A Spoken Afrikaans Language Resource
Designed for Research on Pronunciation Variations

Daan Wissing[1], Jean-Pierre Martens[2], Ulrike Janke[1], Wim Goedertier[2]

(1) Human Language Technology Laboratory (HLT-L), North-West University, Potchefstroom Campus,
Private Bag X6001, Potchefstroom, South Africa. ntldpw@puk.ac.za
(2) ELIS, Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium. martens@elis.ugent.be

Abstract

In this contribution, the design, collection, annotation and planned distribution of a new spoken language resource of Afrikaans (SALAR) is discussed. The corpus contains speech of mother tongue speakers of Afrikaans, and is intended to become a primary national language resource for phonetic research and research on pronunciation variations. As such, the corpus is designed to expose pronunciation variations due to regional accents, speech rate (normal and fast speech) and speech mode (read and spontaneous speech). The corpus is collected by the Potchefstroom Campus of the North-West University, but in all phases of the corpus creation process there was a close collaboration with ELIS-UG (Belgium), one of the institutions that has been engaged in the creation of the Spoken Dutch Corpus (CGN).

1. Introduction

In this paper the design, collection, annotation and distribution of a new spoken language resource of Afrikaans is described. The main goal was to construct a corpus that will be valuable to linguists who want to investigate phonological phenomena and the phonetic realization thereof, as well as to engineers who want to develop acoustic and lexical models for e.g. automatic speech recognition (ASR). Consequently, the corpus design and annotation strategy has been developed with these two user groups in mind.

The new Spoken Afrikaans Language Resource (SALAR) is collected and orthographically/phonetically transcribed by linguists at the Human Language Technology Laboratory (HLT_L) at the Potchefstroom Campus of the North-West University, South Africa. However, the project is realized in close collaboration with ELIS-UG (Belgium), one of the groups that was also engaged in the creation of the Spoken Dutch Corpus (CGN) (Oostdijk et al, 2002). ELIS offered advice and computer tools to speed up the selection, segmentation, transcription and formal checking of the material. The HLT_L-ELIS collaboration has also proven to be very fruitful because the two teams look upon the design and exploitation of a speech corpus from two different angles.

The corpus collection and annotation is still going on, but the whole corpus is anticipated to be ready by the end of 2004.

2. Corpus design

First of all, the corpus is to contain only speech of adult (above the age of 18) mother tongue speakers of Afrikaans. Furthermore, the corpus is designed so as to expose the relations between phonetic variation and factors such as regional accent, gender, speech mode (read versus spontaneous) and speech rate (normal versus fast). To that end, the corpus is designed on the basis of the following principles:

(1) The speakers are sampled according to region, and, as far as possible, also according to race[1] so as to attain as rich a pallet of accents as possible (colored people usually have developed another accent than white people of the same region).

(2) Since there is evidence (Wissing, 1996) that some pronunciation phenomena in Afrikaans are gender dependent, the speakers are also balanced per gender.

(3) In order to expose the effects of speech mode, the same subjects are recorded in three different modes. They provide normally read speech, fast read speech and spontaneous speech. To make the effects of speech rate even more apparent, the fast read speech concerns a fragment of a text passage that is already read at a normal rate by the same speaker.

Mainly for practical reasons (we only had limited resources for developing this corpus), about half of the speakers are students between 18 and 26 years of age. However, for the selection of the other speakers, we do use age as a selection criterion on top of region and gender: half of the non-students are older than 47. The students are coming from all regions of South Africa and as such they are representative of the Afrikaans-speaking community at large.

Since ELIS is also interested in lexical pronunciation modeling for non-native speech recognition, and since Afrikaans speaking persons can read Dutch[2], the persons contributing to the corpus are also asked to read a short text passage of about 50 words from a book in Dutch. These noticeably accented Dutch recordings will constitute a separate corpus that is not integrated in the

[1] Afrikaans is spoken as a first language by two distinct groups: whites and coloreds.

[2] Dutch and Afrikaans are very proximate languages

SALAR. This small Dutch corpus will also be transcribed (be it only orthographically) by ELIS.

3. Data collection

When the data collection will be finished, the corpus will contain read speech from 150 subjects, as well as spontaneous speech of 90 of these 150 subjects. The material is organized into two sub-corpora which will in the end have the following characteristics:

(1) The first sub-corpus SALAR-RA (RA = Read Afrikaans) will contain about 3 hours of read speech originating from all the 150 speakers. Per speaker there will be 50 – 60 seconds of normally read speech (about 200 words) plus approximately 12 seconds of fast reading (an extract of about 50 words out of the passage that was already normally read before).

