Computational Linguistics in the Netherlands Journal 6 (2016) 21-38 Submitted 05/2016; Published 12/2016

Sprekend Nederland: a heterogeneous speech data collection

David A. van Leeuwen[1] d.vanleeuwen@let.ru.nl
Frans Hinskens[2] frans.hinskens@meertens.knaw.nl
Borja Martinovic[3] B.Martinovic@uu.nl
Arjan van Hessen[4] A.J.vanHessen@utwente.nl
Stef Grondelaers[1] s.grondelaers@let.ru.nl
Rosemary Orr[5†] r.orr@uu.nl

[1] *CLS/CLST, Radboud University Nijmegen*

[2] *Meertens Institute, Amsterdam*

[3] *Department of Interdisciplinary Social Science/Ercomer, Utrecht University*

[4] *HMI, University Twente*

[5] *UIL/OTS, Utrecht University* [†] *Deceased 15 November 2016*

Abstract

Sprekend Nederland is a large-scale effort to document the variability of Dutch as spoken in the Netherlands anno 2016. A smartphone app was created to record the speech of as many speakers of Dutch as possible, as well as their attitudes (perceptions and evaluations) towards other participants's speech. Initiated by the national broadcast organisation NTR, *Sprekend Nederland* relies on both traditional and modern media to recruit participants. At this point, about halfway through the project, over 7000 participants have recorded over 200 000 utterances, totalling about 375 hours of speech data, and over a million of attitude judgements have been given. In this paper we report the design and implementation of the data collection, we present some preliminary statistics and demographics, and we outline a number of research possibilities that this data collection offers.

1. Introduction

Linguistic databases have been driving computational linguistic research for several decades now, and remain the main source of information for new research questions. This is specifically true for spoken resources, where annotation at a fine level of granularity is often required. Acoustic-phonetic research typically requires segmentation of the speech recording at the phoneme level. Manual annotation of such data is very labour intensive, and hence databases with fine-grained annotations tend to be smaller in size. This limits the coverage of variability introduced by intrinsic and extrinsic factors. An example of an early database is the well-known TIMIT database of US English (Zue et al. 1990). Although relatively small by current day standards (5.4 hours), it is fully annotated at the phone level in terms of a very detailed phone set of 61 phones (including 4 silences). The database is diverse in speakers (630 from 8 dialect regions), and it is still widely used. However, the variability in speaking style (read speech) and recording quality (very clean) is too limited for use in most current day speech technological research challenges.

An example of a speech data collection effort for Dutch is the Corpus Gesproken Nederlands, which covers several speech styles and recording conditions (Oostdijk and Broeder 2003). The database has enabled a wide variety of linguistic research, and has also been used in speech technology (Despres et al. 2009, Huijbregts et al. 2009, Demuynck et al. 2009) The database is manually annotated at the word level. It also contains manual annotation at the phonemic level for a small fraction, and this information is augmented with automatic annotations found with an automatic

speech recognition by forced alignment of the orthographic transcriptions. With 900 hours of speech material and over 4000 speakers, it can still be considered quite a sizeable database for current day standards.

Traditional methodologies for speech data collection have been recruiting paid participants to make recordings (Godfrey et al. 1992), using material from radio and TV broadcast archives, or making recordings during natural speech activities such as meetings or lectures. With the almost ubiquitous possession of smart phones, however, a new method has become available which allows researchers to collect speech material and opinions from almost anybody in almost any context. An example of such a crowd-sourced database utilising smart-phones is "Red-Dots," a (research) community effort to record a text-dependent speaker recognition database (Lee et al. 2015).

There is a number of studies aimed at documenting the variability of language and accents. In 2004 the British Broadcasting Corporation (BBC) collected 300 recordings from 1200 people across the UK by interviewing people in groups. The interviews, which followed a common methodology, were geared towards eliciting lexical variability (Wieling et al. 2013). The subject sample was stratified in order to investigate both regional and social variability. The recordings are stored in the National Sound Archive of the British Library. In 2013, Kolly et al. (2014) used an app *Dialäkt Äpp* for smartphones to document lexical variation in the Swiss German labelling of 16 concepts. They used this in a follow-up app *Voice Äpp* with Automatic Speech Recognition techniques to localize the user's accents and provide a multidimensional profile of their voice. Voice Äpp was further used to collect more data (Leemann et al. 2015b), which resulted in over half a million recordings. In another study, Leemann et al. (2015a) studied the attitude of listeners to particular German accents w.r.t. the speakers' suitability for certain vocations.

