JASMIN-CGN: Extension of the Spoken Dutch Corpus with Speech of Elderly People, Children and Non-natives in the Human-Machine Interaction Modality

Catia Cucchiarini[1]**, Hugo Van hamme**[2]**, Olga van Herwijnen**[1]**, and Felix Smits**[3]
[1]CLST, Radboud University, Erasmusplein 1, 6525 HT Nijmegen, The Netherlands
[2] Katholieke Universiteit Leuven – Dept. ESAT, Kasteelpark Arenberg 10, B3001 Leuven, Belgium
[3]TalkingHome, Institutenweg 40/42, 7521 PK Enschede, The Netherlands
E-mail: c.cucchiarini@let.ru.nl, hugo.vanhamme@esat.kuleuven.be, o.vanherwijnen@let.ru.nl, smits@talkinghome.nl

Abstract

Large speech corpora (LSC) constitute an indispensable resource for conducting research in speech processing and for developing real-life speech applications. In 2004 the Spoken Dutch Corpus (CGN) became available, a corpus of standard Dutch as spoken by adult natives in the Netherlands and Flanders. Owing to budget constraints, CGN does not include speech of children, non-natives, elderly people and recordings of speech produced in human-machine interactions. Since such recordings would be extremely useful for conducting research and for developing HLT applications for these specific groups of speakers of Dutch, a new project, JASMIN-CGN, was started which aims at extending CGN in different ways: by collecting a corpus of contemporary Dutch as spoken by children of different age groups, non-natives with different mother tongues and elderly people in the Netherlands and Flanders and, in addition, by collecting speech material in a communication setting that was not envisaged in CGN: human-machine interaction. We expect that the knowledge gathered from these data can be generalized to developing appropriate systems also for other speaker groups (i.e. adult natives). One third of the data will be collected in Flanders and two thirds in the Netherlands.

1. Introduction

Large speech corpora (LSC) constitute an indispensable resource for conducting research in speech processing and for developing real-life speech applications. The need for such resources is now generally recognized and large, annotated speech corpora are becoming available for various languages. Other than the term “large” probably suggests, all these corpora are inevitably limited. The limitations are imposed by the fact that LSC require much effort and are therefore very expensive. For these reasons, important choices have to be made when compiling an LSC in order to achieve a corpus design that guarantees maximum functionality for the budget available.

In March 2004 the Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN) became available, a corpus of about 9 million words that constitutes a plausible sample of standard Dutch as spoken in the Netherlands and Flanders and contains various annotation layers. The design of this corpus was guided by a number of considerations. In order to meet as many requirements as possible, it was decided to limit the CGN to the speech of adult, native speakers of Dutch in the Netherlands and Flanders.

The rapid developments in Information Society and the ensuing proliferation of computer services in support of our daily activities stress the importance of CGN for developing such services for Dutch at reasonable costs, thus removing the language barrier for many citizens. Familiar examples of Human Language Technology (HLT) applications are dictation systems and call-centre-based applications such as telephone transaction systems and information systems that use automatic speech recognition instead of a keyboard or a keypad. Furthermore, multilingual access interfaces and cross-lingual speech applications in which people can communicate with each other even though they speak different languages are now being developed, i.e. for telephone reservation systems and voice portals. As embedded technology, HLT will have a crucial role in next-generation products and services that replace information processing methods typical of the desktop computing generation. The advent of ambient intelligence will make it possible for humans to interact with ubiquitous computing devices in a seamless and more natural way. Finally, in a world increasingly dominated by knowledge and information, learning will become a lifelong endeavour and HLT applications will become indispensable in favouring remote access and interaction with (virtual) tutors.

2. Potential Users of HLT Applications

The fact that CGN is restricted to the speech of adult, native speakers of Dutch in the Netherlands and Flanders, limits its usability for developing HLT applications that must be used by children, non-natives and elderly people. This is undesirable, as these groups also need to communicate with other citizens, administration, enterprises and services and should in principle be able to benefit from HLT-based computer services that are available for the rest of the population. In addition, all three social groups are potential users of HLT applications specially tailored for children, non-natives and elderly people, which would considerably increase their opportunities and their participation in our society.

In the case of children, HLT applications have an important role to play with respect to education and entertainment (Narayanan & Potamianos, 2002). For certain applications, such as internet access and interactive learning, speech technology provides an alternative modality that may be better suited for children compared to the usual keyboard and mouse access. In other applications, such as Computer Assisted Language

Learning (CALL) or computer-based interactive reading tutors (Hagen et al., 2003), speech and language technology is the key enabling technology.

