Field Evaluation of a Single-Word Pronunciation Training System

Niels Ole Bernsen, Thomas K. Hansen, Svend Kiilerich and Torben Kruchov Madsen

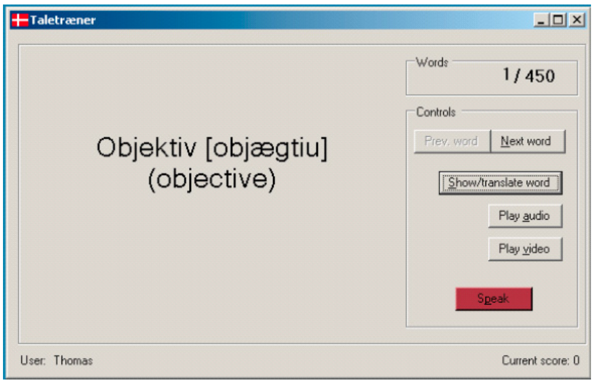
Natural Interactive Systems Laboratory
University of Southern Denmark
Campusvej 55, 5230 Odense M, Denmark
{nob, thomas, kiil, kruchov}@nis.sdu.dk

Abstract

Many learning tasks require substantial skills training. Ideally, the student might benefit the most from having a human expert – a teacher or trainer – at hand throughout, but human expertise remains a scarce resource. The second-best solution could be to do skills training with a computer-based self-training system. This vision of the computer as tutor currently motivates increasing efforts world-wide, in all manner of fields, including that of computer-assisted language learning, or CALL. But, as pointed out by Hincks [2003], along with the growth of the CALL area comes a growing need for empirical evidence that CALL systems have a beneficial effect. This point is reiterated by Chapelle [2002] who defines the goal for Computer Assisted Second Language Research as the gathering of evidence for the effect of CALL and instructional design. This paper presents results of a field test of our pronunciation training system which enables immigrants and others to self-train their pronunciation skills of single Danish words.

1. The Pronunciation Trainer

The prototype of the Danish Pronunciation Trainer (DPT) combines a graphical user interface and an Automatic Speech Recogniser. The main window is shown in Figure 1.1.


Taletræner

Words 1 / 450

Objektiv [objægtiu]
(objective)

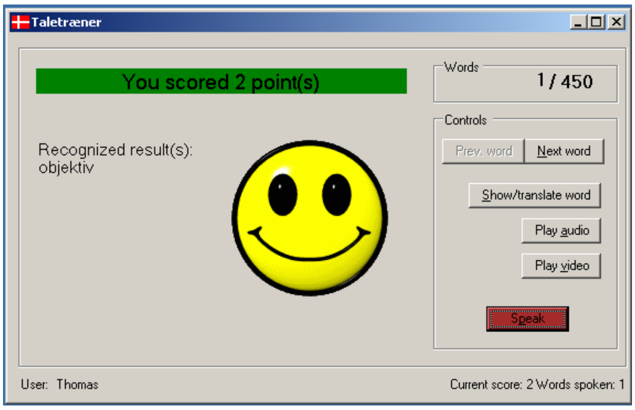
User: Thomas

Current score: 0

Figure 1.1. Main window of the Pronunciation Trainer.

Our aim has been to create an interface which is as simple as possible and intuitively usable by the learner. The window displays the Danish training word, sometimes a simplified 'phonetic' transcription using regular orthography, and a translation into English. Ordinary users are not proficient in standard phonetic representation, but we have found that they are helped by simplified phonetic transcription when training to pronounce Danish words whose pronunciation is irregular. This help is often required, Danish being phonetically highly irregular. What the simplified phonetic transcription does is help the learner avoid endlessly following wrong pronunciation intuitions without getting the pronunciation right. The English translation, we have found, is appreciated because (i) most Danish trainees speak English already and (ii) appreciate knowing the meaning of the word they train to pronounce. Given (i), the interface language of the main window is English.

Having clicked the 'Next word' button and (i) read the written word, its translation and (possibly) its simplified phonetic transcription, the learner may choose to pronounce the word by clicking the 'Speak' button. Alternatively, the learner may (ii) click the 'Play audio' button to listen to a native Danish speaker pronouncing the word, or (iii) both hear and see the word being pronounced by a native Danish speaker on video by clicking the 'Play video' button. The training sound files used in (ii) and (iii) are identical. Having pronounced the word, the learner receives feedback on pronunciation quality. A scoring system was devised for providing immediate feedback to the learner, ranging from 0 to 2, 2 being the highest and 0 the lowest. The learner is given both the numerical score and a graphical indication of success or failure, as illustrated in Figure 1.2.