Recently, an opportunity arose for linguistic researchers in the Netherlands, to take part in the project with ingredients similar to the UK and Swiss efforts described above. The project, entitled *Sprekend Nederland*, is run by the NTR, a national public broadcasting organisation. The project entails participants recording spoken sentences and short word lists, and listening to other participant's recordings and judging them on a variety of attitude factors. The goal of the project is to document the dynamics in present-day Dutch by eliciting speech from a wide diversity of Dutch speakers, and asking them to categorise and evaluate the speech of other participants. By combining speech production and perception, *Sprekend Nederland* wants to track and describe the relevant dimensions of variability and emergent change in spoken Dutch. In view of the fact that *Sprekend Nederland* was initiated by broadcaster NTR, traditional media (radio and television) and new media were used to reach a large audience which is ideally representative of the different regional, social and other demographic groups which constitute the Netherlands. This paper describes the data collection design and some first statistics regarding the demographic information of the participants. The next section describes the original design that was made before the actual implementation, because this may be relevant to other, similar, efforts. In building the app not all design choices could be implemented, and so the third section reports on the actual statistics of the current state of the data collection. In the fourth section, we focus on a number of research questions which can be answered with *Sprekend Nederland*-data.

2. Data collection design

Sprekend Nederland is a project that is conceived and run by the NTR, whose primary goal is to produce radio and television programmes for all people in the Netherlands.[1] For this reason, one of the design criteria of the data collection was that it would be inclusive. We wanted participation from as many Dutch speakers as possible from the widest possible variety of demographic backgrounds, regardless of socio-economic, geographical, ethnic or other factors. In other words, we do not want to exclude any speaker of Dutch from participating.

1. This may be appreciated from the slogan "NTR: speciaal voor iedereen" (NTR: special, for everybody).

The NTR consulted a group of researchers, the present authors, for advice on the data collection design so that the resulting data could be used for a wide variety of research. The data comprises three parts. The primary data is formed by the audio recordings. This material can be used directly for research in (socio)linguistics and computer science. The second part is formed by attitude and perception data. This part consists of perception and attitude data: participants categorize the speech of other participants on a number of parameters (notably regional provenance—on the basis of their accent—and accent strength), but they also evaluate speakers and their speech on prestige and solidarity-related traits. This data is interesting for sociolinguists, social psychologists, sociologists, and various other research fields. Finally there is the metadata, information about the speaker and recording conditions. A significant amount of metadata has been elicited from all the speakers—in addition to all the conceivable demographic data, we elicited information on the speakers' political, religious and ethnic beliefs. These are essential to determine the major dimensions of variability in the production and perception data. We will come back to some of the research possibilities in Section 4.

The NTR further worked with a third-party app-development company, *Alledaags*, to design and implement the software interface used for the data collection. They were also also responsible for the back-end services that are necessary to centrally direct the data acquisition process and store all collected data.

2.1 Legal and ethical aspects of the data collection

Legally the data collection efforts fall under the auspices of the NTR. All legal aspects from the point of view of the broadcasting organisation are covered in the general conditions of using the app. The long-term storage of data is the responsibility of the *Nederlands Instituut voor Beeld en Geluid*, the Dutch national media archive. The consulted researchers recognised early on that the ethical aspects should not be ignored in such a data collection. The Centre for Language Studies (CLS) of the Radboud University Nijmegen was selected for internal consultation regarding the ethical aspects. During the development we remained in continuous discussion with the CLS ethics Institutional Review Board. The tangible effects of their recommendations on the design of the app and the management of the data are the following:

- The app contains an *information document* and a *consent form* based on CLS templates. The information document describes the purpose of the study, the privacy and storage of data, and contains addresses for more information and complaints. The consent form is included in the general conditions, which must be agreed upon before registration. The consent form links to the information document, which is also accessible through the Internet.