The increasing mobility and consequent migration of workers to the Netherlands and Flanders have resulted in growing numbers of non-native speakers of Dutch that have to function in a Dutch-speaking society. For them, HLT applications can be relevant in two respects: to guarantee their participation in the Information Society and to promote their integration in society by facilitating their acquisition of the Dutch language.

When talking about the information society, authorities and policy makers put special emphasis on aspects such as empowerment, inclusion, and elimination of cultural and social barriers. This implies that the information society should be open to all citizens, also those who are not mother tongue speakers of Dutch. To guarantee that also non-native speakers of Dutch can participate in the information society it is necessary that all sorts of services and applications, for instance those mentioned in the previous section, be available for them too.

The teaching of Dutch as a second language (L2) is high on the political agenda, both in the Netherlands and in Flanders, because it is considered to be the key to successful integration. In the last thirty years the Dutch and the Flemish governments have spent billions of euros on Dutch L2 teaching to non-natives. Despite these huge efforts, the results are not always satisfactory and experiments are now being conducted with new methods and new media, to try and improve the quality of Dutch L2 teaching. For example, CALL systems that make use of advanced HLT techniques seem to offer new perspectives. These systems can offer extra learning time and material, specific feedback on individual errors and the possibility to simulate realistic interaction in a private and stress-free environment.

Owing to the increase in average life expectancy, our society has to cope with a growing aged population and government and commercial organizations are concerned about how to meet the needs of this increasing group of older adults and to guarantee independent aging as much as possible. Technology, and in particular, HLT applications, can help in providing assistance to older individuals who want to maintain independence and quality of life. Among the consequences of aging are declines in motor, sensory and cognitive capabilities. HLT can be employed in developing assistive devices that compensate for these diminished capabilities. For instance, it is possible to compensate for motor or sensory deficiencies by developing devices for control of the home environment through spoken commands. Cognitive aging often results in a decline in working memory, online reasoning, and the ability to attend to more than one source of information. Technology can compensate for cognitive dysfunctions either by facilitating information processing or by supporting functions such as planning, task sequencing, managing prescription drug regimens, prioritization and problem solving. The applications can vary from reminder systems to interactive robotic assistants (Takahashi et al., 2002; Hans et al., 2002; Ferguson et al., 2002; Müller et al., 2002).

3. The Need for Dedicated Corpora

Although it is obvious that speech-based services are of social and economic interest to youngsters, seniors and foreigners at the moment such applications are difficult to realize. As a matter of fact, speech recognizers that are optimized for adult speech are not suitable for handling speech of children, non-natives and elderly people (Narayanan & Potamianos, 2002; Raux et al., 2003; Anderson et al., 1999; D'Arcy et al., 2004; Van Compernolle, 2001). The much lower performance achieved with children speech has to do with differences in vocal tract size and fundamental frequency, with pronunciation problems and different vocabulary, and with increased variability within speakers as well as among speakers. In the ASR community, it has long been known that the differences between native and non-native speech are so extensive as to degrade ASR performance considerably (Van Compernolle, 2001). As a consequence, considerable efforts have been spent in trying to understand the reasons for this poor performance and in finding appropriate solutions. Research into automatic speech recognition of elderly speech has shown that performance degrades considerably for people above the age of 70 (Anderson et al., 1999). This deterioration in performance can be ascribed to different spectral and pronunciation patterns that result from a degradation of the internal control loops of the articulatory system and from changes in the size and periodicity of the glottal pulses. Although the performance disadvantage for children, seniors and non-natives can be explained to some extent, there is much that is not well understood. But in the past it has been difficult to conduct research aimed at explaining the difference because of the lack of suitable corpora.

For the time being, the problems in ASR for children, elderly and non-natives are approached with standard adaptation procedures (Narayanan & Potamianos, 2002; Raux et al., 2003; Anderson et al., 1999; Van Compernolle, 2001). Although these do improve performance, straightforward adaptation does not bring the performance to the same level as what can be obtained with adult native speech. Perhaps more importantly, straightforward adaptation does not yield much insight into the fundamental causes of the ASR problems.

An analysis of turn taking and interaction patterns in the face-to-face and telephone dialogues that was carried out within the COMIC project (http://www.hcrc.ed.ac.uk/comic/documents) has shown that these are fundamentally different from the best we can do at this moment in human-computer interaction. Humans handle misunderstandings and recognition errors seemingly without effort, and that capability appears to be essential for maintaining a fluent conversation. Automatic systems have only very limited capabilities for detecting that their human interlocutor does not fully understand prompts and responses. Experience with developing voice operated information systems has revealed a lack of knowledge about the specific behaviour that people exhibit when they have to interact with automatic systems, especially when the latter do not understand what the user says. For instance, it turns out that people do not answer the questions posed by the machine immediately, but first think about what to say and to take time they either start repeating the question, or produce all sorts of hesitations and disfluencies. In addition, if the computer does not understand them, they start speaking more loudly, or modify their pronunciation in an attempt to be more understandable with the result that their speech deviates even more from what the computer expects. The problems experienced in developing spoken dialogs with machines are compounded when the users come from less general sections of the population such as children, non-natives and elderly people (Narayanan & Potamianos, 2002; Raux et al., 2003). Here too, scientific and technological progress is hampered by the lack of appropriate corpora.