(2) The second sub-corpus SALAR-SA (SA = Spontaneous Afrikaans) will contain about 6 hours of spontaneous speech originating from interviews with 90 of the 150 speakers that have contributed to SALAR-RA. More speech will be recorded, but only four minutes per speaker will be annotated for the time being.

At the time of writing of this paper, the data collection is just over half way and is progressing fine.

3.1. Text material

It is common to construct a corpus for phonetic research on the basis of a large list of a phonetically rich sentences (e.g. Lamel et al, 1986). Although there exists a list of 800 such sentences for Afrikaans, we argue that by letting every speaker read a different passage from a large book collection, we can obtain more diverse speech which is (1) representative of written Afrikaans at large, and (2) exhibiting about the same phoneme statistics as those emerging from phonetically rich sentences.

The majority (130) of the selected text passages were taken from 40 books which were randomly chosen from an electronically available corpus of recent publications of the Afrikaans publishing house Protea Boekhuis[3]. The other 20 passages were retrieved from the North-West University news archive.

AST corpus		SALAR corpus	
@	13.5	@	12.3
r	8.2	r	7.1
s	7.8	n	7.0
t	6.8	s	6.9
n	6.5	t	6.5
l	4.9	d	5.0
k	4.4	A	4.3
a	4.1	i	4.2
e	4.1	l	4.1
d	4.0	k	3.8

Table 1: Phoneme frequencies (in %) in the AST and SALAR corpus

Table 1 shows the 10 most frequent phonemes of both the SALAR corpus and the Afrikaans part of the AST corpus[4], a SpeechDat-like corpus containing speech spoken in several South African languages. The SALAR phoneme frequencies were derived from canonical phonetic transcriptions of all the words appearing in the text passages. Generally there is a good agreement between the two frequency lists (the correlation between the two is about 0.9). Both corpora show the same top five phonemes (be it in another order).

Except for the phonemes /z/ (only appearing in loan words), /g/, /J/ and /S/ which appear only 5, 10, 13 and 52 times respectively, all other phonemes appear between 270 and 14000 times. This should be sufficient for the study of pronunciation phenomena with regard to these phonemes, and for the training of reasonable acoustic models (for ASR).

3.2. Spontaneous speech

The spontaneous speech of a speaker is collected during an interview. The interviewer is either the second author or one of three HLT assistants. The topics are chosen on the basis of familiarity of the interviewee with the topic. By only talking about topics that are familiar to the speaker, we have achieved that most of the interviews that have been recorded so far last significantly longer than the targeted four minutes, and this without much interference from the interviewer.

3.3. Recordings

Our aim was to collect high quality broadband speech. If we look at other corpora with a similar scope that were recorded for other languages – such as TIMIT (Lamel et al, 1986) for American English - we observe that these corpora usually contain speech that is sampled at 16 kHz and encoded in 16 bit linear PCM. Due to the invention of the CD however, a new frequency standard of 22.05 kHz is pushing aside the old standard of 16 kHz. In view of this all our recordings are made with professional equipment, and the signals are sampled at 22.05 kHz and stored directly on the hard disc of a computer in 16 bit linear PCM (in wave format).

In order to keep maximal control over the acoustic conditions and the background noise level, we record as much as possible in the sound-treated boot of the HLT-Laboratory. In these cases where we are forced to make the recordings out of the laboratory (especially for obtaining the recordings of persons outside the university community), we carefully select venues with a low background noise level and a low degree of reverberation.

4. Data transcription

4.1. Meta-data

Once a recording is completed and the sound files are created, a set of meta-data is collected per speaker and per recording.

The meta-data of a fragment in SALAR-RA consists of information about the book (title, author, ISBN number, etc.), the selected passage (the text that was read), the word count, and the ID number of the speaker (this ID refers to the speaker database). The meta-data of a

[3] With consent of the publisher

[4] www.ast.sun.ac.za/the_project.htm

spontaneous speech fragment consists of the word count and the ID numbers of the interviewer and the speaker respectively.

The meta-data of a speaker consists of the first three digits of the postal code of his/her home address, gender, place of birth, age at the time of recording, smoker/non-smoker and hearing ability (normal versus affected, and in the case of affected hearing, a short description of the problem). Also recorded are : a list of all the towns where school was attended, information concerning the language most often spoken at home and at work, highest qualification and profession.

4.2. Orthographic transcription

The whole corpus is transcribed orthographically on the basis of a protocol that is largely inspired by the protocol for orthographic transcription of the CGN (Goedertier et al, 2000). The fundamental goals are to adhere as much as possible to written language, and to use different tags to encode spoken language phenomena like abbreviations, miss-pronunciations, etc.

