

Language resources and CALL applications: speech data and speech technology in the DISCO project

Helmer Strik^a, Jozef Colpaert^b, Joost van Doremalen^a, Catia Cucchiarini^a

^a CLST, Department of Linguistics, Radboud University, Nijmegen, The Netherlands

^b Linguapolis, Institute for Education and Information Sciences, University of Antwerp, Antwerp, Belgium

E-mail: H.Strik | J.vanDoremalen | C.Cucchiarini@let.ru.nl; Jozef.Colpaert@ua.ac.be

Abstract

The current paper deals with the relation between language resources and Computer Assisted Language Learning (CALL) systems: language resources are essential in the development of CALL applications, during the development of the system resources are created, and finally the CALL system itself can be used to generate additional resources that are useful for research and development of new (CALL) systems.

We focus on the system developed in the project DISCO (Development and Integration of Speech technology into COurseware for language learning): we describe the language resources employed for developing the DISCO system and present the DISCO system paying attention to the design, the automatic speech recognition modules, and the resources produced within the project. Finally, we discuss how additional language resources can be generated through the DISCO system.

1. Introduction

In the last few years the interest in applying Automatic Speech Recognition (ASR) technology to second language (L2) learning has been growing considerably (Eskenazi, 2009). The addition of ASR technology to Computer Assisted Language Learning (CALL) systems makes it possible to assess oral skills in a second language and to provide corrective feedback automatically. The latter feature appears particularly appealing, since research has shown that usage-based acquisition in the L2 is not as successful as in the L1 (Ellis and Larsen-Freeman, 2006: 571), that L2 learners have difficulty identifying their own errors (Dlaska and Krekeler, 2008), and that they indeed need guidance to improve their language skills (Ellis and Bogart, 2007). Since providing practice and feedback for speaking proficiency is particularly time-consuming, the necessary amount of practice is almost never achieved in traditional teacher-fronted lessons. Against this background, ASR-based CALL systems would seem to make for an interesting supplement to traditional L2 classes.

However, developing ASR-based CALL systems that can provide accurate and useful feedback on oral proficiency is not trivial, because the speech of L2 learners poses special difficulties to ASR technology (Compernelle 2001; Benzeghiba et al. 2007; Doremalen et al. 2009a; Doremalen et al. 2009b). In addition, existing systems in general fail to provide corrective feedback that is detailed enough and accurate, especially on L2 pronunciation which is considered a particularly challenging skill, both for L2 learners (Flege, 1995) and CALL systems (Menzel et al. 2000: 54; Morton and Jack, 2005).

Another problem that has hampered the realization of ASR-based CALL systems, especially for the smaller languages, is that although companies, esp. publishers, are willing to use the technology, many companies do not

have the means to finance the development of such technology. For these and other reasons, in the Netherlands and Flanders a programme was started, called STEVIN (a Dutch acronym that stands for Essential Language Resources in Dutch), which is funded by the Flemish and Dutch governments and aims at stimulating the development of basic language and speech technology for the Dutch language.

Within the framework of the STEVIN programme a project called DISCO (Development and Integration of Speech technology into COurseware for language learning, <http://lands.let.kun.nl/~strik/research/DISCO>) was started that aims at developing a prototype of an ASR-based CALL system for practicing oral skills in Dutch L2. The system addresses different aspects of speaking proficiency (syntax, morphology and phonology), detects errors in speaking performance, points them out to the learners and gives them the opportunity to try again until they manage to produce the correct form.

One of the interesting things about this project is that since it is carried within the STEVIN programme, the technology that is developed for the present project will be publicly made available to interested users (researchers, HLT companies and publishers) through the Dutch HLT Agency.

In the current paper we discuss the relation between language resources and CALL systems: language resources are essential in the development of CALL applications, during R&D resources are created, and finally the CALL system itself can be used to generate additional resources that are useful for research and development of new (CALL) systems.

In section 2 we describe which language resources were employed in the DISCO project. In section 3 we present

the DISCO system paying attention to the design, the automatic speech recognition modules, some preliminary results and the resources produced within the project. In section 4 we discuss how additional language resources can be generated through the DISCO system.

2. CALL applications and the need for language resources

An important requirement for developing ASR-based CALL applications is the availability of language resources such as language and speech corpora and speech technology toolkits.

In order to develop technology that is able to identify errors in oral proficiency we need to know which errors are made by L2 learners in the first place. Part of this information can be found in the literature, but, in general, the information provided in the literature is not complete and not sufficiently quantified to be suitable for developing CALL applications.

