encouragement feedback from the VLT is varied. If for example only one recording is used, expressing the same level of excitement regardless of how well the learner has performed, the learner may not feel that the feedback is genuine, and the emotional effect of the feedback is lost.

Real-time visual feedback loops:

A real-time connection between visual and acoustic modalities, sometimes referred to as transmodal feedback, may hold a real promise in learning low-level pronunciation skills if appropriate, intuitive visualizations are found. Promising experiments have been done in sports psychology, studying the effects of augmented auditory feedback on psychomotor skill-learning (Konttinen et al. (2004). Also in music pedagogy, real-time sound visualisation has been used in

visual feedback loops to train musicians (c.f. Ferguson et al. (2005), Hoppe et al. (2006).

Vowel quality may for example be visualized by creating an immediate feedback loop, as described in chapter 5 and in Wik & Escribano (2009). This is a type of feedback where there is no time for contemplation, or to give explicit remedial information as in pronunciation error detection (PED) feedback described above.

Nickerson & Stevens (1973) developed the first computer-based speech therapy system using visual aids to help hearing-impaired children. Other examples of projects where systems have been using the acoustic signal directly as the source of feedback are SpeechViewer by IBM (Crepy et al., 1983; Öster, 1998), SPECO (Vicsi et al., 2001), and the OLP-method (Öster et al., 2003). This is a type of feedback where technology-enhanced language learning can offer something with a great learning potential, which human teachers cannot possibly give.

Self-monitoring and proprioceptive feedback:

The largest part of a learner's efforts in acquiring good pronunciation in a new language are performed relying on internal feedback mechanisms such as proprioceptive feedback and auditory self-monitoring.

Many errors are 'slip-of-the-tongue' errors, or initial muscular problems that the learner is able to perceive, and all that is required is sufficient exposure and time to exercise, something a CAPT environment can readily offer learners.

Much can be learned by self-monitoring, and when the learner is able to use his or her own perception, external feedback is not only extraneous but can even be perceived as inappropriate and annoying.

Voice modulation is primarily proprioceptive (Oomen & Postma, 2004; Postma, 2000), and when learners are using their own perception and proprioceptive feedback by means of closed-loop motor control in order to consciously adjust muscle movements, motor programs will be strengthened or modified by a tuning process as discussed in sections 2.1.2 and 2.1.5.

The ultimate goal of any learning program must be to make it superfluous once the content is acquired. The aim should be to only give corrective feedback when it is necessary, and one of the challenges in the design of a VLT is hence to find reliable methods for assessing the pronunciation errors of the learner.

Another equally important challenge is to design tasks that are focused on self-monitoring and that encourage the development of self-rehearsal and self-responsibility.

4.4.2 Erroneous Feedback & Pedagogical Overshoot

Erroneous feedback is a common problem in CAPT systems. It is frustrating for the learner if they become aware of it and even more detrimental for the learning process if they don't (Neri et al., 2002a).

There are two cases then the PED is incorrect that should be considered:

- False reject: The learner made no error but the VLT rejected the production
- False accept: The learner made an error, but the VLT accepted the production

If errors are inevitable, is it preferable to give false rejects or false accepts? In other words, if it is possible to tweak the system in either direction should the thresholds of for example a quantity detector be set so that only a very clear, perhaps even exaggerated version of the utterance is accepted? Or should the thresholds be set so that only a clearly wrong version of the utterance is rejected?

It has been argued that it is better to give false accept than a false reject, i.e. to give a green light when a learner is making an error than to give a red light when the utterance was acceptable (Neri et al., 2002b; Eskenazi, 2009), but this is not necessarily the case.

There is no clear border between right and wrong when it comes to pronunciation. First of all it depends on whether the utterance should be judged on a criterion of native-like or comprehensible. Second, there is a large variation also between individual native speakers, and third, many pronunciation errors (for example vowel quantity or vowel quality) are on a gradient with a considerable grey zone, where judgements will vary for different judges based on, the situation, the student, or even their mood.

There are two choices that could be made for the PED thresholds:

"Innocent until proven guilty": Utterances would be accepted, unless the confidence that an error had been done was very high.

"Better safe than sorry": This is pedagogically the opposite strategy to the approach mentioned above, and can be conceived as a "pedagogical overshoot".

When a novice is to learn something it is often customary to change the scale of acceptance somewhat, not accepting something that would have been perfectly acceptable for an expert. Skill-building of this type of events will often be taught by first introducing exaggerated versions of the instances that are to be learnt, or to overshoot as a pedagogical device.

Exaggerating a phonological contrast is common practice in the classroom, both by the teacher to help students better perceive the contrast, and elicited as responses to entice the students to clearly produce the contrast. Tallal et al. (1996) showed that they could remediate children with language impairments when they used a training regime that exaggerated contrasts between plosive stops and other sounds differing by rapid transitions. McClelland et al. (1999) demonstrated a similar effect in adult Japanese learners of English learning to discriminate between /l/ and/r/ using a speech continua ranging from highly exaggerated tokens of "lock" or "load" to highly exaggerated tokens of "rock" or "road". The idea that the use of exaggerated stimuli could induce neural plasticity is also consistent with the findings reported by Merzenich et al. (1996).

One of the strategies that many L2 learners use is "avoidance". If there is a contrast in the L2 that the learner knows is difficult for him or her, rather than learning the distinction people often try to avoid using it. The duration/quantity contrast in Swedish causes much trouble for many immigrants, where many minimal pairs can cause misunderstandings and embarrassment. One common strategy is to try to avoid using such words, but there is such an abundance of them, and the long-short contrast is such an integral part of Swedish phonology that L2 learners end up having to use the words anyway. The next avoidance strategy is then to pronounce the vowel and the complementary postvocalic consonant neither long nor short, thus avoiding saying "the wrong one". This has the unfortunate consequence of making both wrong. The same holds true for other phonological contrasts such as for example lexical stress, where stress on the wrong syllable is wrong - but no stress is also wrong.

Could a pedagogically sound approach to train contrasts that the learners are unable to perceive be to use an 'overshoot paradigm'? Some of the detectors in Ville are designed with the "better safe than sorry" pedagogy in mind (see section 6.2).

4.5 Feedback in Ville

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. One of the great challenges in the construction of a system such as Ville is thus to develop models that in a believable way are able to reflect the complex processes a good teacher is using when choosing what type of feedback to give. As with other modules in Ville, the feedback mechanisms have been built, evaluated, re-designed and re-built in an iterative fashion.

4.5.1 Feedback from pronunciation error detectors

In an early version of Ville, pronunciation exercises on specific phonetic contrasts included verbal feedback, in which Ville commented on the results from the pronunciation error detectors (PED). In a vowel length exercise for example, Ville could say "Good, but your 'e' was a bit too long - try again, say: etta". 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 in line with the findings of Eskenazi (1999). She 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."*

As a consequence, the feedback strategy of this part of the system has been redesigned so that, rather than using verbal feedback, the result from the PEDs is shown as iconic 'traffic lights'. Red or green circles will light up for the active detectors after a student recording.

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.

Since each iconic light belongs to a PED which is designed to detect a specific type of phonetic/phonological error, each light will have a header describing the nature of the error.

In addition many exercises are designed with a focus on a specific phonetic or phonological contrast, and only one light will be shown, so the nature of the error will be apparent to the learner from the context of the exercise.

If the student wishes to know more about 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 text, graphs, or spectrograms, accompanied by verbal feedback. 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 iconic, visual feedback indicates that the error is still occurring, and move on.

As described in section 2.1, the type of feedback a learner needs depends on the learner's developmental stage. More explicit knowledge is required initially, but once that knowledge is grounded, during the associative stage of proceduralization, this information is superfluous. Feedback in the form of traffic lights could be perceived as a faster, more concise type of feedback appropriate for tuning and associative learning.

4.5.2 Verbal feedback from Ville

Even the most fundamental part of the feedback process, such as saying 'correct' and 'incorrect', is a potentially difficult task for a virtual language teacher, and demonstrates the complexity of language use, and the intelligence needed even in relatively mundane tasks.

In perception exercises in Ville, the learner's task is to click on a picture or a button in response to some stimuli that has been presented. The learner's choice is either right or wrong, and Ville will respond verbally by a ***conversational_act*** – which is a combination of recordings and gestures with a common semantic tag, such as 'correct' or 'incorrect' (see section 4.2.3). An effort has been put on creating a considerable number of recordings with different surface utterances, in order to give the user a more varied and interesting ex-

perience, and the impression of a more believable agent. Utterances such as "No, that was wrong", "Sorry, try again", "Nope" and so on, will randomly be called upon.

Constructive criticism during user tests has revealed a preference for a more fine-grained categorization. Utterances with the same semantic tag appear to have different *charge,* and some of the feedback utterances will seem inappropriate or 'odd' as a reply to one specific event, but appropriate if the sequence of events is another. The feedback should somehow reflect the corresponding charge in the learner's action. For example, an incorrect answer will have a different charge the tenth time compared to the first. If this is not conveyed somehow in the feedback it gives the impression that Ville doesn't really know what is going on. It is more important to explicitly acknowledge a correct answer after many erroneous trials than yet another correct answer in a long successful streak of correct answers. A series of correct answers for learner A, who previously has had difficulties with this exercise will have a different charge than the same sequence of answers from learner B, and would by a good teacher (with an internal prediction model of the learner's performance) result in an utterance with a much stronger positive charge etc. This dilemma has not been properly addressed, and to resolve this, a prediction model of the learner's performance is necessary together with methods for how to act when the learner deviates from the predicted. This will be discussed further in chapter 14.

4.6 Lesson management

Each of the isolated CAPT exercises on a particular skill or contrast might consist of several independent components such as:

- An explicit introduction to the exercise, i.e. an explanation of what is being practiced (and why it is important to master this particular subskill)
- Perception training of this particular contrast/skill
- Production training of this particular contrast/skill
- Pronunciation error detection (PED) of this particular contrast
- Feedback

- Performance evaluation/assessment/scoring

It is common to tie together all exercises somehow into a coherent whole on a higher level of abstraction. There are several questions and several choices with regard to this. Should everyone learn the same things? Should everyone learn in the same order? Should everyone spend the same amount of time on each lesson?

Some CALL researchers have criticized the tendency for CALL applications to be technology-driven, rather than pedagogy-driven. In other words the goal has been to make a new application because it is technologically feasible, rather than attempting to put existing theories of language learning into practice (Chapelle, 2001; Hincks, 2005). Here we are standing in front of a crossroads between technology and pedagogy, but it is not obvious which path is preferable.

Linear path:
This is the traditional way of organizing instruction that has been used in schools for centuries. An authority on the subject designs a curriculum which determines what is to be learned and in what order. It is based on the notion that a teacher's job is to transfer knowledge, and that the teacher is the authority who knows best in which order to present subject matter and when it is appropriate to move to the next lesson.

Language specific linear path:
By using for example contrastive analysis, or experience based data it is possible to design a more customized curriculum based on the learner's L1. This way the learner can skip exercises that are deemed as unproblematic for learners with that particular L1 background. Such a strategy does however not take individual differences into account.

User specific linear path:
By some diagnostic means find out the particular strengths and weaknesses of every particular learner and design a user-specific curriculum. The path of progression is still up to the system, but the learners do not have to go through exercises they already master.

Give the learners some autonomy:
Allowing learners to enter places they have already been, but not keep all the content open, i.e. new exercises are opened up as the learner progresses

through the path. The learner can stop and go back (review), but not go forward until some criteria, score or test has been fulfilled.

Give the learners full autonomy:
Give the users the power to design their own training, and only provide a framework and a database of exercises as a 'smorgasbord' for the learners to choose from. Following constructivist notions of how learning occurs, a system should be designed to allow the learner to decide the level, granularity and path to take through the content. The system may advise, and offer suggestions on what exercises to take, but the decision should rest on the learner.

How the learner or the learning process is viewed has an impact on how the content should be presented. If the language learner is viewed from a 'speech doctor' point of view, the user-specific linear path may be a reasonable way to structure the lessons once 'a cure' has been determined. Another alternative is to view the learner as a 'gamer', and model the path as a game of cognitive challenges. Yet another alternative is to try to tailor the training that is presented to the learner based on information such as the learner's preferred learning strategy (Eskenazi & Hansma, 1998), or cognitive, auditory, and linguistic skills (Hazan & Kim, 2010).

The version of Ville described in chapter 10 uses a linear path where Ville presents a set of instructions between each exercise, and the sequence of exercises is fixed. The version of Ville described in chapter 11 is a ‘smorgasbord’ of exercises where the users can choose what to do.

4.7 Learner profile

It is desirable to log the learner’s interaction history with the program for at least two distinct reasons.

Partly it is good for the learner to be able to monitor his or her history and progress as some sort of long-term feedback, and partly it is a way for the program itself to be able to make informed decisions on what to do next, based on the learner’s skill level and progress.

The client-server architecture in Ville enables the program to log every step a learner makes in an exercise and send it to a server at KTH. Long term learner

history such as time spent for each session, and the number of recordings, perception exercises, and writing exercises that are done per session, is saved on the server.

All events such as selecting an item on the grid, making a recording, or pressing the 'listen again' button are chronologically sent to the server together with a timestamp, and saved in each users 'area' as an XML-file, which means that one could in principle replay a learner's session off-line by parsing the individual log files.

The current version of Ville tracks the learner's interaction with the system in several ways. In Ville for SWELL (chapter 7) the login pane visualizes the content divided into 27 topics/lessons and tracks how many lessons the learner has covered. The best score in each category and each type of exercise is visualized as shown in Figure 7.

Short-term learner history is also utilized in Ville. For example during perception exercises, the items that were missed or incorrect are saved as a list and the learner can optionally choose to revise the missed items.

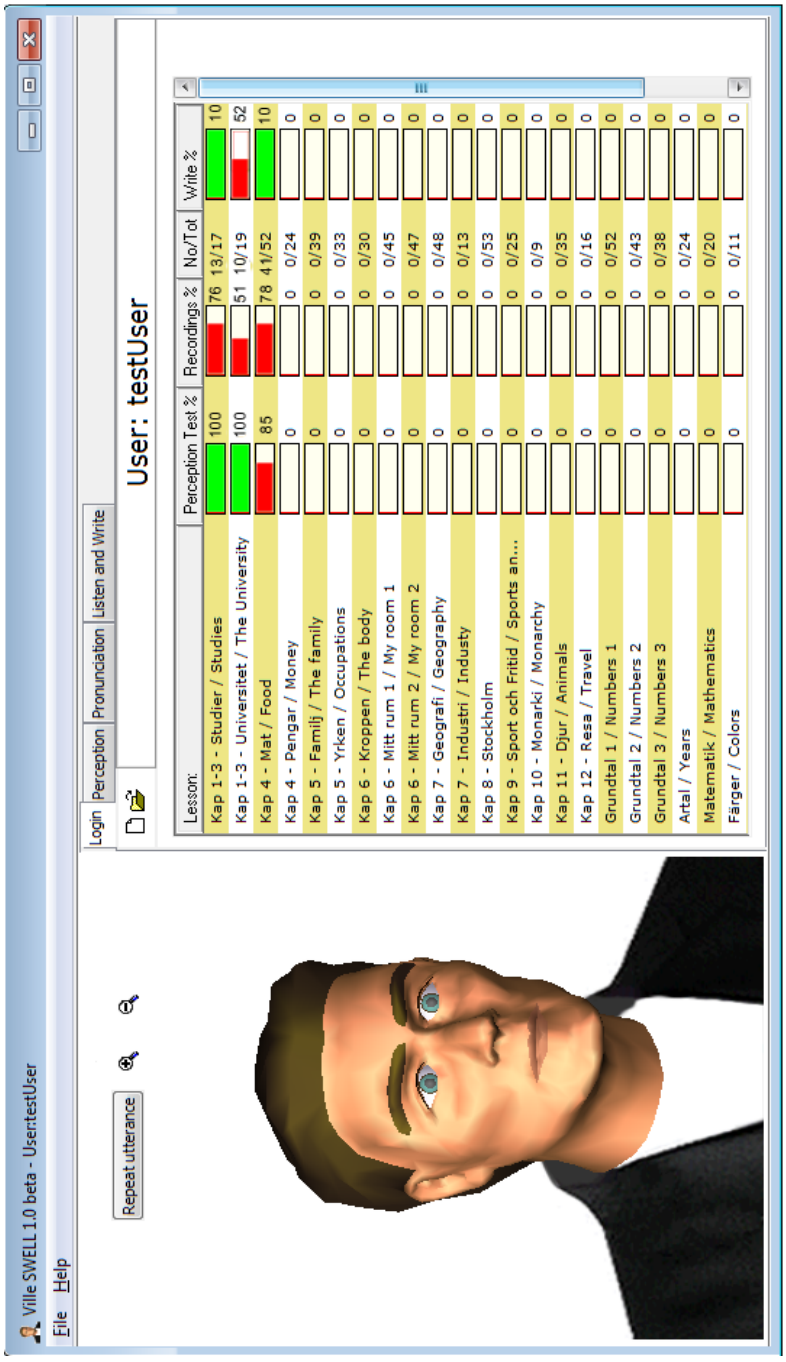

Figure 7 Lessons completed, an aspect of the learner's history is displayed in Ville

-5-

Ville on segmental level

This chapter will demonstrate how some exercises based on low level associative learning can take place in the Ville framework. Examples both in the perception domain and in the production domain are discussed. A production exercise exploring how real-time transmodal feedback can be utilized to help language learners discover new and unknown configurations of their articulators is presented.

Segments are another name for phones, and are commonly referred to as consonants and vowels. More formally, a segment is "any discrete unit that can be identified, either physically or auditorily, in the stream of speech" (Crystal, 2008).

Many language learners have difficulties knowing how to place or move their inner articulators in order to produce the correct sound. Typically learners will have to resort to trial and error, and auditory self-monitoring to figure out how to produce a sound.

A certain level of phonetic knowledge could resolve this by making descriptions using the phonetic nomenclature to inform learners of how to produce novel sounds. Using Ville as a teacher in a crash-course in phonetics and phonology is considered as part of the CALST project described in chapter 13.

Another alternative or parallel track to explore in conjunction with virtual language teachers is to physically demonstrate to the learners how certain sounds are produced by removing parts of the cheek on the models and that way be able to show articulators that are usually hidden.

Figure 8 shows a VLT's possibility to show inner articulatory movements usually not visible.

Models demonstrating articulation, by showing hidden articulators, are not yet a part of the Ville application. This feature should typically be used in conjunction with pronunciation error detectors on segmental level, something yet to be developed for Ville. Such PEDs require training data, something that is not yet available. A database of Swedish spoken with foreign accent is growing through the data collection tool in Ville, and is discussed in section 14.1.1. The availability of such data will make it possible to incorporate hidden articulator demonstrations in the future.

In a Wizard-of-Oz setting, "Artur - the articulation tutor" Engwall & Bälter (2007) demonstrated the feasibility of using virtual agent to teach aspects of pronunciation using a parametric model of a tongue and jaw based on a database of magnetic resonance images (Engwall, 2003). Perception experiments with or without showing hidden articulators as in Figure 8, have also been performed in Wik & Engwall (2008) and Engwall & Wik (2009).

5.1 Perception exercises

As discussed in section 3.1.4, it is important to practice both perception and production in order to master a phonetic contrast in a new language, and this distinction is made throughout the Ville system.

5.1.1 Minimal pairs

Minimal pairs are pairs of words or phrases which differ in only one phonological element and have different meaning. They are often used by linguists studying exotic languages to demonstrate that two phones constitute two separate phonemes in a language. 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 as in Figure 9, and Ville will randomly say one of the words. The learner's task is to identify which word was uttered, and click on it. Ville will then give verbal feedback on the student's choice. The learner can also browse through a set of minimal pairs and click on a card to hear Ville say the word, and that way get exposure to a particular contrast in an otherwise identical setting.

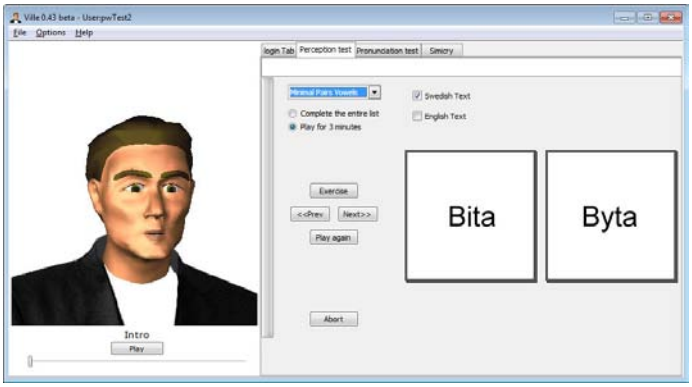

Figure 9 A minimal pair perception exercise in Ville

The same type of exercises but targeted for consonants are also being deployed in the CALST project described in chapter 13

5.1.2 Vowel-grid

Another perception exercise on the segmental level uses a 3x3 grid of cards with the letters of the Swedish vowels on each card, as in Figure 10. When the learner clicks on a letter, a random word (*WordObject)* with that vowel initially will be spoken by Ville. Similarly to the minimal pairs exercise, Ville can also say

a random word, and the task of the learner is to identify the corresponding letter.

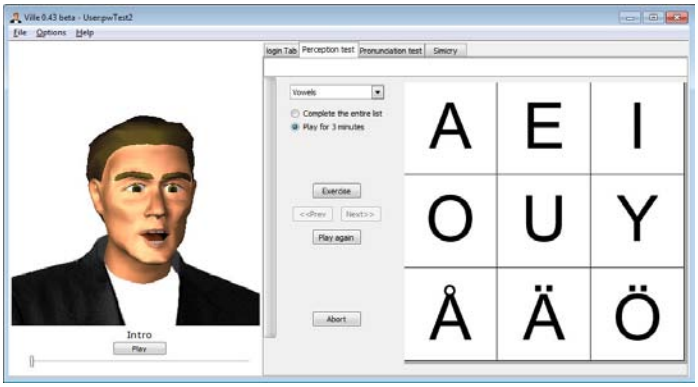

Figure 10 the vowel-grid perception exercise in Ville

5.2 The vowel production game

An exercise for practicing the production of Swedish vowels is presented as an example of how transmodal feedback as described in section 4.4.1 can be used in language learning. Real-time immediate feedback that transforms the audio signal to a visual representation is used in a game scenario.

5.2.1 Formants and vowel charts

Formants are concentrations of acoustic energy around particular frequencies in the speech wave, and are an effect of resonance in the vocal tract. The first formant (F1) corresponds to the front-back dimension and the second formant (F2) to the open-closed dimension of a vowel. They map nicely on the traditional vowel chart (as in Figure 2 on page 35), when F2 is plotted in the negative direction.

Most vowels can be separated in the F1-F2 plane alone, but there are exceptions. Most notably for this work, the distinction between Swedish /ɪ/ and /ʏ/ lies in changing the lips from a widespread position to a pouted. This change will acoustically be noted by a shift of the third formant (F3). To cover the

Swedish vowel inventory, tracking F1 and F2 is thus not enough, but also F3 must in some cases be taken into account.

5.2.2 Implementation

The main part of the software is a 3D canvas with a vowel chart, and a ball as shown in Figure 11. When a language learner speaks into a microphone the ball moves around on the canvas, and will in real-time move to the place on the vowel-chart canvas that corresponds to the vowel uttered by the student, thus giving immediate feedback on the consequences of his/her articulatory movements. The movements of the ball are accomplished by extracting the formants of the acoustic signal, using snack (Sjölander & Beskow, 2000), a sound processing software built into Ville, and using the values of the first and second formant as coordinates on the canvas. To make the movements of the ball smooth, the median of the extracted formants is calculated over a sliding time window. A longer window results in smoother movements, but with the downside of a latency in the movement in relation to the spoken utterance. With a lower latency, and more immediate response, the movement of the ball becomes jerky. A time window of 50 ms, i.e. a refresh rate of 20 frames per second, has been found to give the movements of the ball smooth movements, without a disturbing latency.

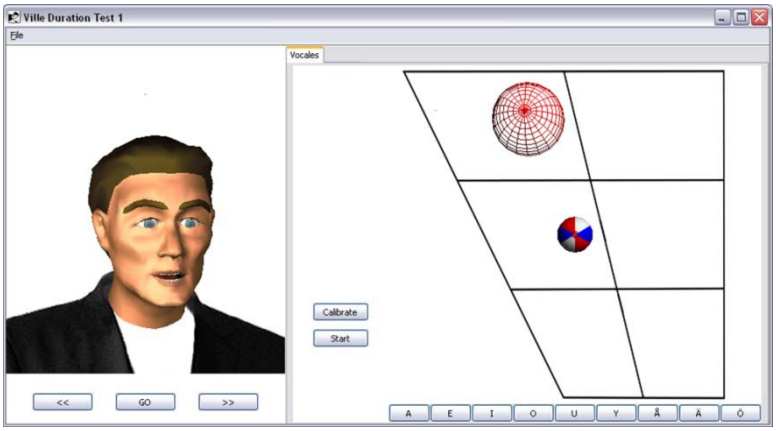

Figure 11 the software, with the moving ball in its resting position, and with one target sphere visible.

5.2.3 Immediate feedback

The direct, immediate feedback of the moving ball is a great facilitator for discovering relationships between configurations of the mouth and tongue, and positions on the vowel chart. By moving the tongue forward and backward in the mouth, the ball moves from left to right on the canvas, and by opening and closing the mouth, the ball moves down and up on the canvas.

Virtually anyone playing around with the software for a few minutes will be able to establish a relationship between articulatory movements and positions on the canvas.

5.2.4 Target spheres

In addition to the vowel-chart canvas and the moving ball, stationary target spheres can be placed at specific pre-determined positions on the canvas. These positions correspond to the locations where the vowels of the target L2 language are found (Swedish in our case).

These positions are determined by having native speakers say the desired vowels and storing the coordinates. The target spheres are a little larger than the moving ball and are, as opposed to the moving ball, not solid but made of a wire-frame mesh, thus making it possible to see the moving ball when it enters the target sphere. A slider is available, allowing the students to change the size of the target spheres, as a way to adjust the difficulty level of the task of getting the moving ball inside the target sphere.

5.2.5 Practice mode and game mode

Two modes are available for the student to choose between. In practice mode the student is free to choose a vowel to practice on, and no time restrictions are given. By clicking on a button with a vowel, the corresponding target sphere will appear on the canvas. When there is no sound input, the moving ball will return to its starting point, which is in the center of the canvas (see Figure 11)

Game mode is a 'catch-the-target-spheres' race against time. Target spheres are placed on the vowel chart, one at the time, and stays until the student has managed to keep the moving ball steadily inside the target sphere for 500 ms. The target sphere then turns green, and is replaced by a new one at another position, corresponding to another vowel. Two versions of the game have been

tried: See how many targets one can get in one minute, alternatively, how long time it takes to get all the targets. For the experiment reported in section 5.2.8, the latter was chosen, to facilitate comparison across subjects and vowels.

5.2.6 /ɪ/-/ʏ/ and the third formant

As mentioned in 5.2.1, the main difference between the Swedish /ɪ/ and /ʏ/, and /i:/ and /y:/ sounds lies in a shift in the third formant (F3). Different ways of visualizing this in an intuitive way that students would be able to understand was explored. Since the vowel chart canvas, the moving ball, and the target spheres are all modeled in 3D, the first attempt was to use the z-axis to represent F3. The standard way of representing the vowel chart is in a plane, where F1 and F2 occupy the x-axis and y-axis respectively. If any movement in the z-axis should be visualized, the vowel chart, now a 3-D cage, must be viewed from an angle. After some initial attempts by students, this idea was abandoned, because it weakened some of the beneficial, intuitive aspects of moving the ball in the traditional x-y plane. Attempts were also made to change the color and size of the moving ball as a representation of shifts in the z-plane. In the end a solution where a binary red/green icon was made visible close to the location of the /ʏ/ in the chart was opted for.

5.2.7 Cardinal vowels as calibration points

The size of the vocal tract affects the formant values so that a man, a woman, or a child saying the same vowel will get different formant values. Fant (1966) drew attention to the fact that the relationship between male and female formant frequencies cannot be described by uniform scaling. This non-uniform scaling of the vocal tract means that if vocalizations of people with different height, gender, or age are to be compared using the formant frequencies, a normalization method must precede the comparison. Making use of the cardinal vowels as calibration points is proposed as a solution to this.

A cardinal vowel is a vowel produced when the tongue is in an extreme position, front-back, or high-low. Since the cardinal vowels are extreme points of articulation, they mark the outer rim of an individual's vowel space and all other vowels are lying within this space. If we are able to elicit some of these cardinal vowels from the users, they can be used as reference points, by scaling the

canvas to fit these points. All target vowels can then be measured in relative distances from them.