- There is specific logic to detect non-adult participants, in which case the consent must be given by a legal guardian. Non-adult participants are allowed to participate, but will be excluded from the storage in the database and distribution thereof.

- Participation is anonymous, i.e., the database exported to the researchers does not contain a traceable identity of the participants. Geographical information given by the participants is distorted with a Gaussian noise of approximately 1 km in the database exports.

- A data management plan has been developed, which covers the policies regarding privacy, security, short and long term storage, access and life cycle of the data collected.

2.2 Acquisition hardware

In view of the fact that the production and use of smartphones surpassing that of traditional personal computers, it was decided that for primary data acquisition the participants's own smartphones

would be used. On-line statistics sites[2] estimated smartphone penetration in the Netherlands at 76 % in 2015, increasing by 12 points per year since 2013. This suggests that in 2016, the year in which the data collection is scheduled, up to 85 % of the Dutch population will own a smart phone. An even higher percentage can be expected to have access to a smartphone through family and friends. Further advantages of using such devices for data collection are that they all come with good microphones, have internet access capability, and seem to become most people's primary access to digital resources by cleverly crafted apps. The use of participant's own devices brings a variety in data acquisition hardware that is usually only found in conversational telephone speech (CTS) collections, but the modern mobile devices allow for higher acoustic bandwidth. The recording format is 44.1 kHz, loosely MPEG4 compressed, with up to 2 channels if the hardware allows for it, averaging at a bitstream of 110 kb/s. The data collection was realised by creating a dedicated app, developed as native applications for the two major smartphone software platforms that are currently used in the Netherlands, Android and iOS.

2.3 Participant interactions and stimulus material

The scientific advisory group had the following approach to the design of the content of the app. We contacted all researchers in the country that we assumed would be interested in collecting data within *Sprekend Nederland*. A questionnaire was used to make an inventory of the type of stimulus material, the way of presenting the stimulus, the way of recording a response, and the type of metadata required for the research of the interested parties. This data was aggregated, after which a priority list was prepared. Because the development of a software component for a particular kind of user interaction is costly, not all the desired functionalities on the priority list could be implemented right away. User interactions eventually implemented included:

- Audio recording. Some stimulus is presented on the screen, and the participant is requested to make a recording of an utterance. Mostly this is self-paced, but for some stimuli the recording is driven by the app. An example is the request *"Please start the recording and describe extensively the area where you are right now"*.
- Single-option questions. A stimulus is presented, and a response from a limited set of options is given by pressing a button on the screen. The most common are yes/no questions. An example is the question *"Would you like to sound like this person?"* when the utterance from another participant is played.
- Multiple-option questions. A stimulus is presented, and a number of values can be selected or deselected. An example is the question *"Which other languages do you speak?"*
- Ordinal value questions. A stimulus is presented, and a response is given by moving a horizontal slider. The slider shows extreme values at either end (lower values on the left), and for numerical values with many options the selected number is shown above the slider. The most common are attitude questions for which responses are recorded on a 7-point Likert scale. An example is the question *"Is this person trendy?"* upon the presentation of another participant's utterance.
- Location questions. A stimulus is presented, and the response is given by placing a marker on an interactive map. The geographical coordinates, as well as the zoom level, are stored. An example is the question *"Where have you lived the longest within the Netherlands?"*

Several modalities for presenting a stimulus were implemented. These are:

- A literal sentence, to be read out aloud by the participant, e.g., *"Zacht knort het varken in de wei; een karakteristiek geluid"* (The pig grunts softly in the meadow, a familiar sound).

2. `http://www.statista.com/statistics/488353/smartphone-penetration-netherlands/`, retrieved 19 May 2016

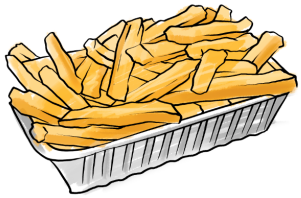

Figure 1: An example of an image used as stimulus

Set	Amount	Description
a	10	sentences, which together with *b* cover all major phonemic variations in the Netherlands.
b	44	isolated words, complementing *a*
c	122	loan words, which should have a higher likelihood to elicit the speaker's natural accent
d	48 M	sentences, to cover maximum lexical variability in the collection. These 8–15 word sentences were taken from the COW corpus (Schäfer 2015).
e	359	additional sentences composed to fit in with the themes that govern the app usage.
f	130	hand drawn pictures, that should elicit regionally varying lexical items.
g	278	words, that together cover all possible consonant-vowel pairs in Dutch.
h	9	description assignments, which should elicit spontaneous speech from the participants.