In our previous research on developing a computer assisted pronunciation training (CAPT) for Dutch, Dutch-CAPT (Cucchiari et al., 2009), we needed to draw up an inventory of pronunciation errors. We discovered that the information on L2 errors provided in the literature was mostly based on observational studies, was often incomplete, and not quantitative in nature. For this reason we had no other choice than conducting L2 error studies ourselves (Neri et al., 2006). However, since a speech corpus of non-native Dutch was not available at the time, we had to resort to the auditory analysis of Dutch L2 speech recordings that had been collected in the framework of previous projects (Neri et al., 2006).

For the DISCO project we had the opportunity of using the results of another STEVIN project that had been completed in the meantime, the JASMIN corpus (Cucchiari et al., 2008).

2.2.1. The JASMIN speech corpus

The JASMIN corpus is an extension of the large Spoken Dutch Corpus (CGN; Oostdijk, 2002). JASMIN contains speech by children of different age groups, elderly people and non-natives with different mother tongues. The JASMIN corpus was collected in the Netherlands and Flanders and is specifically aimed at facilitating the development of speech-based applications for children, non-natives and elderly people. In the case of non-native speakers the applications envisaged were especially language learning applications because there is considerable demand for CALL products that can help making Dutch L2 teaching more efficient.

In selecting the non-native speakers for this corpus, mother tongue constituted an important variable. For the Flemish part, Francophone speakers were selected because they form a significant proportion of the Dutch

learning population. In the Netherlands, on the other hand, a miscellaneous group of L2 learners with various mother tongues was selected because this more realistically reflects the situation in Dutch L2 classes.

Since an important aim in collecting non-native speech material was that of developing language learning applications for education in Dutch L2, various experts were consulted to determine for which proficiency level such applications are most needed. It turned out that for the lowest levels of the Common European Framework (CEF), namely A1, A2 or B1, there is relatively little material and that ASR-based applications would be very welcome. For this reason, speech from adult Dutch L2 learners at these lower proficiency levels was recorded.

The speech collected in the JASMIN corpus was recorded in two different modalities: about 50% of the material consists of read speech material while the other 50% is made up of extemporaneous speech produced in human-machine dialogues. The JASMIN dialogues were collected through a Wizard-of-Oz-based platform and were designed such that the wizard was in control of the dialogue and could intervene when necessary. In addition, recognition errors were simulated and difficult questions were asked to elicit some typical phenomena of human-machine interaction that are known to be problematic in the development of spoken dialogue systems, such as hyperarticulation, restarts, filled pauses, self talk and repetitions.

The speech recordings were annotated at different levels. For the DISCO project, the verbatim transcription and the automatically generated phonemic transcription are particularly relevant.

For all the reasons mentioned above the JASMIN speech material turned out to be extremely useful and appropriate for the development of the DISCO system.

Both read and extemporaneous speech were analyzed to study which errors are made at the level of pronunciation, morphology and syntax. For this purpose the annotations contained in JASMIN were supplemented with extra annotations of the morphological and syntactical errors made by the speakers. The automatically generated phonemic transcriptions were manually verified by trained students and where necessary improved. Subsequently they were used to study which pronunciation errors are made by L2 learners of Dutch with different mother tongues.

The human-machine dialogues were used for conducting experiments for the DISCO system because they closely resemble the situation we will encounter in this CALL application.

2.2.2. The SPRAAK speech recognizer

The speech recognizer adopted in the DISCO project is SPRAAK (Demuyne et al., 2008), a hidden Markov model (HMM)-based ASR package developed for over 15 years by ESAT at the University of Leuven and later enriched with knowledge and code from other partners through the STEVIN project SPRAAK. The availability of a speech recognition system for Dutch was considered to be an important requirement by the whole language and speech technology (LST) community in the Netherlands and Flanders. For this reason a project was started within the STEVIN programme for this specific purpose: the SPRAAK project. The aim of SPRAAK was twofold: a) developing a highly modular toolkit for research into speech recognition algorithms and b) providing a state-of-the-art recogniser for Dutch with a simple interface that could be used by non-specialists. SPRAAK is distributed as open source for academic usage and at moderate cost for commercial exploitation (for further details, see <http://www.spraak.org/>).