It is for this reason that a new project, JASMIN-CGN, was started within the Dutch-Flemish STEVIN program for language and speech technology, which is aimed at the compilation of a corpus of contemporary Dutch as spoken by children of different age groups, elderly people, and non-natives with different mother tongues in the Netherlands and Flanders.

4. JASMIN-CGN: Project Aim

The JASMIN-CGN project is aimed at extending the CGN in three dimensions. First, by collecting a corpus of contemporary Dutch as spoken by children of different age groups, elderly people and non-natives with different mother tongues, we aim at an extension along the age and mother tongue dimensions. In addition, we intend to collect speech material in a communication setting that was not envisaged in the CGN: human-machine interaction. These three dimensions are reflected in the corpus as five user groups: native primary school pupils, native secondary school students, non-native children, non-native adults and senior citizens. For each group, both read and human-machine interaction data will be collected for a total of about 90 hours. One third of the data will be collected in Flanders and two thirds in the Netherlands.

5. Corpus Design

This corpus will contain about 90 hours of speech divided as follows:

In The Netherlands:

- native children between 7 and 11 (12h 21m)
- native children between 12 and 16 (12h 21m)
- Turkish and Moroccan adults (12h 21m)
- Turkish and Moroccan children between 7 and 14 (12h 21m)
- native adults above 60 (9h 26m)

In Flanders:

- native children between 7 and 11 (6h 10m)
- native children between 12 and 16 (6h 10m)
- French-speaking adults (6h 10m)
- French-speaking children between 7 and 14 (6h 10m)
- native adults above 60 (5h 5m)

50% of the material will be read speech and 50% extemporaneous speech recorded in the human-machine interaction modality.

From each speaker we will record about 12 minutes of speech, which means that for obtaining the amount of speech required we will have to record about 70 speakers per group in the Netherlands and about 35 in Flanders. Depending on the specific group, different selection variables will be adopted such as gender, for all groups, region of origin, for natives, mother tongue, and proficiency level in Dutch for non-natives and reading level for children in elementary schools.

6. Speech material

6.1 Read Speech

Half of the material that will be recorded from each speaker in this corpus will consist of read speech. For this purpose we use sets of phonetically rich sentences and stories or general texts to be read aloud. Particular demands on the texts to be selected were imposed by the fact that we have to record read speech of children and non-natives.

Children in the age group 7-12 cannot be expected to be able to read a text of arbitrary level of difficulty. In many elementary schools in the Netherlands and Flanders children learning to read are first exposed to a considerable amount of explicit phonics instruction which is aimed at teaching them the basic structure of written language by showing the relationship between graphemes and phonemes (Wentink, 1997). A much used method for this purpose is the reading program *Veilig Leren Lezen* (Mommers et al., 1990). In this program children learn to read texts of increasing difficulty levels, with respect to text structure, vocabulary and length of words and sentences. The texts are ordered according to reading level and they vary from Level 1 up to Level 9. In line with this practice in schools we have selected texts of the nine different reading levels from books that belong to the reading program *Veilig Leren Lezen*.

For the non-native speakers we selected appropriate texts from books that are intended for learners of Dutch as a second language (Dutch L2).

6.2 Human-Machine Dialogues

A WoZ-based platform was developed for recording speech in the human-machine interaction mode. The human-machine dialogues are designed such that the wizard can intervene when the dialogue goes out of hand. In addition, the wizard can simulate recognition errors to elicit some of the typical phenomena of human-machine interaction that are known to be problematic in the development of spoken dialogue systems. Before designing the dialogues we have drawn up a list of phenomena that should be elicited:

1. hyperarticulation
2. syllable lengthening

3. syllable insertion
4. shouting
5. stress shift
6. restart
7. filled pause
8. silent pause
9. metacommunication
    a. self talk
    b. talking to the machine
10. repetition
11. question repeting
12. paraphrasing

We then considered which speaker's moods could cause the various phenomena and identified three relevant states of mind: (1) confusion, (2) hesitation and (3) frustration.