Table 1: The different sets of stimuli used to elicit speech. 'M' means 'million'.

- A paced set of five words, to be read out aloud, e.g., *"helaas/erwt/zocht/haalt/eis"* (unfortunately/pea/sought/takes/demand).
- A request to describe something in detail, for recording spontaneous speech.
- An image, to elicit a single word or expression, see Figure 1 for an example.
- The playback of an utterance recorded earlier by another participant.
- No particular stimulus, in which case the question is a metadata question about the participant herself.

In principle, any of these stimuli be combined with any of the earlier mentioned kinds of user interaction.

For the first interaction kind, the recording of elicited speech, several sets of stimuli were created. These should allow us to study the different dimensions of variation in present-day spoken Dutch. The material was organised in a number of separate lists, from which the stimuli were drawn. We have tabulated these lists in Table 1.

2.4 App design and interaction flow

For the design of the app, a consensus had to be found among several interests. The basic premise was that using the app should be fun, in order to keep participants motivated to contribute more data that can be used for research. From the research community, a large number of metadata and attitude questions were requested, and a wide variety of speech material. As a broadcast

organisation, the NTR strove for all-inclusive participation, a positive playing experience, thematic organisation and artistic quality. These factors had to be united, and further fit within financial and time limits available for the development of the app, and finally legal and ethical aspects had to be considered.

The concrete app development was carried out following the SCRUM methodology under supervision of the NTR. The scientific advisory group had little involvement in this process, so we will limit ourselves to describing the result of the process here. We can identify the following main parts of the operation of the app:

- An introduction, consisting of several pages to give the participant an overview of the app, and ending with registration of the participant.
- A main menu, from which other parts of the app can be reached, and to which the participant can return.
- Several themes, which can be played in order. Every theme contains a number of interactions, which is a mix of speech recordings, attitude and metadata questions.
- A profile section, where judgements of other participants about the current participant are summarised, and an overview of the participant's metadata is shown.
- An information menu, with additional information about the project and terms and conditions.

The main data collection takes place in several themes: *living, birthdays, flirting, holidays, the canteen, music* and *extra*. Themes are played in a fixed order. The completion of a theme is a natural moment for a participant to take a break, with the option of continuing to the next theme directly, or do this in another session. Each theme contains a number of interactions, and an attempt has been made to fit the questions and prompt text to the theme. For instance, the meta data question *"Where do you come from?"* was placed in the theme "living."

Special care was taken in selecting random utterances from the 48 million sentences from COW (*d* from Table 1). A manual selection was made so that each stimulus would fit one of the themes. This selection limited the lexical variability to 154–500 different utterances per theme.

Each theme contained somewhere between 60 and 130 interactions. In the course of the project it was decided to lower the number of interactions in the first theme to about 40, in order to improve the first experience of new participants. As an incentive for participants to keep 'playing,' they are rewarded with evaluations by other participants, which pertain to their accent but also to their personality characteristics.

2.5 Distribution of the stimuli over participants

The different sets of recording stimuli had different goals with regards to completeness. For instance, the first two sets from Table 1 had to be completed fully for every participant, in order to have any significance for the research question for which they were added. In order to limit the amount of interactions necessary to complete everything for one participant, the 44 isolated words were grouped in groups of five words for a single recording. The five words were presented at about 2 second intervals.

The other stimulus sets (sets *c*–*h*) did not have such completeness requirement. E.g., for speech technology research, such as automatic speech and speaker recognition or accent location (van Leeuwen and Orr 2016), we typically are interested in the spontaneous speech recordings (*h* in Table 1). In this case completing the nine recordings is less important than having several recordings made in different sessions. For these other stimulus sets, stimuli were randomly chosen, up to three stimuli per theme per set.

