Chapter 3
The JASMIN Speech Corpus: Recordings of Children, Non-natives and Elderly People

Catia Cucchiarini and Hugo Van hamme

3.1 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 recognised 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 nine 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

C. Cucchiarini (✉)
CLST, Radboud University, Nijmegen, The Netherlands
e-mail: c.cucchiarini@let.ru.nl

H. Van hamme
ESAT Department, Katholieke Universiteit Leuven, Leuven, Belgium
e-mail: hugo.vanhamme@esat.kuleuven.be

P. Spyns and J. Odijk (eds.), *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-30910-6_3, © The Author(s) 2013

(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, such as in-car navigation systems and miniaturised personal assistants. 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.

3.2 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 in education and in entertainment [13]. 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 [9], 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 30 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 organisations 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, prioritisation and problem solving. The applications can vary from reminder systems to interactive robotic assistants [8, 10, 12, 18].

3.3 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 realise. As a matter of fact, speech recognisers that are optimised for adult speech are not suitable for handling speech of children, non-natives and elderly people [3, 6, 13, 15, 20]. 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 [20]. 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 [3]. 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.

Problems in ASR for children, elderly and non-natives are generally approached with standard adaptation procedures [3, 13, 15, 20]. 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) showed 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 sections of the population not represented in the corpora used for training the ASR systems, typically children, non-natives and elderly people [13, 15]. Also in spoken human-machine interaction, scientific and technological progress is hampered by the lack of appropriate corpora.

3.4 JASMIN-CGN: Aim of the Project

It is for the reasons mentioned above that within the framework of the Dutch-Flemish programme STEVIN [1] the project JASMIN-CGN was started, which was 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. The JASMIN-CGN project was carried out by a Dutch-Flemish consortium made up of two academic institutions (RU Nijmegen, CLST, C. Cucchiarini and KU Leuven, ESAT, H. Van hamme) and TalkingHome, (F. Smits) a company that, at the time, developed speech controlled applications for health care. The JASMIN-CGN project aimed at realising 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.

3.5 Material and Methods

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'.

3.5.1 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).

3.5.1.1 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.

3.5.1.2 Nativeness

In each of the two regions, three groups of speakers consisted of native speakers of Dutch and two of non-native speakers. For native and non-native speakers different selection criteria were applied, as will be explained below.

3.5.1.3 Dialect Region

Native speakers, on the other hand, 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 3 years before the time of the recording.

Within the native speaker categories we strived for a balanced distribution of speakers across the four regions (one core, one transitional and two peripheral regions) that we distinguished in the Netherlands and Flanders in the sense that we organised recruiting campaigns in each of the regions. However, we did not balance strictly for this criterion, i.e. speakers were not rejected because of it.

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).

3.5.1.4 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, the Francophone population has foreign roots and we hence decided to also allow speakers living in a Francophone environment but whose first language is not French.

In the Netherlands, on the other hand, this type of choice 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 constituted two of the four most substantial minority groups [5], 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 because Dutch L2 classes at the time of recruiting contained more varied groups of learners. This was partly induced by a new immigration law that envisaged new obligations with respect to learning Dutch for people from outside the EU. This led to considerable changes which clearly had an impact on the whole Dutch L2 education landscape. As a consequennce, it was no longer so straightforward to imagine that only one or two 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.

3.5.1.5 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 from adult Dutch L2 learners at these lower proficiency levels.

For children, the current class (grade) they are in was maintained as a selection criterion. So although in this case proficiency was not really a selection criterion, it is correlated with grade to a certain extent.

3.5.1.6 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 not represented in the CGN corpus:

- 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.

3.5.1.7 Speaker Gender

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

3.5.2 Speech Modalities

In order to obtain a relatively representative and balanced corpus we decided to record about 12 min 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.

3.5.2.1 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 7–12 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 [26]. A much used method for this purpose is the reading program Veilig Leren Lezen [11]. 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, Codes 1 and 2, from Thieme Meulenhoff Publishers. The texts were selected as to be suitable for learners with CEF levels A1 and A2.

3.5.2.2 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 by saying, for instance: “Sorry, I did not

understand you", or "Sorry, I could not hear you" so as 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 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 characterisation, 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 by asking unclear questions, increasing the cognitive load of the speaker by asking more difficult questions, or by 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.