Taletræner

You scored 2 point(s)

Recognized result(s):
objektiv

Words 1 / 450

User: Thomas

Current score: 2 Words spoken: 1

Figure 1.2. Achieving the maximum pronunciation score.

The scoring system works as follows. Typically, the speech recogniser returns an ordered multi-word n-best list of words in its vocabulary which best match the acoustic signal produced by the learner, with the best acoustic match on top. If the word to be pronounced is on top of the n-best list, the learner receives the top score and only the top scoring word is presented in the feedback to the learner as shown in Figure 1.2. If the word to be pronounced is not on top of the n-best list but is somewhere else in the list, the learner receives the score 1 and the entire n-best list is shown in the feedback. The feedback informs the learner with which other words in the recogniser's vocabulary the pronounced word has been, and hence could be, confused. If the word to be pronounced is not in the n-best list, the learner receives as feedback the score 0 and a Smiley in tears.

Having received the scoring feedback, the learner may address the next word by pushing the 'Next word' button. Counters in the bottom right-hand corner keep track of the number of words which have been pronounced in a particular training session and the learner's accumulating score. A total of 450 of the most common Danish words were chosen for the lexicon. The words were chosen depending on phonetic richness, relevance of practical everyday use and phonotactic combinations. The counter in the top right-hand corner shows the id of the current word in the system's lexicon. This counter is interactive, enabling the user to type in any number between 1 and 450.

The system creates a logfile for the user during each session, which records the number of words spoken, the number of times a word has been (i) read-only, (ii) read and listened to, or (iii) read, listened to and viewed on video, and the scores achieved. In order to keep track of the training logfiles, the system has a separate login window in which the user enters a login. In addition, all logfiles are date and time-stamped.

The logfiles enable us to monitor and assess the progress of each student. The logfiles are sent to NISLab by the system responsible at the training site.

The full description of the pronunciation trainer can be found in the pdf version of the User Manual (in Danish) at *http://www.nis.sdu.dk/projects/CAPT/DUTManual.pdf*

2. Current State of the Art

Presently there are numerous applications available for aiding pronunciation training in a foreign language. These range from research applications under development at different laboratories to full-fledged commercial systems. Although these applications all target the same aspect of foreign language learning, the approach is somewhat different from application to application. Generally speaking, applications can be divided into two categories: (1) those that employ automatic speech recognition (ASR) and (2) those that do not.

Some non-ASR based systems provide information on how speech sounds should be produced or how the articulators should be positioned during pronunciation of certain segments, but only target the learner's perception of sound without offering the possibility of actively pronouncing them and simultaneously receiving feedback on how the learner is doing. Other systems include record and playback features, which intuitively seems to be a step up, but still requires the learner to independently compare the target sound with the produced sound. Some applications attempt to circumvent this problem by allowing the student to record the utterance and send a file to a teacher or upload it for later analysis. Unfortunately, this nullifies the idea of immediate feedback.

Slightly different but still non-ASR based systems, such as WinPitchLTL, display pitch or intonation contours of the students' utterance and allow for comparison with a model utterance. However, this solution, on a par with the record and playback option, requires that the student be able to interpret the contours independently or with the assistance of a teacher, hence almost requiring the learner to possess some knowledge regarding the reading of pitch contours, or at least to spend time deciphering how to read the curves.

The perhaps greatest positive impact on pronunciation training came with the integration of ASR in CALL applications. ASR has been used in CALL applications since the beginning of the 1990s. With its introduction came a promising trend of truly making the computer an instrument with which the learner could speak, interact and receive feedback from, thereby nullifying the need for a teacher to provide constant supervision. The introduction of ASR in CALL systems has enabled the learner to attempt to pronounce words or sentences and have the quality of the pronunciation evaluated instantly.

Some programs employ ASR without necessarily using it for providing the kind of detailed feedback described below. Rather, the user is simply presented with a word or a sentence, or a selection of sentences. Once the pronunciation of the target segment or sentence comes close enough to native-speaker quality as determined by the program, the learner is presented with the next screen and the process is repeated.