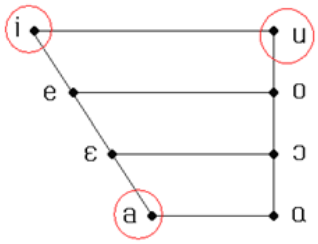

Figure 12 The cardinal vowels, with the three corner vowels used for calibration marked with a circle.

Three cardinal vowels, the corner vowels (see Figure 12) are elicited from the user by an initial interactive calibration phase. Ville starts by giving a short explanation to why this is necessary in order to get accurate measurements. He then proceeds to elicit each individual corner vowel.

The corner vowels are given articulatory definitions. /i/ is produced with spread lips, and the tongue as far forward and as high in the mouth as is possible. /u/ is produced with pursed lips, as in a whistle, and the tongue as far back and as high in the mouth as possible. /a/ is produced with an open mouth, and with the tongue as low as possible, as when going to the dentist, saying Aaaaaa.

There is a bootstrapping problem involved in the calibration phase. A human being can hear if a vowel is mispronounced. This software will measure the formant frequencies, and normalize them relative to a person's corner cardinal vowels. If the cardinal vowels are off, (or from a different person) the analysis of the software will also be off. Since the formant values are based on the size and shape of every individual's vocal tract, we cannot know what the expected values should be. If a user for some reason fails to do the correct articulatory movements, as instructed by Ville, the system could end up with a canvas that is too small, or skewed, and that would affect the quality of the analysis.

Efforts have been made to eliminate this potential problem. First of all by making Ville's explanations as clear as possible, coaching the student into stretching his/her personal vowel canvas as much as possible. After the initial elicitation of the corner cardinal vowels, Ville asks the student to say three easy syllables, /bi:/ /bu:/ /ba:/, containing the easiest, most common vowels. These words are

then run through a forced alignment, the centre piece of the vowel is cut out, and a formant extraction is applied on each of the vowels respectively. If the F1, F2 coordinates fit in the expected areas with a reasonable accuracy, the calibration phase is finished. If not, the whole calibration phase is repeated. Although this method worked successfully on all test-subjects in the experiment described in section 5.2.8, (with some of them doing a second calibration), it is difficult to know if this method is adequate until the system has been tried on a larger set of students.

Making CAPT systems to practice vowel production has been done before (see for example Zahorian & Correal, 1994; Auberg et al., 1998; Paganus et al., 2006). The main contribution of this implementation is thus the calibration technique based on cardinal vowels being elicited from the learner, and using those to normalize the vowel-space canvas, thus allowing all users, regardless of vocal tract size to use the system. Also the third formant, F3 is extracted in order to distinguish between certain vowels in Swedish. The system is not limited to Swedish, as it is fast and easy to make another set of targets, based on the vowel inventory of another language, as long as it is based on formant extraction.

5.2.8 Experiment

10 subjects were enrolled for a user study, to investigate the usefulness of the software as a vowel-learning tool. Five subjects were international language students, and five were native Swedish speakers used as a reference. Among the international students, two were Spanish, two were Italian, and one was from Syria. Both groups had three males and two females.

From the Swedish vowels 10 were selected as part of the experiment. These were the nine long variants (see section 5.2.1) and the open fronted short /a/, which was selected because its vowel quality has a close resemblance to the /a/ sound used in many languages. Since the task in the experiment was to keep the moving ball steadily inside each target sphere for at least 500 ms, it was decided that the long vowels were the most appropriate to try.

The experiments were conducted on a laptop computer with a headset in a quiet private room. Each student performed the experiment on two separate occasions with a few days in between. Each session consisted of a calibration phase, and an initial training period of five minutes, getting acquainted with the

program, before the tests started. On each occasion every student did three consecutive tests, and the times for reaching each target sphere were logged.

5.2.9 Results

To analyze the results, the data was split into four groups: Swedish subjects session one and two, and international subjects session one and two. The distinction between the data from the Swedish subjects and the international subjects is motivated to isolate the effect of getting acquainted with the use of the program under the assumption that all the Swedes already master the Swedish vowels. Comparing first and second session for the Swedish subjects will show the effect of that. Comparing the differences between first and second session for the international subjects and the Swedish subjects is thought to show some learning effects beyond learning to use the program. Inside each of these groups a mean value was calculated for the different Swedish vowels in all the tests.

Learners of Swedish usually exhibit varying degrees of difficulties mastering different vowels. A reasonable assumption would be that they are difficult because they are unfamiliar, and therefore harder to reach. The hypothesis is that the immediate feedback provided by the program would enable students to explore the unfamiliar regions, and that they initially would take a longer time to reach, but that after some training with the program, these areas would not pose a bigger problem than other areas.

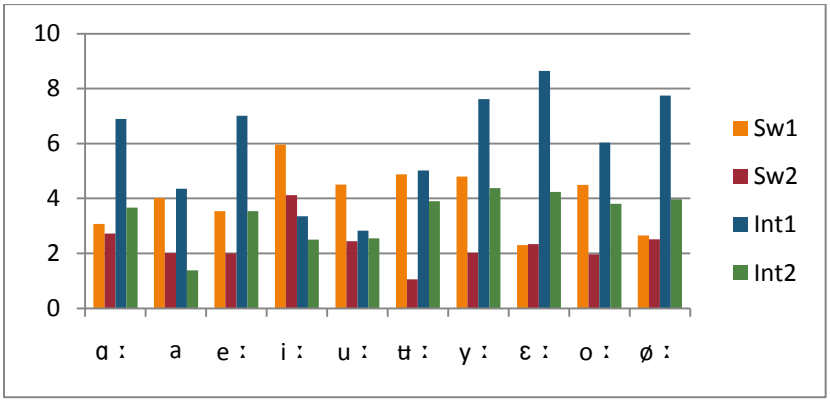

Figure 13 Mean times in seconds for the different vowels divided into four groups: Swedish subjects and international subjects session one and two.

In Figure 13 we see that the 'exotic' /ɛː/, /øː/, /yː/ and /ɑː/ along with /eː/, which is more fronted than in many languages, are the vowels the international subjects spent most time on in the first session.

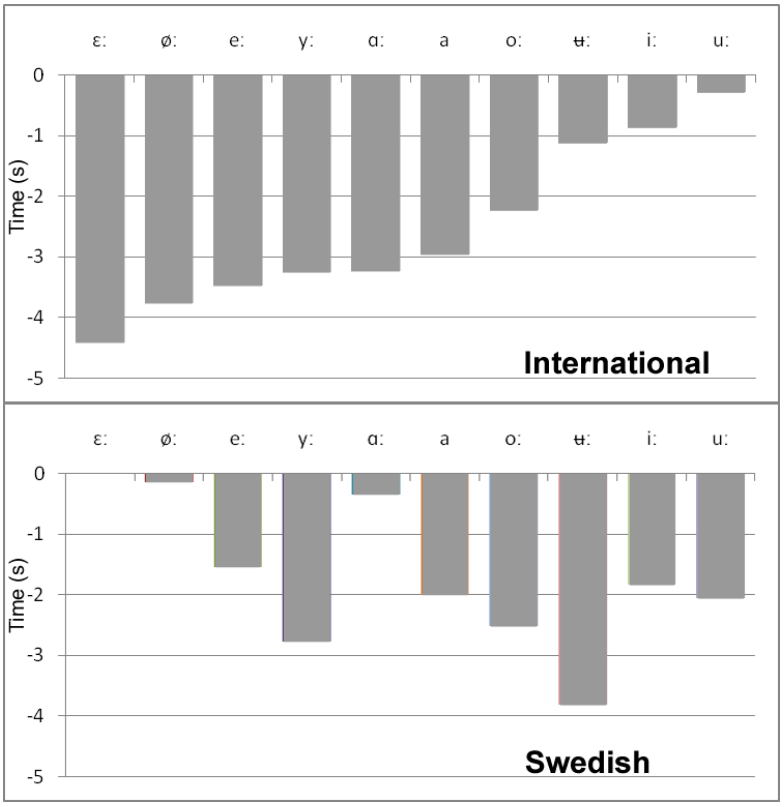

Figure 14 Difference in time (seconds) between session two and one. Top plot: International subjects (sorted), Bottom plot: Swedish subjects.

The top plot of Figure 14 shows the *gain* the international subjects made in time between session one and session two for the different vowels. Here also it is clear that /ɛː/, /øː/, /eː/, /yː/ and /ɑː/ are the vowels the international subjects improved the most on. The gains for the Swedish subjects in the bottom plot of Figure 14 show a very different distribution.

Figure 15 show the results by splitting the data into the six different tests that each participant did (three in each session) and comparing the mean time scores for all the vowels between the international learners and the Swedes.

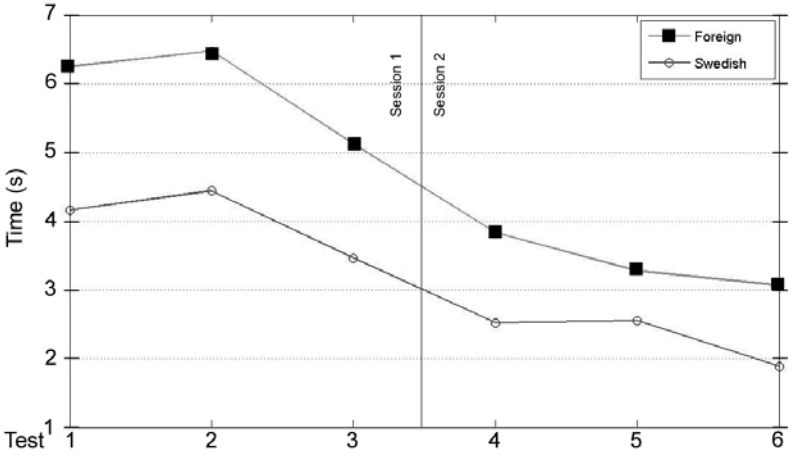

Figure 15 Mean time scores for the foreign subjects on top and the Swedish subjects below, displaying the improvement in time over the six tests.

A one-sample t-test show that both groups improved significantly from test 1 to test 6, ($p<0.001$ for both groups). As can be seen in Figure 15, the average improvement for the foreign subjects was bigger than for the Swedes. In average the difference in response time between foreign and Swedish subjects in session 1 was 1.93 seconds, and in session 2 it was 1.07 seconds. We would like to think that the lines could converge with further tests.

–6–

Ville on syllable level

Above the segmental level in the phonological hierarchy we find the syllable level. As discussed in section 3.2 some of the most serious errors with regard to intelligibility are according to Bannert (2004) lexical stress, quantity, and epenthesis/deletions. These are all suprasegmental aspects of speech in the sense that they cannot be said to belong to a specific phoneme, but coexist over several phonemes, and are thus grouped on a higher level.

These errors also share some characteristics that make the creation of pronunciation error detectors (PEDs) different for them than for segmental errors. The three detectors that are presented in 6.2 are all based on the underlying technology of forced alignment (see section 4.3.3 for description).

6.1 Perception exercises

6.1.1 Lexical stress

Perception exercises on lexical stress are made by a three-by-three grid with symbols representing different stress patterns and different number of syllables, c.f. Figure 16.

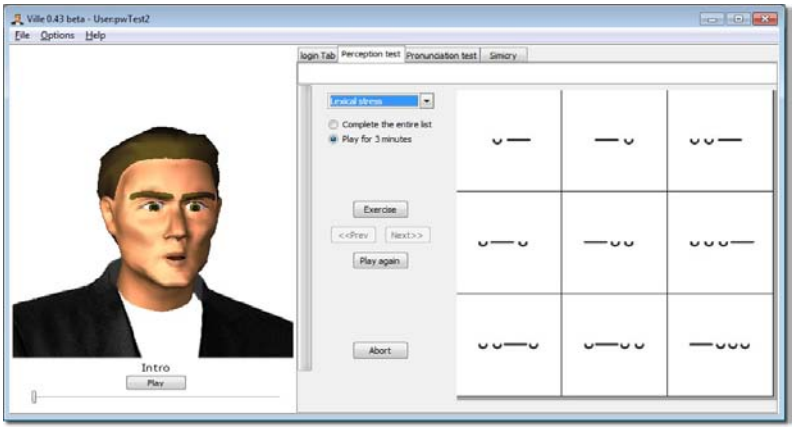

Figure 16 A grid for lexical stress perception exercise: the 'u' shaped symbol is an unstressed syllable, and the — symbolizes a stressed syllable.

For example, a word such as 'telephone' is the same word in both Swedish and English, but will in English be pronounced with stress on the first syllable (—ᴗᴗ **,** i.e. **TE**-le-phone), whereas in Swedish it will be pronounced with stress on the last syllable (ᴗᴗ— , i.e. Te-le **FON**).

When L2 learners' practice or familiarize themselves with the symbols, they can click on a symbol and Ville will search and retrieve all *WordObjects* which are tagged with an appropriate *'lexStressPattern'* element, and pronounce a randomly selected word. This way by clicking consecutively on the same square, a learner will be presented with a number of words that share the same pattern.

Ville will also enhance the perception of the stress by a visual cue. This is done by retrieving the stress marking in the transcription and using this in conjunction with the waveform of the recording to calculate the time of the stress. When Ville pronounces a word he will use this time to trigger an ***event*** (a sequence of ***gestures*** as described in section 4.2.3), at the correct point in time. The ***event*** will be one of several in the library of events which is a nod of the head and quickly lowering and raising of the eyebrows.

6.1.2 Quantity

Perception exercises on quantity can be performed using minimal pairs in the same way as the perception exercises described on segmental level in section 5.1. There are in Swedish many minimal pairs with a quantity/duration contrast,

(such as for example 'byta' vs. 'bytta'). These words are collected as a group and in pairs, in order to allow learners to focus on this phonetic contrast. Ville says a word and the learner's task is to select the correct card.

6.1.3 Estonian duration exercises

In August 2005 the Ville framework was put to use at a summer school, in Palmse, Estonia organized by the Nordic Network on Variation in Speech Production and Speech Perception (VISPP). The aim of using the system at the workshop was to give the participants (organizers, teachers, and PhD students) a hands-on experience of a prototypical VLT and insight into some of the unique aspects of the Estonian language.

In Estonian, a distinction is made between short, long, and over-long vowels and consonants that can be freely combined. Out of the nine combinations possible, seven of them are used in the Estonian language. For example the sound sequence /satɛ/ can be pronounced in seven different ways to create seven distinct words.

In the first exercise the students should learn to perceive the difference between these seven words. The students could click on the squares in the grid as shown in Figure 17 and the VLT would say the corresponding word. The notation used is shown in Table 3.

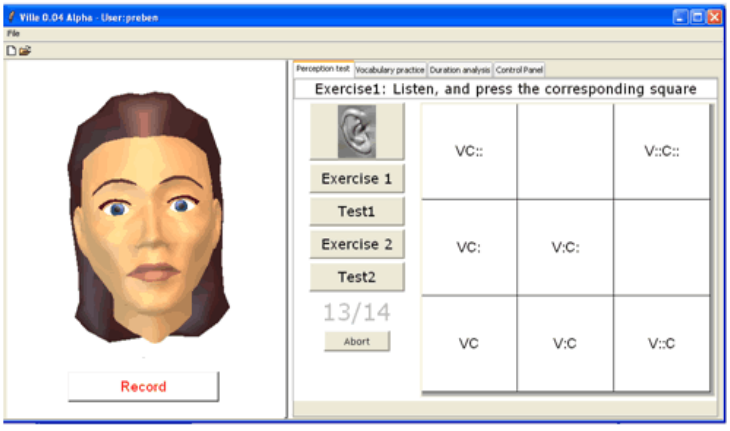

Figure 17 The user interface for a perception exercise on Estonian duration

V = Short vowel	C = Short consonant
V: = Long vowel	C: = Long consonant
V:: = Overlong vowel	C:: = Overlong consonant

Table 3 Notation used for the duration categories in Estonian.

Once they had familiarized themselves with the map, they could do 'Exercises' and 'Tests'. The VLT would say different words in a randomized order, and the student's task was to press the correct corresponding square. Exercise1 and Test1 used the same words as in the map, and Exercise2 used unfamiliar words. During the exercises the VLT gave verbal feedback on whether the students pressed the correct square or not, and they had to continue until they got it right. On the test they only got one chance to get it right on each utterance, and there was no feedback on whether they pressed the correct square or not.

The Estonian phenomenon with a three level quantity distinction is only found in very few languages in the world. According to the contrastive analysis hypothesis, learners with an L1 lacking duration as a phonetic contrast would have a more difficult time with this exercise than learners with an L1 that does have quantity as a phonetic contrast. Our expectations were largely met in that Finns, who do have both long and short vowels and consonants that can be freely combined, found the exercise easier than for example Swedes and Norwegians who have long and short vowels and consonants, but only in stressed syllables and only in complementary distribution (i.e. a long vowel is followed by a short consonant and vice versa). The Scandinavians in turn did better than the Chinese, whose L1 lack duration contrasts.

There was however a puzzling exception with two Chinese students who did surprisingly well on the first exercise. Further investigations in the matter revealed that they had learned to identify the speech samples during the training phase by paying attention to particularities in the recordings, such as puffs of air in the microphone, a creak in the voice, or noise in the recording, and used these cues to link symbols to recordings and thus managing to correctly identify the categories during the test.

The mistake made during preparation of the content was to use only one recording of each word, and using the same recording during training and test. The mistake made by the Chinese students was to attribute saliency to the wrong features in the speech signal, thus not being able to generalize or learn

the right categories. They produced near random on the subsequent test, using the same notation but unfamiliar words.

This is an extreme case of the findings reported by for example McAllister (1998) that it is good to listen to many different speakers in order to achieve better listening comprehension. The probabilistic mechanisms used when learning through associative learning requires abundant repetition within a narrow context, but there must be variation in the samples in order to extract features linked to the target language as opposed to individual features of the speaker, or as in this case features of a single utterance.

This highlights a potential problem with having only one virtual language teacher or only one human teacher as a role model for pronunciation. In the CALST project described in chapter 13, this problem is addressed by having eight speakers with different dialects, four male and four female, in all perception exercises in order to increase the variability of the training material. In order to make this consistent with the person metaphor there are also eight ECAs used in the CALST project. In other versions of Ville where only one speaker voice is used, the problem is addressed by allowing for several recordings of each word used in the training material. Whenever a speech sample is to be presented to the learner, the program looks in the selected *WordObject* and if more than one recording exists it selects a recording randomly.

6.2 Production exercises

6.2.1 Duration/Quantity

The Swedish language has as mentioned earlier, complementary distribution, (i.e. a long vowel is followed by a short consonant and vice versa). 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 the correct distinction between long and short vowels and the following post-vocalic consonant), are detected using forced-alignment (Sjölander, 2003), 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. Due to the complementary distribution in Swedish, the relative duration of the consonant following the stressed vowel must also be taken into consideration.

What is sought in this exercise is to teach the relationship between the duration of the vowel and the following consonant. The rules used for the detectors and feedback in this exercise are thus as follows: For a word with long vowel and short consonant, a learner's utterance is only considered incorrect if the vowel is too short, or the consonant is too long. There are in other words, no rules for giving corrective feedback, if the vowel is very long, or the consonant is very short. The reverse rules apply in a word with short vowel and long consonant: no lower bound for how short the vowel can be, and no upper bound on how long the consonant can be.

When a learner records a word that is determined by the quantity-PED to have incorrect quantity/duration, a red light is shown under the heading 'Duration' as shown in Figure 18. Under the red light a waveform of Ville's pronunciation is drawn, and a waveform of the student's recording is drawn below that. The results from the forced alignment are time markings of each phone start and end. This is used to graphically show the learner what has happened.

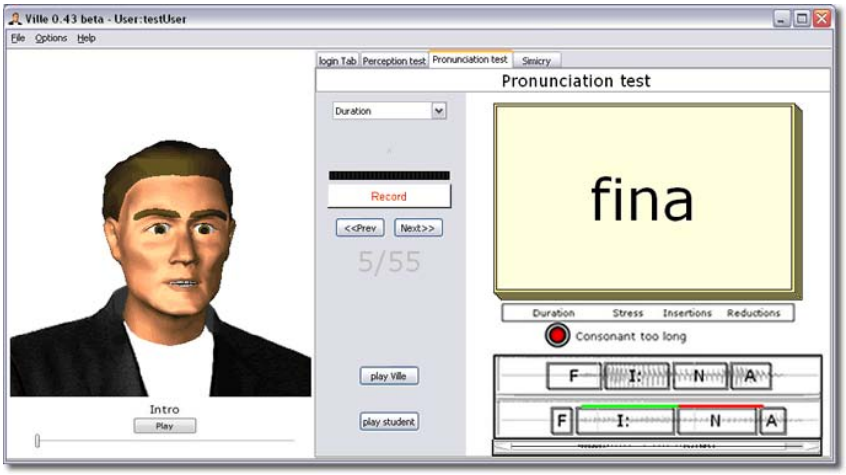

Figure 18 Feedback from the quantity/duration PED if the consonant in the stressed syllable is too long.

A rectangle is drawn over the area that belongs to each phone in the waveform of both the learner's and Ville's utterance respectively. If the consonant is

judged as too long a red line is drawn above the consonant as in Figure 18, and if the vowel is deemed as too short a red line is drawn above the vowel, coupled with the explanatory text as in Figure 19. Two play buttons, 'play Ville' and 'play student', are available for the learner to the left of the waveforms as can be seen in Figure 18.

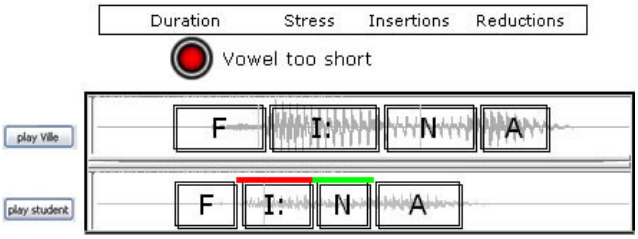

Figure 19 Feedback from the quantity/duration PED if the vowel in the stressed syllable is too short.

6.2.2 Lexical Stress

Lexical stress (giving one of the syllables in a word more prominence) is in Swedish as in English not fixed but varies depending on the word, something that is problematic for students with an L1 with a fixed stress pattern (for example Finnish, Polish, French). Commonly measured acoustic correlates to stress are pitch, intensity and duration. In Swedish, duration has been considered the most important correlate (Fant & Kruckenberg, 1994), 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 PED 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. 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. The visual feedback from the PED is similar to the quantity PED, except the duration of each *syllable* is drawn as in Figure 20 instead of each phone

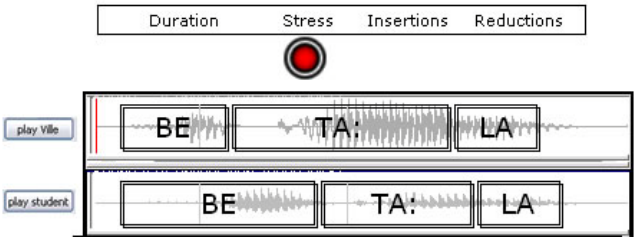

Figure 20 Feedback from the lexical stress PED.

6.2.3 Insertion and deletion detectors

As described in section 3.3.3, the phonological constraints on what sounds can appear in what positions in a student's L1 will often make the student add or remove sounds in L2 words, if they do not follow the same constraints. Insertion (epenthesis) and deletion errors are predictable in the sense that a mispronunciation hypothesis can be created in conjunction with certain consonant clusters. Epenthesis breaks up a consonant cluster or vowel sequence that is not permitted by the phonotactics of a language, and for many L2 learners this carries over to the L2. For example, many native Spanish speakers will when attempting to produce a consonant cluster with an initial /s/ in Swedish insert a vowel in front of the /s/: 'Spanien' (Spain) thus becomes 'Espanien'.

As with the two previous detectors (lexical stress and quantity), the insertion and deletion detectors are based on forced alignment. This PED however is not adjustable, in the sense described for lexical stress and quantity. Both the insertion and the deletion detectors employ forced-alignment in parallel on the original transcription and on one or several modified transcriptions, where a hypothesized insertion or deletion is included. The aligner returns a distance measure using a Viterbi search, where the shorter path means a better match.

By comparing the distance of the original transcription from the hypothesis transcription, the likelihood that an insertion or deletion was made can be determined. This means that one transcription will always be chosen as the most likely representation of what was actually uttered by the learner. If it was the original transcription, the utterance is judged as correct, and if a transcription of a mispronunciation hypothesis is selected as the most likely, then an error is flagged, and the visual feedback to the learner will be something like in Figure 21.

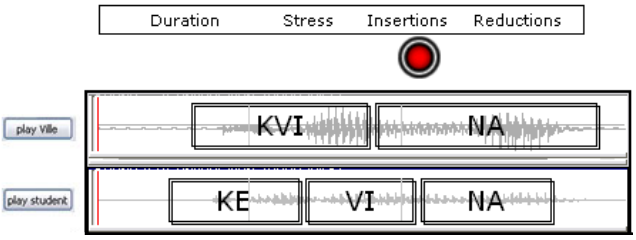

Figure 21 Feedback from the insertion detector.

It should be noted that there are no parameters to tweak here (except possibly setting a threshold measure on how much higher a mispronunciation likelihood must be for it to be selected), thus making this type of PED more difficult to fine tune and more vulnerable. It is possible to imagine adding a threshold on the duration of the measured inserted vowel. If for example the detector returns "Kevina" as the transcription with the best match, a second step could be to look at the measured duration of the inserted /e/, and if it is shorter than some threshold duration the utterance would still be accepted. Samples of L2 utterances from a database could be judged and used to determine such thresholds to further improve this type of PED.

6.2.4 Estonian duration exercises

In the Estonian VLT, described in section 6.1.3, exercises were also done where the participants should produce words that contained the same quantity distinctions that they had first been practicing to distinguish in the perception part.

Since all participants in the workshop were phonetically trained, a more complex user interface was offered containing spectrograms of the recordings, and the possibility to manipulate their recordings. This type of information is otherwise difficult to interpret for non-specialists and has not been used in GUIs for learners who are not phonetically trained.

In the production exercise analysis pane, a spectrogram and a time alignment is shown, both of the teacher's utterance and of the student's utterance. The top part of the pane in Figure 22 displays the teacher's recording, with the duration of each phone marked. The middle part is the student recording where there is, in addition to the spectrogram and the phone borders, a pane displaying the offset ratio between the student's and the teacher's duration on each particular

phone. A rectangle upwards indicates that the student's duration was longer than the teacher's on that particular phone, and downwards means shorter.

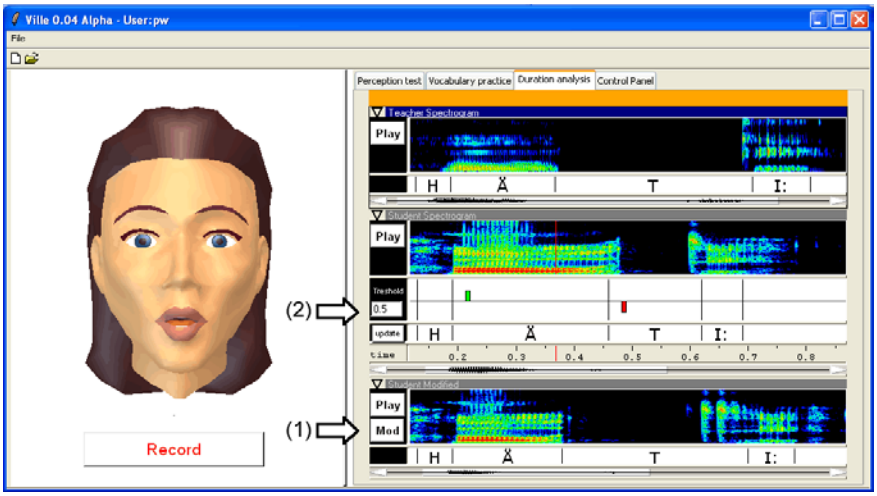

Figure 22 Screenshot of the production part of the Estonian experiment. (1) Modify student (2) Increase difficulty by adusting the threshold

The exercise was to try to match the teacher's duration as close as possible. By reducing the threshold setting to the left of the student pane (2) in Figure 22) the level of difficulty is increased.

The lower part of the pane offers the possibility to create a modified version of the student recording, where the phones in focus are stretched to match the teacher's duration (1).

Investigations on what voice features are preferable for second language learners to imitate when using CAPT systems, indicate that it would be beneficial for L2 students to be able to listen to their own voices producing native accented utterances (c.f. Sundström (1998) Felps et al. (2009). The rationale is that, by removing information that only stems from the teacher's voice quality will make it easier for students to perceive differences between their accented utterances and their ideal accent free counterparts.

–7–

Ville on vocabulary level

The central importance of vocabulary knowledge for language competence is clear. This chapter will demonstrate how the Ville framework is used in a similar way as in the previous chapters to practice both perception and production, with the additional feature of pictures added to the domain model to take the dual-coding theory into account (see 2.2.2 for a description).