3. The DISCO system

3.1 Design of the DISCO system

Within the STEVIN programme a project called DISCO was started on 01-02-2008, in which a CALL system will be developed. The target user group for the DISCO system are immigrants who want to learn Dutch as L2 to be able to work in the Netherlands or Flanders.

The model adopted for designing the system is Distributed Language Learning (DLL), a methodological and conceptual framework for designing competency-oriented and effective language education (Colpaert, 2004). Its starting point is the design of a language learning environment for a specific language learning situation. The design is based on a thorough analysis of all factors and actors in the language learning situation, and on the identification of aspects amenable to change or improvement. The main phases of the design are goal-oriented conceptualization and ontological specification. Goal-oriented conceptualization stands for the formulation of a solution based on the realization of 'practical goals' as a hypothetical compromise between (often conflicting) personal and pedagogical goals, both for teachers and learners. Ontological specification is a detailed description of the architecture of the language learning environment, defined as the network of interactions between learner, co-learner, teacher, content, native, etc. inside or outside the learning place.

In DISCO, we limit our general design space to closed response conversation simulation courseware and interactive participatory drama (IPD), a genre in which learners play an active role in a pre-programmed scenario by interacting with computerized characters or "agents". The use of drama is beneficial for various reasons:

1. it "reduces inhibition, increases spontaneity, and

enhances motivation, self-esteem and empathy" (Hubbard, 2002),

2. it casts language in a social context, and
3. its notion implies a form of planning, scenario-writing and fixed roles, which is consistent with the limitations we set for the role of speech technology in DISCO.

To summarize, this framework allows us to create a rich and communicative CALL application that stimulates Dutch L2 learners to produce speech and experience the social context. On the other hand, these choices are appropriate from a technological perspective, since they make it possible to successfully deploy speech technology while taking into account its limitations (Strik et al., 2009).

To gain more insight into appropriate feedback strategies, pedagogical goals, and personal goals a number of preparatory studies were carried out, such as exploratory in-depth interviews with Dutch L2 teachers and experts, focus group discussions to elicit the personal goals of learners, and pilot studies through partial systems with limited functionality (e.g. no speech technology). The functions of the system that were not implemented (play prompts, give feedback, etc.) were simulated. The results of these preparatory studies were taken into account in finalizing the design of the DISCO system.

The learning process starts with a relatively free conversation simulation, taking well into account what is (not) possible with speech technology: learners are given the opportunity to choose from a number of prompts at every turn (branching, decision tree). Based on the errors they make in this conversation they will be offered remedial exercises, which are very specific exercises with little freedom.

Feedback depends on individual learning preferences: the default feedback strategy is immediate corrective feedback, which is visually implemented through highlighting, and from an interaction perspective by putting the conversation on hold and focusing on the mistakes. Learners that wish to have more conversational freedom can choose to receive communicative recasts as feedback, which let the conversation go on while highlighting mistakes for a short period of time.

The final system will have several parameters that can be changed by the learner or teacher. During development and implementation, we will try to have these parameters behave intelligently (based on error analysis and learner behavior), so that the system can adapt itself to the learner. For future research these parameters offer the possibility of studying different modes of behavior of the CALL system and their effect on language learners.

3.2 The speech recognition modules

First, we provide some technical details about our system. As mentioned above, the human-machine dialogues were

used for conducting experiments for the DISCO system. The material used consisted of speech from 45 speakers who each give answers to 39 questions about a journey. The input speech, sampled at 16kHz, is divided into overlapping 32ms Hamming windows with a 10ms shift and pre-emphasis factor of 0.95. 12 Mel-frequency cepstral coefficients (MFCCs: C1-C12) plus C0 (energy), and their first and second order derivatives were calculated and cepstral mean subtraction (CMS) was applied. The constrained language models and pronunciation lexicons are implemented as finite state machines (FSM).

In the DISCO system feedback on speaking performance is given on three levels: syntax, morphology and phonology. To give feedback, errors on these levels have to be detected automatically. In our system architecture, this task is divided in two modules: (1) the speech recognition module and (2) the error detection module. The first module, speech recognition, determines the sequence of words the student uttered. For each prompt a list of predicted correct and (grammatically) incorrect responses is created beforehand based on errors that are expected on empiric grounds. This list is the basis for a Finite State Grammar (FSG) language model, which is used by an hidden Markov model (HMM)-based speech recognition system. The recognition system is forced to choose among the predicted response from the list.

