Recording Speech of Children, Non-Natives and Elderly People for HLT Applications: the JASMIN-CGN Corpus

C. Cucchiarini[1], J. Driesen[2], H. Van hamme[2], E. Sanders[1]

[1]CLST, Radboud University, Erasmusplein 1, 6525 HT Nijmegen, The Netherlands
[2] Katholieke Universiteit Leuven – Dept. ESAT, Kasteelpark Arenberg 10, B3001 Leuven, Belgium
E-mail: c.cucchiarini@let.ru.nl, joris.driesen@esat.kuleuven.be, hugo.vanhamme@esat.kuleuven.be
e.sanders@let.ru.nl

Abstract

Within the framework of the Dutch-Flemish programme STEVIN, the JASMIN-CGN (Jongeren, Anderstaligen en Senioren in Mens-machine Interactie – Corpus Gesproken Nederlands) project was carried out, which was aimed at collecting speech of children, non-natives and elderly people. The JASMIN-CGN project is an extension of the Spoken Dutch Corpus (CGN) along 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, an extension along the age and mother tongue dimensions was achieved. In addition, we collected speech material in a communication setting that was not envisaged in the CGN: human-machine interaction. One third of the data was collected in Flanders and two thirds in the Netherlands. In this paper we report on our experiences in collecting this corpus and we describe some of the important decisions that we made in the attempt to combine efficiency and high quality.

1. Introduction

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 (Oostdijk, 2002). 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.

However, the fact that CGN does not contain speech of children, non-natives and elderly people limits its usability for conducting research and developing HLT applications. As a matter of fact, these groups of speakers of Dutch 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.

It is for these reasons that in 2005 a project aimed at collecting speech of children, non-natives and elderly people was financed within the framework of the Dutch-Flemish programme STEVIN. This project, called JASMIN-CGN (Jongeren, Anderstaligen en Senioren in Mens-machine Interactie – Corpus Gesproken Nederlands) aimed at extending the Spoken Dutch Corpus (CGN) along 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, an extension along the age and mother tongue dimensions was achieved. In addition, we collected speech material in a communication setting that was not envisaged in the CGN: human-machine interaction.

In this paper we report on our experiences in collecting this corpus and we describe some of the important decisions that we made in the attempt to combine efficiency and high quality.

2. Corpus design

The three dimensions mentioned above 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 all groups of speakers 'gender' was adopted as a selection variable. In addition, 'region of origin' and 'age' constituted variables in selecting native speakers. Finally, the selection of non-natives was also based on variables such as 'mother tongue', 'proficiency level in Dutch' and 'age'.

2.1 Speaker selection

For the selection of speakers we have taken the following variables into account: region of origin (Flanders or the Netherlands), nativeness (native as opposed to non-native speakers), dialect region (in the case of native speakers), age, gender and proficiency level in Dutch (in the case of non-native speakers).

2.1.1 Region of origin

We distinguished two regions: Flanders (FL) and the Netherlands (NL) and we tried to collect one third of the speech material from speakers in Flanders and two thirds from speakers in the Netherlands.

2.1.2 Nativeness

In each of the two regions, three groups of speakers consisted of native speakers of Dutch (native primary school pupils, native secondary school students, and senior citizens) and two of non-native speakers (non-

native children and non-native adults). For native and non-native speakers different selection criteria were applied, as will be explained below.

2.1.3 Dialect region

Native speakers were divided in groups on the basis of the dialect region they belong to. A person is said to belong to a certain dialect region if (s)he has lived in that region between the ages of 3 and 18 and if (s)he has not moved out of that region more than three years before the time of the recording.

Within the native speaker categories we strived for a balanced distribution of speakers across the four dialect regions that we distinguished in the Netherlands and Flanders, but without considering this as a hard demand.

For non-native speakers, dialect region did not constitute a selection variable, since the regional dialect or variety of Dutch is not expected to have a significant influence on their pronunciation. However, we did notice a posteriori that the more proficient non-native children do exhibit dialectal influence (especially in Flanders due to the recruitment).

2.1.4 Mother tongue

Since the JASMIN-CGN corpus was collected for the aim of facilitating the development of speech-based applications for children, non-natives and elderly people, special attention was paid to selecting and recruiting speakers belonging to the group of potential users of such applications. In the case of non-native speakers the applications we had in mind were especially language learning applications because there is considerable demand for CALL (Computer Assisted Language Learning) products that can help making Dutch as a second language (L2) education more efficient.