At KTH there has been a large increase in the number of students from abroad in recent years, and as a consequence, the unit for language and communication at KTH responsible for language instruction has seen a sharp increase in students who wish to learn Swedish. An effort was made to create online resources in order to offer the students alternatives to classroom teaching. Under the working name SWELL (Swedish for Elementary Learners) a set of self-study tools were created at KTH in 2008, and Ville-for SWELL was one of them.

Exercises for vocabulary training were a high priority for the language unit, and were therefore implemented as the main part of what Ville should offer. To take advantage of the Ville architecture and at the same time adhere to the curriculum of the online SWELL course that was developed in parallel, a set of 750 picture words, divided into 27 semantic categories such as 'food', 'colors' or 'occupations' were chosen as the core vocabulary of Ville. These are coded as *WordObjects* containing pictures, recordings and tagged with metadata as described in section 4.1.

7.1 Flashcards

'Flashcards' is 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 learners have 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 compare their own recordings with Ville's.

Figure 23 Picture of the flashcards and recording section of the program

Feedback on their pronunciation, based on automatic pronunciation error detectors was at the moment decided against, because it was too early in the design process, and would jeopardize the robustness of the program. Instead, priority was placed on creating a data collection tool, in the spirit of the *human computation* tools of von Ahn (2006), described on page 43. Students would get free education (possibly with some entertainment value), and we would get data for future research on Swedish L2 pronunciation from a wide variety of L1's.

7.2 Perception and Writing Exercises

In addition to the Flashcards exercises, listen-and-click exercises and listen-and-write exercises were developed using the same *WordObjects*. A grid of pictures that the learner can set to 2x2, 3x3 or 4x4 (see Figure 24) is used as the basis for listen-and-click exercises. In addition to choosing topic, and clicking on the pictures to hear Ville pronounce the words, there are three different exercises that the learner can choose from.

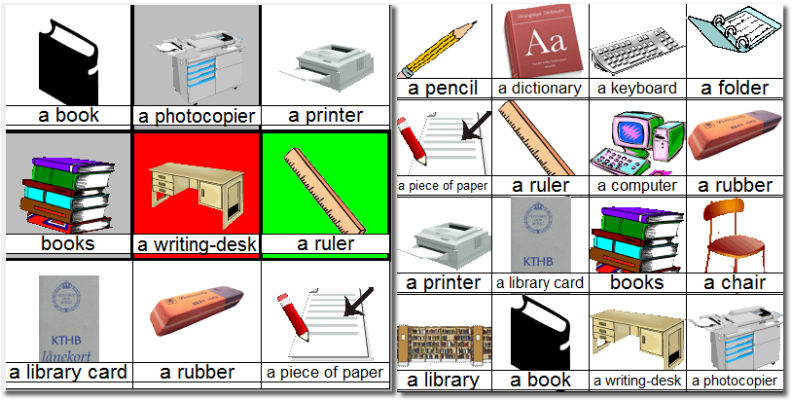

Figure 24 Vocabulary perception grid size 3x3(left) and 4x4 (right). Learners can choose between three sizes 2x2, 3x3, and 4x4

In 'Practice on full list' Ville will randomly say one of the words displayed on the screen in Swedish and the learner's task is to click on the corresponding picture. The learner will get both verbal feedback from Ville and visual feedback from the cards turning green if a correct selection was made, and red if an incorrect selection was made. There is also a (user adjustable) time-out value that adds a level of difficulty to the exercise. If the user has not made any selection within the timeframe, the picture turns grey and Ville continues with the next word. At the end of the exercise the learner can choose to review the incorrect and missed words.

Another way to practice is in 'game-mode', with basically the same functionality as described above, but where the user gets points for correct selections, (and more points if he or she has chosen a higher level of difficulty), and where three incorrect answers renders 'game over'. A personal high-score list is kept to encourage the learner to play more.

A third option for the learner is to 'Do the test', where no feedback is given during the selection process, and no time-out constrains are placed on the user. At the end the final score is given, and the learner can review the words that were incorrect.

Listen and write

In the listen-and-write exercise the learner's task is to correctly spell a word that is being pronounced by Ville. No picture of the object is displayed and it is thus a phoneme to grapheme conversion exercise. If the learner spells the word incorrectly, his or her spelling is highlighted in red and the correct spelling is displayed in green below. The learner has the option to zoom in on Ville's mouth during the exercise (as in Figure 25), and that way get a better picture of what was actually said.

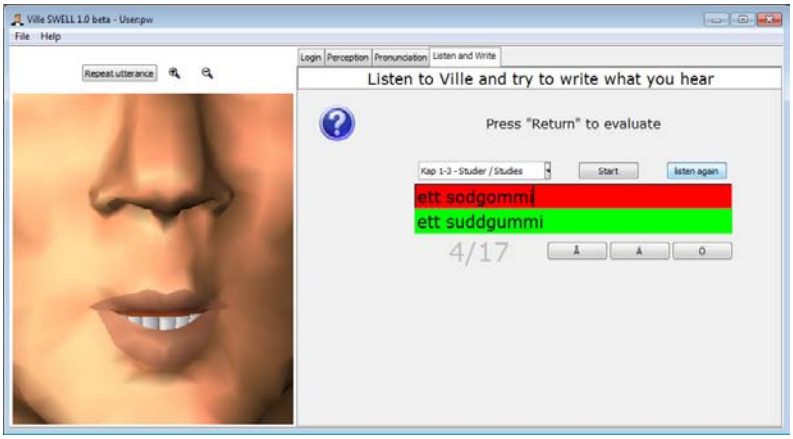

Figure 25 Listen-and-write exercise with Ville's mouth zoomed

It has been reported in the student feedback, that this was a much appreciated option and that the ECA's mouth and lips would reveal valuable information about sounds that the learners were not able to hear (the difference between /i/ and /y/ for example)

7.3 Data collection

A data collection tool has been implemented in the version of Ville being offered to KTH students, which has been successfully sending data to a server at KTH for several years. 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.

Every time a user makes a recording using the flashcards described in Section 7.1, the recording is saved locally on his or her computer, but it is also sent to a server at KTH. 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 the L2 learners of Swedish will eventually contain sufficient amounts of data that it can be used for future pronunciation-error detectors.

7.3.1 Questionnaire

In January 2009 a request was sent out to 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 linking student opinions to their performance was not possible. It focused instead on general questions giving indications on how students perceived the system. Each type of exercise as well as some features such as the talking head was queried. 59 students filled out the questionnaire. A summary of the replies is shown in Table 4.

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

Table 4 gives a summary of the questionnaire sorted by mean score. 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".

There was also a free comment field where students could include their comments: Apart from some informative 'bug-reports' on misspellings and so on, Ville also got some very positive 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.

The number of Ville users and the amount of data is continuously increasing, but still limited (the database is described in detail in 14.1.1). In order to get more data, we need both to increase the number of users, and look at new ways to motivate the existing users to use the program more frequently.

Currently there is only data on word level exercises. It will be possible to extract useful data for future segmental and syllable level detectors from the database, but there is nothing on sentence level in this database so far.

–8–

Ville on sentence level

This chapter presents Simicry, a paradigm that facilitates exposure to large amounts of meaningful input. A different type of feedback, called similarity measures, is presented, which is well suited for practicing prosody. Two different ways to interact with the VLT within the Simicry paradigm are explored.

The diverse set of features that are added on top of the features mentioned in earlier chapters are collectively grouped into what is called prosody. Prosody is a much less understood aspect of language than the other phonological aspects. The higher up in the hierarchy one comes, the more complex and intricate the analyses becomes, as elements and variation from the lower levels cascade upwards and combine in the same acoustic signal.

Prosody may also reflect features such as the emotional state of the speaker, the presence or absence of irony, emphasis, contrast, and focus, in addition to information about whether the utterance is for example a statement, question, or command.

As a way to practice prosody Gabor Harrar introduced in the late 90's something he called "repetitive synchronous imitation" (RSI). The basic notion is to work with target language material that is divided up into phrases of approximately 2 to 5 sec each. A single phrase is put into a loop, and played through headphones. The learner repeats in synchrony with the loop, hearing a 50-50 mix of the target and their own voice through headphones. There was however no evaluation of the student utterances involved in the RSI exercises and no functionality for how or when to move to the next sentence.

Other researchers and SLA teachers have written about techniques for improving student pronunciation by using 'choir practice' as a way to drill learners on the prosodic patterns of the target language (c.f. Kjellin, 2002).

Language consists, to a much larger extent than what was previously thought, of formulaic language, i.e. lexical phrases, idioms, conventionalized expressions, collocations etc. (Wray, 2008). Communicative competence is today considered more a matter of knowing a stock of partially pre-assembled patterns, than a matter of knowing rules. Formulaic frameworks play a much larger role in SLA today than has been the case before (Dörnyei, 2009). These units of language need to be acquired in whole chunks, and given the nature of how sheer statistical frequency affect our acquisition, a mechanism that allows for easy input and output of large amounts of speech may be a vital tool in the acquisition process.

Simicry brings together the acquisition of formulaic language with choir practice, Gabor Harrar's RSI, and the notion of gaming to increase motivation.

8.1 Simicry

As discussed in section 2.1, it is assumed that extended practice, under particular conditions and circumstances, will increase fluency by developing automaticity. Simicry, a term we coined from "Similarity of mimicry", is a framework for practicing L2 pronunciation that builds on a hypothesis that a mimicry-repetition-feedback-loop will facilitate exposure to large amounts of comprehensible, meaningful input, which according to the literature will promote an implicit process of acquisition, and the development of automaticity.

8.1.1 Gameplay

In order to increase the motivation and desire to spend time on the learning task, Simicry has taken inspiration from a subgenre of action games called Rhythm games (games that challenge a player's sense of rhythm). Games in this genre primarily focus either on dancing or simulating the playing of musical instruments, but we hypothesize that a similar form of enjoyment could arise from playing a game focusing on speech prosody instead of melody, or the rhythmical aspects of speech in a foreign language instead of musical beats.

In traditional rhythm games players typically press buttons at a precise time corresponding to a sequence dictated by the game. The buttons come in many variations, from conventional control pads to simulated musical instruments as in Harmonix's Guitar Hero and Rock Band, or dance mats as in Konami's "Dance Dance revolution".

Using a microphone as input device, and the imitation of a sequence of sentences as the challenge of the game, Simicry is reminiscent of Sing Star, a karaoke-like game where players sing along with music in order to score points.

The whole genre is moving towards ever more realistic games, where for example early versions of Rock Band had toy instruments with buttons instead of strings on the guitar, whereas later 'pro' versions are designed to double as actual musical tutorials with authentic guitar fingering and realistic strumming. This is in other words a genre that is seamlessly transforming from toy-game to learning environment, blurring the borders between traditional games and serious games.

Limiting the degrees of freedom

It is in all languages acceptable to vary the pronunciation of an utterance in many different ways, based on differences in dialect, personality, semantic intent, emotional content etc. The pronunciation may vary for example by placing focus on different words, changing the intonation (pitch), slowing down or speeding up parts of the utterance etc. Even though all languages show variation of this kind, every language will have its own regions of variation that are acceptable, and regions of variation that are unacceptable. Such language-specific restrictions are something an L2 learner will initially be unaware of, and due to the fact that they cannot easily be described or understood in explicit terms, may be best taught as implicit knowledge.

We see a pedagogical point in limiting the student's degrees of freedom in variation in this learning setting, and suggest that learning to reproduce, i.e. to mimic a native speaker, is an appropriate first step to acquire native-like pronunciation. Although being able to mimic well is not something a native speaker necessarily is good at, it can nevertheless be something that a language student can benefit from in order to get a good pronunciation in a new language.

We hypothesize that trying to produce an utterance that as closely as possible corresponds to that of a model utterance, and getting feedback on similarity of mimicry, rather than giving students all the degrees of freedom that a native speakers have in their production, will implicitly force the student to adjust their pronunciation, and learn to pay attention to aspects of the language that would otherwise go by unnoticed.

In Simicry one might for example have one analysis method for measuring similarity in rhythm, and another for similarity in melody (pitch). In an implementation of this where a student is consistently weak in one aspect, say rhythm, one could turn off all other aspects of similarity measures, and only focus on the rhythm. In such a scenario the student could say whatever he or she wanted, using any phonemes and any pitch contour without affecting the score, as long as the rhythm was similar. If the nature of the analysis was such that one region of an utterance could be determined as problematic, one could zoom in on that region and loop over that part until the student had mastered it, then zoom back out again and loop over a larger region or the whole utterance.

8.1.2 Simicry in Ville

The Simicry module in Ville, shown in Figure 26, has the following features: The interaction is such that once the play button is pushed a mimicry-repetition-feedback loop is entered, and the students only interact through speech. Students may select two different ways to do the actual mimicry interaction. Say-after, where Ville says the sentence first, and the students repeat it, and Shadowing, where Ville and the student speak at the same time. In Shadowing, a countdown tick is preceding every utterance to help students start at the same time as Ville. Students may select level of difficulty: If the student's performance is above a certain threshold, Ville moves on to the next sentence; otherwise, the same sentence is presented again, until the student has repeated it successfully. An acceptance-rate slider enables the students to choose this threshold themselves. This adds an aspect of gameplay to the exercise, where students should find a level on the acceptance-rate slider that makes it hard, but not too hard for them to pass, and then try to increase the level of difficulty over time.

Figure 26 The Simicry module in Ville

Packages

Students (or a teacher) may select what packages to work with. A package may be a collection of sentences with a common semantic topic, such as: At the bar, or at the airport, or it may be a collection of sentences with a common prosodic structure, or some other unifying attributes. Packages use the same underlying XML-structure that other word-objects in the Ville framework use, with top nodes being *SentenceObjects*.

Similarity measures - Salient features

Exactly which similarity measures that are relevant or salient for Simicry is yet to be determined. Even though the pitch, and duration of an utterance in absolute terms do not carry linguistic meaning (but changes in both pitch and duration might), there could be pedagogical benefits in telling the students to focus on getting the length of the whole utterance similar to the target utterance in order to make them focus and concentrate on other aspects of speech than they usually do. The framework is intended to be open for different kinds of similarity measures and allow for different researchers to experiment on the effect that feedback of a new similarity measure could have on student performance. Even though we see the feedback as an important part, it is the massive exposure to meaningful input and the implicit learning generated by the looping construction that is the main focus of the paradigm.

Measuring similarity

The version of Simicry in Ville implemented so far puts focus on prosody, to investigate if the students will notice or change the prosodic realizations of their utterances. The score presented to the student as that which is compared with the acceptance-rate slider (which determines whether to loop or move to the next sentence), is calculated as an average of four measures:

- Psyllabicity, a measure of pseudo syllabic units, resembling that of the convex hull algorithm (Edlund & Heldner, 2006). This is a measure of how many syllables there are in each utterance and the score is calculated by comparing the number of syllables in the native (target) utterance and the student utterance.
- Length (the total length of the utterance).
- Timing (The mismatch in duration of each syllable).
- Melody (a normalized F0 correlation score).

All scores result in a measure between 0-100 and the final score is the average of these four scores. This final score is then compared with the acceptance-rate threshold, which determines if the student should move on to the next sentence, or loop (mimic the same utterance again).

8.1.3 Discussion

Although the Simicry implementation presented in this chapter only gives feedback on aspects of prosody, there is nothing prosody-specific about the Simicry paradigm, and we would like to explore it further with other aspects of similarity measures in mind. As discussed above, which measurements of similarity that are the most relevant for a student to get feedback on needs to be investigated further.

The measures used in this experiment need to be optimized, and several proposals from the literature have been suggested, for example prominence detection (Al Moubayed & Beskow, 2010) or measures of prominence as proposed in Tepperman et al. (2010).

It should be noted however that good methods for classification of a linguistic feature do not necessarily entail good pedagogical feedback. An interesting question to investigate is: Should similarity measures be language-specific, or

even phrase, sentence, or word-specific, or are there measures that are able to capture language specific variations using some universal features? However, we believe that the Simicry-feedback loop paradigm might be pedagogically valuable even with pseudo arbitrary acoustic similarity measures, leaving the learning and abstraction implicit for the language learner herself.

Another, related, issue in order to achieve maximum efficiency and retention from working with Simicry is how often, and when to introduce new material to the student. Whether it is vocabulary, or variants of pronunciation that is being trained, the frequency of repetition is not arbitrary. Practice should be scheduled according to some optimal spaced repetition algorithm in order to maximize learning and retention (for example Pavlik & Anderson, 2008).

–9–

Ville on discourse level

The aim of this chapter is to pursue the CLT methodology which promotes language learners to practice conversational skills and learn to communicate through interaction in the target language rather than learning structures, sounds or words.

9.1 Spoken dialogue systems for CALL

An introduction is first given on how spoken dialogue systems can be put to use in CALL applications in order to cater to other teaching methodologies putting focus on exercises on a higher level in the linguistic hierarchy.

Research in second language acquisition has since the origin of the methodology known as communicative language teaching (CLT) in the 1970s, been divided both in theory and in research. It is a debate regarding the benefits of focus on function versus focus on form. Focus on form requires the learner to focus on phonological errors, vocabulary or the grammatical correctness or incorrectness of the L2, whereas focus on function requires the learner to focus on the meaning or message being conveyed by the L2.

Pursuing the CLT methodology is however something much more difficult for a CALL system than focusing on form since it ultimately requires the computer to have the communicative competence of a human being in order to be a full-

fledged conversational partner, something we are nowhere near achieving. Natural language understanding is by many considered an AI-complete problem requiring artificial intelligence that matches human intelligence to be done.

The use of spoken dialogue systems for CALL would be a way to approach CLT. Some work in this direction has been attempted, and a brief introduction to some of them follows.

9.1.1 Alterations of existing dialogue systems

Considering the fact that building a dialogue system from scratch is a huge effort, one strategy adopted by several researchers has been to modify existing dialogue systems to make them useful for CALL purposes.

The Spoken Language Systems group (SLS) at MIT has used this strategy extensively. SLS has created an architecture for conversational speech systems where speech recognition, speech synthesis, language understanding, language generation, and discourse and dialogue modeling modules can communicate with one another through a programmable hub called Galaxy (Seneff et al., 1998). This system is used in several well known projects such as the Mercury airline flight planning system, the Penates restaurant information system, the Voyager city guide system, and the Jupiter weather information system. The same core technology has been used for second language learning (Seneff et al., 2004).

The Jupiter weather information system has been adapted to a CALL system for English-speaking students learning Mandarin. A student can practice by attempting to communicate with the dialogue system over the telephone. If they don't understand a response, they can ask for a translation into English. For the language-learning application, the system is configured so that if an utterance is spoken in English by the student, it is translated into Mandarin, whereas if an utterance is spoken in Mandarin it is answered by the system in Mandarin, and the dialogue continues. This means that if the student cannot remember how to say a certain phrase in Mandarin he can get a translation, which he can then attempt to repeat to the system in order to complete the task (Seneff et al., 2004). Several other CALL systems have been created at SLS using the Galaxy architecture, for example LanguageLand (Lee, 2004b) ISLAND (McGraw & Seneff, 2007), and a restaurant information system (Gruenstein et al., 2006).

The spoken dialogue system "Let's Go" from Carnegie Mellon University (CMU) is working in the domain of bus information, with the objective to create a basic dialogue system that groups such as the elderly and non-native speakers can access. A discussion on how to use this system also as an environment for language learning is presented in Raux & Eskenazi (2004).

9.1.2 Contrast to an information-seeking dialogue system

There are however several differences between dialogue systems used for information retrieval and for CALL, and the design of a CALL dialogue system should differ from that of an information retrieval system.

Metaphors used in dialogue systems

Some dialogue systems may be perceived metaphorically as an interface, whereas other dialogue systems are best viewed using a human metaphor. Edlund et al. (2006) discuss the distinction between the interface metaphor and the human metaphor. Some dialogue systems can be perceived as an alternative interface to a web form or something similar. In such a system, users who understand how to talk to a system will use something they call "menu-speak" Saying 'London' rather than 'I would like to go to London please'. Conversely, users with little experience with dialogue systems, sometimes display a behavior that suggests that they view the system as a person, and that they will have expectations on the system's conversational abilities based on what they would expect from a human. They talk freely and at length, and commercial systems today are not able to handle such users very well.

It is not only a matter of personal choice for the user to decide what metaphor to choose in his or her interaction with the dialogue system. Some types of applications seem better predisposed to use the human metaphor than others. The interface metaphor is good for some purposes and the human metaphor for others. If all you want to do is find out what time a bus is departing, the interface metaphor may actually be preferable, and the designers of such systems should keep that in mind.

A dialogue system for CALL is designed to improve human-human communication, and an interface based on such criteria could therefore successfully be designed around the human metaphor.

Reasons for using the system affect the design

There is a difference between dialogue systems designed for information retrieval, and dialogue systems for language learning. Dialogue systems used for information-seeking applications, such as train-schedule or weather-information systems need to be polite, unobtrusive and efficient in terms of task completion.

Dialogue system for CALL however, needs to be none of these things. Quite the contrary actually, the longer the conversation takes, and the more turns between the user and the system, the better. Rather than being polite, such a dialogue system should try to create an interesting conversation where politeness is not necessarily a good thing. Perhaps something in line with the 'drama manager' developed in Façade (Mateas & Stern, 2003) would increase the chances of an interesting conversation in a CALL application, whereas a drama manager to intentionally increase the amount of drama would be out of the question in an information-seeking dialogue system.

The reason for talking to a dialogue system for CALL differs from an information-seeking application. Both the design criteria and the evaluation criteria will be dramatically different as a consequence of this. 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 designed as a serious game, the interaction would be so entertaining that the user would consider it to be a combination of the second and the third example: A game where the reason for playing would be just because it is fun (and at the same time getting useful exercise).

User history in dialogue systems

Another distinction that comes into play is the level of intimacy/privacy that different types of dialogue systems require. In an information seeking application, for example a bus-schedule service, customers are completely anonymous. Every new caller is treated as a new user, regardless of the fact that some callers may perhaps use the system every day. In such an application it may be convenient both for the system designers and for the users to keep it that way. That is perhaps one of the landmarks of a system that will benefit from the interface metaphor.

A CALL application on the other hand, which is designed to be used more than once, would benefit from taking the history of the users into account, in order to tailor the lessons based on this history.

A game-based CALL application also requires a history in order to create adaptive gameplay, i.e. to increase the level of difficulty to match the skill level of the player/learner. This is done not only to increase the level of enjoyment, but in order to find the best mixture of challenge and support to promote learning in a person. In constructivist terminology, the aim is to find the zone of proximal development, and in gaming terminology to try to achieve a state of 'flow'.

User perception of system performance

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 a dialogue system in CALL, 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 quite different from efficiency in task completion. Apart from creating a good story, the social competence and personality of the character may be considered impor-

tant 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 thinks 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 or she is able to communicate with the system without communication breakdown, it can be seen as a confirmation of his or her ability to communicate in the new language.

If this line of reasoning is turned around, would be possible to use it as a design criterion: Is it possible to take advantage of the limitations/imperfections of current ASR? Could the difficulties of ASR in combination with foreign accent 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 could become part of a challenge of passing a gate-keeper. This way of motivating learners 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 learner's communicative skills, where the challenge is to be able to negotiate with an ECA. If the learner 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 learner with the exercises and feedback she needs in order to try again, and hopefully pass the gate-keeper at a later time.

This is what has been the basis for a game-based learning application called DEAL, which is introduced in the next section.

9.2 DEAL

The goal of DEAL has been to create an entertaining meta-level for the Ville framework exploring how spoken dialogue systems can utilized in CALL to simulate a real conversation for language learning purposes.

The objective is thus similar to that of the Tactical Language and Culture Training System (TLCTS) (Johnson & Valente, 2009), in the sense that both systems are simulation games for the acquisition of language and cultural skills. However, where TLCTS places focus on realism (teaching US military appropriate manners and phrases to be used on foreign ground), the aim of the DEAL system is to focus more on entertainment.

In that respect the 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).

The objective is also similar to that of the Nice project (Gustafson et al., 2004), in that the goal is to create a game in which spoken dialogue is not just an add-on, but is the primary means for game progression. The ultimate goal of DEAL is however on language learning.

While the Ville character provides exercises on isolated speech utterances, i.e. phone, syllable, word, and sentence level, DEAL is intended to add the possibility of practicing these skills 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, with whom you need to communicate using your new language.

9.2.1 Domain

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, and the 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).

9.2.2 Scenario

The basic teaching plan is for the language learner to use Ville in conjunction with DEAL. First Ville will teach the rudimentary vocabulary that is associated with for example 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 shopkeeper 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 28)
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

9.2.3 DEAL architecture

The dialogue system in DEAL is based on Higgins, a spoken dialogue system developed at KTH (Skantze, 2005a) with modules for semantic interpretation and analysis. A schematic picture of the architecture is shown in Figure 27.

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 to address this problem. Grammar rules are automatically relaxed to handle unexpected, ungrammatical and misrecognized input robustly.

The discourse modeler Galatea, (Skantze, 2005b) 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*** described in 9.2.5, a ***communicative manager***, modules for ***text generation*** and ***text-to-speech*** generation, all described in 9.2.6, and a ***user interface*** described in 9.2.4

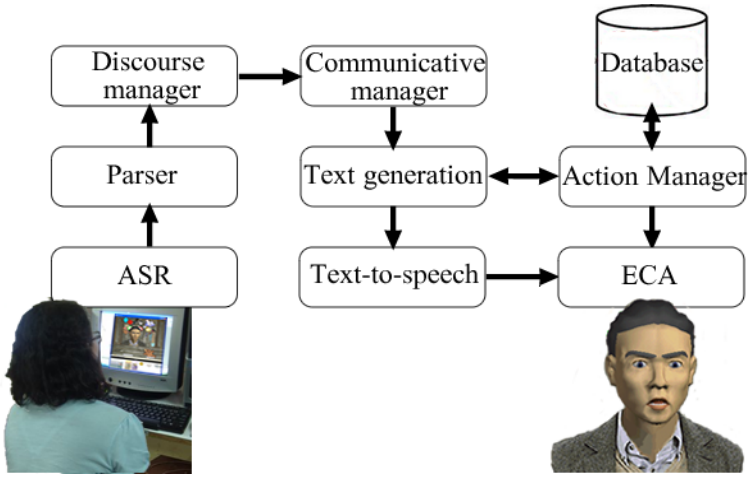

Figure 27 Schematic picture of The DEAL architecture

9.2.4 The DEAL user interface

The user interface in DEAL, as shown in Figure 28, is divided into six parts. The top part contains the shopkeeper, an ECA with the same expressive abilities as Ville, 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.

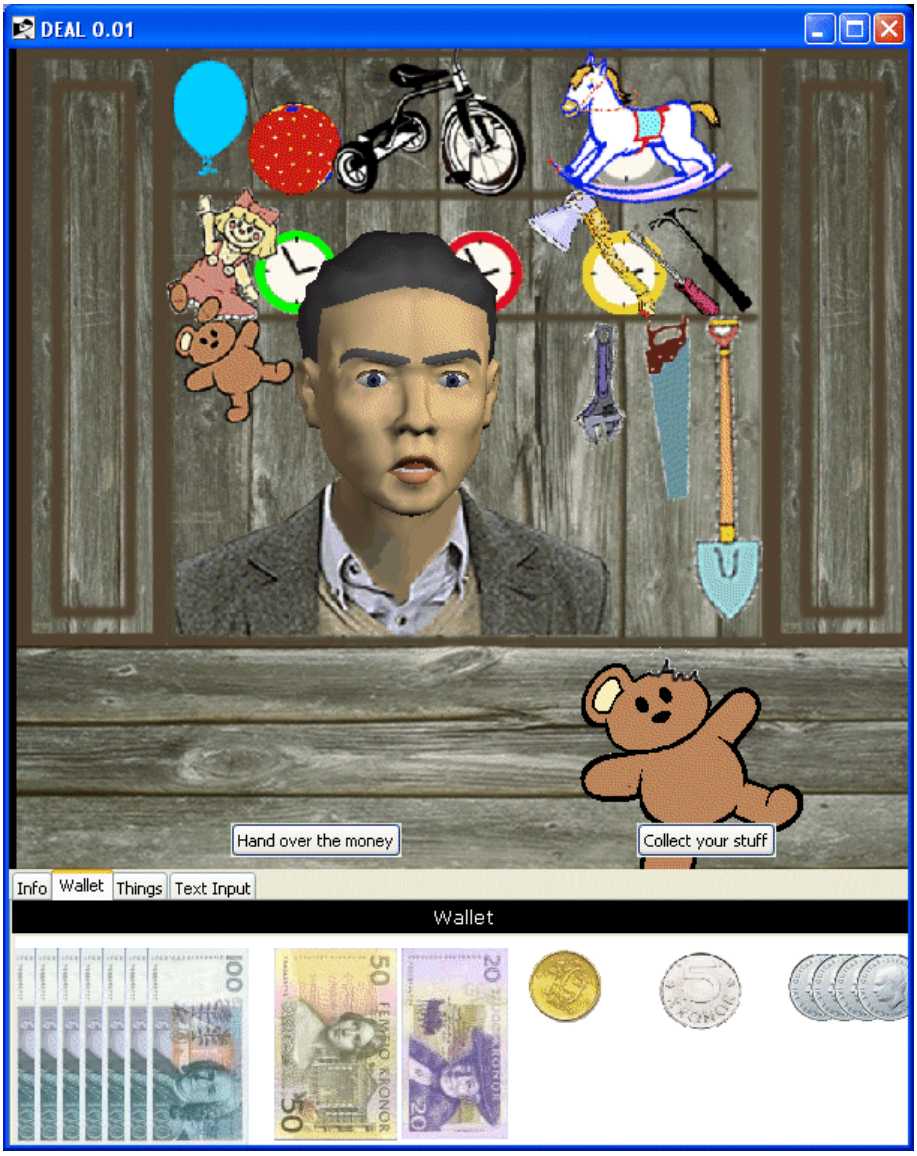

Figure 28 The DEAL interface with a stingy shopkeeper trying to sell a teddy-bear with a missing ear.