The transcription is performed in the open software Praat (Boersma & Heuven, 2001). It is time aligned with the signal at the level of complete sentences (for the read speech part) or chunks of just a few seconds long (for the spontaneous speech part).

In the case of read speech, the speech is first manually segmented into speech intervals corresponding to the subsequent sentences of the text passage and non-speech intervals representing clearly distinctive pauses between sentences. The successive sentences of the text passage are then put into the successive speech intervals so as to produce an initial orthographic transcription that is then to be corrected by the transcriber where necessary.

In the case of spontaneous speech, the segmentation in chunks is completely left to the transcriber. According to the protocol the transcriber should preferably insert a chunk boundary at the end of a pause (if the pause is short) or at both ends of a pause (if the pause is long enough). If the so obtained chunks are too large (e.g. more than 10 seconds), extra chunk boundaries are inserted at places where there seems to be at least some break between two words.

At present, we have orthographically transcribed most of the recordings we have made so far.

4.3. The lexicon

The SALAR corpus comes with a lexicon containing all the words appearing in the corpus, together with their canonical phonetic transcription. How this lexicon is constructed in an incremental way is discussed below.

4.4. Phonetic transcription

Once the orthographic transcription is available, a phonetic transcription of each utterance is created according to rules and procedures outlined in a protocol that is very similar to the CGN protocol for phonetic transcription[5]. The envisaged transcription is basically a broad phonetic transcription on the basis of a restricted symbol set, namely, a version of SAMPA[4] for Afrikaans. That version was developed by the first author and is very similar to the version for Dutch that was adopted in the CGN project. The only fundamental difference between the CGN and SALAR protocols is that in the SALAR protocol allows the transcriber to annotate the nasalization of any vowel by inserting a diacritic symbol /~/ whereas in the CGN, this symbol can only occur in combination with a restricted and predefined set of vowels.

Just as in the CGN, the phonetic transcription is kept synchronized with the orthographic transcription at the level of the words. For that purpose, the protocol mentions the word junction conventions to use (e.g. how to indicate that a phoneme is shared by two words, and how to annotate the insertion of a phoneme between two words). Since preference is given to the orthographic transcription, only a very few files have been transcribed phonetically thus far.

4.4.1. Canonical transcriptions

In order to speed up the phonetic transcription, a tool for the automatic generation of an initial canonical phonetic transcription was developed. The tool searches in a lexicon for the words encountered in the orthography of a chunk, and it puts the phonetic transcriptions of these words (as found in that lexicon) in the canonical phonetic transcription of that chunk. In the case a word cannot be found in the lexicon, this word is added to an out-of-lexicon word list and its orthographic transcription, put between square brackets, is inserted in the canonical phonetic transcription. This generated transcription must then be manually verified and corrected by the transcriber.

When processing a new list of files, we first run our tool so as to produce the corresponding out-of-lexicon list, and we then supply this list to a grapheme-to-phoneme (g2p) converter for Afrikaans (see section 4.4.2) that produces phonetic transcriptions of these words. We manually verify them and add the words with their verified transcriptions to the SALAR lexicon. By then re-running the canonical transcription generation tool with the new lexicon we can then construct the final canonical transcriptions of the files and ask the transcribers to modify these.

By working in the outlined way, the construction of the SALAR lexicon is achieved incrementally. For producing the canonical phonetic transcriptions of the lexicon entries, we use the g2p PATANA.

4.4.2. Afrikaans grapheme-to-phoneme converter

PATANA is a rule-based g2p that was developed by the first author and that was previously used successfully in the AST corpus annotation process. It comprises three modules working on the basis of three types of rules: (1) syllable boundary insertion rules, (2) stress assignment rules and (3) rules for translating graphemes into phonemes. The modules are called in the order from (1) to (3), which is exactly the opposite of what is proposed in (Daelemans & Van den Bosch, 2001). It is our experience that PATANA usually produces good initial guesses of the canonical phonetic transcriptions of the isolated words we want to add to the lexicon.

5. Potential users of the corpus

A first category of users are phonologists. It is generally claimed, be it in an impressionistic way, that the

[5] http://lands.let.kun.nl/cgn

realization of phonological rules is especially sensitive to speech rate: the faster one speaks the more likely it is that a number of phonological processes will surface. Precise details concerning this supposition are still lacking for Afrikaans, and, as a matter of fact, also for most if not all other languages. The hope is that this corpus will help phonologists to provide more clarity on this issue.