3.5.3 Collecting Speech Material

3.5.3.1 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, schools were contacted directly and recruiting children turned out to be much more time-consuming than we had envisaged.

In Flanders, most recordings in schools were organised 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) were invited to participate. Through these schools we managed to contact non-native speakers with the appropriate levels of linguistic skills. Specific organisations 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.

3.5.3.2 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 consisted 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 was integrated in the soundcard (M-audio) and contained all options for adjusting gain and phantom power. Resolution was 16 bit, which was considered sufficient according to the CGN specifications. The microphone and the amplifier were 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 were recorded.

Table 3.1 Amount of speech material and number of speakers per speaker group. The numbers between round brackets are the number of female participants in each group

Speaker group	NL	FL	NL(F)	FL(F)
Native primary school pupils between 7 and 11	15 h 10 min	7 h 50 min	72 (35)	43 (23)
Native secondary school students between 12 and 16	10 h 59 min	8 h 01 min	63 (31)	44 (22)
Non-native children between 7 and 16	12 h 34 min	9 h 15 min	53 (28)	52 (25)
Non-native adults	15 h 01 min	8 h 02 min	46 (28)	30 (19)
Native adults above 65	16 h 22 min	8 h 26 min	68 (45)	38 (22)
Total	70 h 06 min	41 h 34 min	302 (167)	207 (111)

3.6 Results

3.6.1 Speech Files

In total 111 h and 40 min of speech were collected divided over the different speaker groups as shown in Table 3.1. The corpus documentation contains further details about the speakers (exact age, native language, proficiency in Dutch, gender, place of birth, ...). The samples were stored in 16 bit linear PCM form in a Microsoft Wave Format. The sample frequency is 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.

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

3.6.2 Orthographic Annotations

All speech recordings were orthographically transcribed manually according to the same conventions adopted in CGN and using the same tool: PRAAT [2]. Since this corpus also contains speech by non-native speakers, special conventions were required, for instance, for transcribing words realised 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 [14]. 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 but

was not further disseminated, 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.

3.6.3 Annotations of Human-Machine Interaction Phenomena

A protocol was drawn up for transcribing the HMI phenomena that were elicited in the dialogues. This document can be found in the corpus documentation. 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. The following phenomena were transcribed: hyperarticulation, syllable lengthening, shouting, stress shift, restarts, filled pauses, silent pauses, understanding checks, self talk, repetitions, prompt/question repeating and rephrasing. In addition, examples were provided of the manifestation of these phenomena, so as to minimise subjectivity in the annotation.

As for the orthographic transcriptions, the HMI transcriptions were also made by one transcriber and checked by a second transcriber who listened to the sound files, checked whether the transcription was correct and, if necessary, improved it. 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 prevent inconsistencies in the transcription, cross checks were carried out.

3.6.4 Phonemic Annotations

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 [19]. 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, we chose the second option and decided to adopt an automatically generated broad phonetic transcription (using Viterbi alignment).

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 Normalisation) [7]. 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 ten categories we distinguished in the data, we used four initial acoustic models: Dutch native children (trained on roughly 14 h 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.

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.

Quality checks of the automatically generated phonemic transcriptions were carried out by verifying the proposed transcription 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). Lexicon and cross-word assimilation rules were adapted to minimise 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.

3.6.5 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 [21]. The tagset consists of 316 tags and is extensively described (in Dutch) in [25]. Manual correction of the automatic POS tagging was not envisaged in this project.

3.6.6 External Validation and Distribution

The JASMIN speech corpus was validated by BAS Services at the Phonetics Institute of Munich University against general principles of good practice and the validation specifications provided by the JASMIN consortium. The validation had the following aims:

1. Assess the formal correctness of the data files
2. Assess the correctness of the transcriptions and annotations, specifically the orthographic transcriptions, the automatically generated phonemic transcriptions and the HMI annotations.
3. Indicate to what extent transcriptions and annotations were in line with the guidelines laid down in the corresponding protocols.
4. Determine whether the protocols provided adequate information for users of the corpus.