For the purpose of collecting attitude data, recordings of other participants were used as stimuli. This entails that speakers have to be linked in some way to listeners, and that for these speakers,

particular sentences must be chosen. With the expected number of participants being in the thousands, complete experimental designs where every participant judges every other participant are clearly impossible. Because geography is an important factor in the development of someone's accent, the coupling of speakers to a particular participant was carried out on the basis of geographical distance. Five distance ranges were defined (under 10 km, 10–20 km, 20–40 km, 40–80 km and over 80 km), and speakers were chosen with equal probability from these categories. Similarly, the choice of sentences to be used as playback stimulus was designed to be random, but in equal proportions, from the sentence stimuli and spontaneous speech, *a*, *d*, *e* and *h*.

Because there is no control over when each participant uses the app, these assignments between speakers and listeners had to be carried out dynamically, by a central server that also stores the audio, metadata and attitude data. The decision logic about these assignments and utterance choice are together known as the *business rules* of the server. In principle, these allow dynamic changes in the way the stimuli are distributed over the participants. In fact, all content, including the questions, text prompts, images and video material, are stored and controlled by this server. This is quite a desirable design, especially in the early development stages. However, it was found after the first introduction of the app to the general public, that the implementation of the business rules was computationally too expensive. As a result, most of the rules had to be removed, yielding a more random assignment of speakers to listeners and a random choice of utterances from speakers. At a later stage, the original, overly expensive business rules were replaced by frequency-based rules which give precedence to speakers with no or few evaluations for presentation as stimulus material.

2.6 Recruitment of participants

Because of the inclusive design, any person can participate in the project. There is one exception, and that is for potential participants under 18 years old. They can only participate with an approval from a parent or guardian, and for ethical reasons their speech data will not be stored in the databases used for research.

People register using an email address, which is used solely for registration, and is not available to anyone, including the present authors, using the database. Participants are given a numeric ID which is further used internally to identify speakers and listeners. In the app, no information about any other participant is shown to the user.

Recruitment is carried out through social media and radio and TV. As an incentive to participate, the app shows feedback about the participant's own accent when enough judgements have been made. Several media events (radio and television) are planned throughout the year that the project runs. These range from scientific documentaries to popular game shows. Everybody in the world can participate, but the media events have been limited to the Netherlands, and the texts and spoken material are in Dutch. From the areas in the world were Dutch is spoken (the Netherlands, Belgium, the Antilles and Suriname) we therefore expect the most contributions from the Netherlands.

3. Data collection statistics

In this section we will report on some of the data collection statistics in the first five and a half months of the project, starting from 30 November 2015 up to 18 May 2016.

3.1 Participation

In Figure 2, the rate at which recordings are received at the central servers is plotted over time. We have fitted a model that assumes impulses at three particular dates, with a common exponential decay factor. The impulse dates correspond to specific media events in which the project is mentioned, namely the launch on 2 December 2015, with several radio and TV programmes, and 28 January and 11 March, when a TV programme was aired. The fitted decay rate is $d = 5.2\,\%$ per