HUGO, as described in Tsubota [2004], is an English pronunciation trainer developed for Japanese students of English at the Academic Center for Computing and Media studies at Kyoto University in Japan and makes use of ASR. HUGO focuses specifically on pronunciation error detection, categorisation and correction. While using the program, the learner is asked to do role-play conversation as well as single-word pronunciation practice. Once the learner has gone through a practice session, s/he is presented with an overall estimation of intelligibility from the point of view of a native English speaker, ranging from *very hard to understand* to *perfectly understandable*. The learner is also given an overview of the most frequently occurring segmental errors, what s/he should focus on in the corrective phase and then asked to repeat all the words that were pronounced wrongly during the role-play

The Swedish Virtual Language Tutor, Ville, Granström [2004] which is being developed By Preben Wik at the Royal Institute of Technology in Stockholm, is an application intended to be used by foreigners wanting to learn Swedish and also integrates ASR. Ville, in addition to other functionality, includes a single-word pronunciation function where the learner is prompted to pronounce a specific word which is then analysed. The learner can listen to a target pronunciation by a teacher, record his or her own pronunciation and receive verbal feedback on the quantitative level, such as "I think your *a* is a little too long." The program also includes a feature which allows the learner to time-align pronunciation with that of the target pronunciation and then listen to the compressed or expanded version of his/her own attempt.

The ISLE 1.4 project [Menzel et. al. 1999] is another example of a language tutor attempting pronunciation correction at the segmental level. The learner is asked to speak a pre-selected sentence and have it evaluated. All the committed errors are then highlighted in colour-coding and the learner is asked to practice these specific words. Unfortunately, the program turned out to yield too much erroneous feedback. The problem of erroneous feedback described in the ISLE 1.4 project reflects the, perhaps, most severe challenge to using ASR currently found in the

CALL area. Namely, how to provide perfect and precise feedback to a learner using less than perfect technology?

DPT in its current form was developed as a prototype with the intention of examining how an ASR system would cope with foreigners attempting to learn Danish. The initial aim was to create a rather modest application with the ability to provide the learner with sufficient and relevant material to acquire vocabulary items and learn their pronunciation at the same time. The idea was further to provide the learner with stimuli focusing on different areas, such that the learner's ears, eyes and mouth became involved in the process. Hence, both audio- and video files as well as written words were provided. The option of having the word translated further provides the learner with a chance of increasing his or her vocabulary. At the same time, the DPT points out the major pitfalls between orthography and pronunciation by providing an adapted phonetic transcription using the regular alphabet. The learner is pointed to the fact that sometimes there are severe differences in the way a word is written and the way it is pronounced.

A competitive element was added to the application by displaying the number of pronounced words and the accumulated score for the current session, hence, prompting the learner to *do better than last time*, or simply to maintain a high score (Figure 1.2).

The underlying database records everything which is done by the learner, thereby providing an important tool for evaluation, either by the learners themselves or by a teacher using the application as an aid. The logfiles display when and at what time the user started training; which words were pronounced; how many times did the learner attempt to pronounce a specific word; how many times did the learner read, read and hear, or read, hear and see the word pronounced before attempting to pronounce it and, finally, how many pronunciation attempts were made and what were the individual scores. Given the described functionality of the DPT, there are still certain shortcomings, in particular the current lack of specific, corrective segmental feedback, but a strong foundation has been laid for further development of the prototype.

3. Field Trials

Following in-house user testing of the Danish pronunciation trainer during 2004 with, among others, Chinese and Finnish students, the system was installed at 9 language schools and similar institutions across the country. Installation began in November 2004 and was completed by end of January 2005. The training data to be analysed in this paper were produced between November 2004/January 2005 and end of April 2005, at which point data collection was "frozen" in order to assess student progress with DPT.

An appropriate metaphor for the DPT's role at the test sites in this first field test is perhaps that of an unexpectedly demanding visitor: someone must have invited the DPT because otherwise it wouldn't be there. Given the novelty of the technology, even that someone might not know what to do with it nor be able to find others at the site who could, or had the time to, install DPT in close collaboration with us, motivate students and others to use it systematically, send us the training logfiles, etc. In brief, the result was that only four of the training sites managed to send us logfiles which satisfied our criteria for being able to assess student progress by means of them (Section 4). The other five training sites either did not manage to send us any logfiles or sent us files which did not satisfy our criteria. It was practically impossible for us to have systematic access to the students during the field trial.