9.2.5 Action management in DEAL

Action management in DEAL, i.e. deciding what to say next, is currently done according to a set of rules based on 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 based on the amount of the offer the user is suggesting in relation to the "retail price" which is a price stored in the system database. All objects in the database have obvious visual defects (e.g. the missing ear as in Figure 28) and if detected and pointed out by the language learner, the agent reduces the price of the object, 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 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 becomes bored and refuses to discuss that object anymore. Instead, he suggests a new one.

9.2.6 Human-like language generation

DEAL is still under development and has yet to be evaluated. Since the agent's behavior is crucial for how the system is perceived, much effort so far has been placed on trying to build a system that can generate utterances in a human-like manner.

In a CALL system based on the CLT methodology, where focus is placed on conversation and negotiation of meaning, it is crucial that the system acts as human-like as possible. 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). As humans, 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 low, the dialogue system needs to be capable of grabbing the turn, holding it while the system is producing the rest of the message, and releasing 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 labeled as belonging to cue phrases. Frequently used strategies, to buy time while processing input, were to generate filled pauses (e.g. eh, ehm), repetitions, or prosodic phrase-final lengthening.

Incremental language production also requires that the system knows 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 communicative manager

The generation task in DEAL is distributed over different modules. The communicative manager 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 communicative manager 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 communicative manager generates a clarification CA without passing the message on to the action manager, 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 S1b).

```
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 communicative manager generates neutral feedback such as "yes" or "ok". Since the discourse manager not only keeps track of the user's CAs but also its own previous CAs, the communicative manager can modify new responses from the action manager 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 S3b).

The action manager

The action manager 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 action manager has generated a response it is passed on to the communicative manager, which is responsible for modifying the response based on the previous dialogue context. For example, the communicative manager decides how entities should be referred to, e.g. determining 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 communicative manager forwards its message to Ovidius, the module responsible for realizing the textual representation. Ovidius takes a system CA as input and generates a text that is subsequently realized acoustically by a text-to-speech generator.

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 realization 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-synthesized 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 semantic tag (e.g. neutral-feedback, filled pauses and so on).

Whether the functions of these elements are interpreted by users as intended is, however, still to be evaluated.

–10–

User study 1

This chapter presents the first of two user studies where some of the exercises, PEDs and feedback strategies developed within the Ville framework were evaluated by 'real' language learners. This study is an attempt to investigate whether students are able to understand the instructions given by the virtual language teacher, and whether they find the exercises useful or not. To try to shed some light on how students perceived the system, closed and open response questionnaires were given and the students' interactions with the system was analyzed.

In the spring of 2009 a group of twenty-seven second-term Swedish learners at KTH (13 men and 14 women) were recruited to participate in a user study to investigate how difficult and useful learners would find a selection of new capabilities in Ville. Qualitative feedback was collected and performance was monitored while the learners familiarized themselves with the program for approximately one hour. Everyone completed a pre-experiment questionnaire with demographic and language experience questions before the exercises started.

Twenty-three students were between the ages of 20 and 30. They had been in Sweden for an average of just over a year, and all reported using computers every day. The largest single native language represented was French (6), followed by German (4) and Chinese (3). There were two speakers each of Italian, Turkish and Russian, and individual speakers of eight other languages. All spoke English as a second language, most of them (reported) fluently, and eight spoke a third language as well – excluding Swedish. All were taking their second

Swedish course at KTH, classified as 'Advanced Beginners.' They reported using little Swedish outside of the classroom: only about 50% reported using Swedish with friends or watching Swedish TV, while about 20% heard lectures in Swedish or listened to Swedish music, and 10% spoke Swedish at home or went to Swedish films. Most of them found speaking Swedish to be the hardest linguistic skill to master, closely followed by understanding spoken Swedish, and then writing and reading in that order. Likewise, they found learning pronunciation and word recognition to be more difficult than learning vocabulary and grammar.

10.1.1 Experimental design

All the instructions for each exercise were given by Ville. Pre-recorded utterances, switching of panes in the program, and highlighting of buttons or areas of the screen were collected into scenarios, and a collection of these, tailor-made for the experiment, was prepared. In addition to exercising Ville's ability to display scenarios to the students, this approach had the advantage of ensuring that all students were given the same, unbiased information before and after each exercise.

10.1.2 Exercises

Perception exercises

Perception exercises were presented for Lexical stress, Quantity, and Vowel Quality. Minimal pair exercises were presented for the quantity and vowel quality exercises and a grid with stress patterns for the lexical stress exercise. The learner's task is to select which word was said by clicking on the corresponding word or symbol on the screen. All exercises are explained in section 5.1 and 6.1.

Production exercises

Carefully selected words that contained specific pronunciation difficulties were grouped together and presented as three production exercises where individual words were targeted. The categories were quantity, lexical stress, and insertion/deletion. The learners recorded words using the flashcards paradigm described in section 6.2. Finally, two production exercises were provided where

learners produced whole sentences using the Simicry paradigm described in detail in section 8.1

10.1.3 Questionnaires: closed questions

Students completed an identical questionnaire after each of the eight exercises of the experiment. Each questionnaire asked five closed and two open questions. The closed questions (shown in Table 5), elicited a response on a five-point Likert Scale, where (perhaps unintuitively) 1 was the most positive response and 5 the most negative. Written descriptors (i.e. 'very good' — 'very bad') accompanied the numbers. The five closed questions were:

Q1	How easy was it to understand how to use Ville for this exercise?
Q2	How useful do you think this exercise was?
Q3	How good were the examples in the exercises?
Q4	How useful was the presence of the animated agent (Ville) in this exercise?
Q5	How likely would you be to do more exercises like this if they were available?

Table 5 Questions given to the users after each exercise

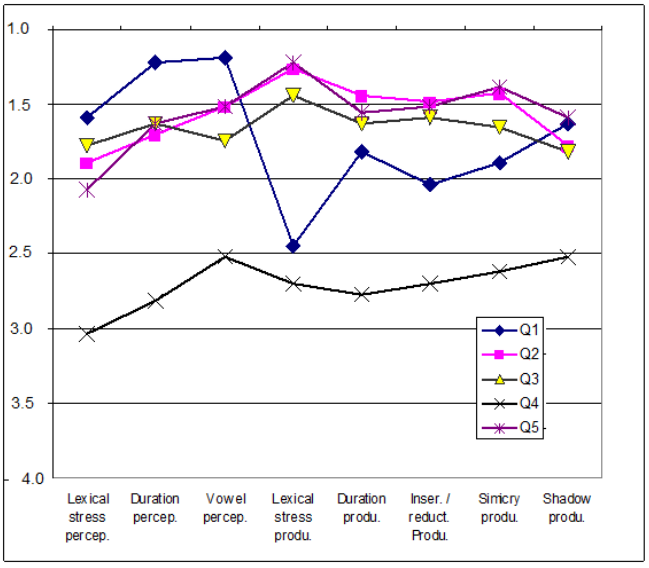

Figure 29 Mean responses to the five closed questions in each of the eight exercises, where 1= most positive and 5= most negative

Responses overall were very positive: the means to all eight Q2s and Q5s regarding usefulness and desire to use again were both 1.56. Figure 29 plots responses to the five closed questions. The first three exercise sections of the experiment tested perception only, while from the fourth section on students were also producing speech. Responses to Questions 2, 3, and 5 are similar to each other across the eight sections, where subjects were most satisfied with the section testing lexical stress production, although they at the same time found it difficult to understand how to do the exercise (Q1). This is possibly because this was the first exercise to test production, and therefore students were simultaneously excited about vocalizing instead of only listening, and unsure of how to follow prompts and interpret feedback. In general, they found it harder to understand how to perform the production than the perception exercises. Subjects were less convinced of the usefulness of the presence of the animated agent. Responses to Q4 deviated strongly from the other very positive answers, though they are better than the neutral point of 3.

The subjects completed a final questionnaire when they had finished working with Ville. Here they answered questions about the usefulness of the system overall. Mean responses to questions regarding usefulness and desire to use again were now even slightly improved: both a mean of 1.37. Responses to the question regarding the usefulness of the agent were also least positive: 2.77. Subjects were asked to rank the usefulness of the eight different exercise sections, a task which may have been difficult since they had been introduced to a large amount of new material at once. The resulting ranking favored SayAfter production (1), followed by Shadowing production (2), Duration production (3), Lexical stress production (4), Vowel perception (5), Duration perception (6), and Lexical stress perception (7). Subjects thus preferred the production to the perception exercises.

10.1.4 Animated agent

The presence of the animated agent is a unique feature for a CAPT system, and therefore subject response to the agent is of interest. As can be seen in Figure 29, subjects gave the agent poorer ratings than the other parts of the system. However, ratings of the agent improved over the course of the experiment. A Pearson correlation of 0.68 was found between answers to Q4 and time. It may be that, like a person or a human teacher, the agent's presence grew on the students and they came to appreciate him more as they became more familiar

with him. One student explained her score of 3 (moderately useful) in the final questionnaire with the added comment that "but I liked that he was there." Another student wrote "I never looked on the face before but in this section, I recognized that I didn't make so many mistakes when I had watched how the animation pronounced the words." In fact, female subjects rated the animated agent significantly more positively than male subjects over all ($t(7) = 6.89$, $p<0.001$, two tailed).

10.1.5 Performance scores

Performance scores were computed for each of the exercise sections individually by calculating percentage of tasks performed correctly vs. all attempted tasks, and computed overall means for the whole study, and means for the perception and production sections from these. The overall mean for the study was 0.65 accuracy, with a standard deviation of 0.07. There was a significant difference between the perception and production scores, with the former significantly higher than the latter (0.77 vs. 0.59; tstat=7.17, p=0.001). Overall score showed main effects for several demographic factors, based on one-way ANOVAs. Younger subjects (18-30) did significantly better than older ones ($F(1,25)=9.57$, $p<0.005$). There was also an effect for native language type ($F(1,25)=7.91$, $p<0.01$), with students whose native language was Germanic or Romance performing better than students from other language backgrounds. Curiously, subjects who reported speaking Swedish at home did more poorly than others ($F(1,25)=8.04$, $p=0.009$). The number of non-native languages (excluding Swedish) that subjects reported knowing showed a tendency to influence overall score but was not significant ($F(1,25)=3.29$, $p=0.081$).

Mean scores for each exercise set allow us to rank exercises from least to most difficult: Insertion (0.92), Minimal Pair Vowels (0.81), Minimal Pair Duration (0.81), Lexical Stress Perception (0.67) Lexical Stress Production (0.62), Shadowing (0.61), Reduction (0.58), SayAfter (0.50) and Duration Production (0.30). Performance and post-exercise responses showed some weak correlations (using Pearson's correlation coefficient); recall that since '1' is the most positive rating in each case, a negative correlation is a positive ranking: Subjects who performed better rated Ville easier to understand how to use (Q1) for the Lexical Stress Perception exercises ($r=-0.56$) but were less likely to do more similar exercises on their own (Q5) ($r=0.33$). Subjects who performed better on the Minimal Pairs Duration perception study rated the presence of the agent

(Q4) as less important (r=0.31) although those performing well on the Duration Production exercises said that they would be more likely to do similar exercises on their own (r=-0.40). Finally, there was a correlation between rating the agent as less important and performance on the SayAfter exercises, with higher performance correlating with worse scores (r=0.45).

In general, there is a weak correlation (r=0.29) between overall performance and agent ratings for the post-questionnaire, i.e., subjects who performed better rated the agent as less useful. None of the other questions show a correlation with the performance scores.

10.1.6 Open response questions

The open responses to the questionnaires provided useful and varied suggestions regarding Ville's interface design, content, and feedback. The most common comment expressed some frustration at practicing words whose meanings subjects did not know, and requested that words and phrases be presented in writing and translated into English so students could learn new vocabulary. This would be easy to provide in the system, but it may be that, by leaving out the semantics, students can more easily place the cognitive focus on the intended phonetic aspects of these exercises. One of the most successful studies regarding the acquisition of L2 pronunciation deliberately refrained from letting learners know the meaning of the sentences they were learning to say (Neufeld, 1978).

The feedback on production of lexical stress and duration production was met with some skepticism by some students, who made comments such as "sometimes I feel me and Ville are pronouncing the word the same way, but it's red=wrong". This raises the question of whether the students were unable to themselves perceive the differences in the pronunciations, or whether the feedback was inaccurate in some way. Duration distinctions are not binary in reality, so that a student could have produced a long vowel that was almost long enough to receive a green light, but still have received a red light. A design option to adjust for this could be to add a third feedback alternative such as a yellow light for borderline cases. The analysis tool in the system, with its visual representation of vowel length, was there to help students diagnose their problems, but while some students found it useful, others complained that they did not understand the scoring system used in the analysis. Feedback for other sec-

tions was sometimes seen as too lenient; some students felt they had not done a good job but were still rewarded by green lights or high scores. Suggestions were made for summative feedback, showing how well one had done in a section, and adaptive exercises, where subjects were given more examples to practice items they had gotten wrong.

Students were asked for constructive criticism, and that is to a large extent what was received; however, many enthusiastic comments expressing appreciation of the interface and feedback, and gratitude for the opportunity to improve oral and aural skills were also received. The few comments regarding the agent suggested that he be friendlier and more encouraging, and perhaps female. Only a few subjects realized the usefulness of the visual information the agent provided regarding the articulation of Swedish. It is possible that students need to be explicitly guided to look at the agent; it is also possible that if they were to use the system for more than an hour, they would be able to do so naturally because they would be more familiar with the interface.

Students did not hesitate to personify the agent: it is consistently referred to as 'he', 'him', 'the character' or 'Mr. Ville' and ascribed abilities such as 'thinking', 'liking' 'approving' or 'disapproving.'

The complete set of replies to the open response questions is provided in appendix 16.1 on page 193.

The results from user study 1 have shown that Ville is able to capture the interest of the L2 learners. Learners generally at the end of the short-term study believed that Ville and its exercises were useful and met their expectations well.

However, a short term study directed towards the first impressions of students, such as Study 1, might be affected by factors such as novelty and does not reflect the long term ability of Ville to keep the students engaged and interested.

–11–
User study 2

This chapter presents a more elaborate user study, targeted towards evaluating Ville and its different characteristics as a long term teaching companion, where students were given the opportunity to download and utilize Ville at home whenever they want.

The study targeted testing the different characteristics of Ville by different groups of L2 learners for a period of four weeks. Different groups received different versions of the program to investigate the effectiveness of specific aspects of the system. Pre and post tests were carried out on all students to be able to measure the differences of specific language skills before and after the learning period, and between the different versions of exercises (e.g. learning lexical stress with or without feedback).

In addition to the language skills test, questionnaires were presented to the students who used Ville at home after they had completed the post test, in order to capture the qualitative experience and impressions of the software after four weeks.

11.1.1 Methodology

All the subjects participating in the study underwent a first test before using Ville, and the same test was repeated after four weeks of using Ville at home. The subjects as described in the next section were divided into three groups.

One group was a control group that underwent pre and post test without having Ville at home. This condition was made to capture the natural change in language skills for students living in Sweden and being exposed to the Swedish

language over a period of one month. Note that all students in the study were enrolled in a Swedish course over the period of the study.

Ville guided the users from beginning to end during each of the tests, playing the role of a language examiner, explaining each of the tests as the users progressed through the exercises, thus ensuring that all students would get identical instructions. As shown in Figure 31, users were able to 'rewind' the instructions and listen again to parts of the explanations, as a measure to ensure that all users were given a chance to understand the instructions.

11.1.2 Pre and Post tests

The language skill tests that were introduced in the pre and post tests were split into three categories: perception tests, production tests, and Simicry tests.

The perception test contained three different items.

- *Duration/Quantity:* As explained in section 6.1, users were presented with two words in a minimal pair, where both were two syllable words and the duration of the stressed vowel and the post vocalic consonant was the contrastive feature (e.g. mata – matta). One word was pronounced by Ville, and the user's task was to choose the word that matched the pronunciation. Twelve minimal pairs were presented.
- *Vowel quality:* In the vowels test, (explained in section 5.1.1), minimal pairs were presented to the user, where the contrastive feature was the type of vowel. Twelve minimal pairs were presented.
- *Lexical stress:* In the lexical stress test (described in section 6.1), users were presented with a 3x3 grid where each cell in the grid represents a different lexical stress pattern, as shown in Figure 30. The 'u-like' symbol represents unstressed syllables and the '–' symbol represents stressed syllables. A word was pronounced by Ville and the user's task was to choose the appropriate cell in the grid. Twelve words were presented.

Figure 30 The grid used in the lexical stress exercises

The production test: Here the students should make their own recordings of words after hearing Ville pronounce them. Users were provided with a push-to-talk button, a '*play Ville*' button (to hear Ville pronounce the word again), and a '*play student*' button (to hear their own last recording), as shown in Figure 31. The user was allowed to try a maximum of three times and listen to the recordings, before deciding to 'submit' a recording by clicking 'next'.

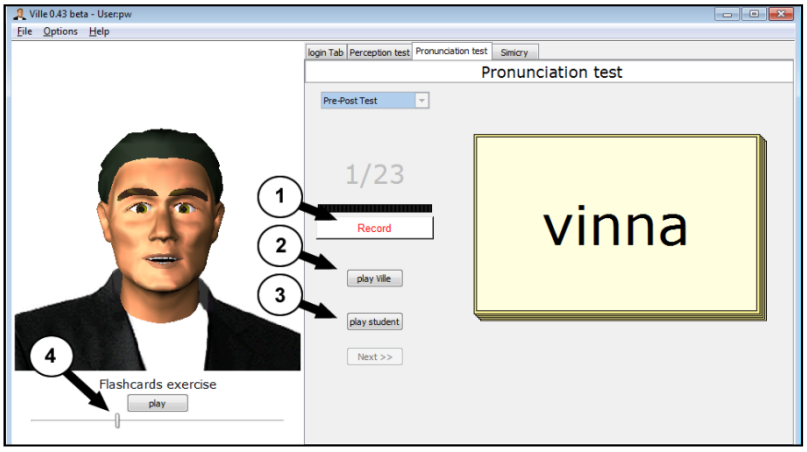

Figure 31 The simplified user interface for the production part of the pre and post test. 1 a push-to-talk button, 2 'play Ville' to hear the word again, 3 'play student' to hear their own last recording, 4 a slider to 'rewind' the instructions given by Ville.

The words were chosen to elicit four types of pronunciation errors: Quantity, lexical stress and insertion and deletion errors, as explained in sections 6.2.1 and 6.2.3.

All in all 23 words were recorded, 7 with quantity detectors, 8 with lexical stress detectors, 7 with insertion detectors, and 6 with reduction detectors.

The Simicry test: In the Simicry test a sentence was pronounced by Ville and the user's task was to imitate the same sentence. A set of 56 phonetically balanced sentences was presented in the test. The system recorded the pronunciations of the users and scored them against Ville's pronunciation, using the similarity scoring functions described in section 8.1.2.

As mentioned before, aside from these three groups of tests, the students were presented with questionnaires at the end of the post test.

11.1.3 Subjects

A group of 41 foreign students (16 female and 25 male) at KTH, studying Swedish as a second language at elementary and advanced elementary levels, participated in the study. The language background (L1) of the students is shown in Table 6

L1	Participants	L1	Participants
Persian	6	Korean	2
English	5	French	1
Chinese	5	German	1
Spanish	5	Italian	1
Russian	4	Lithuanian	1
Turkish	3	Urdu	1
Greek	2	Vietnamese	1
Polish	2	Arabic	1

Table 6 Language background (L1) of the 41 participants.

From Table 6 it is clear that there was a big variability in L1 of the participants. In one way this variability is positive for this study since it does not provide results that are biased towards one specific L1. However, this variability comes with the disadvantage that clear results about correlated effects between the exercises and the L1 cannot be drawn due to the fact that the number of students per L1 is too small to provide any significant statistical differences.

Users were split randomly into three equally sized groups. The choice of splitting randomly rather than with regard to some property (e.g. L1, gender, current language skills), is motivated by the fact that the number of students is too small to provide any balanced groups. The three groups and distribution of

gender and L1 can be seen in Table 21 and Table 22 on page 207 in the Appendix.

Out of the three groups, group 3 was the control group that performed only pre and post test, as mentioned earlier. The other two groups were provided with different versions of Ville where different approaches to the same exercises were given.

11.1.4 Exercises

The pre and post tests were made to reflect the training situation, except that different words and sentences were given at the test instances compared to the training sessions, to see if any generalizations would spill over from the work made at home. The following exercises were provided to the students in group 1 and 2 for learning Swedish at home.

Perception exercises: Four different perception exercises were provided. Three of these were identical to the duration, lexical stress, and vowel minimal-pair exercises used in the pre and post tests explained in section 11.1.2, except that the students received feedback from Ville on whether their answers were right or wrong. The fourth exercise is a vowel grid exercise of the type described in section 5.1.2. Both group 1 and 2 received the same perception exercises.

Production exercises: Three production exercises were provided to the students. These were identical to those provided in the pre and post test, i.e. duration, lexical stress, and insertion/deletion.

Group 1 received the three exercises with feedback from pronunciation error detectors (PED's) as described in section 6.2. Group 2 received the exercises without any feedback from Ville, so students could listen to their own production and compare it with Ville's pronunciation, leaving the students with self-monitoring as the only means of judging their pronunciation.

Simicry exercises: The Simicry exercises provided to the students consisted of semantic categories on 11 topics such as "at the bar", "at the train station", or "telephone conversations". Group 1 only worked in the *'shadowing'* mode (speaking at the same time as Ville) whereas group 2 worked only in *'say after'* mode (having Ville speak first and the learner repeating after), all explained in detail in section 8.1.2.

Both groups also had access to an acceptance slider in order to adjust the level of difficulty for moving on to the next sentence with respect to the similarity measure.

11.2 Analysis and results: perception exercises

11.2.1 Pre and post test

Since the subjects of group 1 and group 2 received the same perception exercises there is no reason to differentiate between them, and their results and answers have thus been merged in the graphs.

By comparing the improvements shown in Figure 32 between the pre and post tests one can see that all groups improved, but the general improvement is larger in the groups who received Ville compared to group 3 (the control group) on all three types of exercises. A one-sample t-test was applied on the improvements (to factor out the within-subject effect). The improvements are significant for all three categories for group 1+2, and significant for group 3 in lexical stress but not in the duration and vowels exercises as seen in Table 7.

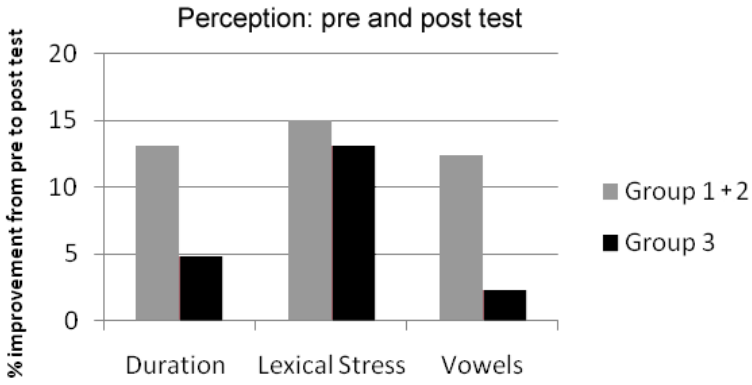

Figure 32 Improvement in % between the pre test and the post test on group 1 and group 2 compared to the control group (group 3).

	Group 1+2	Group 3
Duration	$p<0.02$	$p>0.5$
Lexical stress	$p<0.005$	$p<0.005$
Vowels	$p<0.002$	$p>0.5$

Table 7 improvement was significant for group 1+2, who had used Ville at home in all three exercises, and for group 3, the control group in lexical stress only.

11.2.2 Questionnaires Perception exercises

Students in group 1 and group 2 were after the post test given a set of open questions and closed questions on a five-point Likert Scale, to elicit answers on how they had experienced the work with Ville at home during a month. Three questions on each of the perception exercises regarding the usefulness, ease, and enjoyment of the perception exercises were given on a five-point Likert scale:

	Questions on the perception exercises
Q1	Rate each perception exercise based on usefulness: 1 for not useful at all - 5 for very useful
Q2	Rate each perception exercise based on ease: 1 for very difficult - 5 for very easy
Q3	Rate each perception exercise based on enjoyment: 1 for boring - 5 for fun

Table 8 Mean responses from questions regarding usefulness, ease, and enjoyment on the perception exercises

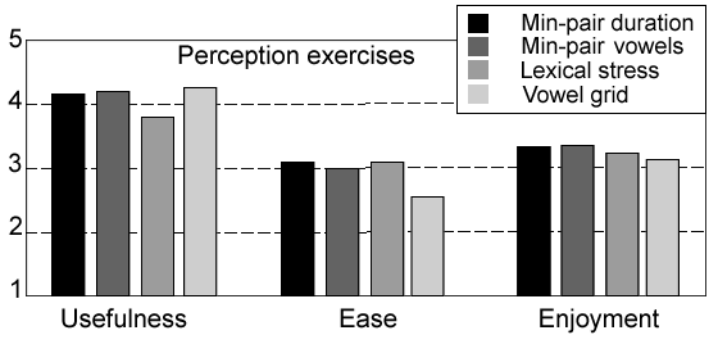

Figure 33 Average answers to questions regarding usefulness, ease, and enjoyment on the perception exercises.

Over all it is worth noting that the students believed all perception exercises were very useful. At the same time the students' rating on usefulness was not positively correlated with how easy the exercises were. There are no significant differences between any of the exercises in regard to the same questions. All the four exercises were similar as to how easy, useful, and fun they were.

According to the theory of flow described in section 2.2.5, a pre-requisite for achieving the state of flow is that a task is neither too easy nor too difficult. When the task is too easy it is boring, and when it is too difficult it is frustrating. When the difficulty is average the learning is optimal. From this view the results from the questions regarding difficulty being "at the middle" might be considered optimal, rather than intuitively regarding easy as the optimal case.

Open response questions perception exercises

The comments in Table 9 highlight some of the individual differences users feel for the different exercises, based on native language. A Russian learner (C1) finds the lexical stress exercise obvious and hence useless, whereas another learner finds the same exercise hard (C2). Several of the learners with Spanish L1 expressed that the vowel grid was difficult (as for example in C3 and C4). More comments on the perception exercises can be found in appendix 16.5.1 on page 213.