In selecting non-native speakers, mother tongue constituted an important variable because certain mother tongue groups are more represented than others in the Netherlands and Flanders. For instance, for Flanders we opted for Francophone speakers since they form a significant fraction of the population in Flemish schools, especially (but not exclusively) in major cities. A language learning application could address the school's concerns about the impacts on the level of the Dutch class. For adults, CALL applications can be useful for social promotion and integration and for complying with the bilingualism requirements associated with many jobs. Often, Francophone speakers are immigrants from other countries and have other languages as their mother tongue. Such speakers were also allowed in the sample.

In the Netherlands, on the other hand, the choice of mother tongue groups turned out to be less straightforward and even subject to change over time. The original idea was to select speakers with Turkish and Moroccan Arabic as their mother tongue, to be recruited in regional education centres where they follow courses in Dutch L2. This choice was based on the fact that Turks and Moroccans constitute two of the four most substantial minority groups, the other two being people from Surinam and the Dutch Antilles who generally speak Dutch and do not have to learn it when they immigrate to the Netherlands.

However, it turned out that it was very difficult and time-consuming to recruit exclusively Turkish and Moroccan speakers. In addition, the introduction of a new immigration law that envisages new obligations with respect to learning Dutch for people from outside the EU, led to considerable changes in the whole Dutch L2 education landscape, in particular to a significant decrease in the proportion of L2 learners from outside the EU. Moreover, in this modified context, it was no longer so straightforward to imagine which mother tongue groups would be the most obvious candidates for using CALL and speech-based applications. After various consultations with experts in the field, we decided not to limit the selection of non-natives to Turkish and Moroccan speakers and opted for a miscellaneous group that more realistically reflects the situation in Dutch L2 classes.

2.1.5 Proficiency in Dutch

Since an important aim in collecting non-native speech material is that of developing language learning applications for education in Dutch L2, we consulted various experts in the field to find out 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, we chose to record speech form adult Dutch L2 learners at these lower proficiency levels. For children, the current class (grade) they are in was used as a selection criterion.

2.1.6 Speaker age

Age was used as a variable in selecting both native and non-native speakers. For the native speakers we distinguished three age groups:

- children between 7 and 11
- children between 12 and 16
- native adults of 65 and above

For the non-native speakers two groups were distinguished:

- children between 7 and 16
- adults between 18 and 60.

2.1.7 Speaker gender

In the five groups of speakers we strived to obtain a balanced distribution between male and female speakers.

2.2 Speech modalities

In order to obtain a relatively representative and balanced corpus we decided to record about 12 minutes of speech from each speaker. About 50% of the material would consist of read speech material and 50% of extemporaneous speech produced in human-machine dialogues.

2.2.1 Read speech

About half of the material to be recorded from each speaker in this corpus consists of read speech. For this purpose we used 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 had to record read speech of children and non-natives.

Children in the age group 7-11 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 selected texts of the nine different reading levels from books that belong to the reading programme *Veilig Leren Lezen*.

For the non-native speakers we selected appropriate texts from a widely used method for learning Dutch as a second language, *Code 1* and *Code 2*, from Thieme Meulenhoff Publishers. The texts were selected as to be suitable for learners with CEF levels A1 and A2.

2.2.2 Human-machine dialogues

A Wizard-of-Oz-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 drew up a list of phenomena that should be elicited such as hyperarticulation, syllable lengthening, shouting, stress shift, restarts, filled pauses, silent pauses, self talk, talking to the machine, repetitions, prompt/question repeating and 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 and self talk 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 self talk 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.

Different dialogues were developed for the different speaker groups. To be more precise, the structure was similar for all the dialogues, but the topics and the questions were different.

3. Collecting speech material

3.1 Speaker recruitment

Different recruitment strategies were applied for the five speaker groups. The most efficient way to recruit children was to approach them through schools. However, this was difficult because schools are reluctant to participate in individual projects owing to a general lack of time. In fact this was anticipated and the original plan was to recruit children through pedagogical research institutes that have regular access to schools for various experiments. Unfortunately, this form of mediation turned out not to work because pedagogical institutes give priority to their own projects. So, eventually, we had to contact the schools ourselves and recruiting children turned out to be much more time-consuming than we had envisaged.