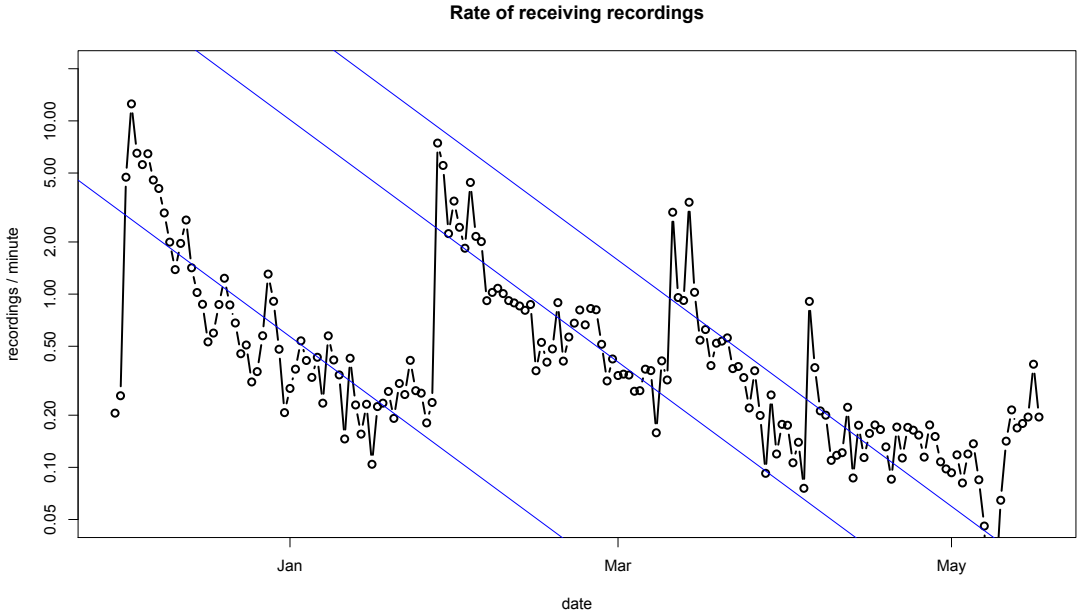

Figure 2: Data collection rate, expressed in terms of the number of recordings per minute averaged over a day, from just before the launch of the project (30 Nov 2015) to 18 May 2016. The fitted lines represent exponential decays with impulses on 2 Dec, 28 Jan and 11 Mar.

day, and the nice thing about an exponential decay model is that it makes it possible to associate a total amount of expected participation to a single event (van Leeuwen and Orr 2016). If the rate of recordings at the day of the event is R_0, then the total amount follows from the geometric series

$$N_{\text{tot}} = N_0 R_0 \Big(\sum_{i=0}^{\infty} (1-d)^i \Big) = \frac{N_0 R_0}{d}, \tag{1}$$

where $N_0 = 24 \times 60$, the number of minutes per day.

We can do a similar analysis for the number of new participants per day, or the rate at which responses to the questions are recorded. They follow the same qualitative trend as the rate of recordings, with decay rates of 5.7 % and 4.9 % per day, respectively.

In the reported period, 9526 people have registered. Of these, 76.5 % have given at least one response, and 74.0 % have made at least one recording. 57.0% of the registered people have answered at least one metadata item about themselves.

The pattern of the number of recordings made by each participant is shown in Figure 3. Again, we see rapid decays as a function of the number of recordings made, but there are clear peaks that correspond to finishing a theme. The total recording duration is 377 hours, an average of 3.2 minutes per participant. Finally, the participants have given 1.12 million attitude judgements about other participant's utterances, an average of 154 per participant.

3.2 Age and Gender distribution

In Figure 4 the distribution of age and sex is shown, for the 57 % of the participants that provided this information. Ignoring the 0.4 % "other" specifications on this variable, the current data show a 58.9 % female vs. 40.7 % male gender distribution.

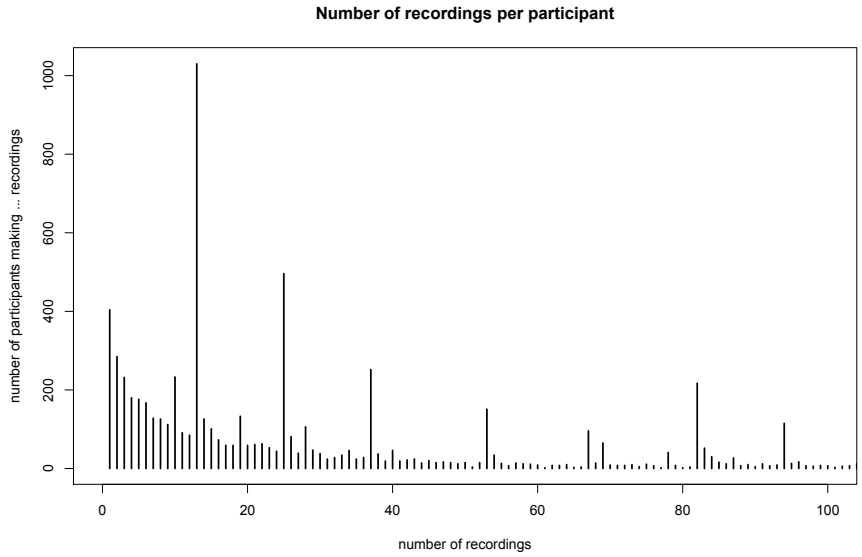

Figure 3: The distribution of the number of recordings per participant.