	Free comments on the perception exercises
C1	*The part contains different aspects which could be difficult for people (as depends on the native language). the difficult part for me as for Russian speaking is Vowels duration. Even though it is kind of boring to repeat all the time the same words - I find it important to me to practice. Lexical stress part, on the other hand, is very obvious for me and thus useless for me.*
C2	*In lexical stress part, it is sometimes hard to understand where the stress is.*
C3	*Especially the exercises with vowels are rather hard. However, I consider vowels as the hardest part in Swedish, so I find these exercises very useful and teaching. Getting them false often annoys me, but this was normal for beginning.*
C4	*I found the first exercise with 9 vowels quite difficult and not always clear the difference between A and O.*
C5	*It is designed well and very useful for me to improve my perception part, and I really made big progress, the result can say.*
C6	*Good explanations, I like this part.*

Table 9 A selection of free comment responses on the perception exercises the subjects did during four weeks at home.

11.3 Analysis and results: Production exercises

The subjects in this study had been divided into separate groups to test the effect of a number of variables. One of these variables was the effectiveness of the PEDs described in section 6.2. During the training at home, subjects in group 1 had received a version of Ville with feedback from PEDs, whereas subjects in group 2 had done the same type of exercises without PED feedback. The expected outcome was that group 1 would perform better than group 2 in the post test, and that group 2 would in turn outperform Group 3, the control group.

Figure 34, show the score from the pre and post tests on the three groups as calculated by the PEDs (that were on in the background without revealing their score to the learners during the pre and post tests).

Comparing the results between the pre and post tests in the production exercises reveal no clear difference between any of the groups, and we must conclude that the expected outcome of a transfer effect from the production exercises was not confirmed.

It is also interesting to note that the scores from the duration exercises were considerably lower than the other exercises for all groups. This may be a consequence of the 'overshoot' paradigm described in section 4.4.2, since the duration PED was set so that only very clear or exaggerated utterances are accepted. The hope that the subjects in group 1 had learned to produce utterances resulting in positive PED feedback at home, and carry that knowledge over to the post test, was not met.

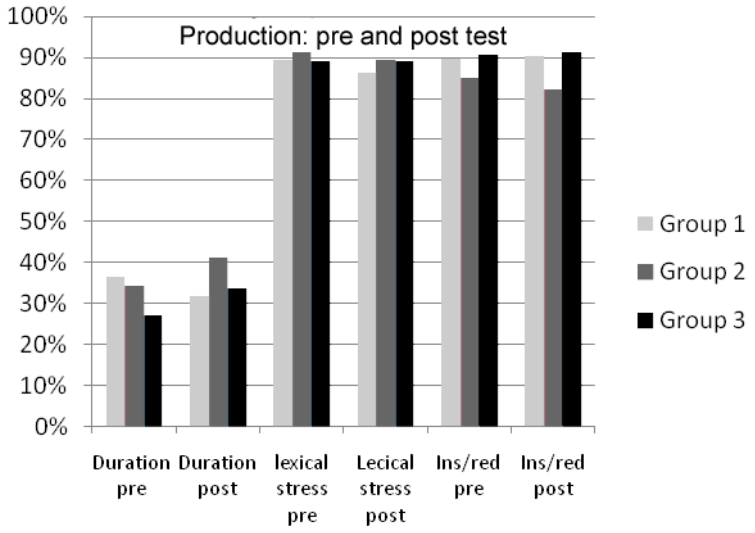

Figure 34 Pre and post test results in the production exercises.

11.3.1 Human listener vs. PED ratings

One may question the accuracy of the production scores since they were calculated using the score from the PEDs. The accuracies of the PEDs are not yet validated and are thus not necessarily correct.

In an attempt to validate the PEDs, human judges were asked to listen to the utterances and rate them in the same way as the PEDs had done. This was done in order to resolve two questions. Would the improvement of the user groups be the same if scored by human judges as they were by the PEDs, and if there was enough agreement among the human listeners, would they be able to validate the performance of the PEDs?

A group of seven naive listeners (native Swedish speakers with little or no phonetic training), and four phonetically trained listeners (three Swedish L2 teachers and one phonetician) were recruited to listen to and rate a balanced subset of the recorded utterances from the pre and post-tests.

The judges were asked to act in the same manner as the PEDs did, i.e. to press a green or a red button after hearing an utterance. They were given the same information as the PEDs, that is, which word the subjects were supposed to say, and what aspect of the utterance that should be judged (quantity, lexical stress, insertion, or deletion). The judges were not given any information about the subjects, such as group or L1, and did not know whether an utterance was from the pre or post test. The 'naive' judges were expected to give more of a global score of the pronunciation, whereas the phonetically trained judges were expected to be able to separate the different aspects and make judgments similar to a PED.

A comment that was given by several judges was that the task was very difficult, because the subjects were so good. The kappa agreement among the four phonetically trained judges was 0.34 (where 0 means chance agreement, and 1 means perfect agreement on all utterances) and among the naïve judges -0.02 (below chance level) indicating that the task was indeed a very difficult one.

Considering the fact that the agreement among the human judges was so low, calculating the accuracy of the PEDs based on such a gold standard does not reveal much. A majority vote among the phonetically trained judges was used as the gold standard to calculate the accuracy, precision and recall as shown in Table 10.

- Accuracy (% of times PED was correct compared to gold standard)
- Precision (% of times PED says 'error' and it is really an error, compared to gold standard)
- Recall (% of times there was an error and PED said there was an error, compared to gold standard)

	Duration	**Lexical stress**	**Insertion/deletion**
Accuracy	0.42	0.73	0.61
Precision	0.26	0.67	1
Recall	0.90	0.10	0.05

Table 10 Accuracy, precision, and recall of the duration, lexical stress, and insertion/deletion PEDs when compared to a majority vote from the human judges.

The difference between the precision and recall values for the duration PED and the insertion/deletion PEDs is interesting in that it may say something about the design of the PEDs.

As explained in section 6.2.1, the duration detector can be set to be more or less forgiving in its judgment, and has in these tests been set to follow the 'overshoot paradigm', and only accept very clear or exaggerated utterances. This is reflected by a high recall value, i.e. when the judges said there was an error, the duration PED also said there was an error (but there were also many instances where the PED said there was an error and the judges were more forgiving). The insertion detector on the other hand is based on a limited set of insertion hypothesis, which is reflected by the high precision score, i.e. every time the PED flags for an error, the judges agree (but the low recall value indicates there are instances that the PED misses).

11.3.2 Questionnaires production exercises

Three questions on a five-point Likert scale were given on each of the production exercises regarding the usefulness, ease, and enjoyment:

Also on the production exercises, users scored very high on usefulness and less so on ease, as shown in Figure 35. The difference between group 1 and group 2 on the production exercises was expected to have a bigger impact on the rating of the exercise than it did (group 1 received feedback from PEDs on their pronunciation, whereas group 2 did not).

	Questions on the production exercises
Q1	Rate each production exercise based on usefulness: 1 for not useful at all - 5 for very useful
Q2	Rate each production exercise based on ease: 1 for very difficult - 5 for very easy
Q3	Rate each production exercise based on enjoyment: 1 for boring - 5 for fun

Table 11 Questions regarding usefulness, ease, and enjoyment on the production exercises focusing on Duration, Lexical stress, and Insertion/Deletion.

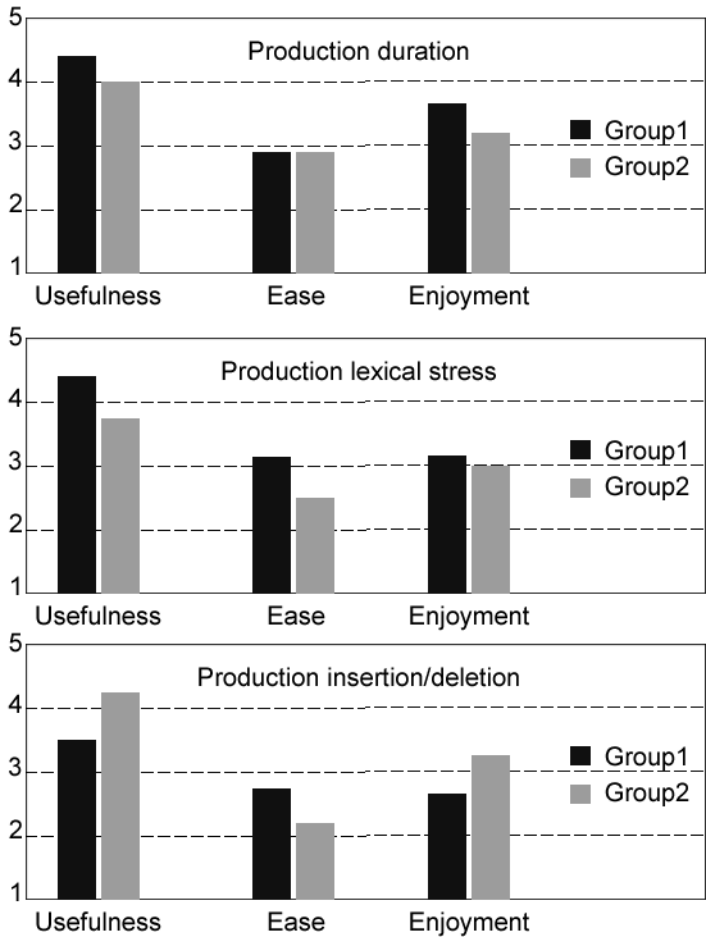

Figure 35 Ratings for usefulness, ease, and enjoyment of the three production exercises

Open response questions

Below are some examples of comments from some subjects on the production exercises. The complete list of responses from all subjects can be found in appendix 16.5.3 on page 214.

Comment C1 from a person in group 1 showing appreciation for the PED feedback, and C2 from a person in group 2, who did not have PED feedback from Ville, are both indications that PEDs can add to the user experience. It is however difficult to read out any clear differences regarding the same matter from the closed questions presented in Figure 35. Comment C3 highlights the importance of PEDs to give reliable and correct feedback, or alternatively the difficulties learners might have to judge their own production, and perceive salient features that they do not have in their L1 (C3's L1 was Mandarin).

	Free comments on the production exercises
C1	*Very helpful as well to clear out pronunciation matters. More words would be good to be added in the database. Quite accurate most of the times and good feedback!*
C2	*It would have been nice to get some feedback about how I did after each word, instead of relying on my own perception of how I sounded….*
C3	*At times I could not hear any difference between my recording and Ville but was told that I was quite wrong.*

Table 12 Selected user comments on the production exercises

Group1 on Detector feedback

Group 1 who had received feedback from the pronunciation error detectors, were given some additional questions with regards to how they experienced the usefulness and accuracy of the PEDs. As can be seen from Table 13, most thought the feedback was very useful, even though they expressed some doubts about the transparency and accuracy of the feedback.

	Questions on the feedback	Mean	St.dev
Q1	Was the feedback useful? 1 for not useful at all - 5 for very useful	4.43	0.51
Q2	Was the feedback easy to understand? 1 for difficult - 5 for easy	3.64	1.34
Q3	Was the feedback in your opinion accurate? 1 for not accurate 5 for very accurate	3	1.11

Table 13 Mean responses from questions regarding usefulness, ease, and accuracy on the detector feedback

11.4 Analysis and results: Simicry exercises

On the Simicry exercises, group 1 used the 'shadowing' mode exclusively while working at home, and group 2 used the 'say-after' mode. As can be seen from Figure 36 there is an improvement based on the Simicry score in all the three groups between the pre and the post test. A one-sample t-test applied on the improvements show that the improvement is significant for group 1 ($p<0.01$) and for group 3 ($p<0.05$), but not for group 2 ($p>0.4$).

The hope that a transfer effect resulting in a noticeable improvement in pronunciation from the exercises users in group 1 and group 2 made at home on a completely different set of sentences was not fulfilled, and was perhaps in retrospect overly optimistic.

The exercises were nevertheless very well received as can be shown from the replies in Table 14 below.

In addition, the Simicry paradigm has proven itself as a very effective data collection tool. The pre and post test alone provided more than 4,500 sentences of a phonetically balanced corpus.

Additionally, the data collected from the work users did at home contains more than 10,000 sentences (~2 hours) of automatically aligned speech data labeled with L1, gender, age etc.

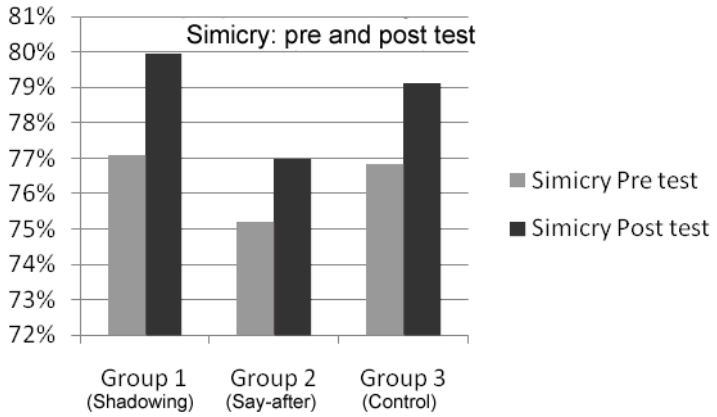

Figure 36 Results from the pre and post test on the Simicry exercises. The score is the mean of the four similarity measures over all sentences for all participants in each group.

11.4.1 Questionnaires Simicry exercises

Three questions on a five-point Likert scale were given on the Simicry exercises regarding the usefulness, ease, and enjoyment:

	Question	Group 1	Group 2
Q1	Rate the Simicry exercises on usefulness? 1 for not useful at all - 5 for very useful	4.5	4.83
Q2	Rate the Simicry exercises on ease? 1 for difficult - 5 for easy	2.29	1.92
Q3	Rate the Simicry exercises on enjoyment? 1 for boring 5 for fun	3.72	3.92

Table 14 Mean responses from questions regarding usefulness, ease, and enjoyment on the Simicry exercises.

It is interesting to note that although both groups rated the Simicry exercise as the most fun and the most useful when compared with all the other exercises, they also thought it was the most difficult.

Say-after or Shadowing?

The students in group 1 also had a question regarding the preference for say-after or shadowing, since they had tried both types of interfaces, (say-after in the pre and post tests, and shadowing while working at home). The question was stated as: You have tried two types of Simicry: Which do you prefer?

Shadowing: 1 for 'I don't like it - 5 for 'I love it'

Say-after: 1 for 'I don't like it - 5 for 'I love it'

The answers for these two questions are presented in Figure 37. The graph clearly show a preference for the say-after version of Simicry compared to shadowing.

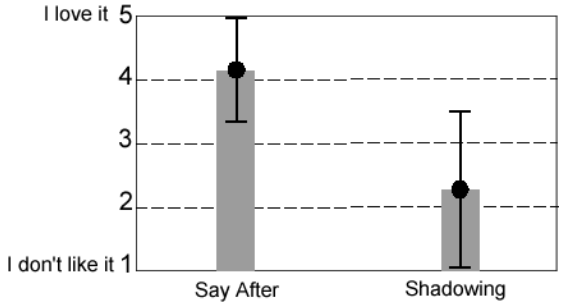

Figure 37 Replies on preference for say-after or shadowing for group 1

Open response questions Simicry

	Free comments on the Simicry exercises
C1	*The best part of Ville software.* *Really helpful and enjoyable! It is more difficult but far more productive...*
C2	*This part is very useful. Especially the packages show examples from everyday life. It is great to hear how Swedish is actually spoken by native speakers. Also trying to mimic them constantly made me memorize sounds and words, which is good.*
C3	*Sometimes I feel I "catch" the melody but I miss pronunciation and yet my result is accepted (even with high acceptance rate). Is melody so important?*

Table 15 User responses regarding the Simicry exercises

Both from the closed questions in Table 14 and the open questions in Table 15 it is clear that the Simicry part was the most appreciated part of the exercises. The constructive criticism on the Simicry part is mainly concerned with the scoring aspect of it, where C3 is interesting in that it is explicitly questioning the importance of 'melody' (prosody).

It is explained by Ville in the instructions the learners received at home that the feedback on this exercise focuses exclusively on the rhythm and melody of the utterances. Even so, some learners found it strange that they get a high score when they in their own opinion pronounced the sentence poorly. It highlights the importance of raising the awareness of prosody as an important aspect to practice.

11.4.2 Feedback Questions

Questions were also given to investigate how the learners experienced the amount of instruction and the amount of feedback they were offered.

	Question	Yes%	No%
Q1	Did you miss getting explicit instructions from Ville on what to do next?	23%	77%
Q2	Would you have liked to get more feedback on your performance from Ville?	85%	15%

Table 16 User replies on questions regarding feedback

Q1 gives relevant input to the discussion on what type of lesson management system is optimal for a framework such as Ville. In section 4.6 a few different options are described regarding the amount of autonomy one wishes to give the learners on how to traverse the content space. The version of Ville used in this study is very open and content is presented more as a 'smorgasbord' than as structured lessons. 77% of the users in this study did not wish for more explicit instructions. Although some users did ask for more structured instruction, this is perhaps an indication that, if resources are limited, priority could be given to other things than the building of an elaborate lesson manager.

The feedback requested in Q2 is, contrary to Q1, something a large majority of the users wanted to have more of. 85% of the respondents wanted more feedback on their performance from Ville. There was a follow up question: If

you answered yes on the previous question, what type of feedback would you like? Several of the answers from group 2 were requesting the kind of short-term feedback on the pronunciation of individual words, which group 1 was given, but there were also requests from both groups about long term feedback on progression, and more encouragement. The complete set of answers can be found in the appendix, section 16.5.9

11.4.3 Ville compared to other CALL systems

A final set of questions was given to get a picture of the overall impression the subjects had of Ville, after having spent a month using the system at home.

A surprisingly high rating was given to Ville on Q3, by users who reported to have used other CALL systems when asked how they would rate Ville compared to other systems, considering the fact that Ville is not a commercial product. Regarding the usefulness of the talking head, Q4 gave a clear answer in favor of the speculation in chapter 10 about the possibility that the talking head would be more appreciated after people got more accustomed to him. In the questionnaire given on the short-term user test, as shown in Figure 29 on page 127, the questions regarding the talking head were seen as low considering the central role of the talking head in the system. After a month of usage however, people's comments were generally very positive as can be seen in appendix 16.5.8, stating among other things that the head added to the entertainment value and helped clarify pronunciation issues.

The final question Q5, asking whether people felt they had learned something from using Ville, received the second highest score of all the closed questions, clearly indicating that a framework such as Ville has a very promising potential, not only as a research tool but also as a learning tool.

	Question	Yes%	No%
Q3	Have you used other computer assisted language-learning applications?	50%	50%
		Mean	St.dev
Q4	If yes on the question above: How would you rate Ville in comparison? 1 for very bad 5 for very good	4.27	0.70
Q5	Do you think the talking head was a useful part of the application? 1 for not useful - 5 for very useful	4.15	0.83
Q6	Do you think that you learned something from using Ville? 1 for nothing - 5 for a lot	4.5	0.58

Table 17 Questions regarding the overall impression of the system and on the usefulness of the talking head.

Open response questions on the Talking head (ECA)

Some users also had comments on the usefulness of the talking head. More comments can be found in the appendix, section 16.5.7

	Free comments on the talking head
C1	*I found that I could learn a lot by watching while I listened. It was very strange when I accidentally rotated the head back and I could see the inside of the eye-balls and the tongue through the neck. Maybe that rotation feature should be changed a bit or removed.*
C2	*Some words are hard to pronounce or hard differenciate. The talking head is useful in that cases.*
C3	*very useful, it provides a "friendly environment" and at least to me it was useful to distinguish vowels easily*

11.5 Discussion

It is interesting to note that although the users in their subjective impressions of the software were very positive and felt that they had learned a lot from using Ville, the objective measures did not reveal this. Several possible explanations to this could be considered:

11.5.1 Easy or difficult

The respondents were asked if they thought any of the exercises were 'very difficult' and 'very easy' respectively. The questions were presented with check marks for every type of exercise and users could hence select more than one exercise. Obviously, no one selected the same exercise as both easy and difficult, but rather some people thought for example the lexical stress perception exercise was very easy whereas other people found it very difficult. Users' responses can be seen in Figure 38.

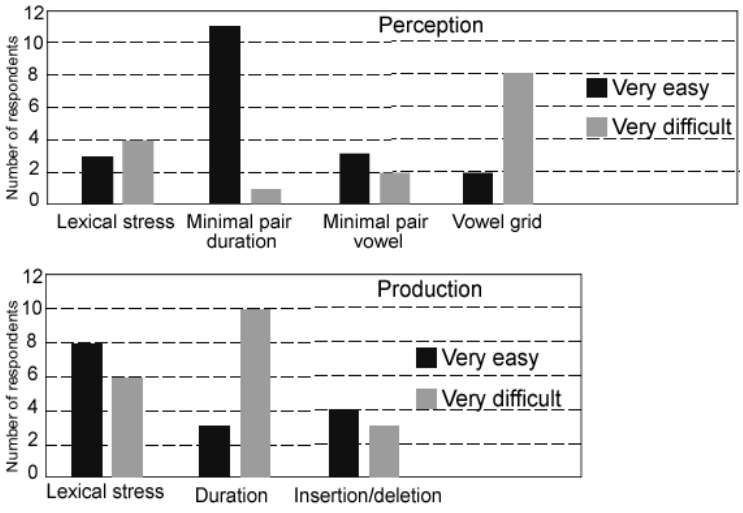

Figure 38 Some exercises were perceived by some users to be very easy, and the same exercises were by other users perceived to be very difficult.

It is interesting to note that the perception exercise considered 'very easy' by most people was the minimal pair duration exercise, while the most difficult exercise was the corresponding production duration exercise.

Although every exercise received at least one vote in each category, the most 'polarized' type of exercise seems to be the lexical stress with an almost equal but high (10 vs. 11) number of respondents having an opinion in one direction or the other.

–12–

CALL on mobile devices

This chapter looks at some portability issues and investigates how well the Ville framework can be adapted to also be utilized in mobile devices. Langofone, a mobile phone application for language learning using the Ville framework is presented.

The advent of mobile phones has opened up for new and interesting ways in which to employ CALL and CAPT applications. With small hand-held devices connected to the internet, and with processing power approaching a computer, the technical limitations that were present only a few years ago have vanished.

Compared with traditional learning in a classroom, many of the advantages of CALL on a mobile phone are similar to those of CALL on a PC. Learners can study at their own pace, at home, whenever they want, in a stress-free environment etc. The size of the device makes the learner even more independent of time and place, and free to learn anywhere and at any time, studying on the bus while commuting or while walking or driving.

The size of the device is however also one of the biggest drawbacks. The small display and the small keyboard create restrictions in terms of usability. A new environment for language learning, calls for new interfaces, functionality, pedagogy and technological solutions.

12.1 Langofone

A mobile phone implementation of the Ville framework was initiated in 2008. It was of interest to see how versatile the architecture was, and how much of the structure made for PC was portable to a mobile phone.

Figure 39 Langofone, utilizing parts of the Ville architecture for a mobile phone application for both iPhone and J2ME phones.

Language learning content is delivered as *packages* using the same domain model and architecture as Ville (see section 4.1). A package is in Langofone a semantic group such as "at the restaurant", or "at the train station", reminiscent of the way content is structured in traditional phrasebooks. Compared to a book made of paper the advantages are however many. It fits in your pocket, in a device most people carry with them anyway, it contains recordings of all the phrases, and the content can be explored and practiced in several interactive ways through the player. Content has so far been created in Swedish, Spanish, Italian, English, French, German, and Portuguese.

The Langofone player consists of a content manager, which allows users to download new packages from within the application, and three language learning features: the phrasebook, quiz, and translate.

The **PhraseBook** is a way to browse, listen, memorize and record phrases or words from a package. The learners choose a package to work with, and listen to recordings of native speakers of each phrase and word within a package.

When the learners feel confident with their own pronunciation, they can record their own utterance and send it to a server for analysis. The recording is then compared to the original recording based on the same similarity measures as used in the Simicry part of Ville (see section 8.1.2). Feedback is then returned to the mobile phone, consisting of four individual scores, and a weighted average as seen in Figure 40.

The **Quiz** is a multiple choice perception exercise for practicing listening or reading comprehension. The learner chooses a package to work with, and the system selects a random phrase from that package. The selected phrase is presented to the learner either as audio only or as audio and text (L2). The learner's task is to select one out of five possible translations. Langofone will give feedback and if an incorrect answer was chosen the correct answer will be shown. The performance is tracked and the phrases with incorrect answers will be queued and presented again.

A **translation service** is also integrated in Langofone where learners can type in any word or sentence, and get a translation into a chosen language, using the Google translate API.

Figure 40 Screenshots from the iPhone GUI in the phraseBook part of Langofone. Listen (left), record (middle), analysis results (right)

WAP and Smartphones

The development of Langofone started shortly before the arrival of smartphones, and was initially developed using Java J2ME (which was the predominant mobile development platform at the time the project started).

The technical solution was based on WAP (Wireless Application Protocol) requests, which required the handheld device to call up a number to transfer data between a server and the handset. The feature specification of the application was constrained by the WAP communication protocol, and a complicated dataflow was developed to circumvent the limited connectivity of WAP (Figure 41).

The more advanced computing ability and connectivity of the Smartphone has made WAP use largely disappear, since all modern handsets support full HTML and do not need the WAP markup.

Another drawback of the old generation mobile phones was the frequency that new handsets came out on the market with different size screen, and device-specific ways to handle playback and recording of sound, downloading of content, communication with the server etc.

WAP pushes (the way the server pushed WAP content to the mobile handset) were also handled differently by different handsets, often requiring user intervention that imposed restrictions in the design.

The Langofone application was actually developed a little before its time, and is in need of a redesign based on the increased possibilities given by current handsets. It has been ported to iPhone, but with the limited functionality given by the original design based on another type of device.

The potential for CALL and CAPT in handheld devices is however clear, and the architectures of the Ville framework has been shown to handle the transition to another medium well. Work on extending the functionality of the application to also incorporate an ECA, and other parts of the Ville framework are being addressed in future work.

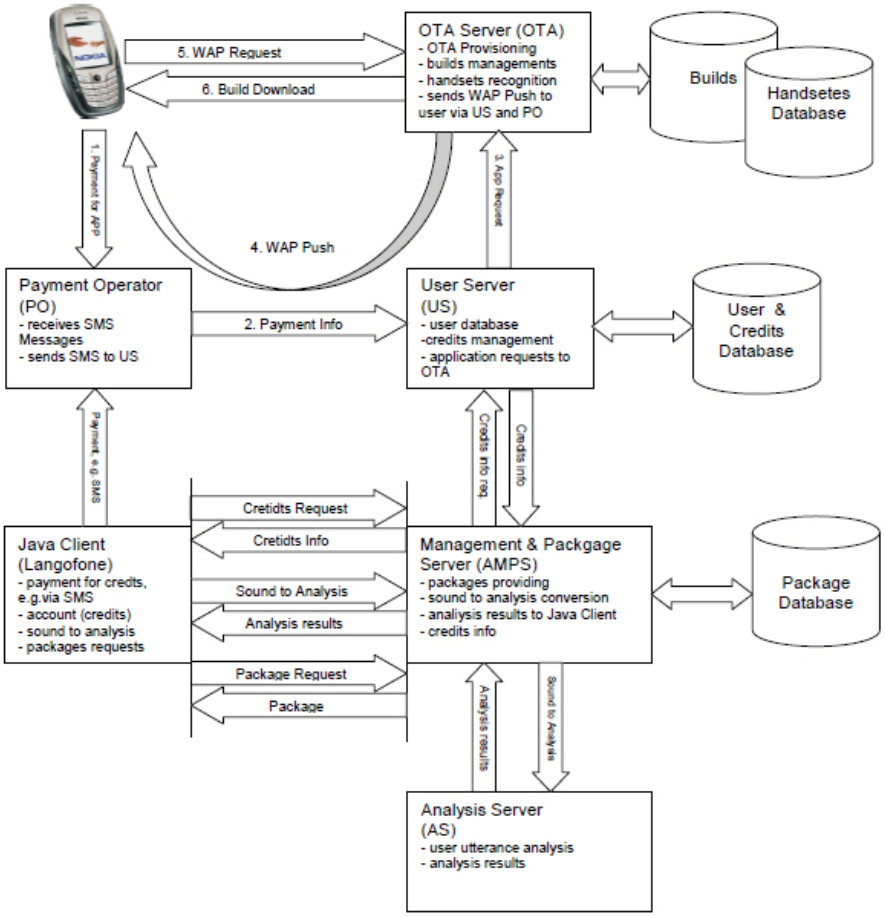

Figure 41 A schematic picture of the dataflow in the Langofone application using WAP.

–13–

Portability to another L2: The CALST project

This chapter presents the CALST project which aims at creating a Norwegian version of the Ville framework. Portability issues concerning localization are addressed, and differences in target audiences as well as some language-and dialect-specific differences with Swedish are discussed.

The CALST (Computer Assisted Listening and Speaking Tutor) project is headed by the Norwegian University of Science and Technology (NTNU) in cooperation with the University of Oslo (UiO), The Centre for Adult Education (EVO) in Trondheim, and KTH. The project has two main goals:

- To develop a Norwegian CAPT system that is able to cater for L2-learners with a wide range of communicative abilities, ranging from foreign university students at NTNU and UiO who are proficient in English, to illiterate users at EVO with no English skills. The program should be used with minimal requirements on instruction or supervision.
- To create a database and tools for contrastive phonetic and phonological analysis for all relevant L1-L2 pairs, partly as a foundation for future research in SLA and partly in order to better tailor exercises used in VilleN.

As elsewhere, there is generally not enough time to teach L2-students pronunciation or to offer extensive listening training in Norwegian courses for foreign-

ers at the participating institutes (and elsewhere), since this will be at the expense of more basic language learning needs.

The system will be used to complement pronunciation teaching in the Norwegian courses both for foreign students and employees at NTNU and UiO, as well as in several courses for teachers of Norwegian as a Second Language. In addition, the collaboration in the project with EVO widens the target group to also include L2-learners that are illiterate and from other language and social backgrounds than what is found at the universities.

The system is based on the Swedish implementation of Ville described in this thesis, but several extensions to the Swedish version have been made in order to accommodate the different needs of the Norwegian system.

13.1 Norwegian dialects

The Norwegian language situation is quite different from that of Sweden and many other language communities in that there is no accepted pronunciation standard in Norwegian. Although there is a common form taught in adult second language classrooms called Bokmål (Urban East Norwegian, UEN), different dialects are used both in formal and informal situations. This creates a serious problem for L2-learners of Norwegian since there are large pronunciation differences in the various dialects and often different words are used to express the same meaning. This dialectal variation cannot easily be addressed in standard language courses, and a need has been identified to better equip L2-learners to deal with everyday communicative situations where variation in the speakers' dialect is typical.

To address this problem recordings of multiple speakers have been made. One male and one female speaker of the dialects in the following regions of Norway have been recorded: Northern Norway, Trøndelag, Western/Southern Norway and Southeastern Norway (UEN). All in all eight different voices will be used in the program, accompanied by 8 different ECAs to give each voice a personality.

The learner can select an ECA from the GUI and listen to and practice one specific dialect, or train across dialects in the same exercise and let the program select the target voice.

Several researchers, including McAllister (1998) have reported that it is good to listen to many different speakers in order to achieve better listening comprehension. Multiple-talker models have been reported as particularly effective to improve perception of novel contrasts (c.f. Logan et al., 1991; Probst et al., 2002), as the inherent variability allows for induction of general phonetic categories or other L2-specific salient features.