4. Evaluation Procedure

We received 821 training logfiles from six training sites. Table 4.1 shows a logfile excerpt. Column 1 shows the word id. Column 2 shows a typed version of the word to be pronounced. Words which have highly irregular pronunciation are supplemented with a simplified phonetic representation. Columns 3 through 5 show how many times the subject has looked at the typed word, listened to it, and perceived its audio-visual pronunciation on video, respectively, before pronouncing it. Column 6 shows the subject's pronunciation score. Note the gap in the score for "Pile", meaning that the subject did not pronounce this word.

ID	Word	Seen	Audio	Video	Score
1	Objektiv [objægtiu]	1	3	1	2
2	Fortælle	1	1	1	2
3	Begynde [begøne]	1	1	1	2
4	Føtex [føtæks]	1	1	1	0
5	Arbejde [Abaide]	1	2	1	2
6	Betyde	1	1	2	0
7	Lunge [långe]	1	2	1	2
8	Betale	1	3	1	0
9	Forsøge	1	2	1	0
10	Pile	0	0	0	
11	Fortsætte	1	3	2	0

Table 4.1. Training logfile excerpt.

The User manual suggests that students train the DPT's 450 words in consecutive series in order to make sure that they get through all the words several times with some suitable time interval in-between, enabling both them and us to measure their progress. Many students did not follow this procedure, however. Instead, for instance, student Sn might start by training words 205-267 followed by words 55-96, never repeating the same word sequence but, at most, repeating, e.g., words 228-267 when the student trained with words 228-301. Some students trained very little, or they did substantial training but only for less than a week; some training logfiles include scoring gaps, such as Row 10 in Table 4.1, and some logfiles include just a few words which have been pronounced many times in succession.

Since progress evaluation depends on comparison of *repeated* test sequences produced by students who have trained for a minimum period of time and who have trained at least a minimum number of words in total, we specified a set of criteria for identifying those students whose logfile corpora conformed to the criteria. The result was that 22 out of the 88 students who had used DPT at one of the 9 training sites did produce results suitable for progress evaluation. The criteria were as follows.

We define a *test cycle* as a series of training sessions in which the student has trained all 450 words once. We then (1) find all students who have done at least one test cycle and has started on the second cycle; (2) find a sequence of at least 25 words which have been pronounced at least twice, preferably with a test cycle interval of 6 weeks in-between; (3) clean up the logfiles by removing repeated pronunciations of the same word, noting unpronounced words in the sequence to avoid flawed statistics, etc.; (4) compute the average score in % for the sequence when pronounced the first time, the nth time and the last time in the data. This is done by taking the actual score in % of the maximum score, i.e., ((number or words in the sequence) x 2); (5) plot the student's progress in a joint graph for all students at this particular test site; and (6) repeat for the next test site. In the graph, each student is thus plotted at most three times, with a first score, and intermediate score, and a final score. In a few cases during data analysis, we compromised condition (1) by including students who had repeated a particular test sequence but who had not completed a full test cycle,

Whilst the *primary evaluation procedure* just described takes training time into account, it does not fully take into account the *training amount* put in by each student in order to arrive at the final evaluation score. For instance, students S1 and S2 might both have improved from a start average score of 45% to a final score of 60%, but S1 might have done this with a relatively light training load over 5 weeks whereas S2 spent, say, four times as much training over 15 weeks to achieve the same progress. We therefore defined the notion of a *full training curriculum* as that of training 4500 words and measured, for each student, the percentage of the full curriculum they had done between the first and the last measured average score.

Finally, we did three *control measurements* (*secondary evaluations*) of each student's primary performance score as described above. These were done at test fragments from the first training day, sometime in the middle of the student's training, and at the final training day. These fragments were chosen such as to all be different from those in the primary performance evaluation and, clearly, they would typically concern the pronunciation of words other than the sequences analysed in the primary performance evaluation. These three additional measurements per student gave us useful control points with respect to the validity of the primary evaluation scores. The combination of primary and secondary evaluation means that the performance of each student was measured at least five times in the data.

5. Test Conditions

This section describes what we hypothesise to be a genuine discovery made during student progress evaluation. We have not found similar observations elsewhere.

Before proceeding, let us establish two points. Early in DPT development, we base-lined the system with native Danish speakers who scored between 78% and 94% on average. Thus, we may define three broad *scoring bands* as follows. A *high score* is >70%, a *middle score* is 50-70%, and a *low score* is <50%.