To avoid false accepts, for example when an utterance is uttered that is not in the list of predicted responses, utterance verification (UV) is carried out. Using a combination of acoustic and durational similarity measures it is determined whether the response chosen by the speech recognizer reflects what has been said. If it is rejected the user is asked to try again; if it is accepted, the system will proceed to error detection (Van Doremalen et al. 2009a, b).

Note that once the chosen response is accepted by the utterance verifier we can already detect errors on the syntactic level because the system is confident enough that the student uttered a specific sequence of words and it also knows what the student was supposed to say.

Detecting errors on the morphological and phonological levels requires another, more detailed analysis of the speech signal. The starting point of this analysis is a segmentation of the speech signal into a sequence of phones obtained from the speech recognition module. Using a variety of spectral and temporal features a confidence measure (CM) is calculated for each of these phones. Based on this CM the system decides to mark the hypothesized phone in the segmentation as correctly pronounced or incorrectly pronounced (Van Doremalen et al. 2009c).

In the way described above, phonological errors can be detected. Since some phonemes are critical for certain

morphological constructions, the approach used for detecting phonological errors will be used also for detecting some of the morphological errors, for instance those concerning regular verb forms. Irregular verbs, on the other hand, may require an approach that is more similar to that adopted for detecting syntactic errors. Once the system arrives at this final stage, the system has a detailed overview of all the errors on the different levels and based on this overview the system can provide feedback to the student.

3.3 The resources produced in the project

The resources mentioned above are employed to develop the DISCO system which consists of various parts. First of all, a blue-print of the design and the speech technology modules for recognition (i.e. for selecting an utterance from the predicted list, and verifying the selected utterance) and for error detection (errors in pronunciation, morphology, and syntax). In addition, the following resources have been developed: an inventory of errors at all these three levels, a prototype of the DISCO system with content, specifications for exercises and feedback strategies, and a list of predicted correct and incorrect utterances.

The fact that DISCO is being carried out within the STEVIN programme implies that its results, all the resources mentioned above, will become available for research and development through the Dutch Flemish Human Language Technology (HLT) Agency (TST-Centrale; www.inl.nl/tst-centrale). This makes it possible to reuse these resources for conducting research and for developing specific applications for ASR-based language learning.

3.4 Evaluation

A system that gives meaningful feedback must operate in a manner that is similar to what a competent teacher would do. Therefore, for the final evaluation of the whole system we intend to use a design in which different groups of students of Dutch as a second language (DL2) at the University of Antwerp and at the Radboud University in Nijmegen use the system and fill in a questionnaire with which we can measure the students' satisfaction in working with the system.

Teachers of DL2 will then assess all sets of system prompt, student response and system feedback for the quality of the feedback on the level of pronunciation, morphology and syntax. For this purpose, recordings will be made of students who complete the exercises developed to test the DISCO system.

Given the evaluation design sketched above, we consider the project successful from a scientific point of view if the DL2 teachers agree that the system behaves in a way that makes it as useful for the students as a teacher is, and if the students rate the system positively on its most important aspects.

4. Generating additional language resources

Above we described which resources we used in developing our CALL system, and which resources become available during development of the system. In this section, we describe which additional resources can be collected by using the CALL system.

After the CALL system has been developed, language learners can use it to practice oral skills. The system has been designed and developed in such a way that it is possible to log details regarding the interactions with the users. This logbook can contain, e.g., the following information: what appeared on the screen, how the user responded, how long the user waited, what was done (speak an utterance, move the mouse and click on an item, use the keyboard, etc.), the feedback provided by the system, how the user reacted on this feedback (listen to example (or not), try again, ask for additional, e.g. meta-linguistic, feedback, etc.).

Finally, all the utterances spoken by the users can be recorded in such a way that it is possible to know exactly in which context the utterance was spoken, i.e. it can be related to all the information in the logbook mentioned above. An ASR-based CALL system, like DISCO, can thus be used for acquiring additional non-native speech data, for extending already existing corpora like JASMIN, or for creating new ones. This could be done within the framework of already ongoing research without necessarily having to start corpus collection projects.

Such a corpus and the log-files can be useful for various purposes: for research on language acquisition and second language learning, studying the effect of various types of feedback, research on various aspects of man-machine interaction, and of course for developing new, improved CALL systems. Such a CALL system will also make it possible to create research conditions that were hitherto impossible to create, thus opening up possibilities for new lines of research.