In Flanders, most recordings in schools were organized in collaboration with the school management teams. A small fraction of the data were recorded at summer recreational activities for primary school children ("speelpleinwerking").

The elderly people were recruited through retirement homes and elderly care homes. In Flanders older adults were also recruited through a Third Age University.

In the Netherlands non-native children were recruited through special schools which offer specific Dutch courses for immigrant children (Internationale Schakelklassen). In Flanders the non-native children were primarily recruited in regular schools. In major cities and close to the language border a significant proportion of pupils speak only French at home, but attend Flemish schools. The level of proficiency is very dependent on the individual and the age. A second source of speakers was a school with special programs for recent immigrants.

Non-native adults were recruited through language schools that offer Dutch courses for foreigners. Several schools (in the Netherlands: Regionale Opleidingscentra, ROCs – in Flanders: Centra voor Volwassenen Onderwijs, CVOs). Through these schools we managed to contact non-native speakers with the appropriate levels of linguistic skills. Specific organizations for

foreigners were also contacted to find enough speakers when recruitment through the schools failed.
All speakers received a small compensation for participating in the recordings in the form of a cinema ticket or a coupon for a bookstore or a toy store.

3.2 Recordings

To record read speech, the speakers were asked to read texts that appeared on the screen. To elicit speech in the human-machine interaction modality, on the other hand, the speakers were asked to have a dialogue with the computer. They were asked questions that they could also read on the screen and they had received instructions that they could answer these questions freely and that they could speak as long as they wanted.
The recordings were made on location in schools and retirement homes. We always tried to obtain a quiet room for the recordings. Nevertheless, background noise and reverberation could not always be prevented.
The recording platform consists of four components: the microphone, the amplifier, the soundcard and the recording software. We used a Sennheiser 835 cardoid microphone to limit the impact of ambient sound. The amplifier is integrated in the soundcard (M-audio) and contains all options for adjusting gain en phantom power. Resolution is 16bit, which is considered sufficient according to the CGN specifications.
The microphone and the amplifier are separated from the PC, so as to avoid interference between the power supply and the recordings.
Elicitation techniques and recording platform were specifically developed for the JASMIN-CGN project because one of the aims was to record speech in the human-machine-interaction modality. The recordings are stereo, as both the machine output and the speaker output are recorded.
The samples were stored in 16 bit linear PCM form in a Microsoft Wave Format. The sample frequency was 16 kHz for all recordings. Each recording contains two channels: the output from the TTS system (dialogues) and the microphone recording. Notice that the microphone signal also contains the TTS signal through the acoustic path from the loudspeakers to the microphone.
In total 111 h and 40 m of speech were collected divided as follows:
In The Netherlands:

- native children between 7 and 11 (15h 10m)
- native children between 12 and 16 (10h 59m)
- non-native adults (15h 01m)
- non-native children between 7 and 16 (12h 34m)
- native adults above 65 (16h 22m)

In Flanders:

- native children between 7 and 11 (7h 50m)
- native children between 12 and 16 (8h 01m)
- non-native adults (8h 02m)
- non-native children between 7 and 16 (9h 15m)
- native adults above 65 (8h 26m)

About 50% of the material is read speech and 50% extemporaneous speech recorded in the human-machine interaction modality (HMI).

4. Annotations

Given the limited budget available, the annotations were limited to a verbatim transcription, a transcription of the human-machine interaction (HMI) phenomena, POS tagging of the words, and an automatic phonetic transcription.

4.1 Orthographic annotation

All speech recordings were orthographically transcribed manually according to the same conventions adopted in CGN. Since this corpus also contains speech by non-native speakers, special conventions were required, for instance, for transcribing words realized with non-native pronunciation. Orthographic transcriptions were made by one transcriber and checked by a second transcriber who listened to the sound files, checked whether the orthographic transcription was correct and, if necessary, improved the transcription. A spelling check was also carried out according to the latest version of the Dutch spelling (*Woordenlijst Nederlandse Taal*, 2005). A final check on the quality of the orthographic transcription was carried out by running the program 'orttool'. This program, which was developed for CGN, checks whether markers and blanks have been placed correctly and, if necessary, improves the transcription.
The speech material recorded in the Netherlands was also transcribed in the Netherlands, whereas the speech material recorded in the Flanders was transcribed in Flanders. To avoid inconsistencies in the transcription, cross checks were carried out.