When analysing student progress, we sometimes found an unexplained large *drop* in performance. Such drops may appear at two different locations in the scoring space. In the first case, the non-native Danish speaking student has a high (<70%) start average score and then drops 5-15% at the next evaluation point. Such a *high-drop* does not mean much. The student is close to being able to pronounce Danish fluently anyway and native Danish speakers vary in their scorings as well. However, it was surprising to find *middle-to-low* drops and even *high-to-low* drops in the data. We believe that our data firmly establishes the basic point that *you don't score high at random*. If you score high once, you are close to pronouncing words like the native Danes do, period. So, why did those students drop like that?

Let us define another term, that of *test condition*. A test condition is the priming you receive before pronouncing a particular word through either (i) reading the word, (ii) reading and listening to the word, (iii) reading and (listening and seeing) the word pronounced, or (iv) reading and listening and (listening and seeing) the word being pronounced, cf. Table 4.1. What we found was that all those drops could be explained by the fact that the student, from test T1 to test Tn had changed the test conditions *upwards* from (iv) to (iii), (ii) or (i). Our data does not support the full story yet. However, what we have found is that, in addition to reading the word and its primitive phonetic transcription, if any - viewing the video of a native Danish pronunciation is crucial to the average scoring of most students. It is only the initially high scoring students who can dispense with video priming, and they actually do dispense with it. The initially middle scoring and low scoring students all use the video before pronouncing each word. And then, at some point at which they might feel that they are doing well enough primed by the video, they stop using video priming and their performance drops significantly as described above. The data shows that the students who do not have a high start score are perfectly aware of the importance of the video to their pronunciation quality, so they use it consistently until the point at which some of them, equally consistently, stop using the video.

The implications of this finding for pronunciation training would seem rather important. Let us just mention two points here in addition to calling upon others in the field to test and refine the finding. Firstly, a student pronunciation score is next to worthless unless we know the test conditions, and so is student score comparison. A student who scores 47% without video priming could easily be better at pronouncing language Ln than a student scoring 57% using video priming. Secondly, it seems likely that the appropriate test condition for evaluating a student's real ability to pronounce words in language Ln must be one without video priming, such as one in which only the written word is being presented or, alternatively, one in which only the spoken word is being presented as priming factor.

6. Results

In this section, we illustrate the findings made in the 22-student corpus and present the general progress results found. We discuss the dependencies of the results upon student training effort, student differences and differences in test conditions.

6.1. Training Effort

Let us first look at the training amount put in by the students. Assuming a full training curriculum of 10x450 words over 10 weeks, the 22 evaluated students performed 19.1% of the full curriculum on average between their first and last measured test score (primary evaluation), ranging from 2% to 43% of the curriculum. We had hoped to see at least a handful of students complete the full curriculum, which would have provided invaluable information on the effects on pronunciation progress of having trained with the system as recommended. Unfortunately, we had no direct access to the students and had no way of providing them with the additional motivation needed to complete the full curriculum. The discussion below of the evidence of learning progress which we actually found in the data is limited by the lack of complete training data just described.

6.2. Some Student Progress Cases

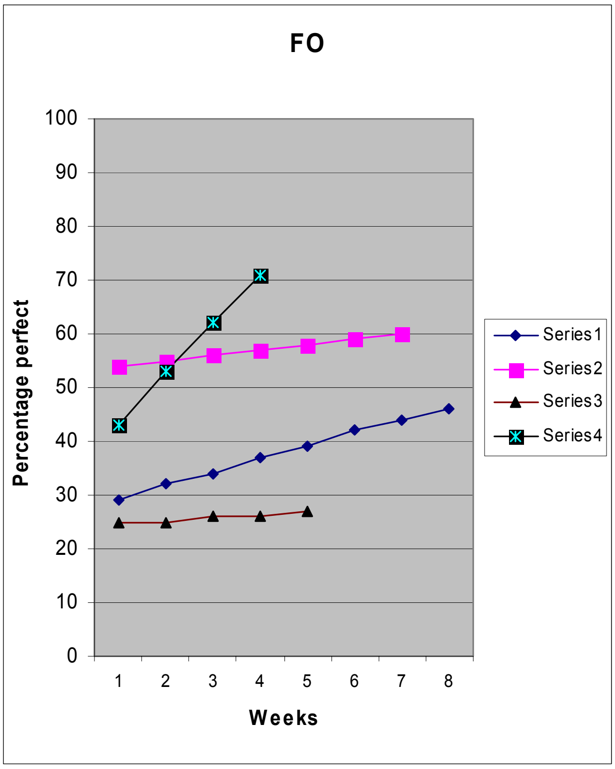


Figure 6.1. Student progress graph.

