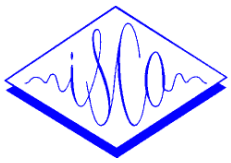

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

CREATION AND ANALYSIS OF THE DUTCH POLYPHONE CORPUS

M. Damhuis[1], T. Boogaart[1], C. in 't Veld[2], M. Versteijlen[2], W. Schelvis[1], L. Bos[2], L. Boves[1,2,3]

[1]PTT Research, [2]SPEX, [3]Nijmegen University
e-mail M.H.Damhuis@research.ptt.nl

ABSTRACT

In this paper the linguistic design, speaker selection, and the recording and transliteration procedures for the Dutch POLYPHONE corpus are described in some detail. Over 5,000 have been recorded. The paper gives details of the distributions of the speakers according to regional and socio-economic background, sex and age of the speakers. Also, first results of the analysis of the linguistic contents of the recordings are reported.

I. INTRODUCTION

Large corpora of speech recorded over the public switched telephone network have become of crucial importance for the progress of Research & Development in speech technology. In the fringe of ICSLP-'92 COCOSDA defined guidelines for such a corpus, which should be recorded for as many different languages as possible. In the course of 1993 these guidelines were made more concrete by the Linguistic Data Consortium, who specified a recording protocol for American English and Spanish. Taking those protocols as a point of departure, PTT Research and the Speech Processing Expertise Centre SPEX set out to record the Dutch POLYPHONE corpus.

This paper starts with a summary description of the Dutch POLYPHONE corpus. First, information is given about the distribution of speakers over cells defined according to regional, social and personal characteristics. It appears that we have not been able to reach our goal of essentially uniform distribution and we explain why this happened. Next we give data about the linguistic contents of the corpus. One of the aims that we had in mind was research into the way in which speakers of Dutch express concepts, like answers to *yes/no*-questions, telephone numbers, dates and times, etc. In this paper first results of in-depth analysis of the spoken replies are presented. The paper ends with some conclusions and recommendations for future speech corpus recording projects.

II. THE DUTCH POLYPHONE CORPUS

The recording workstation used for POLYPHONE was based on an Aculab MVIP/PEB E1/G703 PC Card with 1TR6 isdn-30 signaling (i.e., a primary rate German ISDN 2MB/s connection) for the telephone interface, a Rhetorex Voice Card and driver software, Show-'n-Tel application development software, and a 16 port operational license, in an OS/2 PC. Data files were copied to a Unix network for transliteration and archiving.

The recording platform is set up to store the speech signals in 8 bit A-law coded samples at a sampling rate of 8 kHz. In the Dutch PSTN it is guaranteed that a speech signal with an ISDN connection as its destination remains in an A-Law coded digital form after the first major network switch that it encounters. Thus, the acoustic quality of the recordings is completely determined by the characteristics of the caller's local loop and the background noise in the caller's location. Post-Processing was done at the Dept. of Language & Speech, Nijmegen University, using software running on a PC under MS-Windows, equipped with a Pro-Audio board. The software supporting transliteration was developed jointly by PTT Research and SPEX.

When post-processing for a new speaker begins, the transcriber must first listen to the speaker's identification code, and enter that code into the program. The server maintains a data base with information about the prompts which each speaker has answered. Whenever the answer is predictable (i.e., in all cases where the caller is supposed to read preprinted material), the expected answer is displayed on the screen.

III. SPEAKER SELECTION

Prospective callers were sent a personalized letter. We contracted with a direct mail company, that made an initial random selection of addresses from their data base. In principle, this company would be able to select callers from neighbourhoods which are known to be occupied by a specific socioeconomic group with overrepresentation of certain age groups. However, we never got round to asking them to select such neighbourhoods.