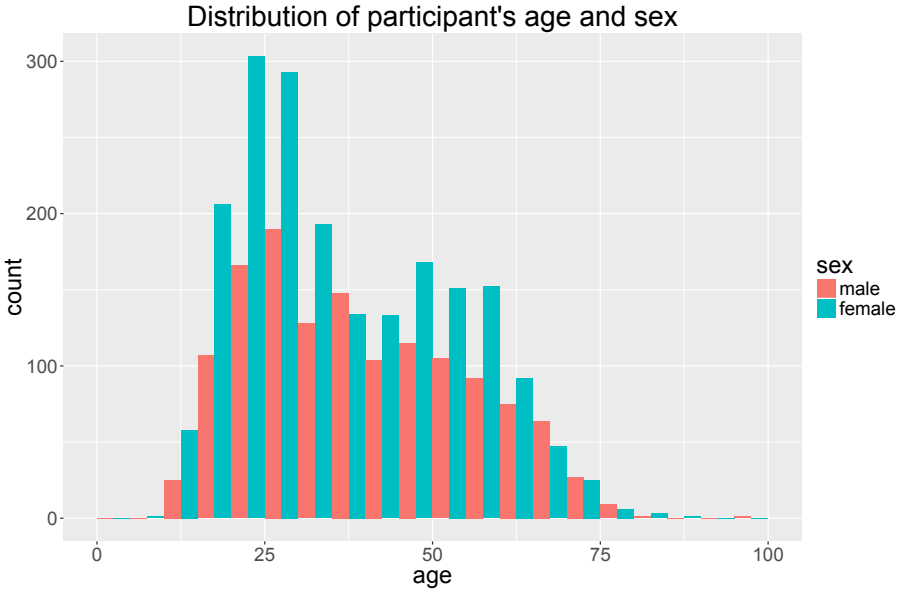

Figure 4: The distribution of age for male and female participants, in 5-year bins.

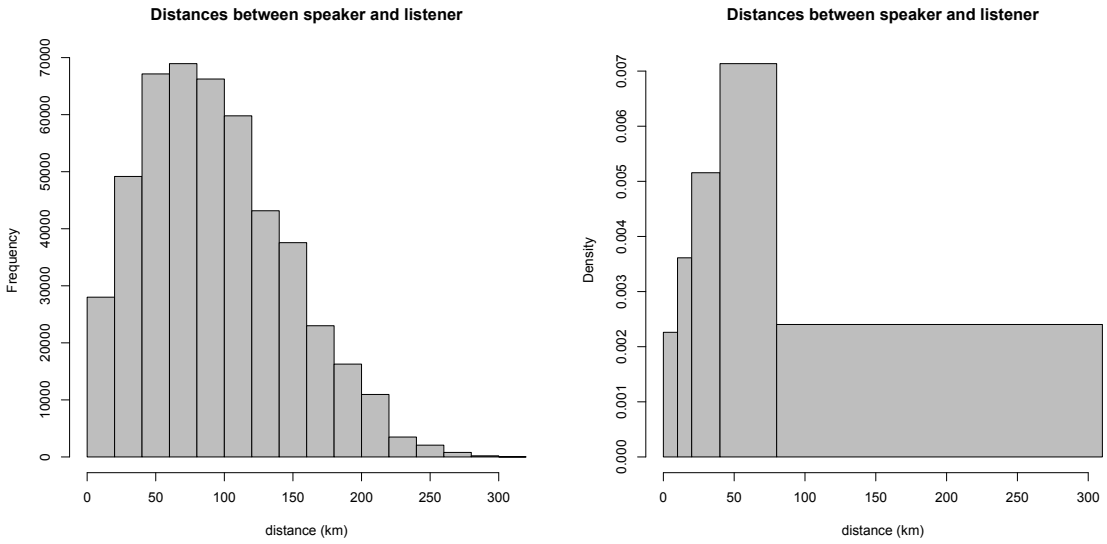

Figure 5: Histograms of the distribution of speaker-listener distances. The left pane shows the normal histogram in 10 km bins, the right pane shows the distributions when the bins correspond to the originally requested distance ranges 0–10, 10–20, 20–40, 40–80 and 80–310 km. The heights in the right-hand pane are scaled such that the area of the bar is proportional to the counts in the bin.