Learners will also be able to run the exercises with one specific dialect in mind, which is useful for example when selecting a role model for pronunciation in production exercises. The advantage of having both male and female voices in each dialect becomes apparent in this case to allow learners to choose a role model with the same gender as themselves.

The native speakers who helped us in the recordings were instructed to speak as they normally do with normal speed, reductions and coarticulations in order to offer the learners spoken utterances that are as close to authentic speech as possible. It is however difficult for beginners to assimilate reductions and coarticulations in the early phases of learning, and it is often common practice to speak slower and clearer when speaking to L2 learners who are in an early phase of development.

To incorporate such considerations in the program, two options were considered: to make it possible for the learners within the program to slow down the speech samples by manipulating the acoustic signal, or to make double recordings, one normal and one slow hyper-articulated version of all the recordings. The latter was in the end chosen even if it would include more work and a larger set of recordings. The difference between slow and normal speech is not uniformly distributed, as for example plosives are not stretched in the release burst but only in the occlusion. Long vowels are typically exaggerated, coarticulations will be reduced or removed, and other aspects of the speech such as the lexical stress will be affected differently, with stressed syllables being stronger and more emphasized in a hyper-articulated version. It was hence decided that to ensure the best possible quality in the learning material, exaggerated, hyper-articulated versions of all recordings should be made and stored in the *Word-Objects* with a special tag. The learners have the option to choose to do the exercises using either type of recording, or a learner can choose to use normal recordings and select a: "Say again" button, to get a repetition of the last spoken word in a hyper-articulated version.

13.2 Wordlists

A wordlist of basic vocabulary from the course books used by the participating institutions has been created and categorized into semantic categories. Approximately 1000 words were selected from the aggregated wordlists and divided into 43 categories. The criterion for the selection of the base vocabulary was also that it should satisfy the A1 and A2 vocabulary range of CEFR, (Common European Framework of Reference for Languages)

- A1: Has a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations.
- A2: Has a sufficient vocabulary for the expression of basic communicative needs.

  Has a sufficient vocabulary for coping with simple survival needs.

  Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics.

In addition to the wordlist 30 useful phrases have been selected and placed in a separate category.

All words have then been visualized. Approximately 30% of the images are taken (with permission) from "UVic's Language Teaching Clipart Library". For words where no appropriate image was found, complementary drawings have been made by a local artist in the same artistic style as the drawings from the internet, in order to get a consistent and coherent set of images.

English translations, transcription and the inflection of words have been added as part of the *WordObjects*, and as mentioned above, sound files have been recorded for the aggregated wordlist, in four dialects, with one male and one female speaker for each dialect, and in both normal and hyper-articulated versions.

13.2.1 Homonyms

The perception exercises that the learners are working with are structurally the same as in the Swedish version described in section 5.1. VilleN says a word and the learner's task is to identify the corresponding picture within a grid of pic-

tures. Some changes in the architecture have however been necessary as a result of the added complexity that the dialect recordings have created.

If there are synonyms, hyponyms, or homonyms among the words, this needs to be specified in the underlying *WordObject* XML-structure, otherwise it can create confusing and undesirable situations. If for example, VilleN says 'bird', and both a picture of a parrot and a picture of a generic bird are present on the screen, the learner must be admitted to click on either of the two pictures and get positive feedback. Anything else would be confusing.

There are however some words that are homonyms in one dialect and not in another, for example 'yours' and 'theirs' are both 'deres' in Urban East Norwegian, but 'dokkers' and 'dis' in Southwestern Norwegian. This means that a restructuring of the domain model has been necessary.

13.3 Minimal pairs

In addition to the basic vocabulary, focus has been placed on introducing the basic phonology of the Norwegian language.

The system will be able to offer an explicit introduction with recordings of every sound alone, and recordings of every sound within words, and for consonants in initial, medial, and final position. A set of contrasts (minimal pairs) has been determined from an experience-based analysis of which contrasts are difficult for (at least some) to learn.

Consonants

A set of ten word pairs per contrast category have been made in each position (initial, medial, final) The minimal pair definition has in some cases been somewhat expanded, i.e. sub-minimal pairs, where more than one feature differs, have in some cases been made where no 'pure' minimal pairs have been found.

Plosives	Fricatives	Nasals	Liquids
/p/ vs. /b/	/h/ vs. /k/	/n/ vs. /ŋ/	/r/ vs. /ʃ/
/t/ vs. /d/	/h/ vs. /j/	/m/ vs. /n/	/l/ vs. /ɭ/
/t/ vs. /ʈ/	/f/ vs. /ʋ/	/m/ vs. /ŋ/	/r/ vs. /l/
/ʈ/ vs. /ɖ/	/f/ vs. /p/	/n/ vs. /ŋ/	/r/ vs. /ɭ/
/ʈ/ vs. /d/	/ʋ/ vs. /p/	/ŋ/ vs. /ŋ/	
/k/ vs. /g/	/ʋ/ vs. /b/		
	/f/ vs. /b/		
	/s/ vs. /ʃ/		
	/s/ vs. /ç/		
	/ʃ/ vs. /ç/		

Table 18 the 25 consonant contrasts that are selected for minimal pair exercises.

Vowels

96 contrasts have been selected among the vowels and diphthongs, and a set of ten word pairs per contrast category have been made. The 34 pairs presented in Table 19 are the basic segments that should be expanded so that long and short segments of the same vowel are contrasted against each other, and for many vowels also long vs. short in another category are contrasted against each other, i.e. the first cell (/ɑ/ vs. /æ/) in Table 19 can be expanded into: /ɑː/ vs. /ɑ/, /æː/ vs. /æ/, /ɑː/ vs. /æː/, /ɑː/ vs. /æ/, /ɑ/ vs. /æː/, and /ɑ/ vs. /æ/.

Vowels & diphtongs				
/ɑ/ vs. /æ/	/e/ vs. /i/	/ø/ vs. /o/	/æʉ/ vs. /e/	/æi/ vs. /ɑi/
/ɑ/ vs. /o/	/e/ vs. /æ/	/ø/ vs. /ʉ/	/æʉ/ vs. /æ/	/æi/ vs. /e/
/ɑ/ vs. /ə/	/e/ vs. /ø/	/y/ vs. /ʉ/	/æʉ/ vs. /o/	/æi/ vs. /i/
/ɑ/ vs. /ɑi/	/e/ vs. /o/	/y/ vs. /ø/	/æʉ/ vs. /ø/	/æi/ vs. /oy/
/ɑ/ vs. /ə/	/i/ vs. /y/	/u/ vs. /ʉ/	/æʉ/ vs. /u/	/æi/ vs. /æʉ/
/æ/ vs. /ø/	/i/ vs. /ʉ/	/u/ vs. /ø/	/øy/ vs. /ø/	
/æ/ vs. /o/	/o/ vs. /ʉ/	/u/ vs. /o/	/øy/ vs. /æi/	
			/øy/ vs. /oy/	

Table 19 The base segments that expands into 96 vowel contrasts selected for minimal pair exercises.

All in all 250 minimal pairs for consonant exercises and 960 minimal pairs for vowel exercises are thus offered within the system. Perception exercises along the same lines are also being prepared for supra-segmental aspects such as stress, consonant clusters, tone, and assimilations.

Not all of these contrasts are useful exercises for all learners, and so some selection criteria is required, as will be discussed in the next section.

13.4 Contrastive analysis in CALST

As described in section 3.1.2, it is generally accepted that the claims made by the contrastive analysis hypothesis (CAH) as presented by Lado (1957) are too strong, and that there are other factors which determine the difficulty language learners have with acquiring new sounds (Eckman, 1977; Odlin, 1989; Flege, 1995; Major, 2001; Abrahamsson, 2004).

This does however not mean that CAH should be completely rejected or abandoned. As stated by Ellis (1994):

"*The problem with CAH is that it is too simplistic and too restrictive. The solution as many researchers have come to recognize, lies not in its abandonment but in careful extension and revision"*

13.4.1 L1-L2map

The second task in the CALST project is to design and evaluate a revised and extended contrastive analysis tool called the L1-L2map.

As discussed in section 3.1 different languages have different ways of encoding phonological salience. Adult L2 learners have already acquired a number of features from their L1, and will thus only need to acquire a sub-set of the phonological inventory of the L2.

As it is difficult for an L2 learner to guess, or in other ways deduce which the novel features of the L2 are, these must be presented to the learner somehow. In Ville, a language learner could thus receive an explicit presentation of all phonological categories in the L2 that are new or novel for the learner. Learning the novel features is an activity suitable for individual instruction, and a very appropriate task for a CALL/CAPT application such as Ville. This type of information is also inherently multimodal, (sounds, pictures, lip-movements, recordings of words and sentences containing every feature), and hence appropriate to encode in a CAPT application such as Ville.

Which features are new and different depends on the L1-L2 pair, hence a mechanism for automatically selecting the relevant features for each individual learner is needed, which is the task for the L1-L2map.

The L1-L2map will serve as a platform for researchers with a phonetic background to encode language data and make it available in a format that can be used by CAPT creators. L1-L2map is a wiki with two levels of users.

It is a generally accessible tool, where any user can access the data and browse and compare the phonological features of different languages. Each language is encoded individually by an expert in that particular language, and the contrastive analysis is performed by 'superimposing' the data for two languages on top of each other.

As shown in Figure 42 for vowels and Figure 43 for consonants, a color scheme is employed in the visualization of the data. The first language chosen is blue and assumed L1 status, the second language chosen is red and assumed L2 status, and the features they have in common (where there is an overlap) are green. This way it becomes apparent which features are different (red), and thus need attention since they are absent in the L1.

The first version of the L1-L2map is built around data from the UPSID database (Maddieson, 1980), and a script made available to the CALST project by Henning Reetz.

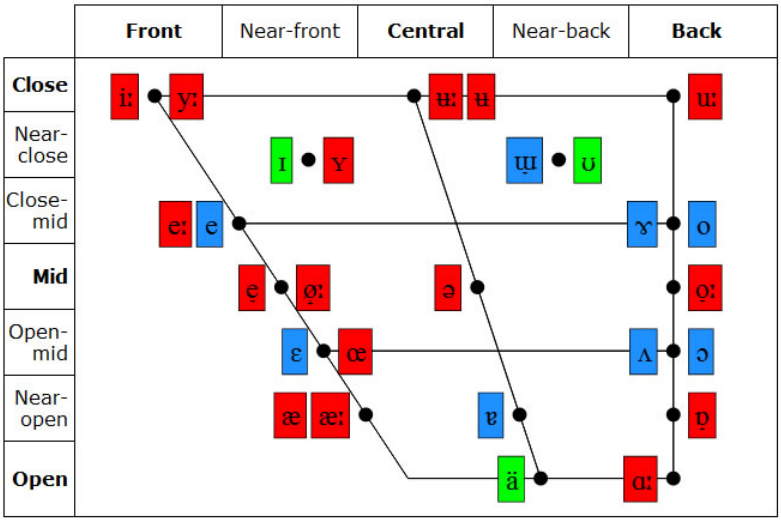

Figure 42 Vietnamese (blue) and Norwegian (red) vowels as documented in UPSID and visualized in L1L2-map. Green indicates vowels shared by both languages.

13.4.2 Extensions to UPSID

The UPSID database only lists sounds that are distinct phonemes in any given language. It is not enough to do a simple comparison of which sounds constitute phonemes in each language. Extensions that are added to the L1-L2map are: positional restrictions, syllable structure (phonotactics), tone, stress, and timing.

Positional phonemic restrictions

As pointed out by for example Setter & Jenkins (2005) the positions in which sounds occur in syllables must be taken into account. As described in section 3.1.2 difficulties that could be predicted from CAH if position is part of the description, will otherwise go by unnoticed. For example in Mandarin, only two consonants ([ŋ] and [n]) are allowed at the end of a syllable, even if many other consonants appear in syllable-initial position. A consequence of this is that many consonants which can occur at the end of Norwegian syllables present a difficulty for learners of Norwegian with Mandarin as L1.

The L1-L2map presented in Figure 43 is the complete set of consonants in both Norwegian and Mandarin, taken from the UPSID data, where positional

restrictions are not taken into account. In the current version this pane is divided into three separate panes, with each pane displaying only the phonemes that are allowed in initial medial, and final position respectively.

Syllable structure

Not only positional restrictions, but also the phonotactic constraints of languages (restrictions on the permissible combinations of phonemes) will be encoded in the L1-L2map. Norwegian has a relatively complex syllable structure comparable to that of Swedish (described in section 3.3.3), and constitutes a difficult part of acquiring the language for many learners.

Permissible consonants and consonant clusters in onset and coda will be encoded into separate lists in the database, and when a contrastive analysis is performed between the learners L1 and the L2, clusters that are allowable combinations in the L2 but missing from the learners L1 will be extracted.

13.4.3 Level of phonetic detail

The problem can be described in general terms as:

If the description has *'too little detail'* important differences between languages will be lost. For example, discarding Russian palatilazation and saying that the Russian /n/ and Norwegian /n/ are the same nasal would for a Russian learner of Norwegian simplify things, but would if ignored render him or her with a strong Russian accent.

If the description has '*too much detail*' unimportant differences between languages will be highlighted and emphasized. For example, the difference between a dental or alveolar /n/ is acoustically small and phoneticians may perhaps have varied opinions on whether a /n/ in a language is dental or alveolar. In the UPSID database some languages are labeled as having a dental /n/ and some as having an alveolar /n/. In Mandarin /n/ is for example categorized by Husby & Jin (2006) as dental and by UPSID as alveolar. Figure 43 is based on the UPSID description of Mandarin whereas Figure 44 is based on the description by Husby & Jin (2006). In the first description there are 7 consonants that overlap with Norwegian, whereas in the second description there are 13 consonants that overlap.

The variation in overlap between the two descriptions in Figure 43 and Figure 44 exemplifies the problem well. If phoneticians do not agree on how to en-

code a language, we are not able to reliably compare two languages described by two different phoneticians using different standards.

This potential problem is addressed by having a set of guidelines for the contributors and a moderator that is able to assist and if necessary change a description if conflicts are discovered. This will be an iterative process where inconsistencies will be discussed and addressed underway in the process.

13.4.4 Lack of ranking

The two descriptions in Figure 43 and Figure 44 also touch upon a separate but related difficulty with this type of description, namely the lack of ranking between differences.

If for example the description of Mandarin from Figure 43 were used in a learning scenario, both the dental /t/ and /d/ and the retroflex /ʈ/ and /ɖ/ would be seen as new phonemes that should be brought to the learner's attention.

The lack of retroflexion is a much more serious error, with a potential of misunderstandings (minimal pairs such as /fut/ vs. /fuʈ/ ('foot' vs 'fast'), compared to the /t/not being dental. There is however no way for the learner to know which ones, in the list of new phonemes that are potentially creating a communicative hindrance if pronounced incorrectly, and which ones are not.

To address this issue, a priority listing is needed which states that if the contrastive analysis generates an item of kind X it should be of higher priority than an item of type Y.

13.4.5 Alternative usage of the templates

The template designed in the CALST project is intended to be used to code the phonology of languages independently of each other. It is however possible to use the same template, and the same editing and input mechanisms to create charts of language difficulties, instead of phonological features. Based on the experience of phonetically trained teachers of second languages, or from analysis of the data collections the Ville framework will provide, it will be possible to create another type of chart, where the elements to insert are language difficulties a learner from a particular L1 will have in the target L2 instead of phonological features in the L1.

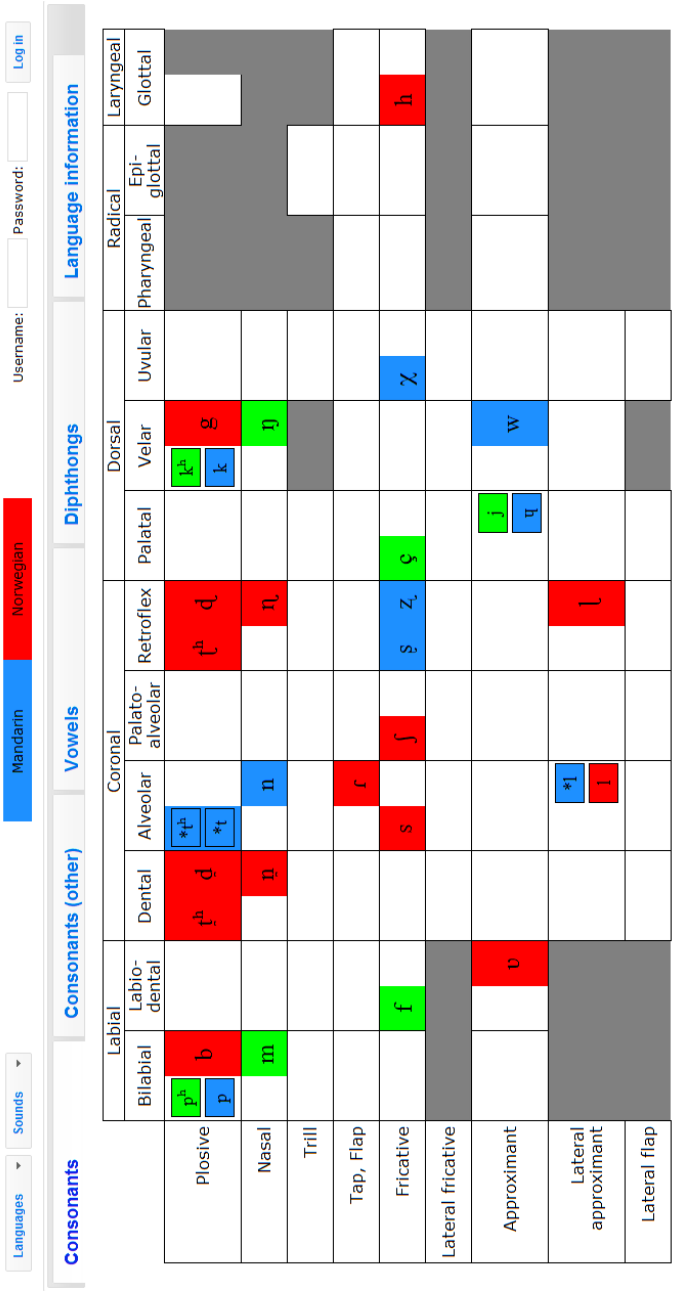

Figure 43 Contrastive description of Mandarin (blue) and Norwegian (red) consonants from UPSID. Green squares indicate shared segments and red and blue respectively that the other language is missing these segments.

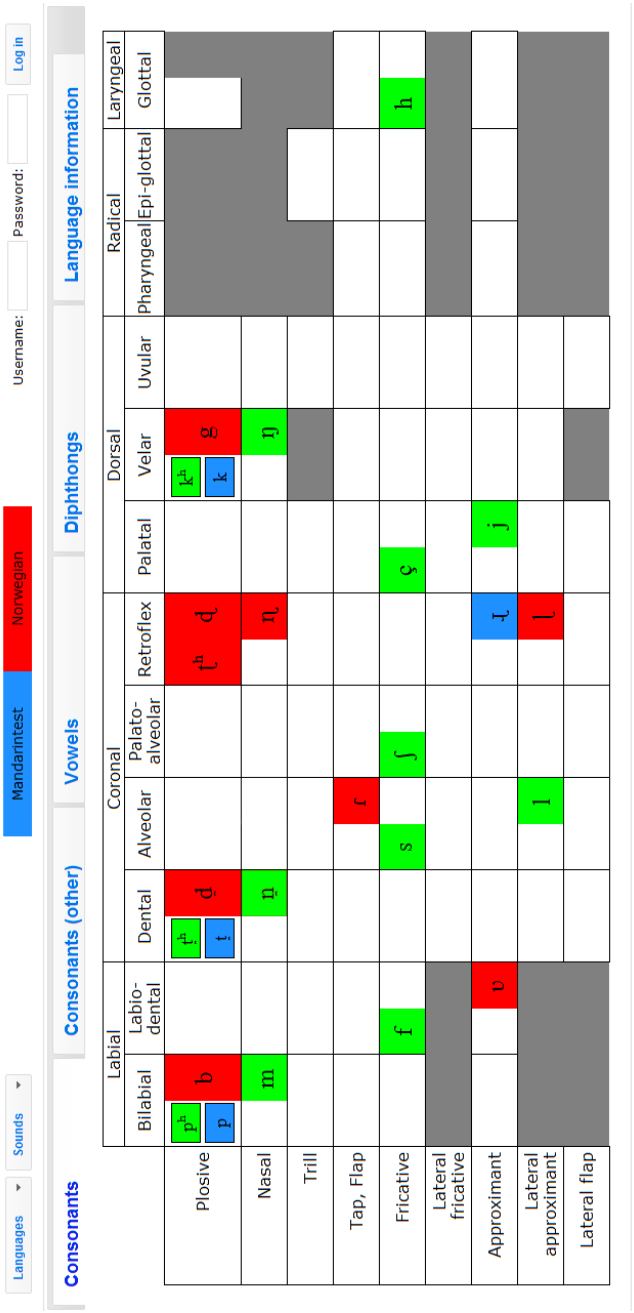

Figure 44 alternative description of Mandarin to that of Figure 43 taken from Husby & Jin (2006).

–14–

Future work and conclusions

The ambition of this thesis has been to build a framework for language learning based on a virtual language teacher. The goal has been that it should be fully functional in the sense that it can be used by 'real' language learners. In addition the framework was envisioned not only as a new type of language learning application, but that it should also serve as a data collection tool and a research platform to improve future CALL and CAPT applications.

14.1 Future work

14.1.1 Data analysis and new PEDs

Ville for SWELL presented in chapter 7 is currently collecting data autonomously, fulfilling the initial goal set out for the project as discussed in the introduction, section 1.1.1.

As of March 23 2011, 1655 students from 99 different countries have used the version of Ville that is available for KTH students. The distribution of students grouped by continents, and the number of perception exercises, writing exercises and recordings that are logged is shown in Table 20. For a detailed listing of which countries and number of users per country see appendix 16.6.

Continent	*Students*	*Countries*	*Perception Exercises*	*Writing Exercises*	*Recordings*
Asia	824	33	2068	493	5678
Europe	484	37	2319	786	8237
Africa	54	11	296	36	520
North America	18	2	148	18	118
South America	34	7	292	28	264
Central America	21	7	109	35	456
Oceania	11	2	61	4	51
Unknown	209	0	349	99	1032
Total	1655	99	5642	1499	16356

Table 20 The number of exercises and recordings made by the students (grouped by continent).

Table 20 shows that a substantial amount of users have tried the program, and thus contributed to the data collection. Individual variation is however large, and the distribution of how much data is generated per user resembles a Zipfian distribution where the number of exercises per user are very low in many cases, whereas some people have used the program extensively.

The database currently being created by the L2 learners of Swedish will eventually contain many instances of the same words spoken by different users, and since many of the words contain particular Swedish difficulties, it 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 demographic 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 difficulties with and where to put the focus in future tool development. It will also be possible to determine which words are most difficult with regard to spelling or listening comprehension.

Currently there is only data on word level exercises. While it is possible to extract useful data for future segmental and syllable level detectors from the database, adding specific exercises with such data in mind will be a valuable addition.

There is nothing on sentence level in the database so far, since the Simicry exercises are not included in the publicly available version of Ville. The inclusion of Simicry exercises is likely to enhance both the user satisfaction of the program (since the Simicry exercises were the highest rated exercises in both user study 1 and user study 2) and enhance the database. The Simicry paradigm was designed with exposure to large amounts of sentences in mind, and the results from user study 2 have indeed shown that the Simicry exercises generate a lot of data.

A key factor in the future development of Ville is considered to be to utilize this database for development of new PEDs. Data generated by the Ville framework has already been used in Koniaris & Engwall (in press), and a new type of PED, using support vector machines as described in Picard et al. (2010) and modeled on Ville data, is currently under investigation.

14.1.2 Content:

Many pronunciation difficulties are due to the learners' lack of knowledge regarding the phonetics and phonology of the target language. Both text-to-sound confusions (subconsciously using the grapheme to phoneme rules of the L1), as well as the absence in the L1 of certain phonemic contrasts that are present in the L2 are common causes of errors for the L2 learners. To make learners aware of such contrasts and phonological rules, Ville could offer a crash-course in phonology coupled with explicit introductions to every salient feature of the L2. Since many language teachers lack training in the phonetic sciences this is seen as a way to cover the necessary declarative knowledge in the domain.

A complete description of all the phonological features of the L2, and descriptions of the difficulties language learners might have in acquiring these, could serve as a knowledge base and guideline for the creation of exercises. The L1-L2map and extended contrastive analysis in the CALST project, described in chapter 13 is an effort in this direction, where not only the features of the L2 are described, but also a description of the learners L1, in order to extract the features of the L2 that are absent in the learners L1, and bring them to the learner's attention.

14.1.3 Make Ville a web application

It would for several reasons be beneficial to make Ville a web application, embedded in a browser, rather than a stand-alone application. It would make the application platform independent (Ville currently only runs on Windows), and it would allow the application to also run on handheld devices such as smartphones and PDAs. A thin client and a server connection would greatly facilitate updates, and allow new content to be brought to the users seamlessly. It would also encourage the creation of community activities such as user-generated content. Also for the research, development and deployment of new PEDs, a web-based application would open up for new possibilities. The ongoing work in the CALST project with the L1-L2map (section 13.4) also calls for an architecture where content is stored in a separate database and is dynamically loaded into Ville, based on the availability of content, matching the L1-L2 contrastive analysis. This also entails a shift towards client-server architecture to a larger extent than what is already available.

14.1.4 Gameplay

Although some effort and much thought has been put into it, Ville still lacks much of 'the fun part'. Considering the fact that motivation is such an important factor for successful learning, a meta-level of gameplay is a much needed feature in Ville, in order to tie the different exercises together to a complete experience with a common goal. To create a reward system in order to increase the motivation of the learners to continue using the program, references can be sought in for example various types of action-adventure games, involving long-term obstacles that must be overcome, in combination with smaller obstacles such as item gathering, or simple puzzle solving.

This requires a more sophisticated tracking of student progress and a consistent scoring mechanism. It will also have an impact on the lesson management issues discussed in section 4.6. Dynamic difficulty adjustment is a key factor for many games and should get much more focus also in game-based learning.

14.2 Conclusions

We may conclude that a fully functional virtual language teacher has been built, and tested on a large number of language learners, and that a substantial amount of data (both speech data and data on user behavior) has already been collected.

A number of functionally separate but interacting units in the VLT have been discussed. ***Structural elements*** such as the 3D model of the head, audio-visual speech synthesis, feedback modules, the learner profile, and a lesson manager, have been described. A separation between structure and content has been sought and the domain model for the language-specific information has been presented as a way to handle that (section 4.1).

A number of ***feedback strategies*** have been explored, including explicit feedback in the form of explanations, corrections and encouragement from the virtual language teacher Ville, as well as implicit feedback in the form of comments, negotiation of meaning, or communication breakdown from the DEAL character. Several types of visual feedback have been demonstrated, such as iconic traffic-lights from PEDs, real-time transmodal feedback, and scores from the similarity measures in the Simicry game.

Exercises for several different ***aspects of language*** have been developed and shown possible to teach within the Ville framework: both perception and production exercises on segmental level as shown in chapter 5. Suprasegmental aspects such as quantity, lexical stress, epenthesis and deletion at syllable level (chapter 6). Vocabulary acquisition and writing exercises at word level (chapter 7), prosody exercises and acquisition of formulaic language at sentence level (chapter 8), and communicative skills at discourse level (chapter 9), have been demonstrated.

A number of ***pronunciation error detectors*** (PEDs) have been developed and tried out within the framework. PEDs focusing on specific phonological

features such as quantity, lexical stress, insertion and deletion errors have been presented in chapter 6.

Several different ***teaching methodologies*** have been shown possible to deploy within the Ville framework. Constructivist principles of learning where users should play an active role in the learning process and be given the power to design their own training fit well with the overall approach of the framework. Possibilities to incorporate the communicative language teaching methodology have also been explored in DEAL (chapter 9). By using automatic speech recognition and spoken dialogue system technology, negotiation of meaning is pursued in a game where the learners are confronted with a shop keeper in a flea-market. Methods praised by the audio lingual method, relying on low level associative learning, has also been explored as for example in minimal pairs exercises in (5.1.1). In addition to these well established but very different methodologies, the new paradigm of game-based learning has been explored and several attempts to add an aspect of gameplay have been demonstrated.

Portability and localization issues have been discussed and the architecture has been shown to be flexible in terms of adaptation to new target languages. Apart from the Swedish version discussed throughout the thesis, a Norwegian version (chapter 13) and an Estonian version (chapter 6) have been built with minimal changes to the domain model.

Portability has also been demonstrated in terms of utilizing the Ville framework on handheld devices. Langofone, a mobile phone application, was presented in chapter 12. This is so far a scaled down version of Ville in terms of functionality, but not in terms of L2 languages where content has been created in Swedish, Spanish, Italian, English, French, German, and Portuguese.

Last but not least, the framework presented in this thesis has demonstrated its viability as a ***research tool***. The data collected by Ville has already been put to use in the development of new PEDs, and other ways to make use of the data are being investigated. As new exercises are added to the program, and more L2 learners download and use the program, the data collection will continue to grow, both in size and in manner, to offer scientists in the CALL and CAPT community much needed L2 speech data for future research.

References

Abrahamsson, N., & Hyltenstam, K. (2004). Mognadsbegränsningar och den kritiska perioden för andraspråksinlärning.. In Hyltenstam, K., & Lindberg, I. (Eds.), *Svenska som andraspråk – i forskning, undervisning och samhälle* (pp.221-258. Lund: Studentlitteratur.,.

Abrahamsson, N. (2004). Fonologiska aspekter på andraspråksinlärning och svenska som andraspråk. In Hyltenstam, K., & Lindberg, I. (Eds.), *Svenska som andraspråk – i forskning, undervisning och samhälle* (pp. 79-116). Lund: Studentlitteratur.

Al Moubayed, S., & Beskow, J. (2010). Prominence Detection in Swedish Using Syllable Correlates. In *Interspeech'10*. Makuhari, Japan.

Anderson, J. R. (2002). ACT: A Simple Theory of Complex Cognition John R. Anderson. *Cognitive modeling*, 49.

Auberg, S., Correa, N., Rothenberg, M., & Shanahan, M. (1998). Vowel and intonation training in an English pronunciation tutor. In *STiLL-Speech Technology in Language Learning*.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556.

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device.. *Psychological Review*, *105*(1), 158.

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

Begley, S. (2009). *The plastic mind: new science reveals our extraordinary potential to transform ourselves.* Constable.

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

Beskow, J., Karlsson, I., Kewley, J., & Salvi, G. (2004). SYNFACE - A talking head telephone for the hearing-impaired. In Miesenberger, K., Klaus, J., Zagler, W., &

Burger, D. (Eds.), *Computers Helping People with Special Needs* (pp. 1178-1186). Springer-Verlag.

Beskow, J., Edlund, J., Granström, B., Gustafson, J., Skantze, G., & Tobiasson, H. (2009). The MonAMI Reminder: a spoken dialogue system for face-to-face interaction. In *Interspeech 2009*. Brighton, U.K.

Bongaerts, T., Van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in second language acquisition, 19*(04), 447-465.

Bongaerts, T., Mennen, S., & Slik, F. (2000). Authenticity of pronunciation in naturalistic second language acquisition: The case of very advanced late learners of Dutch as a second language. *Studia linguistica, 54*(2), 298-308.