Figure 6.1 shows the progress made by four students at one test site according to the primary evaluation metrics. Note that, at most, the start and end points and, possibly, a middle point on each curve are data points.

The students (Sn) trained for 7 (S1), 6 (S2), 4 (S3), and 3 (S4) weeks, respectively, putting in a training effort of 29%, 21%, 2% and 4%, respectively. Given that, *S4's* progress is remarkable. However, there is a relatively large gap in S4's final test logfile. We cannot exclude that S4 deliberately omitted the most difficult words in that test in order to get a high average score. S4 used video priming throughout. *S3's* modest 2% progress might be attributed to S3's small training effort of 2% of the curriculum. However, S3's progress is somewhat larger than that when we add that S3 used video in the first test but only used a small amount of video priming in the final test. *S1*, using video priming throughout, makes a solid 17% progress over 7 weeks based on 29% of the full training curriculum. Compared to S1, *S2's* progress of only 6% based on 21% of the curriculum might seem modest indeed. However, S2 used video priming in the first test but only used typed text and audio priming in the final test, making the test harder than it would have been otherwise.

S1, S2, S3 and S4 illustrate the 18 *standard cases* found in the data. These students all make larger or smaller progress under similar, although not always identical, training conditions. Otherwise, the range of these cases is rather broad. There is a case of high-drop from 76% to 70% (Section 5). There is nice progress from 63% to 88% with no video priming and 9% of full-curriculum training. There is substantial training, such as 38% and 43% of curriculum, resulting in progress from 14% to 52% and 38% to 57%, respectively. In both cases, video priming was partly used in the start test but not in the final test. The case which limits this group of standard cases is a student who puts in 39% of full curriculum training between the start test and the final test in the primary evaluation and who makes no progress at all but stays at 54% average score. In fact, this student put in 60% of total curriculum training overall. However, the secondary evaluation did not show more than a few % progress even if it spanned far more training effort. This student simply is a very slow learner, if able to learn to pronounce Danish at all.

Common to the four *non-standard cases* is a strong drop in performance from start test to final test. M1 dropped from 85% to 49%, M6 from 56% to 36%, M7 from 89% to 73%, and M9 from 77% to 42%. However, M1 is the only real anomaly here, as we shall see. M1's test condition was typed text-only, so it seems clear that M1 is fully able to pronounce Danish words. To explain M1's drop, we only see hypotheses, such as insincerity in the final test or even that someone else logged in for M1. M6 seems to be a clear case of priming with video in the first test and testing without video priming in the final test. The same applies to M9. M7, finally, is a clear case of high-drop which might also happen to native Danish speakers.

6.3. Progress in General

Given the uncertainty inherent to our progress results and due to the unexpected and quite strong effect of training with and without video priming, respectively, the 18 standard cases would seem the best basis on which to form a general picture of student progress. From these 18 cases, we subtract the two cases of students who perform above 70% in the tests. These students probably performed maximally already in the first test and are clearly capable of pronouncing Danish words. On average, the 16 remaining students made an average progress of 16.4%. As already pointed out, progress was achieved with rather different training effort in each case, so we need to factor in the training effort between first and final test for each student. When we do that, we find that the effort needed for *10% progress* in pronunciation skills amount to *12%* of the full training curriculum. Moreover, except for the student who made no progress at all (Section 6.2), none of the 16 students made their final test under *easier* test conditions than those used in the first test, and some made their final test under more difficult conditions, notably without using video priming. So, the cost of 10% progress in terms of training effort stated above is a *maximum* figure.

We consider this result a promising one even though we are talking about the *first* 16.4% progress on average for the 16 students. Clearly, we argue, progress in Danish pronunciation is not a *linear* phenomenon from the first % progress until full mastery of Danish pronunciation. If it were, our figures suggest that the main body of students would be proficient Danish speakers before having completed the full curriculum, at least if they were to use video priming in their final test. Experience shows that the pronunciation performance of most adults who learn a second language tends to level out at some point at which there is still a recognisable accent.