It appeared that the response in this way of approaching prospective respondents was only 15%. Since each letter costs approximately $ 1, we decided that we could not stay within the project budget unless we could increase the response rate considerably. Therefore, we approached a number of acquaintances, who were asked to provide us with address lists of people they knew, so that we could send these persons extra personalized invitations to participate. Although the request for addresses was successful, the plan failed, because the direct mail company that processed these addresses did not include the invitation letter signed by the person who provided the names.

In the final stage of the project subjects were recruited from the personnel of the Dutch PTT. Thanks to intensive publicity and wide press coverage, this increased the response rate to well over 30% in the last group of subjects. In the very last stage of the sampling process we were able to oversample the regions which until then were heavily underrepresented.

Originally, we aimed at collecting 5000 speakers, uniformly divided over a large number of cells, defined according to four criteria, viz. (1) geographical region, (2) socioeconomic status, (3) sex, and (4) age. Geographical region, operationalized as the province in which the speaker lives, is the best practically feasible approximation to regional accent and dialect background. By sampling provinces, we sidestep the unsolved problems of how many different regional accents should be distinguished and how these should be defined. However, the Dutch population is quite mobile; this adds

Table 1: Distribution of respondents over geographical regions

Province	inhabitants (*1,000)	respondents males	females
Groningen	555.2	139	212
Friesland	601.8	135	164
Drente	445.6	140	120
Overijssel	1,032.4	160	169
Flevoland	232.8	80	88
Gelderland	1,828.8	302	291
Utrecht	1,037,3	215	206
Noord Holland	2,421.7	349	311
Zuid-Holland	3,271.5	464	379
Zeeland	359.2	157	110
Noord Brabant	2,225,3	291	259
Limburg	1,115.5	184	125
Totals	15,128.1	2,616	2,434

to the difficulty of predicting dialectal background of callers from their present location. Because we think that knowledge about dialectal background is important, we decided to ask the callers to tell us in what part(s) of the country they grew up. Due to the very uneven distribution of the population over provinces it appeared to be practically impossible to get equal numbers of speakers from each province.
Socioeconomic status is difficult to define, and even more difficult to assess reliably from what respondents are willing to say. We decided to approximate status on the basis of the highest education level of the respondents. To avoid complicated and confusing questions, we settled for a division into three education classes, only elementary school, secondary school, and college/university level education. Using hindsight, this decision proved to be somewhat unfortunate: since the mid-fifties youngsters in the Netherlands are obliged to follow classes up to the age of 16, so that practically everybody under the age of 55 has had more than just elementary school. Thus, it is not surprising that we were able to recruit only 256 speakers who said that they had no more than elementary school. 2600 callers said that they have secondary level education, while 2194 claimed college level education.
We distinguish four age classes, i.e., under 20, between 21 and 40, between 41 and 60, and 61 and older. Information about age is acquired by asking the respondents for their year of birth. Since we also set a minimum age of 16 for participation, the under 20 group is necessarily much smaller than the other groups (168 respondents). For similar reasons, the group of 61 and older is underrepresented with 457 respondents. The group between 20 and 40 comprises 2686 speakers, whereas the group between 40 and 60 consists of 1739 callers.
Information about speaker sex is obtained by asking the respondents to say whether they are male or female. The distribution of the speakers with respect to sex and regional background, based on a total of 5,050 respondents, is shown in Table 1.

IV. THE SPEECH MATERIAL

The speech material recorded in the POLYPHONE project consists of 32 read items, 14 extemporaneous answers to printed questions, and 4 extemporaneous answers to questions not printed on the response sheet.

The material **to be read** consists of the following items:
5 digit strings (one telephone number, two bank accounts or credit card numbers, one string of isolated digits, and the participation number)
3 natural numbers
3 guilder amounts
2 city names
4 application words
3 spelled words
1 date
1 time
1 amount
1 postal code
4 sentences with an application word
5 phonetically rich sentences

The following list of *printed* questions is asked:
Is Dutch your native language?
Did you ever live outside the Netherlands?
Would you willing to participate in another study like this one?
What is your last name?
What is your house number?
What is the name of the street you are living?
What is your postal code?
In which city do you live?
In which cities did you grow up?
Are you a man or a woman?
What is your age?
Which code (1, 2, or 3) represents your education level (1=primary school, 2=high school, 3=college/university)?
Please, say a familiar phone number.
Please, give your comments about this recording session.