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 linguistics. In *Proceedings of the 38th Annual Meeting of the Association for Computational Lingusitics*.

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. (1996). The Waxholm spoken dialogue system. *Phonetica Pragensia IX. Charisteria viro doctissimo Premysl Janota oblata. Acta Universitatis Carolinae Philologica 1, 1996.*

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.

Chapelle, C. A. (2001). *Computer Applications in Second Language Acquisition.* Cambridge Univ Press.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* Harper & Row, Publ., New York.

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *Modern Language Journal, 80*(2), 183-198.

Crawford, C. (1984). *The art of computer game design.* Osborne/McGraw-Hill.

Crepy, H., Denoix, B., Destombes, F., Rouquie, G., & Tubach, J-P. (1983). Speech processing on a personal computer to help deaf children. *9th World Computer Congress, Paris, France*, 669-671.

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics.* Blackwell.

Csikszentmihalyi, M. (1991). *Flow: The psychology of optimal experience.* Harper-Perennial New York.

Cucchiarini, C., Strik, H., & Boves, L. (2000). Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication, 30*(2-3), 109-119.

Cunningham-Andersson, U. (1996). Native speaker reactions to non-native speech. *Second-language speech. Structure and process*, 133-144.

De Saussure, F. (1986). *Course in general linguistics.* Open Court Publishing Company.

DeKeyser, R. M. (1997). Beyond explicit rule learning. *Studies in Second Language Acquisition, 19*(02), 195-221.

DeKeyser, R. (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology.* Cambridge Univ Pr.

Deroo, O., Ris, C., Gielen, S., & Vanparys, J. (2000). Automatic detection of mispronounced phonemes for language learning tools. In *Proceedings of ICSLP* (pp. 681-684).

Doughty, C., & Williams, J. (1998). *Focus on form in classroom second language acquisition.* Cambridge University Press.

Dörnyei, Z. (2009). *The psychology of second language acquisition.* Oxford University Press.

Eckman, F. R. (1977). Markedness and the contrastive analysis hypothesis. *Language Learning, 27*(2), 315-330.

Edlund, J., & Heldner, M. (2006). /nailon/ - software for online analysis of prosody. In *Proc of Interspeech 2006 ICSLP* (pp. 2022-2025). Pittsburgh PA, USA.

Edlund, J., Heldner, M., & Gustafson, J. (2006). Two faces of spoken dialogue systems. In *Interspeech 2006 - ICSLP Satellite Workshop Dialogue on Dialogues: Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems.* Pittsburgh PA, USA.

Elert, C-C. (1995). *Allmän och Svensk Fonetik.* Norstedts.

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

Ellis, N. C. (2006). Cognitive perspectives on SLA: The associative-cognitive CREED. *AILA Review, 19*(1), 100-121.

Engstrand, O. (1999). Swedish. In *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet* (pp. 140-142). Cambridge: Cambridge 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. (2009). Are real tongue movements easier to speech read than synthesized?. In *Proceedings of Interspeech*.

Engwall, O. (2003). Combining MRI, EMA & EPG measurements in a three-dimensional tongue model. *Speech Communication, 41*, 303-329.

Eskenazi, M., & Hansma, S. (1998). The fluency pronunciation trainer. In *Proceedings of the STiLL Workshop*.

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

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication, 51*(10), 832-844.

Falk, G. (2001). *Stigma: How we treat outsiders.* Prometheus Books.

Fant, G., & Kruckenberg, A. (1994). Notes on stress and word accent in Swedish. *STL-QPSR, 35*(2-3), 125-144.

Fant, G. (1966). A note on vocal tract size factors and non-uniform F-pattern scalings. *STL-QPSR, 7*(4), 022-030.

Fant, G. (2004). *Speech acoustics and phonetics. Selected writings.*. Kluwer Academic Publ.

Felps, D., Bortfeld, H., & Gutierrez-Osuna, R. (2009). Foreign accent conversion in computer assisted pronunciation training. *Speech Communication, 51*(10), 920-932.

Ferguson, S., Moere, A. V., & Cabrera, D. (2005). Seeing Sound: Real-Time Sound Visualisation in Visual Feedback Loops Used for Training Musicians. In *Proceedings of the Ninth International Conference on Information Visualisation* (pp. 97-102).

Fitts, P. M. (1964). Perceptual-Motor Skill Learning1. *Categories of human learning*, 243.

Flege, J. (1987). The production of 'new'and 'similar'phones in a foreign language: evidence for the effect of equivalence classification. *Journal of Phonetics, 15*, 47-65.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, 233-277.

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

Gluszek, A., & Dovidio, J. F. (2010). The Way They Speak: A Social Psychological Perspective on the Stigma of Nonnative Accents in Communication. *Personality and Social Psychology Review, 14*(2), 214.

Goertzel, B. (2006). *The hidden pattern.* BrownWalker Press.

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

Gruenstein, A., Seneff, S., & Wang, C. (2006). Scalable and portable web-based multimodal dialogue interaction with geographical databases. In *Ninth International Conference on Spoken Language Processing*.

Gustafson, J., Lindberg, N., & Lundeberg, M. (1999). The August spoken dialogue system. In *Proc of Eurospeech 99* (pp. 1151-1154).

Gustafson, J., Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D., & Wirén, M. (2000). AdApt - a multimodal conversational dialogue system in an apartment domain. In Yuan, B., Huang, T., & Tang, X. (Eds.), *Proc. of ICSLP 2000, 6th Intl Conf on Spoken Language Processing* (pp. 134-137). Beijing: China Military Friendship Publish.

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

Hazan, V., & Kim, Y. H. (2010). Can we predict who will benefit from computer-based phonetic training?. In *Proceedings of L2WS*.

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech communication, 47*(3), 360-378.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory.* Lawrence Erlbaum.

Hincks, R. (2005). *Computer Support for Learners of Spoken English*. Doctoral dissertation, KTH.

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

Hoppe, D., Sadakata, M., & Desain, P. (2006). Development of real-time visual feedback assistance in singing training: a review. *Journal of computer assisted learning, 22*(4), 308-316.

Husby, O., & Jin, F. (2006). *Innføring i kinesisk for nordmenn, innføring i norsk for kinesere.* Tapir akademisk forlag.

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

Johnson, W. L., & Valente, A. (2008). Tactical language and culture training systems: using artificial intelligence to teach foreign languages and cultures. *Proceedings of IAAI 2008*.

Johnson, W. L., & Valente, A. (2009). Tactical Language and Culture Training Systems: using AI to teach foreign languages and cultures. *AI Magazine, 30*(2), 72.

Jurafsky, D., & Martin, J. (2000). *Speech and Language Processing- An introduction to natural language processing,*..... Prentice Hall.

Kjellin, O. (2002). *Uttalet, språket och hjärnan. Teori och metodik för språkundervisningen.* Hallgren& Fallgren Studieförlag AB.

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

Koniaris, C., & Engwall, O. (in press). Perceptual Differentiation Modeling Explains Phoneme Mispronunciation by Non-Native Speakers. To be published in *2011 IEEE Int. Conf. on Acoust., Speech, Sig. Proc. (ICASSP)*. Prague, Czech Republic.

Konttinen, N., Mononen, K., Viitasalo, J., & Mets, T. (2004). The effects of augmented auditory feedback on psychomotor skill learning in precision shooting. *Journal of Sport and Exercise Psychology, 26*(2), 306-316.

Koster, R. (2004). *A theory of fun for game design.* Paraglyph press.

Lado, R. (1957). *Linguistics across cultures.* University of Michigan Press.

Larsen-Freeman, D., & Long, M. H. (1991). An introduction to second language acquisition research. *London, New York*.

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied linguistics, 22*(1), 1.

Lee, C. H. (2004a). From knowledge-ignorant to knowledge-rich modeling: a new speech research paradigm for next generation automatic speech recognition. In *Proc. ICSLP*.

Lee, V. C. (2004b). *Languageland: A multimodal conversational spoken language learning system*. Master's thesis, Massachusetts Institute of Technology.

Lepper, M. R., GREENE, A., & Nisbett, R. E. (1973). Undermining Children\'s Intrinsic Interest with Extrinsic Reward: A Test of the \"overjustification\" hypothesis. *Journal ol Personality and Social Psychology, 28*(1), 129-137.

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.

Levis, J. (2007). Computer technology in teaching and researching pronunciation. *Annual Review of Applied Linguistics, 27*, 184-202.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English/r/and/l: A first report. *Journal of the Acoustical society of America, 89*(2), 874-886.

Long, M. H. (1990). Maturational constraints on language development. *Studies in second language acquisition, 12*(03), 251-285.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake. *Studies in second language acquisition, 19*(01), 37-66.

Maddieson, I. (1980). *UPSID: UCLA phonological segment inventory database*. Phonetics Laboratory, Department of Linguistics.

Major, R. C. (2001). *Foreign accent: The ontogeny and phylogeny of second language phonology*. Lawrence Erlbaum.

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.

McAllister, R. (1998). Second language perception and the concept of foreign accent. In *STiLL-Speech Technology in Language Learning*.

McClelland, J. L., Thomas, A. G., McCandliss, B. D., & Fiez, J. A. (1999). Understanding failures of learning: Hebbian learning, competition for representational space, and some preliminary experimental data. *Progress in brain research, 121*, 75-80.

McClelland, J. L., Fiez, J. A., & McCandliss, B. D. (2002). Teaching the/r/-/l/discrimination to Japanese adults: Behavioral and neural aspects. *Physiology \& Behavior, 77*(4-5), 657-662.

McGraw, I., & Seneff, S. (2007). Immersive second language acquisition in narrow domains: A prototype ISLAND dialogue system. In *Proc. of the Speech and Language Technology in Education Workshop*.

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).

Menzel, W., Herron, D., Bonaventura, P., & Morton, R. (2000). Automatic detection and correction of non-native English pronunciations. *Proceedings of InSTILL 2000*, 49-56.

Merzenich, M. M., Jenkins, W. M., Johnston, P., Schreiner, C., Miller, S. L., & Tallal, P. (1996). Temporal processing deficits of language-learning impaired children ameliorated by training. *Science, 271*(5245), 77.

Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: do students learn more deeply when they interact with animated pedagogical agents?. *Cognition and Instruction, 19*(2), 177-213.

Mostow, J., & Duong, M. (2009). Automated Assessment of Oral Reading Prosody. In *Proceeding of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling* (pp. 189-196). Amsterdam, The Netherlands, The Netherlands: IOS Press.

Mote, N., Johnson, L., Sethy, A., Silva, J., & Narayanan, S. (2004). Tactical language detection and modeling of learner speech errors: The case of Arabic tactical language training for American English speakers. In *Proceedings of InSTIL/ICALL2004-NLP and Speech Technologies in Advanced Language Learning Systems-Venice* (pp. 19).

Murphy, J. M. (1991). Oral communication in TESOL: Integrating speaking, listening, and pronunciation. *Tesol Quarterly, 25*(1), 51-75.

Nakatani, C., & Hirschberg, J. (1994). A corpus-based study of repair cues in spontaneous speech. *JASA, 95*, 1603–1616.

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.

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, W., Solé, M., Recasens, D., & Romero, J. (2003). Automatic speech recognition for second language learning: how and why it actually works. In *Proc. ICPhS* (pp. 1157-1160).

Neufeld, G. (1978). On the Acquisition of Prosodic and Articulatory Features in Adult Language Learning.. *Canadian Modern Language Review, 34*(2), 163-74.

Nickerson, R., & Stevens, K. (1973). Teaching speech to the deaf: Can a computer help?. *Audio and Electroacoustics, IEEE Transactions on, 21*(5), 445-455.

Odlin, T. (1989). *Language Transfer*. Cambridge University Press.

Oomen, C., & Postma, A. (2004). Effects of Time Pressure on Mechanisms of Speech Production and Self-Monitoring. *Journal of Psycholinguistic Research, 30*(2), 163-184.

Öster, A-M. (1998). Spoken L2-teaching with contrastive visual and auditory feedback. In *Proc of ICSLP98, 5th Intl Conference on Spoken Language Processing* (pp. 2663-2666). Sydney, Australia.

Öster, A-M., House, D., Hatzis, A., & Green, P. (2003). Testing a new method for training fricatives using visual maps in the Ortho-Logo-Paedia project (OLP). In *Proc of Fonetik 2003, Umeå University, Dept of Philosophy and Linguistics PHONUM 9* (pp. 89-92).

Paganus, A., Mikkonen, V. P., Mantyla, T., Nuuttila, S., Isoaho, J., Aaltonen, O., & Salakoski, T. (2006). The Vowel Game: Continuous Real-Time Visualization for Pronunciation Learning with Vowel Charts. *Lecture Notes in Computer Science, 4139*, 696.

Paivio, A., & Clark, J. .. (1991). Dual coding theory and education. *Educational Psychology Review, 3*(3), 149-170.

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied, 14*(2), 101-117.

Picard, S., Ananthakrishnan, G., Wik, P., Engwall, O., & Abdou, S. (2010). Detection of Specific Mispronunciations using Audiovisual Features. In *International Conference on Auditory-Visual Speech Processing*. Kanagawa, Japan.

Piller, I. (2002). Passing for a native speaker: Identity and success in second language learning. *Journal of Sociolinguistics, 6*(2), 179-208.

Postma, A. (2000). Detection of errors during speech production: a review of speech monitoring models. *Cognition, 77*(2), 97-131.

Prensky, M. (2001). *Digital game-based learning.* McGraw Hill.

Prensky, M. (2002). The motivation of gameplay. *On the Horizon, 10*(1), 5 - 11.

Probst, K., Ke, Y., & Eskenazi, M. (2002). Enhancing foreign language tutors-In search of the golden speaker. *Speech Communication, 37*(3-4), 161-173.

Raux, A., & Eskenazi, M. (2004). Using task-oriented spoken dialogue systems for language learning: potential, practical applications and challenges. In *InSTIL/ICALL Symposium 2004*.

Robinson, P. (1997). Generalizability and automaticity of second language learning under implicit, incidental, enhanced, and instructed conditions. *Studies in Second Language Acquisition, 19*(02), 223-247.

Schmidt, R. A., & Lee, T. D. (2005). *Motor control and learning: A behavioral emphasis.* Human Kinetics Publishers.

Schneider, W., & Chein, J. M. (2003). Controlled \& automatic processing: behavior, theory, and biological mechanisms. *Cognitive Science, 27*(3), 525-559.

Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P., & Zue, V. (1998). Galaxy-II: A reference architecture for conversational system development. In *Proc. of ICSLP '98* (pp. 931-934). Sydney, Australia.

Seneff, S., Wang, C., & Zhang, J. (2004). Spoken conversational interaction for language learning. In *InSTIL/ICALL Symposium 2004*.

Setter, J., & Jenkins, J. (2005). State-of-the-Art Review Article. *Language Teaching, 38*(01), 1-17.

Siniscalchi, S. M., Svendsen, T., & Lee, C. H. (2008). Toward a detector-based universal phone recognizer. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on* (pp. 4261-4264).

Sjölander, K., & Beskow, J. (2000). WaveSurfer - an open source speech tool. In Yuan, B., Huang, T., & Tang, X. (Eds.), *Proceedings of ICSLP 2000, 6th Intl Conf on Spoken Language Processing* (pp. 464-467). Beijing.

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. (2005a). Exploring human error recovery strategies: implications for spoken dialogue systems. *Speech Communication, 45*(3), 325-341.

Skantze, G. (2005b). Galatea: a discourse modeller supporting concept-level error handling in spoken dialogue systems. In *Proceedings of SigDial* (pp. 178-189). Lisbon, Portugal.

Sluijter, A. (1995). *Phonetic Correlates of Stress and Accent*. Doctoral dissertation, Rijksuniversiteit te Leiden.

Squire, K., & Jenkins, H. (2003). Harnessing the power of games in education. *Insight, 3*(1), 5-33.

, W. & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America, 26*, 212-215.

Sundström, A. (1998). Automatic prosody modification as a means for foreign language pronunciation training. In *Proc of STiLL98, ESCA-Workshop on Speech Technology in Language Learning* (pp. 49-52). Marholmen, Sweden.

Sweetser, P., & Wyeth, P. (2005). GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE), 3*(3), 1-24.

Tallal, P., Miller, S. L., Bedi, G., Byma, G., Wang, X., Nagarajan, S. S., Schreiner, C., Jenkins, W. M., & Merzenich, M. M. (1996). Language comprehension in language-learning impaired children improved with acoustically modified speech. *Science, 271*(5245), 81.

Tang, M., Seneff, S., & Zue, V. W. (2003). Modeling linguistic features in speech recognition. In *Eighth European Conference on Speech Communication and Technology*.

Tepperman, J., Stanley, T., Hacioglu, K., & Pellom, B. (2010). Testing Suprasegmental English Through Parroting. *Proceedings of Speech Prosody, Chicago, USA*.

Thorén, B. (2008). *The priority of temporal aspects in L2-Swedish prosody: Studies in perception and production*. Doctoral dissertation.

Tomasello, M. (2005). *Constructing a Language: A usage-based theory of language acquisition.* Harvard University Press.

Ullman, M. T. (2001). The neural basis of lexicon and grammar in first and second language: The declarative/procedural model. *Bilingualism: Language and Cognition, 4*(02), 105-122.

van Mulken, S. A., & Andre, E. (1998). The Persona effect: How substantial is it?. In *Proc. of HCI98* (pp. 53-66).

Vicsi, K., Roach, P., Öster, A., Kacic, Z., Csat‡ri, F., Sfakianaki, A., & Veronik, R. (2001). SPECO: A multilingual, multimodal speech training system, education arena. In *Proc of Eurospeech 2001*. Aalborg Scandinavia.

von Ahn, L. (2006). Games with a Purpose. *COMPUTER*, 92-94.

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).

Webster, J., Trevino, L. K., & Ryan, L. (1993). The dimensionality and correlates of flow in human-computer interactions. *Computers in human behavior, 9*(4), 411-426.

Wik, P., & Engwall, O. (2008). Can visualization of internal articulators support speech perception?. In *Proceedings of Interspeech 2008* (pp. 2627-2630). Brisbane, Australia.

Wik, P., & Escribano, D. (2009). Say 'Aaaaa' Interactive Vowel Practice for Second Language Learning. In *Proc. of SLaTE Workshop on Speech and Language Technology in Education*. Wroxall, England.

Wik, P. (2004). Designing a virtual language tutor. In *Proc of The XVIIth Swedish Phonetics Conference, Fonetik 2004* (pp. 136-139). Stockholm University.

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*.

Wray, A. (2008). *Formulaic language: Pushing the boundaries.* Oxford Univ. Press.

Zahorian, S. A., & Correal, N. S. (1994). Vowel training experiments with a computer-based vowel training system. *The Journal of the Acoustical Society of America, 95*, 3014.

–16–

Appendix

16.1 User study 1: Replies from questionnaire

Section 1: Lexical stress perception

More 'Swedish' i.e non-loan words: reasoning what the stress 'should' be from e.g. English is too tempting otherwise
i didn't like the time counted because it made me hurrying instead of really remembering every pronunciation. i took the time to listen to the words but not to put them in my head. to evaluate ourselves on this exercise, it would be better to have a certain number of words and we can try a pronunciation only once at the end, the number of correct answers is counted instead of the number of answers/amount of time thus we would concentrate more.
maybe the lines and the half circles are not very explicit for stressed and unstressed syllables *the symbol____ meant for me the opposite meaning before listening the instructions . to interpret the symbols is practice*
mr ville's face can be a little more encouraging… or voice be a little more kind. *maybe it's better if ville can be more encouraging by saying 'you are doing great' or 'try again' (Ville sounds a bit stern)*

start a part with words of 2 syllables, then 3 and 2, then 4,3, and 2
i had difficulties to distinguish how many syllables the word had
i want to see the word when ville is pronouncing them because i did not know most of the words and i could not understand how many syllables the word includes
it would be good if the section is followed by explanations of how to find the stressed syllables of the words
the pronunciation is very clear and the stressed syllables are easy to identify
rather difficult. many time i click and its pure chance.
Perhaps the spelling of the words could be shown after clicking the right answer
show the correct scheme with the word
the square in which one has to click could become green or red for a good/bad answer
it didn't say there is a time limit at the beginning (or i didn't get it).
i particularly liked the small comments of ville when i was right or false. it is more interactive to change the comments
this exercise is very good, but it just took a while that i could hear the differences in the words. i have never done this kind of exercise before.
instructions should be more clear
the different number of syllables would be easier spotted if in different rows. too much time was spent looking for the right pattern. use different colors on the stress points would be helpful.
The clock on time was too present, it should be smaller, it became a distraction.
having to look for the repeat utterances button was not helpful. should be closer to the stress boxes.
it was confusing to know if the selected boxes were correct because ville said right or wrong in different ways. just right or wrong would be helpful.
it would be better if the explanation of the words are given.

Section 2: Duration perception

pronunciation of 'äta' seems to have been recorded differently
good exercise! made me understand things!
it could be interesting to hear also the pronunciation of the other word in order to hear the difference (after you have answered of course) *maybe that have a way to compare the two different pronunciation could be interesting*
when it is wrong, is that possible to show the pronunciation of the other word to 'feel' the difference
mr ville should look more cheerful. he looks too serious
repeat more examples with vowel that were wrong
give the written words when giving the first example
some words came 3 times maybe more *more exercises*
i learned a rule about this exercises so it was easier for me than the first exercise. maybe it could be better if ville gives a little bit information about the rule.
it might be a good idea to read the two similar words eg 'tak' and 'tack' together once so it becomes easier to identify which is which
i felt quite uncertain about some of the words and din't know if they were long or short
i don't remember the exact rule about when it's a long or short syllables *i don't feel like it is so useful at the beginning and that it is more a subtlety. the context will make me understand which word we are talking about*
showing the translations of the words
give some rules about long or short vowels *maybe its better to have a introduction to know how the pronunciation is! i figured it out myself at the end, but its nice to know it from beginning.*
explanation of some common examples in the quiz and life
i particularly liked the fact that some words were tested again. but i haven't noticed if it was the word i was wrong. it can be useful to repeat the test for the words which are wrong.

now i saw the clock. but the color is good, because otherwise you work too much against the clock.

it would be better if the repeat utterance button was closer to the exercise area

i found it difficult to hear the real differences between the long and short pronounced words even though i can see the word and know when a sound should be spoken long or short

Section 3: vowel perception

maybe some examples with å, ö would be nice

there are more vowels similar to each other. i tihink it depends on where you come from. to me 'u' and 'y' is easy but 'y' and 'i' is difficult. maybe it can take up more samples with more variations after some survey.

give an example with u and y.

some vowels seem to cause resonance / noise in the audio, making them 'easier' y notably

it would be good to hear the pronunciation of the word we clicked on when we fail. to be able to compose and to understand our mistake

it can be interesting to hear the pronunciation of the other word when we make a mistake in order to hear the difference

it would be nice if i can listen to 2 words together

I want to listen to the difference between the words in word pairs at the end of the test. i want to listen to the word i chose because at the end of the test i still have some uncertainty about some of the words.

change the words after the first half of them, it is possible just to remember which one we clicked previously.

show the translations of the words

it is better to give the reslt at the end. which vowels usually are not distinguished

maybe some examples at the beginning, to learn better

more specific exercises should be introduced

i think it'll be useful to renew the voice of ville. he is a bit stern (it feels like he chides whenever a mistake is made)
expression using the similar words
possibly not show the words but instead the vowel and select the vowel. if a student already knows some Swedish they will select the correct word just by hearing it.
possibly use more complicated words or longer to make it more realistic

Section 4: Lexical stress production

if you don't like to listen to ville, you can't stop him and just start the exercises, he continues speaking when the exercise has already started. and he begins to repeat the rules again i didn't get why
ville did, by default, not pronounce the word. at least for me his pronunciation was a major help, having to click 'repeat utterance' every time breaks focus. *he should pronounce the words per default. i had to click on repeat utterance every time i wanted to listen to him even the first time.* *i'd like ville to really say the word once first and give me a feedback of what he hears when i get wrong otherwise i have no didea how to correct it and the exercise helps for nothing*
the analysis tool didn't work very well for me. maybe i didn't understand how correctly to use it
have to click 'repeat' it starts speaking…show the analysis. 2 curves are the same when red light on…maybe it's nice you can click to give together 'accurate' and 'students voice'
i like that you can hear your own voice to compare
it would be nice that the correction comes automatically when it's wrong. i dislike to have to keep pressing during registrating since is sometimes click too early. click to start and click to stop would be nicer. it would also be good if ville switches to the next word when the answer is correct
the recording procedures could be more simple
i didn't remember which button to click if i want to see which word ville thought i did
it would be nice to have a translation

the reaction time of the program is a bit slow in my mind
i couldn't hear a difference between my pronounced word and the word from ville and it was still wrong
more time—i felt stressed
it's actually quite distracting to have ville in this section because one tends to look at him for 'approval' or 'disapproval' i.e. his expression but his nodding and shaking of head are quite confusing
stress production for more common expression
move the repeat utterance closer to the exercise area
the analysis should be seen in the same window as the exercise
useful to have the instructions written as well as spoken due to the length and steps of the exercise

Section 5: Duration production

again i didn't get how to stop ville and start the exercise before he ends his speech. when i clicked red button in case of my mistake, it appears the diagram , i also didn't reallyunderstand how to use it.
i didn't really get what are the percentages in the analysis
the analysis tool was very useful *it is very useful to find some part in uttal that myself not notice*
you have to click red light then click analysis to see current analysis
sometimes red light even though pronunciation should be right (according to supervisor)
it's impossible to understand what's done wrong *i got almost all wrong and i didn't really understand why that was quite frustrating* *sometimes we don't understand why we are wrong, even if we look at the matching graph*
translation would be good
i feel like what i say is nevertheless understandable even if not perfect and even if ville doesn't

like it
more introductory analysis page
'update' button could be more visible
it was good that it had some instruction at the beginning
the exercise was very interesting
i like this kind of exercises although it is difficult for me to hear what i have done wrong. very important to practice that
it is not possible to understand the duration production details/patterns so it may not be useful to show the users
button to listen the instructions, for this part, while the tutor give the instruction the mouse can move showing the buttons on the screen
could be two presentations for the exercise, one where show the letter that i need to pronounce longer, for example gl---a----ss
what about using any source to compare? for example supervisor could pronounce correctly, and then student would try to repeat
it was a bit complicated to have to go to another tab in order to see the analysis. since its timed it should more intuitive to move from one action to another
sometimes i feel me and ville are pronouncing the word in the same way, but its red =wrong

Section 6: Insertion/deletion

there was green (not red) light under insertion or reduction, or sometimes both. or sometimes it was red. if it's green, in different places, what does it mean?