At the other end of the spectrum is the sceptical view that, following 10-20% absolute progress, such as from 30% to 45% or from 50% to 65%, students tend to stop progressing. We believe that our data contradicts this view, at least, because we have already seen a number of students making 30% progress or more with a relatively modest training effort. Our data, in fact, also contradicts another scepticist view, i.e., that fast progress is limited to those students who perform poorly in their first test.

The truth probably lies somewhere in-between those two extremes. To this uncertainty, which is quite considerable, we need to add another, i.e., the fact that a high score obtained with video priming may only translate into a mediocre score without video priming. Unfortunately, our data does not support the forming of any idea of the magnitude of this "translation factor".

6.4. Student and Other Differences

At the time of writing, we have no access to data on (i) the individual students, their first language, other languages spoken, age, education, time spent in Denmark prior to training with the DPT, motivation, etc., and (ii) the physical, social and instructional conditions under which their training took place at the training sites. It is probably due to factors, such as those, that we cannot find any clear correlations between individual progress in pronunciation and training effort. For instance, we found that the largest individual progress of 43%, from 0% to 43%, was achieved with a training effort of only 15% of the full curriculum. The five students who made +20% progress in absolute % terms started from 0, 14, 43, 24 and 63% in their first test, one of them having the word gap noted in Section 6.2. None of this is surprising in the least. Just consider the difference between, say, a highly educated young German student with flair for languages and mastery of several, and a elderly Somali with no education and little sense of language learning.

How representative is a 22-student corpus given the fact that 88 students actually began to use the DPT? The assumption seems fair that the students who actually continued to use the DPT so as to provide us with usable data are among the most motivated to use the DPT. We don't know the extent to which motivation implies or presupposes ability but a connection would seem likely.

6.5. Secondary Evaluation

As described in Section 4, we did three control measurements per student in addition to the primary evaluation measurements. The control measurements tend to cover longer training spans and more training effort than the primary evaluation. In brief, the control tests showed (i) the *same* results (progress) patterns as the primary evaluation and (ii) that students tended to make *larger progress* in the control tests if these cover more training effort than the primary evaluation.

7. Conclusions

In this paper, we have presented main results of the field test of the Danish Pronunciation Trainer, DPT. We analysed the progress made by 22 students from four training sites and found a consistent pattern of progress which, arguably augurs well for using speech recognition technology for self-training in second-language pronunciation. This is despite the facts that (i) no student completed the recommended training curriculum and (ii) we have not been able to correlate student progress with data on the individual students as well as on their physical, social and instructional training environments.

In addition to the general progress trend reported, we unexpectedly found a consistent correlation between student performance and their test conditions in terms of the priming they used prior to pronouncing each word. There is a strong correlation between a student's performance in a test and whether or not the student had used video prompts before pronouncing the test words. Performance is clearly higher when using video prompts than when using just the written word or audio as a prompt. This finding, if confirmed, carries important implications for the future use of pronunciation self-training technology. A student's pronunciation score, or progress rate, must be considered relative to the test condition used by the student for achieving that score or progress rate. And "real" pronunciation proficiency probably cannot be measured on the basis of video priming which makes the pronunciation task too easy.

8. Acknowledgements

We gratefully acknowledge the support for the research reported by the Danish Ministry of Refugees and Immigrants.

9. References

Chapelle, C.: Computer Applications in Second Language Acquisition, Foundations for Teaching, Testing and Research. Cambridge University Press, 2002.

Granström, B.: Towards a Virtual Language Tutor. In NLP and Speech Technologies in Advanced Language Learning Systems. Proceedings of InSTIL/ICALL2004 Symposium on Computer Assisted Language Learning, 2004, 1-8.

Hincks, R.: Speech Technologies for Pronunciation Feedback and Evaluation. ReCALL 15 (1), Cambridge University Press, 2003, 3-20.

Menzel, W. Atwell. E. Herron, D. Howarth, P. Morton, R. Wick, H.: Pronunciation Training: Requirements and Solutions. ISLE deliverable 1.4. Available online at https://nats-www.informatik.uni-hamburg.de/~isle/public/D14/D14.pdf 1999.

Tsubota, Y., Dantsuji, M., Kawahara, T.: Practical Use of Autonomous English Pronunciation Learning. In NLP and Speech Technologies in Advanced Language Learning Systems. Proceedings of InSTIL/ICALL2004 Symposium on Computer Assisted Language Learning, 2004, 139-143.