The validation concerned completeness, formal checks and manual checks of randomly selected samples. Data types covered by this validation were corpus structure, signal files, orthographic, phonetic, POS, HMI events annotation and all English documentation files. Manual checks were carried out by native Dutch and Flemish speakers for the orthographic transcript, the phonetic transcript and the HMI event labelling.

The validation results indicated that the JASMIN corpus was of sufficient quality and received a relatively high score (16/20). In addition, minor errors or inaccuracies signaled during validation were subsequently redressed by the JASMIN consortium

before transferring the JASMIN corpus to the Dutch-Flemish HLT Agency, which is now in charge of its management, maintenance and distribution.

3.7 Discussion

Eventually, the realisation 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 realised 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 realise 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.

3.8 Related Work and Contribution to the State of the Art

Since its completion in 2008, the JASMIN corpus has been employed for research and development in various projects. At CLST in Nijmegen the non-native part of the JASMIN corpus appeared to be particularly useful for different lines of research, as will be explained below. Within the STEVIN programme, the JASMIN corpus has been used in the DISCO project (cf. Chap. 18, p. 323 on the DISCO project).

In DISCO the adult non-native JASMIN speech material was used in combination with the SPRAAK toolkit (cf. Chap. 6, p. 95) in research aimed at optimising automatic speech recognition of low-proficient non-native speakers [23]. In addition, the same JASMIN subcorpus was employed in research on automatic detection of pronunciation errors [22] and in research aimed at developing alternative automatic

measures of pronunciation quality [23]. For these purposes the automatically generated phonemic transcriptions of the adult non-native speech material were manually verified by trained transcribers.

Furthermore, the adult non-native part of the JASMIN corpus also appeared to be particularly suited for studying possible differences in pronunciation error incidence in read and spontaneous non-native speech [24] and for investigating fluency phenomena in read and spontaneous speech samples of one and the same non-native speaker [4]. Finally, the JASMIN adult non-native dialogues were successfully employed to the benefit of research on automatic detection of syntactical errors in non-native utterances [16, 17].

Recently, new research started at the Max Planck Institute in Nijmegen which is aimed at studying aging and the effects of lexical frequencies on speech production. For this purpose the elderly speech of the JASMIN corpus will be employed.

Although the above list indicates that the JASMIN speech corpus has already been used for different investigations, it is clear that its use has so far been relatively limited to researchers that had been involved in its realisation. In a sense this is obvious, because the compilers of the corpus know it in detail and are more able to gauge its potential. However, it seems that more effort should be put in raising awareness among researchers of the availability of these speech data and the possibilities they offer for research and development. This is a challenge for the Dutch-Flemish HLT Agency, which is now in charge of the JASMIN speech corpus and its future lifecycle.

Acknowledgements 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, to all the speakers as well as institutions that participated and thus made it possible to collect this corpus and to the people who, at different stages and for different periods, were part of the JASMIN team: Leontine Aul, Andrea Diersen, Joris Driesen, Olga van Herwijnen, Chantal Mülders, August Oostens, Eric Sanders, Maarten Van Segbroeck, Alain Sips, Felix Smits, Koen Snijders, Erik Stegeman and Barry van der Veen.

References

1. http://taalunieversum.org/taal/technologie/stevin/
2. http://www.fon.hum.uva.nl/praat/
3. Anderson, S., Liberman, N., Bernstein, E., Foster, S., Cate, E., Levin, B., Hudson, R.: Recognition of elderly speech and voice-driven document retrieval. In: Proceedings of the ICASSP, Phoenix, USA (1999). Paper 2060
4. Cucchiarini, C., van Doremalen, J., Strik, H.: Fluency in non-native read and spontaneous speech. In: Proceedings of DiSS-LPSS Joint Workshop, Tokyo, Japan, pp. 15–18 (2010)