4.2 Annotation of Human-Machine Interaction (HMI) phenomena

A protocol was drawn up for transcribing the HMI phenomena that were elicited in the dialogues. The aim of this type of annotation was to indicate these phenomena so that they can be made accessible for all sorts of research and modeling. As in any type of annotation, achieving an acceptable degree of reliability is very important. For this reason in the protocol we identified a list of phenomena that appear to be easily observable and that are amenable to subjective interpretation as little as possible. In addition, examples were provided of the manifestation of these phenomena, so as to minimize subjectivity in the annotation.

4.3 Phonemic annotation

It is common knowledge, and the experience gained in CGN confirmed this, that manually generated phonetic transcriptions are very costly. In addition, recent research findings indicate that manually generated phonetic transcriptions are not always of general use and that they can be generated automatically without considerable loss of information (Van Bael et al., 2003). In a project like JASMIN-CGN then an important choice to make is

whether the money should be allocated to producing more detailed and more accurate annotations or simply to collecting more speech material. Based on the considerations mentioned above and the limited budget that was available for collecting speech of different groups of speakers, the second option was chosen to adopt an automatically generated broad phonetic transcription (using Viterbi alignment).

4.3.1 Acoustic models

Given the nature of the data (non-native, different age groups and partly spontaneous), the procedure requires some care. Since the performance of an automatic speech aligner largely depends on the suitability of its acoustic models to model the data set, it was necessary to divide the data into several categories and treat each of those separately. Those categories were chosen such that the data in each could be modelled by a single acoustic model, making a compromise between intra-category variation and training corpus size. Both for Flemish and Dutch data we therefore made the distinction between native children, non-native children, native adults, non-native adults and elderly people.

Deriving an acoustic model for each category was not a straightforward task, since the amount of available data was not always sufficient, especially for the Flemish speakers. In all cases, we started from an initial acoustic model and adapted that to each category by mixing in the data on which we needed to align.

For children, however, both native and non-native, this solution was not adequate. Since vocal tract parameters change rather drastically during childhood, a further division of the children data according to age at the time of recording was mandatory. We distinguished speakers between 5 and 9 years old, speakers between 10 and 12 years old, and speakers between 13 and 16 years old. These sets of children data were then used to determine suitable vocal tract length warping factors, in order to apply VTLN (Voice Tract Length Normalization) (Duchateau et al, 2006). Because of this, data from speakers of all ages could be used in deriving suitable acoustic models for children data.

To end up with an acoustic model for each of the 10 categories we distinguished in the data, we used four initial acoustic models: Dutch native children (trained on roughly 14 hours of JASMIN data), Flemish native children (trained on a separate database), Dutch native adults (trained on CGN) and Flemish native adults (trained on several separate databases). For each category of speakers, a suitable model was derived from one of these initial models by performing a single training pass on it. For instance, to align the Flemish senior speech, a single training pass was performed on the model for Flemish native adult speech using the Flemish senior data.

4.3.2 Lexicons

The quality of the automatic annotation obtained by the speech aligner depends on the quality of the lexicon used. These lexicons should contain as many pronunciation variants for each word as possible for the Viterbi aligner to choose from. For instance, the “n” at the end of a Dutch verb or plural noun is often not pronounced, especially in sloppy speech. The omission of this “n” should be accounted for in the lexicon.

The base lexicons were Fonilex for Flemish and CGN for Dutch. Additionally, two pronunciation phenomena, which were not present in CGN, were annotated manually in the JASMIN database: pause in a word, (typically in hesitant speech by non-natives, which was annotated orthographically with “*s” following the word) and foreign pronunciation of a word (marked by a trailing *f). The lexicon for these words was created manually in several iterations of inspection and lexicon adaptation. In general, this leads to an increase in the options the Viterbi aligner can choose from. Further modelling of pronunciation variation is in hard-coded rules as in the CGN. An example of such a rule is vowel substitution due to dialectic or non-native pronunciation.

4.3.3 Quality check

The quality of the automatically generated phonemic transcriptions was manually verified for three randomly selected files per Region (FL/NL) and category (non-native child, non-native adult, native child and senior) (a total of 24 recordings) by inspection of the proposed transcription. Lexicon and cross-word assimilation rules were adapted to minimize the number of errors. Most of the required corrections involved hard/soft pronunciation of the “g” and optional “n” in noun plurals and infinitive forms.