For instance, at the moment a project is being carried out at the Radboud University of Nijmegen, which is aimed at studying the impact of corrective feedback on the acquisition of syntax in oral proficiency (<http://lands.let.kun.nl/~strik/research/FASOP>). Within this project the availability of an ASR-based CALL system makes it possible to study how corrective feedback on oral skills is processed on-line, whether it leads to uptake in the short term and to actual acquisition in the long term.

This has several advantages compared to other studies that were necessarily limited to investigating interaction in the written modality: the learner's oral production can be assessed on line, corrective feedback can be provided immediately under near-optimal conditions, all interactions between learner and system can be logged so that data on input, output and feedback are readily

available for research.

5. Conclusions

In this paper we have discussed the importance of language resources for CALL application development on the basis of our experiences in the DISCO project in which speech data and speech technology are employed to develop a system for practicing oral skills in a second language.. We have seen that language resources are actually indispensable for developing sound CALL applications. Once developed, such applications can also be employed to produce new valuable language resources which can in turn be used to develop new, improved CALL systems.

6. Acknowledgements

The DISCO project is carried out within the STEVIN programme funded by the Dutch and Flemish Governments (<http://taaluniversum.org/taal/technologie/stevin/>).

7. References

- Benzeghiba, M., Mori, R. D., Deroo, O., Dupont, S., Erbes, T., Jouviet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., Wellekens, C. (2007). Automatic speech recognition and speech variability: a review. *Speech Communication*, 49, 763–786.
- Colpaert, J. (2004). Design of Online Interactive Language Courseware: Conceptualization, Specification and Prototyping. Research into the impact of linguistic-didactic functionality on software architecture. (Doctoral dissertation). University of Antwerp, 2004.
- Cucchiari, C., Neri, A., and Strik, H. (2009). Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback. *Speech Communication*, 51, 853-863.
- Cucchiari, C., Driesen, J., Van hamme, H., and Sanders, E., (2008). Recording speech of children, non-natives and elderly people for HLT applications: the JASMIN-CGN corpus. In *Proceedings of LREC-2008*.
- Demuyne, K., Roelens, J., van Compernelle, D., and Wambacq, P. (2008) SPRAAK: an open source SPEECH Recognition and Automatic Annotation Kit. In *Proceedings of ICSLP-2008*, p. 495.
- Glaska, A. and Krekeler, C. (2008). Self-assessment of pronunciation. *System*, 36, pp. 506-516.
- Ellis, N.C., Bogart, P.S.H. (2007). Speech and Language Technology in Education: the perspective from SLA research and practice. In *Proc. SLATE*, Farmington PA, pp. 1-8.
- Ellis, N. and Larsen-Freeman, D. (2006). Language emergence: implications for applied. *Linguistics, Applied Linguistics*, 27.4: 558–89.
- Eskenazi, M. (2009). An overview of Spoken Language Technology for Education, *Speech Communication*.

- Flege, J. (1995). Second language speech learning: theory, findings and problems. In W. Strange (Ed.) *Speech perception and linguistic experience*, Baltimore: York Press, pp. 233-272.
- Hubbard, P. (2002). Interactive Participatory Dramas for Language Learning. *Simulation and Gaming*, vol. 33, pp. 210-216.
- Morton, H., Jack, M. (2005). Scenario-Based Spoken Interaction with Virtual Agents. *Computer Assisted Language Learning*, 18, 171-191.
- Oostdijk, N. (2002). The design of the spoken dutch corpus. In *New Frontiers of Corpus Research*, P. Peters, P. Collins, and A. Smith, Eds. Rodopi, pp. 105–112.
- H. Strik, Cornillie, F., van Doremalen, J., Cucchiari, C. (2009). Developing a CALL System for Practicing Oral Proficiency: How to Match Design and Speech Technology. In *Proc. SLATE*, Wroxall Abbey.
- Van Compernelle, D. (2001). Recognizing speech of goats, wolves, sheep and ... non-natives. *Speech Communication*, 35, 71-79.
- Van Doremalen, J., Cucchiari, C., Strik, H. (2009a). Optimizing automatic speech recognition for low-proficient non-native speakers. Accepted for publication in *EURASIP Journal on Audio, Speech, and Music Processing*, to appear.
- Van Doremalen, J., Strik, H., Cucchiari, C. (2009b). Utterance Verification in Language Learning Applications. In *Proc. SLATE*, Wroxall Abbey.
- Van Doremalen, J., Cucchiari, C., Strik, H. (2009c). Automatic Detection of Vowel Pronunciation Errors Using Multiple Information Sources. *Proceedings of the biannual IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*.