easier to have a word correct than in the previous exercise. that is good.
the part analysis is really well done. we knew exactly what we have to change
some insertion / reduction sentences are easily mistaken too. maybe its good to include that.
have an option to see translations

the differences between that we sometimes get one and sometimes two
it seems that the system has difficulties to recognize some sounds
i could get a green light without registering anything *i said some words wrongly, but it was ok, even for just part of a word.*
the section which showed what parts are missing/extra could be easier to understand. it should be more obvious if i have pronounced the whole word right, for example say 'det är rätt'
it wasn't easy to see where i made mistake. i didn't really understand when i was totally right
bug in analysis for skratta
i did not understand if i had to fulfill in green all the point. maybe an example could be useful.
i never looked on the face before but in this section i recognized that i didn't make so many mistakes when i had watched how the animation pronounced the words
the english word in the corner will be nice, so you can learn s ome new words (true for all the other exercises)
there is a tendency to look at ville's lips movements to learn how to pronounce a word. however, ville many times isn't articulating the word's pronunciation. can ville's lip movements be refined to teach users how to pronounce?
my voice is hard to hear?
the instructions were long and required several steps, but no examples.
too many steps from pronouncing to seeing the problems. Also, there is a tendency to push the wrong tabs when moving from the analysis to the exercise
i like the possibility to see where the mistake was and i appreciate the fact that i could correct myself and answer properly

Section 7 Simicry Say-after

I didn't find the score in percentage, i couldn't move to next phrase at all. so i have no idea how good were the exercises. what do figures on the left side mean? and i didn't get until the instructor said that i can move to another sentence. it would be nice to see this possibility when

you start the exercise
feedback /s score disappears too quickly if 'good enough'
lots of time 'the volume was too low' but i was talking loudly. louder than for the previous exercises. or even the software didn't recognize i talked. but good exercise!
i had some trouble with my microphone so i had to speak very loud
i was not loud enough
should have more detailed survey and show difference from student and ville
i like the score part. it is good and encouraging
i didn't get to see the percentage of how much i said right. i was confused if i got the sentence right or wrong
have a translation! it's frustrating to learn to say sentences without knowing what they mean
it would be helpful to show the user when ville is going to start his sentences
i think it would be better to have a 'record' button
i liked this section very much. but when you have to repeat a sentence 5 or 6 times it is a little bit boring. it would be better when a new sentence appears and then the old sentence afterwards again
it needs some improvements with matching results
exercise not easy to do but it is very useful
the ranking must be higher than previous, i think i did not pronounce correct some words in the expression and it shows that they were good
indicator in percent would be better. i would like to see statistics, or average result for the test
it crashed a few times but was easy to start again
ability to say each word separately, especially for the more difficult words. after which the sentence could be said in whole
the speaking is very fast

Section 8 Simicry Shadowing

everything was nice until the expression i göteborgs skärgård when i tried to repeat it, it was written that the volume was too low even if i spoke loudly. so i couldn't move till next phrase

maybe a time to know when ville starts to speak

maybe a countdown for the start of his speaking, longer sentences/parts of sentences

it is difficult to know when to begin to speak. it can be useful to include a timer

loopback does not work for last sentence

good training. instead of just repeating, to try at the same time. i really liked this exercise. it is going a bit fast but it is good. it is stimulating

it's nice

it is easier than the former one. better controlled

when you don't catch the right time to speak, they say you speak not loud enough

the program could have run more smoothly

again, i didn't see how well i did. and ville just kept repeating the same thing again and again and again and that was driving me crazy

it's nice to speak with ville, but translation please

a lot of bugs: 'divide by zero'

once again the recording was too low

i had often over 80% although id haven's spoken the right words because the text was so fast

there should be more exercises and more commonly used phrases

at first i was wondering how such shadowing exercises could help. then i realised that such exercises could help to improve or perfect one's pronunciation. it was frustrating at first to just try to catch up with ville's speed of reading the line

use karaoke way—highlight words which are pronounced at the moment

i would like to adjust threshold

sometimes it was possible to cheat speaking different, or just noise

would be good to have ability to slow down the sentences in order to learn to say each word

Question 7: suggestions for changes or additions

add possibility to see text descriptions of the rules before the exercise, if you miss something ville said to clarify some options we can see on the screen
there are a lot of words in the examples which i did not know, perhaps it would be more to have the possibility to see the translation in english if the user wants it
a mixed exercises option would be good
it can be interesting to see own results at the end
mr ville can be more friendly
more analysis would be nice
sometimes a cleaner sound, without cracking at the beginning
the program could have run more smoothly
make some conclusion about user's mistakes
an advanced tool where every criteria is credited in the same sentence without sound assistance first
the overall interface could be more user friendly
its good to have more examples from the book, like the last exercises
different type of voice (feminine one) could also be interesting to have, but it is not a priority
more commonly used phrases and words should be introduced
it would be helpful to have ville's lips move in tandem with the pronunciation of a word so

users can watch his lips and learn how to pronounce better
for each exercise identify which are the most problematic point and suggest extra exercises.
try to limitate the numbers of exercise is good because the student doesn't feel tired using the program
it would be good to be able to add additional exercises, and/or compose them by myself.
online version of ville?
videos of speaking people (to me, and between them)
more flexibility to adjust settings like duration, thresholds etc.
in the perception exercises, show the whole sentences and not just words

Question 8. is there anything you particularly liked?

the idea is very good and program is really useful.
the movements of his mouth were realistic and helpful
the possibility to practice the pronunciation! in class we don't always dare, in front of a computer we feel more confident. good training the time to be confident.
the layout is very attractive
the comparison between our pronunciation and the true pronunciation is very accurate
the speaking part is really good, it should be more explanation of the errors
it is very easy to use it.
the production exercises part could have been very useful if the program could run more smoothly. it's a great tool to help improve listening and speaking Swedish
i really think it improves my pronunciation by mimicry with what ville says
it is very good that you can replay your own voice
comparing my voice with ville is good
the simicry production part was good
distinguish different problems with different sections

it is without doubt efficient to improve one's use of oral Swedish
i liked the kind of coaching provided by the character ville: the analysis given, the relevance of the vocabulary, the repetitions of the words when we are wrong, the learning of typical pronunciation
i think the animation is very good, although i didn't use it so much. i was more concentrated on the sound and the written word but i recognized that i made less mistakes, when i concentrated on the face. i just have to get used to it
last exercise (repeating sentences with ville) and ville itself
sound is soothing though stern
pronunciation exercises are exactly what i need
useful tools and exercises to learn. i liked the mimicry production part
the sound is very good

Question 9: is there anything you disliked?

some bugs, but i hope they will be fixed
i don't like when i can't correct and get wrong all the time, but it might be me more than ville
i would like to see clearly how well/bad i did in each section
on the previous version there were some writing exercises and some learning words exercises too and that was really useful to learn new vocabulary
even with a black jacket as background, i find the floating head of ville3 a bit creepy
it's a bit slow, you cant make the guy stop talking easily
i disliked that we sometimes don't understand all the possibilites of the application (but it is already shortly explained).
sometimes we don't understand why we are wrong
sometimes the application does not recognize the voice
in the mimicry production i didn't like it to repeat the same sentence until you reach 80%. it is better to have some kind of success between it because you are very frustrated and bored when you don't manage a phrase. so it would be better, when the sentence appears again after 1-2 different sentencse in between
ville looks and sounds a bit stern. it will help if he looks and sounds friendlier althought there is nothing seriously wrong with the present ville. not sure if a woman character might work better?
interface could be more convenient
the munus and the layout seem a bit complicated at times. i think only those buttons or tabs needed for a particular exercise should be present only
the repeat utterance needs to be closer to the exercise areas
last two exercises were maybe too fast for my level

16.2 User study 2: Individual results from Perception

	Group1			Perception exercises improvement %		
Id	Gender	L1	Hours homework	Duration	Lexical Stress	Vowels
1	M	Turkish	2.7	16	0	8
2	M	Turkish	4.0	23	-9	34
3	M	Spanish	0.9	23	8	41
4	M	Greek	2.2	8	9	17
5	F	Spanish	2.6	-53	0	58
6	M	Italian	9.1	8	34	17
7	F	Arabic	3.8	8	67	25
8	F	Chinese	4.3	0	25	0
9	F	Persian	0.7	8	0	8
10	M	English	1.0	-8	17	8
11	M	Greek	2.8	7	25	16
12	M	French	2.1	46	16	25
13	M	Persian	8.6	47	33	9
Tot	13	Average	3,4	10,2	17,3	20,4

	Group2			Perception exercises improvement %		
Id	Gender	L1	Hours homework	Duration	Lexical Stress	Vowels
14	M	Spanish	1.5	-23	16	8
15	F	Chinese	2.3	46	8	-8
16	M	German	3.3	8	-17	16
17	F	Persian	3.6	-8	17	9
18	F	English	2.0	47	0	16
19	M	Turkish	2.7	24	25	9
20	M	Russian	6.0	16	0	0
21	F	Chinese	5.4	23	58	-8
22	M	Urdu	1.0	0	-42	9
23	F	Persian	4.9	0	25	-16
24	F	English	4.5	62	9	-17
25	M	Chinese	7.2	0	50	25
Tot	12	Average	3,9	16,2	12,4	3,6

Table 21 Improvement from pre to post test for the perception exercises in group 1 and 2

Group3				Perception exercises improvement %		
Id	Gender	L1	Hours homework	Duration	Lexical Stress	Vowels
26	M	Spanish	0.0	8	0	9
27	M	English	0.0	-8	17	-42
28	F	Polish	0.0	16	0	25
29	M	Chinese	0.0	38	34	-17
30	M	Korean	0.0	30	17	17
31	M	English	0.0	38	0	-8
32	M	Vietnamese	0.0	8	42	17
33	F	Lithuanian	0.0	23	0	0
34	F	Korean	0.0	-69	17	9
35	M	Persian	0.0	0	-9	0
36	F	Russian	0.0	7	8	-8
37	F	Polish	0.0	-46	25	17
38	M	Ibo	0.0	-8	25	9
39	F	Russian	0.0	16	17	0
40	M	Spanish	0.0	8	17	0
41	M	Persian	0.0	16	0	9
Tot.	16	Average	0,0	4,8	13,1	2,3

Table 22 Improvement from pre to post test for the perception exercises in group 3 (control group)

16.3 User study 2: Individual results from Production

Group1				Production exercises improvement %		
Id	**Gender**	**L1**	**Hours homework**	**Duration**	**Lexical Stress**	**Insertion/ Deletion**
1	M	Turkish	2,7	-14	13	-8,5
2	M	Turkish	4	-28	0	8,5
3	M	Spanish	0,9	0	0	1
4	M	Greek	2,2	28	-13	8,5
5	F	Spanish	2,6	0	12	17
6	M	Italian	9,1	28	0	-25
7	F	Arabic	3,8	0	-25	8,5
8	F	Chinese	4,3	0	0	0
9	F	Persian	0,7	-29	0	-8,5
10	M	English	1	-43	-12	0
11	M	Greek	2,8	X	X	X
12	M	French	2,1	0	0	0,5
13	M	Persian	8,6	0	-13	7,5
Tot	13	Average	3,4	-4,83	-3,17	0,79

Group2				Production exercises improvement %		
Id	**Gender**	**L1**	**Hours homework**	**Duration**	**Lexical Stress**	**Insertion/ Deletion**
14	M	Spanish	1.5	28	-25	-41
15	F	Chinese	2.3	14	13	8
16	M	German	3.3	-15	-25	0
17	F	Persian	3.6	14	-13	-8,5
18	F	English	2.0	43	13	-9,5
19	M	Turkish	2.7	14	0	0
20	M	Russian	6.0	0	0	8,5
21	F	Chinese	5.4	0	0	-1,5
22	M	Urdu	1.0	-14	13	16
23	F	Persian	4.9	0	13	7,5
24	F	English	4.5	0	-13	0
25	M	Chinese	7.2	0	0	-16
Tot	12	Average	3,7	7,00	-2,00	-3,04

Table 23 Improvement from pre to post test for the production exercises in group 1 and 2

Group3				Production exercises improvement %		
Id	Gender	L1	Hours homework	Duration	Lexical Stress	Insertion/ Deletion
26	M	Spanish	0.0	-14	0	0
27	M	English	0.0	28	13	-16
28	F	Polish	0.0	-15	0	0
29	M	Chinese	0.0	0	0	-7
30	M	Korean	0.0	-14	0	8
31	M	English	0.0	14	13	0
32	M	Vietnamese	0.0	42	13	8
33	F	Lithuanian	0.0	-14	-25	-32
34	F	Korean	0.0	15	-13	17
35	M	Persian	0.0	29	0	23
36	F	Russian	0.0	28	-13	8
37	F	Polish	0.0	0	12	17
38	M	Ibo	0.0	-14	0	-17
39	F	Russian	0.0	14	0	0
40	M	Spanish	0.0	X	X	X
41	M	Persian	0.0	0	0	0
Tot	16	Average	0,0	6,60	0,00	0,63

Table 24 Improvement from pre to post test for the production exercises in group 3 (control group).

User 11 and user 40 had some errors in the log of the pre-test, and their scores have therefore been excluded from the results in the production part. User number 11 completed however all the tests and questionnaires, and his scores in the perception part, and comments are included.

16.4 User study 2: Homework data from group 1 and group 2:

Type 1-8 corresponds to:
1) Simicry: number of recorded sentences
2) Production Duration: number of recordings
3) Production Lexical stress: number of recordings
4) Production Insertion\deletion: number of recordings
5) Perception Vowel grid: number of nodes (questions, not exercises)
6) Perception Lexical stress: number of nodes (questions, not exercises)
7) Perception Minimal pair Vowel: number of nodes (questions, not exercises)
8) Perception Minimal pair Duration: number of nodes (questions)

Group 1	Simicry	Production			Perception			
Id\Type	1	2	3	4	5	6	7	8
1	724	225	174	58	116	78	117	126
2	316	153	193	82	193	76	238	213
3	103	212	0	0	53	45	21	52
4	428	167	157	102	142	60	97	167
5	196	270	0	49	174	79	119	122
6	1106	509	281	203	430	173	425	534
7	313	398	385	76	189	54	231	257
8	764	537	338	173	450	159	239	329
9	265	55	0	0	38	0	0	0
10	0	138	0	6	104	29	49	63
11	420	288	161	52	154	129	221	197
12	391	207	173	83	43	29	50	56
13	1179	349	335	133	837	143	88	127
Sum	6205	3508	2197	1017	2923	1054	1895	2243

Table 25 Amount of work done at home for participants in group 1.

Group 2	Simicry	Production			Perception			
Id\Type	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
14	144	11	1	1	136	64	181	216
15	139	78	158	37	282	67	94	122
16	419	95	0	48	151	0	94	107
17	0	140	6	2	150	101	168	121
18	866	16	17	69	1622	201	230	307
19	60	8	7	13	36	120	114	52
20	484	56	203	86	38	37	72	90
21	502	207	382	112	260	61	147	273
22	21	8	4	3	78	69	56	35
23	687	184	351	63	72	95	51	52
24	106	53	118	36	606	128	137	201
25	438	208	195	53	218	200	237	233
Sum	**3866**	**1064**	**1442**	**523**	**3649**	**1143**	**1581**	**1809**

Table 26 Amount of work done at home for participants in group 2.

16.5 User study 2: Replies from questionnaire

This is the complete set of comments users from group 1 and 2 had on the various parts of the program after one month of using Ville at home.

16.5.1 The perception part – Group 1

It would be easier if Ville explain more with written examples and may be also add translation..

I found the first exercise with 9 vowels quite difficult and not always clear the difference between A and O.

Very Good effort so that the student understand some basic pronunciation matters in Swedish language. Very useful...

Maybe it will be useful if English translator will be put on it

"Especially the exercises with vowels are rather hard. However, I consider vowels as the hardest part in Swedish, so I find these exercises very useful and teaching. Getting them false often annoys me, but this was normal for beginning.

Lexical stress part is very fun to work with."

Good explanations, I like this part.

As a Spanish speaker due to we do not have some vowels Swedish has is difficult to tell the different between them, not that difficult in the minimal par vowels. It should be more exercises to differentiate between ä å and a and the same with o and ö.

16.5.2 The perception part – Group 2

"it is designed well and very useful for me to improve my perception part, and I really made big progress, the result can say."

Some of the words in the "Vowels" part sounded very similar and it was very difficult to tell the difference between for example å and o sometimes. Also, a lot of the words were very quiet

and hard to hear, compared to the other exercises and sometimes there were cracks in the sound when the words were played. It was a little distressing that no matter how many times I tried I just didn't seem to be improving my score in the Vowels part.

It was difficult to differentiate between ä and e, å and o, u and ö. The exercises helped a bit, but I still get confused often.

"The part contains different aspects which could be difficult for people (as depends on the native language). the difficult part for me as for Russian speaking is Vowels duration. Even though it is kind of boring to repeat all the time the same words - I find it important to me to practice. Lexical stress part, on the other hand, is very obvious for me and thus useless for me.

As a general comment I find it useful if you will write a translation of the words you use in the card so to say. In this case there will be an additional motivation for me to do the exercise."

I think the "lexical stress" it is a little bit confusing with the line and semicircle this can be improved, i got confuse in how many syllabels has the word, maybe if you write the word in syllabals with the different stress option for that word, it is easier to find the stress in the word

I cannot tell the difference between o and u. but i think that is my problem :(and the stress part i hope that the word may show up. Then I can know which word it is and why it is like that.

16.5.3 Pronunciation part – Group 1

for me it was easy when the word has long vowel, the feed back was accurate.. But for short vowels even when I feel I I got it correct sometimes it give me wrong feedback..

It was quite difficult to understand where the mistake was since in many times I had the felling of being exactly on the same level of speech of Ville.

Very helpful as well to clear out pronunciation matters. More words would be good to be added in the database. Quite accurate most of the times and good feedback!

At times I could not hear any difference between my recording and Ville but was told that I was quite wrong.

"The duration part is hard, it gets even harder if the word is long and hard to pronounce. The other two parts are easier and more fun.

The feedback is very useful, though sometimes I think I say the words correct (when I listen to

the example) the feedback doesn't accept it. I always checked the feedback and tried to understand what I was doing wrong. It might not be 100% accurate but that could also be because of my microphone."

The feedback is too sensitive, sometimes even if I speak the word almost the same way as Ville did, the system will pop out red light. This is especially the case for the pronunciation duration part. (Said by Chinese…)

16.5.4 Pronunciation part – Group 2

"the study in the field helped me in the previous part, it is pretty useful and good designed in recording.

It would have been nice to get some feedback about how I did after each word, instead of relying on my own perception of how I sounded. Also, it was a bit tedious to have to press the record button each time. It would have been better if it recorded each word automatically, like in the test, whether the word was played first by the program or not.

"The same comment- if there will be more words with translation and, may be, devided by category (as it is not always the time to perform lets say 57 words in a row).

Additional comments which I already have written earlier as: ability to choose Ville to play words auto, another arrangement of the buttons... "

16.5.5 The Simicry part Group 1 (Who used shadowing)

The best part of Ville software. Really helpful and enjoyable! It is more difficult but far more productive...

"it is not very accurate for long sentences bécause sometimes i said ""bla bla bla"" and i obtained 0.9

but i think it is difficult to create an accurate software because the speaking rythm is fast"

"Ville talks so fast.. I felt that speed is propotional to the length of the sentence.. so for me it was very diffecult to cope with him when sentence is long and fast..

It would be even easier if we donot record in the same time!"

Was useful to learn how to speak more like a Swedish speaker but had some real difficulties

in trying to get the phrases right with the right lengths. At times it seemed like no matter what I did the result did not change, and then at others I would say something completely wrong and still get it "right".

"This part is very useful. Especially the packages show examples from everyday life. It is great to hear how Swedish is actually spoken by native speakers. Also trying to mimic them constantly made me memorize sounds and words, which is good.

An example package to add could be a simple lecture with a teacher and couple of students in discussion."

this part should be reviewed carefully because you can get a good punctuacion even if you do not pronounce correctly or even if you do not say anything and there is some noise in the environment you can get a punctuation

Same problem with the Pronunciation Duration part, where the system is too sensitive. In this part, the shorter sentences are extremely hard to pass, even if I spoke quite all right. And the longer sentences are easy to pass, even if I do not complete the sentence.

16.5.6 The Simicry part Group 2 (Who used 'say after')

really difficult, the ville is speaking fast to me, I know it is normal speed, but to me it seems to be a little fast, but I believe that will be very useful when my level is higher than now.

The simicry part helped me a lot I think, but I couldn't finish some of the packages because the program would return an error message when recording and it wouldn't go past that sentence in the package. This happened with a few (2 or 3) packages. After adjusting the recording settings (which I think should have more documentation), it was a lot better, but sometimes the program didn't detect silence and just kept recording. Some of the sentences were very long and complex and very difficult for a beginner, and it would be nice if those were split. Many times the program didn't seem to detect the words correctly.

It was interesting to learn various phrases that were useful when I speak Swedish in Sweden.

Very fun and useful. I put the major comments before (1st questionary). Sometimes it listening very long (like minute or so) - that makes me crazy :) But overall very good

I really like that i have feedback of pronunciation just when i finish my sentence

16.5.7 The talking head (ECA) Group 1

it is useful for concentration

it makes things a little bit more funny and less of a serious environment of learning.. it makes you more relaxed.

It is quite helpful for me. There could be some improveménts in the detail of his mouth. Even so, it helped to understand differences in pronunciation in i and y vowels, as for instance when comparing the words byta and bita (because my ears were hearing the same sound for both vowels).

At times the talking head was useful to see the mouth shapes and helped me to be more accurate with what I said. In some cases it helped me to say correctly words/phrases that I could not say correctly.

For pronunciation part It really help me specially on difference between: Y and I; ö and å

Understandable and motivating. Explains everything very detailed.

very useful, it provides a "friendly environment" and at least to me it was useful to distinguish vowels easily

16.5.8 The talking head (ECA) Group 2

I like the talking head, sometimes I can learn how to speak 'y' according to the talking head, the shape of month.

Some words are hard to pronounce or hard differenciate. The talking head is useful in that cases.

I found that I could learn a lot by watching while I listened. It was very strange when I accidentally rotated the head back and I could see the inside of the eyeballs and the tongue through the neck. Maybe that rotation feature should be changed a bit or removed.

It helped me to differentiate y and i better.

I tried to tell the difference among vocals using the talking head. i think it is useful here. However I makes this program interesting.

Brings some entertainment. Very good. But it easily could be rotated by user by mouse. May-

be you should fix it..

16.5.9 More feedback - Group 1 (who had PEDs)

All of them

It depend on the parts of the test but it would be good if he can tell why answer was wrong, can give tips and/or answers!

"long term feedback would be good. also some hints after a lot of unsuccessful attempts."

Some more long term feedback would be good to see generally what things I needed to improve. Adaptation of the program to focus on problem areas would also be useful.

Short term feedback is nice; but somehow I like long term feedback too, if it possible to continue working with Ville.

"More encouragement when doing good, suggestions for corrections when I'm doing the same mistake many times.

How about a 'swedo-meter' that increases with every consequent correct answer and goes back to zero with every wrong answer. Just a fun factor, telling the user how much he resembles a native Swedish guy or girl."

long term feedback showing my process and if I am doing good or bad when compared to other participants.

"yes, it could be good idea a long term feedback in order to see your progress and which area you have to work more on.

Maybe it is also useful if ville could detect your personal most common mistakes and then we can create a personal folder with exercises in order to improve them.

i would have liked to have the program you sent us in the beginning, i think is very useful and you can see your progress.

"hardness or difficulty of the test could be changed internally, for example if a person is not doing well, ville should start to give easier exercises when person begins to do less mistakes, ville starts to increase difficulty. For those who are O.K at swedish can improve more if ville gives difficult exercises than easy ones. So ville acts depended on user swedish level."

immediate feedback would be quite a good feature.

16.5.10 More feedback– Group 2 (who worked without PEDs)

long term feedback, telling me when it was not so good
when I was doing the test in perception test, once I made the same mistakes in one question, ville would say wrong again and again, it makes people a little boring, I'd like to hear some tip or encouragement or a smile.
"short term feedbacks would be useful, specillay in case of making mistakes that could be so helpful to know that one is making mistake and far from the acceptale level so that can try harder and of course encouragements could be nice to receive that one would be sure that is doing well and can know that the way of (for example my pronunciation) is generally good, so i can work and improve it."
I would have liked feedback on speaking individual words, and graphs of my progress in each of the exercises so I could determine what I needed more work on.
Less discouragement when you get an answer wrong. Maybe a chart to show the progress in various tests over the weeks.
To a what degree I improved myself and maybe some statistics
In the flash cards just record my voice and hear by myself it is not that useful i would rather to be grade in every word.
I prefer to have long term feedback on my progress to see how I can have improvement respecting the time spend. Like simicry part, it can remind me to pay more attention to which point. For example, if i record it once and did not pass it. It will be nice if he can tell me which word that I pronunced different.

16.6 Number of users per country using Ville-SWELL 2011-03-30

Africa	
Burundi	1
Congo	1
Ethiopia	28
Ghana	3
IvoryCoast	2
Marocco	1
Nigeria	8
SouthAfrica	3
Sudan	2
Tunisia	3
Uganda	2
NorthAmerica	
Canada	1
USA	18
CentralAmerica	
Barbados	1
Haiti	1
Jamaica	1
Mexico	14
DominicanRepublic	1
Nicaragua	1
SouthAmerica	
Argentina	7
Bolivia	1
Brazil	6
Chile	6
Colombia	10
CostaRica	2
Guatemala	1
Venezuela	3
Oceania	
Australia	10
NewZealand	1

Asia	
Afghanstan	1
Armenia	5
Azerbaijan	2
Bangladesh	53
China	233
Egypt	8
Georgia	4
HongKong	4
India	63
Indonesia	8
Iran	155
Iraq	3
Israel	3
Japan	11
Jordan	6
Kazakhstan	1
Korea	1
Kyrgyzstan	4
Lebanon	4
Malaysia	1
Mongolia	2
Nepal	15
Pakistan	88
Palestine	3
Singapore	38
SriLanka	6
Syria	10
Taiwan	6
Tajikistan	1
Thailand	22
Turkey	54
Uzbekistan	1
Vietnam	8
Unknown	209

Europe	
Andorra	1
Austria	15
Belarus	8
Belgium	6
Bosnia	1
Bulgaria	2
Croatia	5
CzechRepublic	6
Estonia	5
Finland	4
France	71
Germany	76
Greece	28
Hungary	4
Iceland	3
Ireland	1
Italy	30
Kosovo	2
Latvia	9
Lithuania	4
Luxembourg	1
Macedonia	11
Moldova	2
Montenegro	2
Netherlands	12
Norway	19
Poland	14
Portugal	1
Romania	7
Russia	10
Serbia	22
Slovakia	1
Spain	35
Sweden	43
Switzerland	4
UK	8
Ukraine	11