5. Dagevos, J., Gijsberts, M., van Praag, C.: Rapportage minderheden 2003; Onderwijs, arbeid en sociaal-culturele integratie. SCP-publicatie 2003–13, Sociaal en Cultureel Planbureau, The Hague (2003)
6. D'Arcy, S.M., Wong, L., Russell, M.J.: Recognition of read and spontaneous children's speech using two new corpora. In: Proceedings of ICSLP, Korea pp. 588–591 (2004)
7. Duchateau, J., Wigham, M., Demuynck, K., Van hamme, H.: A flexible recogniser architecture in a reading tutor for children. In: Proceedings of ITRW on Speech Recognition and Intrinsic Variation, Toulouse, France, pp. 59–64 (2006)
8. Ferguson, G., et al.: The medication advisor project: preliminary report. Technical Report 776, CS Department, U. Rochester (2002)
9. Hagen, A., Pellom, B., Cole, R.: Children's speech recognition with application to interactive books and tutors. In: Proceedings of ASRU, St. Thomas, USA, pp. 265–293 (2003)
10. Hans, M., Graf, B., Schraft, R.: Robotic home assistant care-o-bot: past-present-future. In: Proceedings of the IEEE ROMAN, Berlin, pp. 380–385 (2002)
11. Mommers, M., Verhoeven, L., Van der Linden, S.: Veilig Leren Lezen. Zwijsen, Tilburg (1990)
12. Müller, C., Wasinger, R.: Adapting multimodal dialog for the elderly. In: Proceedings of the ABIS-Workshop 2002 on Personalization for the Mobile World, Hannover, Germany, pp. 31–34 (2002)
13. Narayanan, S., Potamianos, A.: Creating conversational interfaces for children. IEEE Trans. Speech Audio Process. **10**(2), 65–78 (2002)
14. Nederlandse Taalunie: Woordenlijst nederlandse taal (2005). http://woordenlijst.org/
15. Raux, A., Langner, B., Black, A., Eskenazi, M.: LET'S GO: improving spoken dialog systems for the elderly and non-natives. In: Proceedings of Eurospeech, Geneva, Switzerland, pp. 753–756 (2003)
16. Strik, H., van de Loo, J., van Doremalen, J., Cucchiarini, C.: Practicing syntax in spoken interaction: automatic detection of syntactic errors in non-native utterances. In: Proceedings of the SLaTE-2010 Workshop, Tokyo, Japan (2010)
17. Strik, H., van Doremalen, J., van de Loo, J., Cucchiarini, C.: Improving asr processing of ungrammatical utterances through grammatical error modeling. In: Proceedings of the SLaTE-2010 Workshop, Venice, Italy (2011)
18. Takahashi, S., Morimoto, T., Maeda, S., Tsuruta, S.: Spoken dialogue system for home health care. In: Proceedings of ICSLP, Denver, USA (2002), pp. 2709–2712
19. Van Bael, C., Binnenpoorte, D., Strik, H., van den Heuvel, H.: Validation of phonetic transcriptions based on recognition performance. In: Proceedings of Eurospeech, Geneva, Switzerland (2003), pp. 1545–1548
20. Van Compernolle, D.: Recognizing speech of goats, wolves, sheep and . . . non-natives. Speech Commun. **35**(1–2), 71–79 (2001)
21. Van den Bosch, A., Schuurman, I., Vandeghinste, V.: Transferring pos-tagging and lemmatzation tools from spoken to written dutch corpus development. In: Proceedings of LREC, Genoa, Italy (2006)
22. van Doremalen, J., Cucchiarini, C., Strik, H.: Automatic detection of vowel pronunciation errors using multiple information sources. In: Proceedings of ASRU, Merano, Italy, (2009), pp. 80–585
23. van Doremalen, J., Cucchiarini, C., Strik, H.: Optimizing automatic speech recognition for low-proficient non-native speakers. EURASIP J. Audio, Speech, Music Process. (2010). http://www.hindawi.com/journals/asmp/2010/973954/
24. van Doremalen, J., Cucchiarini, C., Strik, H.: Phoneme errors in read and spontaneous non-native speech: relevance for capt system development. In: Proceedings of the SLaTE-2010 Workshop, Tokyo, Japan (2010b)
25. Van Eynde, F.: Part of speech tagging en lemmatisering van het corpus gesproken nederlands (2004). http://lands.let.ru.nl/cgn/doc_Dutch/topics/version_.0/annot/pos_tagging/tg_prot.pdf
26. Wentink, H.: From graphemes to syllables. Ph.D. thesis, Nijmegen (1997)