The following *unprinted* questions are asked:
Please, spell your name.
Are you calling from your home phone?
Are you using a cordless telephone?
What time is it now?

4.1 Construction of the texts

The text material to be read was carefully constructed in order to maximize the coverage in linguistic sense. Numbers have been constructed in such a way that all digits appear approximately with equal frequency, with one restriction: in order to balance the number of *teen'* and *ty* forms, the digit *1* had to be overrepresented. Care has been taken to prevent unreasonable combinations of amounts and units, like *1,567,329 mm.*
In order to obtain approximately equal numbers of tokens for all letters of the alphabet, a greedy search algorithm was used to selects words from an electronically readable dictionary (provided by CELEX) in such a way that the least frequent letters would occur at least 120 times, while the frequency of occurrence of more frequent would be minimized. In doing so, we ended up with a list of 797 words. Only the words shorter than 11 and longer than 5 characters are selected.
We have created a list containing the names of all train stations in the Netherlands, to which the names of all Dutch communities with more than 5,000 inhabitants were added. The lists of city names was completed by adding all European capitals, and the biggest cities on other continents. Foreign cities are represented by their Dutch names, whenever such a form exists.
We have designed a list of over 1000 words that may be used

in applications based on isolated word recognition systems. To enable us to study the effects of context, all these words are also be embedded in sentence. All callers read the application words twice: once isolated and once embedded in a sentence.
There are many ways to write and express dates in Dutch. We want to catch all variations. Therefore, we have printed the dates using a range of different notations, viz. (a) (Monday) 1 August 1996; (b) (Monday) 01-08-96; and (c) (Monday) 1 August '96.
Times can also be printed and expressed in different ways, for example as (literally translated into English): 10:15 a quarter past ten, ten hours fifteen, fifteen past ten; 18:40 eighteen hours forty, six hours forty, ten past half seven, twenty to seven. Here too, the response sheets show a number of different formats.
For each application word a short sentence (minimally four words, totalling less than 80 characters, including blanks) was constructed. Care has been taken to create sentences that would seem reasonable in an Interactive Voice Response application.
We decided to try to record all phonemes of Dutch in as many different phonetic contexts as possible. At the same time, we want to record all phonemes from each speaker. To that end, a large number of sets of five sentences was constructed, in such a way that each set contains all phonemes at least once. Since the frequency of occurrence of phonemes in Dutch is heavily skewed, large numbers of sentences had to be scanned in order to fulfil the requirement. In fact, we have not yet succeeded in constructing enough such sets to be able to offer a different set to each caller. By scanning an electronic newspaper (Trouw) and adding by hand sentences containing the least frequent phonemes, we succeeded so far in collecting 2500 x 5 = 12.500 sentences.
All sentences consist of at least four words, with a maximum of 80 characters. Moreover, all sentences have been checked for possibly offending contents or words. The resulting set of sentences was processed by a grapheme-to-phoneme converter, in order to be able to compose sets of five with all phonemes. The output of our grapheme-to-phoneme converter was checked by hand and corrections were made when necessary.

V. POSTPROCESSING

Postprocessing consists of four steps, viz. (1) word-by-word transliteration of all items, (2) transliteration of extra sounds and noises, (3) collecting demographic data, (4) assessing the quality of all items. The students who carry out the work are instructed to do the tasks in exactly this order.

5.1 Transliteration

The transliterators are presented with a best guess (in most cases the prompt text printed on the response sheet) of what the speaker has spoken. If the real speech deviates from that text, an efficient editor can be invoked to make all necessary corrections. In transliteration only lower case spellings of conventional lexicalized forms are used. No attempt is made to represent pronunciation differences on a phonetic or phonemic level.