Concerning pronunciation variation modeling for ASR, it is known that a considerable amount of phonetic variation can be captured in context-dependent phoneme acoustic models, but some researchers (e.g. Cremelie & Martens, 1999) argue that a number of phenomena like the frequent deletions of phonemes in fast and spontaneous speech, may better be handled at the level of the lexicon. However, doing so requires the development of distinct lexical models for normal and fast speech. A corpus like the SALAR – putting in opposition read speech at a normal and a fast rate, and contrasting read speech with spontaneous speech of the same speakers - could be very helpful in this respect.

Given that the corpus is phonetically transcribed, the read speech part can also represent a useful corpus for engineers who want to use it for the training of good initial acoustic models for the automatic recognition of broadband continuous speech in Afrikaans.

6. Data distribution

Since the corpus collection was sponsored with public money, and since we have the consent of the publisher for using the short text passages appearing in SALAR-RA, and since we also have the consent of the speakers that we can use their speech, there are no known legal obstructions for distributing the corpus as soon as it will be completely ready.

The present idea is to start negotiations with ELDA about the conditions of distribution of the corpus at a fair price. In the event we would be unable to come to an agreement, some kind of distribution via the worldwide web is the other option we would consider.

In the mean time, a restricted set of sound files with annotations, as well as copies of the protocols, the meta-data and the lexicon can be freely accessed on the HLT-L website (http://www.puk.ac.za/HLT_Resources).

7. Future work

By means of an automatic alignment tool it must be possible at the end of the project to align the canonical phonetic transcriptions of the chunks with the manually verified phonetic transcriptions of these chunks. Such a tool can be based on Dynamic Programming, and a good cost function for penalizing discrepancies between the aligned phonetic transcriptions. Our plan is to build such a tool as a first step towards the development of an exploitation tool that can further analyze the results of this alignment so as to test proposed theories of phonological phenomena. By making available the sources of that tool to people who want to enhance or extend it, and by demanding that code derived is made publicly available as well, we can hope that better and better exploitation of the data by linguists becomes possible.

8. Comparison with other data collections

We already referred to African Speech Technology project (see section 4.4.) that also collected Afrikaans speech. However, that project differed considerably from the present one in that (1) it focused on the role of idiosyncratic linguistic and pragmatic features of all the languages spoken in South Africa, and how these features are to be accommodated within the creation of applicable speech corpora, and (2) it collected mostly well prepared simple speech utterances over the telephone. The AST project aimed at creating a corpus for the development of voice-driven tele-services, such as a hotel booking service. The design and annotation adhered to the international SpeechDat conventions (http://www.speechdat.org/).

9. Conclusion

Although the project is still running, there are enough reasons to believe that by the end of 2004, there will be a corpus for spoken Afrikaans that has the potential to become a valuable tool for phonetic research and research on pronunciation variation in speech. The corpus will incorporate about 9 hours of broadband speech that will be fully phonetically transcribed. As such, the corpus may also become an interesting resource for the development of good initial acoustic models of broadband continuous speech.

10. Acknowledgement

The work described in this paper is carried out in the context of a project entitled “Lexicon development for ASR of native and non-native speakers” which is granted under the Flemish – South African bilateral agreement. An additional HLT-L grant from the Research Focus Area "Languages and Literature in the SA context" is hereby acknowledged as well.

11. References

Boersma, P. and Heuven, V. van, 2001. Speak and un-Speak with Praat. *Glot International*, 5:341-347.

Cremelie N., Martens J.P. (1999). "In search for better pronunciation models for speech recognition," Speech Communication 29 (2), 115-136.

Daelemans, Walter, and Antal van den Bosch. 2001. "TreeTalk: Memory-based word phonemisation" In: Damper (Ed.) Data-Driven Techniques in Speech Synthesis. Kluwer, 149-172.

Goedertier W., Godijn S., Martens J.P. (2000). "Orthographic transcription of the Spoken Dutch Corpus," Proceedings LREC (Athens), 909-914.

Lamel L., Kassel R., Seneff S. (1986). “Speech database development: design and analysis of the acoustic-phonetic corpus”, Proceedings DARPA Speech Recognition Workshop (edt. Baumann), 100-109.

Oostdijk N., Goedertier W., Van Eynde F., Boves L., Martens J.P., Moortgat M., Baayen H. (2002). "Experiences from the Spoken Dutch Corpus," Proceedings LREC (Las Palmas), 340-347.

Wissing, D.P. (1996). “Regressiewe stemassimilasie in die Afrikaans van Tswanamoedertaalsprekers” Suid-Afrikaanse Tydskrif vir Taalkunde / South African Journal of Linguistics, Suppl 33, 14(4): 150-153