4.4 Part-of-speech tagging

For all (orthographic) transcriptions, a part of speech (PoS) tagging was made. This was done fully automatically by using the POS tagger that was developed for CGN at ILK/Tilburg University. Accuracy of the automatic tagger was about 97% on a 10% sample of CGN (Van den Bosch et al. 2006). The tagset consists of 316 tags and is extensively described (in Dutch) in Van Eynde (2004). Manual correction of the automatic POS tagging was not envisaged in this project.

5. Conclusions

Eventually, the realization of the JASMIN-CGN corpus has required much more time than was initially envisaged. The lion share of this extra time-investment was taken up by speaker recruiting. We had anticipated that speaker recruiting would be time consuming because, owing to the diversity of the speaker groups, we had to contact primary schools, secondary schools, language schools and retirement homes in different dialect regions in the Netherlands and Flanders. In addition, we knew that schools are often reluctant to participate in external projects. Nevertheless, speaker recruiting turned out to be more problematic than we had expected. Anyway, one lesson we learned is that while

talking to user groups one should not only ask them about their wishes, but also about the feasibility of what they suggest.

Another thing that we realized along the way is that very often, especially in schools, various forms of research or screening are carried out for which also speech recordings are made of children or non-native speakers. These speech data could be used not only for the studies for which they were originally collected, but also for further use in HLT. The only problem is that, in general, the researchers in question do not realize that their data could be valuable for other research fields. It would therefore be wise to keep track of such initiatives and try to make good agreements with the researchers in charge to ensure that the recordings are of good quality and that the speakers are asked to give their consent for storing the speech samples in databases to be used for further research, of course with the necessary legal restrictions that the data be made anonymous and be used properly. This would give the opportunity of collecting additional speech material in a very efficient and less expensive way.

6. Acknowledgements

The JASMIN-CGN project is carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments (http://taalunieversum.org/taal/technologie/stevin/). We are indebted to the publishers Thieme-Meulenhoff and Zwijsen who allowed us to use their texts for the recordings, to A. van den Bosch who allowed us to use the POS tagger and to all the speakers who participated and thus made it possible to collect this corpus.

7. References

Demuynck, K. Laureys, T. Wambacq, P. and Van Compernolle, D. (2004) Automatic Phonemic Labeling and Segmentation of Spoken Dutch. In Proc. 4th International Conference on Language Resources and Evaluation, 61--64, Lisbon, Portugal.

Duchateau J., Wigham, M., Demuynck, K. and Van hamme, H. (2006) A Flexible Recogniser Architecture in a Reading Tutor for Children. In Proc. ITRW on Speech Recognition and Intrinsic Variation, 59-64, Toulouse, France.

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

Oostdijk, N. (2002) The design of the Spoken Dutch Corpus, in Peters P., Collins P., Smith A. (Eds) New Frontiers of Corpus Research, Rodopi, Amsterdam, 105-112.

Van Bael, C., Binnenpoorte, D. Strik, H. and van den Heuvel, H. (2003) Validation of Phonetic Transcriptions Based on Recognition Performance, Proceedings of Eurospeech, Geneva, Switzerland, 1545-1548.

Van den Bosch, A., Schuurman, I. and Vandeghinste, V. (2006) Transferring PoS-tagging and lemmatization tools from spoken to written Dutch corpus Development, Proceedings LREC 2006, Genoa, Italy.

Van den Bosch, A., Busser, G.J., Canisius, S., and Daelemans, W. (2007). An efficient memory-based morpho-syntactic tagger and parser for Dutch. In P. Dirix, I. Schuurman, V. Vandeghinste, and F. Van Eynde (Eds.), Computational Linguistics in the Netherlands: Selected Papers from the Seventeenth CLIN Meeting, Leuven, Belgium, 99-114.

Van Eynde F. (2004). Part of Speech Tagging en Lemmatisering van het Corpus Gesproken Nederlands. Technical Report, Center for Computational Linguistics, Katholieke Universiteit Leuven, Belgium. http://www.ccl.kuleuven.be/Papers/POSmanual_febr2004.pdf.

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

Woordenlijst Nederlandse Taal (2005) Nederlandse Taalunie, The Hague, http://woordenlijst.org/.