5.2 Extra Sounds

A closed set of extra sounds has been defined, following the guidelines in the American MACRPHONE project. Extra sounds that can be located in time relative to the words are placed accordingly in the transcript. Under this heading background noise accompanying the speech is also marked, as long as the speech is clearly audible; if that is not the case, the response in classified as *Noise* (cf. 5.5).

Table 2: Distribution of responses for telephone numbers

Answer type	read	spontaneous
Only digits	34%	23%
Digits plus		
Numbers	58%	68%
Hundreds	8%	8%

5.3 Quality Assessment

The last thing the transliterators do is to select a quality indication for each item from a fixed menu, offering the options *O.K., Other, Garbage,* and *Noise.* The verdict *O.K.* is given to each response that contains only relevant speech, without overt hesitations, etc. *Other* is assigned to relevant responses that do contain hesitations, self-repairs, stutters, etc. The verdict *Garbage* is given when the caller did respond, but with meaningless speech. Finally, *Noise* was assigned to the items which contained only background noise.
In total, 96.55 of all items was judged *O.K.*, 3.15% got the label *Other*, 0.2% was judged as *Garbage*, and 0.09% as *Noise.* Thus, it can be seen that on average the quality of the recordings is quite high.

VI. RESULTS

In this section a number of results of analysing the material actually spoken are presented.

6.1 Telephone Numbers

Two items related to telephone numbers were analysed. The first pertains to numbers read from the response sheet. All these numbers were printed in the same format, i.e., area code, dash, subscriber number (e.g. 020 - 5252183). The second item consists of answers to the question *Please, say a familiar telephone number.* In discussing the results we will use the term *digit* for the words *zero, one, ..., nine*; the term *number* will denote numbers between 10 and 99.
Presently, the Dutch PTT's number plan has two groups of area codes, one comprising three digits (like 020 in the example above) and one comprising five digits (e.g. 08894). The initial 0 is equivalent to the 1-prefix for area codes in the American telephone network. Subscriber numbers can have from four (only with five digit area codes) to seven digits (with a small number of three digit area codes). Because transliteration does not include intonation markers, it is not possible to discriminate between three and five digit area codes. It is not evident that the transliterators would have been able to disambiguate all answers to the request to give a familiar number.
The format of the read numbers is quite different from the format of spontaneously produced familiar numbers, as can be seen from Table 2. Table 3 gives an overview of the size and type of material that could be used for training a connected digit recognizer on the telephone numbers only. In addition to the response types shown in that table a lot of other words were used, but none of these words occurred sufficiently often to make it worthwhile to explicitly account

Table 3: Frequency of occurrence of digit types in the telephone numbers

Number type	read	spontaneous
Digits	29,440	22,977
Numbers	7,593	8,806
Hundreds	529	439
"Double n"	41	36

for it in a telephone number recognizer.
It is worth mentioning that 18% of the read and 23% of the spontaneous numbers contain extra sounds, far more often preceding the number than following it.

6.2 Yes/No Expressions

For this paper we analysed the responses to four *yes/no* questions; for two we expected affirmative, and for the remaining ones negative responses. There was a large difference between the two putative affirmative items: Almost 93% of the subjects used a single word *ja, jawel, jazeker* to confirm the fact that Dutch was their native language; that proportion dropped to 75% for the question whether the caller was willing to participate in another recording session. Very few callers said "no" to the latter question, but the way in which they expressed their confirmation was much more varied. Almost all persons who said that Dutch was not their native language explained that their first language was Frisian or a regional dialect. We decided to accept these talkers as effectively native speakers.
83% of the subjects used a single word *nee, neen* to say that they never lived abroad for an extended period of time. The most obvious explanation for this relatively low number is the large proportion of callers who gave an affirmative answer to this question. 80% of the callers used a single word to deny that they were using a cordless phone; here too, the low proportion of single word answers is mainly due to the large proportion of affirmative replies (over 13% of the callers said they were using a cordless phone).
Since the transliterators had to code the answers to three of the four *yes/no* questions we could check how often an affirmative answer contained a negation and how often the reverse was true. In our data these cases that make the life of a recognizer very difficult were virtually absent: on a total number of 10,702 affirmative responses eight contained a single negation, another eight responses contained two "no" expressions and a single reply contained no less than three negative expressions. Out of 4,525 clearly negative responses only three contained an affirmative expression.
Another observation that is worth mentioning is that politeness forms like *yes, sir; no ma'am* were virtually absent. This may be due to the fact that the *yes/no* questions were located in the last part of the recording session, when the caller should be fully aware that they were talking to a recording machine.