If the speaker is confused or puzzled, (s)he is likely to start complaining about the fact that (s)he does not understand what to do. Consequently, (s)he will probably start talking to him/herself or to the machine. Filled pauses, silent pauses, repetitions, lengthening and restarts are likely to be produced when the speaker has doubts about what to do next and looks for ways of taking time. So hesitation is probably the state of mind that causes these phenomena. Finally, phenomena such as hyperarticulation, syllable lengthening, syllable insertion, shouting, stress shift, metacommunication probably result when speakers get frustrated. As is clear from this characterization, certain phenomena can be caused by more than one state of mind, like metacommunication that can result either from confusion or from frustration.

The challenge in designing the dialogues was then how to induce these states of mind in the speakers, to cause them to produce the phenomena required.

We have achieved this in different ways such as asking unclear questions, increasing the cognitive load of the speaker by asking more difficult questions, or simulating machine recognition errors of different types. The dialogs are not specific for an application. At the same time, we could not but make choices regarding topic of interaction, vocabulary, grammar, etc. One the one hand, this limits the scope of the JASMIN-CGN corpus, but on the other hand, the dialogs are designed in such a manner that the resulting data will allow future research to perform acoustic and linguistic modelling studies, the results of which can be generalized. As a matter of fact, we expect that the basic knowledge gathered from these data can be generalized to developing interactive systems for adult natives.

7. Annotations

Given the limited budget available, we decided to limit the annotations to a verbatim transcription, POS tagging of the words, and an automatic phonetic transcription. The rationale behind this is that it is preferable to spend money in collecting a sufficient amount of material than in obtaining detailed annotations that are not always of general use and that can be generated automatically without considerable loss of information. Therefore, we opted for an automatically generated broad phonetic transcription. This involves building/adapting acoustic models for children and elderly speech based on the orthographic transcriptions. For the non-natives, we anticipate one or two iterations of the cycle consisting of automatic segmentation, manual corrections, adjustment of the pronunciation lexicon and acoustic models. The CGN project has shown that the speech modelling and segmentation tool used were well suited for this task.

8. Dissemination

The results of this project constitute a valuable basis for conducting research and for developing different sorts of HLT applications. The ultimate results will be made available through the Dutch-Flemish HLT Agency.

9. Acknowledgement

This research is financed by the *Nederlandse Taalunie* through the STEVIN subsidy programme, website www.taaluniversum.org/stevin.

10. References

Anderson, S., Liberman, N., Bernstein, E., Foster, S., Cate, E., Levin, B., and Hudson, R. (1999). Recognition of elderly speech and voice-driven document retrieval. In *Proceedings of the ICASSP*, Phoenix.

D'Arcy, S.M., Wong, L.P. and Russell, M.J. (2004). Recognition of read and spontaneous children's speech using two new corpora. In *Proceedings of ICSLP'04*, Korea, October 4-8, 2004.

Ferguson G. et al. (2002). *The medication advisor project: Preliminary report.* Technical Report 776, CS Dept., U. Rochester.

Hagen, A., Pellom, B. and Cole, R. (2003). Children's Speech Recognition with Application to Interactive Books and Tutors. In *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, St. Thomas, USA

Hans M., Graf B. and Schraft R.D. (2002). Robotic home assistant care-o-bot: Past-present-future. In *Proceedings of the IEEE ROMAN*, Berlin, pp. 380-385.

Mommers, M.J.C., Verhoeven, L. and Van der Linden, S. (1990) *Veilig Leren Lezen*, Tilburg, Zwijsen.

Müller, C., and Wasinger, R. (2002). Adapting Multimodal Dialog for the Elderly. In *Proceedings of the ABIS-Workshop 2002 on Personalization for the Mobile World*. Hannover, Germany. Oct. 9-11, 2002.

Narayanan, S. and Potamianos, A. (2002). Creating conversational interfaces for children. *IEEE Trans. Speech and Audio Processing*, 10(2), pp. 65-78.

Raux, A., Langner, B., Black, A. and Eskenazi, M. (2003). LET'S GO: Improving Spoken Dialog Systems for the Elderly and Non-natives. In *Proceedings Eurospeech 2003*, Geneva, Switzerland.

Takahashi S., Morimoto T., Maeda S. and Tsuruta S (2002). Spoken dialogue system for home health care, In *Proceedings of the ICSLP*, Denver, pp. 2709-2712.

Van Compernolle, D. (2001). Recognizing speech of goats, wolves, sheep and ... non-natives. *Speech Communication*, 35(1-2), pp. 71-79.

Wentink, H. (1997). *From Graphemes to syllables*, Doctoral dissertation, University of Nijmegen.