6.3 ZIP-Codes

In the Netherlands ZIP-codes consist of four digits followed by two letters. In principle, all letter combinations can appear. Specifically, a large number of letter pairs which are known as acronyms (e.g. NS: "Nederlandse Spoorwegen" *Dutch Railways*, ME: "Mobiele Eenheid" *Riot Police Squat*) or for which ad hoc acronyms or spelling alphabets are easily invented occur. It was not known in which ways people express ZIP-codes.
For POLYPHONE two ZIP-codes were recorded, one arbitrary which was read and the ZIP-code of the respondent which was given in reply to the prompt *Please, say your ZIP-code."* It appears that the way in which ZIP-codes are read differs considerably from the way in which people say familiar ZIPs. In the read ZIPs slightly over 65% of the numeric parts are expressed in the form of two numbers; for the familiar ZIPs this proportion is 74%. Only in approximately 11% of the read forms people use expanded acronyms (e.g. *Nederlandse Spoorwegen* for *NS*), while this proportion is as high as 30% in familiar ZIPs. Less than 1% of the read forms had two letters followed by an expanded acronym, while this proportion was over 2% for familiar ZIPs. Somewhat surprisingly, the number of disfluencies does not differ between read and spontaneous ZIPs.

6.4 Sentences

The nine sentences recorded in POLYPHONE are divided into two groups, i.e., four sentences constructed around application words and five constructed so as to contain all phonemes of Dutch. The first observation that must be made is that the number of *Other* judgments for both sets of sentences is over 7%, whereas that proportion is less than 3% for all other items. Apparently, our subjects had considerable difficulty in reading the sentences aloud, even though they had been encouraged to study the texts before calling the recording platform. The proportion of disfluencies in reading sentences seems to be larger than what was obtained in "Voice Across America" (Wheatly & Picone, 1991). Yet, considerable effort has been spent in selecting and designing sentences for easy readability. Some 12% of the sentences contained extra sounds preceding the first word.
The modal length of the application sentences was 12 words; for the phonetically rich sentences the modal number of words was 11. There were approximately 1,100 different application sentences containing 4000 different words, and 12,500 different rich sentences containing 17,000 different words. The relatively large number of different words is certainly due to the way in which the sentences were constructed or selected.

REFERENCES

Godfrey, J., Graff, D. & Martin, A. (1994) "Public databases for speaker recognition and verification." *Proc. ESCA Workshop Automatic Speaker Recognition, Identification and Verification,* Martigny, April 5-7 1994, pg. 39-42.

Bernstein, J., Taussig, K. & Godfrey, J. (1994) "Macrophone: An American English Telephone Speech Corpus for the Polyphone Project." *Proc. ICASSP-'94,* Adelaide 19-22 April 1994, pg. I-81 - I-83.

Boves, L., Boogaart,T. & Bos, L. (1994) "Design and recording of large data bases for speaker verification and identification." *Proc. ESCA Workshop Automatic Speaker Recognition, Identification and Verification,* Martigny, April 5-7 1994, pg. 43-46.

Wheatly, B. & Picone, J. (1991) "Voice Across America." *Digital Signal Processing,* Vol. 1, pg. 45-63.