3.3 Distribution of stimuli

As indicated in Section 2, it was not always possible to implement the original design for distributing stimuli across participants. In this section we analyse what the effect has been of some of the practical decisions that had to be made in the implementation.

The first thing we look at is how the speakers are selected for playing stimuli to a particular participant, whom we will call the listener in this context. In the original design the speakers were to be chosen based on distance to the listener, with an equal probability of being chosen from one of five logarithmically increasing ranges. In Figure 5 the distance distribution is shown, where distances are computed as the crow flies, and the origins of speaker and listener are taken from the response to the question *"Where have you lived the longest within the Netherlands?"* For 57.6 % of the 1.12 million responses, the origin of both speaker and listener were known, since they were provided in the metadata. The right-hand pane in Figure 5 shows a version of the histogram where the bins correspond to distance ranges from the original design. In this histogram actual counts correspond to surface areas of the bars, so if the original design had been followed, the areas of the bars should have been roughly equal. Because the business rule that selects a speaker based on distance had to be removed for operational reasons, we have ended up with relatively many responses where the distance between speaker and listener is large. This will have some consequences for potential research. On the one hand, there will be relatively many judgements where the accent of the speaker is likely to be different from the listener, which may be interesting for the study. On the other hand, we will get very few cases where listeners will be able to recognise a nearby accent. For instance, of the responses to the question *"How for away from you do you think the speaker lives?"* only 1 % of the cases the speaker was located less than 10 km from the listener.

set	a	b	c	d	e	f	g	h
Number of unique stimuli in design	10	44	122	48 M	359	130	278	9
Number of unique stimuli implemented	10	44	48	2071	359	130	204	9
Occurrence in recordings	25 k	136 k	27 k	52 k	34 k	50 k	29 k	16 k

Table 2: Recording statistics for the various sets of stimuli.

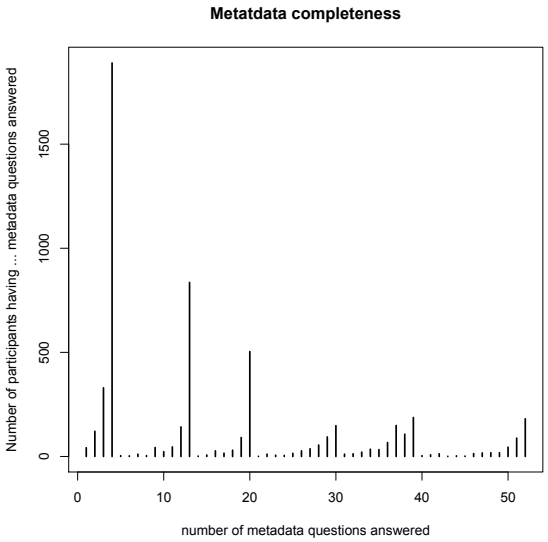

Figure 6: Number of participants completing a certain number of metadata questions.

As a second aspect of the distribution of stimuli, we look at how the various stimulus sets are distributed over the recordings. In Table 2 the statistics of recordings are analysed for each of the stimulus sets. The goals of lexical variability for *d* (web sentences) were not reached, because it was decided at the operational level that each of the sentences should fit in one of the themes. In addition, the risk of profane or insulting content in the COW sentences was considered too big to present these without manual filtering. Also in other sets, *c* (loan words) and *g* (consonant-vowel pairs), there is less variability than in the original design.

3.4 Completeness of metadata

Different analyses of the data require different metadata, so in total, 53 metadata questions were defined in order to accommodate the research interests of the various investigators. In order to make the use of the app a little more appealing, the metadata questions were scattered over all six themes. However, as we have seen in Figure 3, not all participants completed the different themes. This would then lead to incomplete metadata. In Figure 6 the theme structure is clearly visible again. At each local peak (at 4, 13, ... questions answered) a theme session has been completed. The order of the metadata questions is the same for every participant, so the first metadata question (*"Have you always lived in the same place?"*) has the highest yield, at 96.8 %.

Because the metadata can be important for analysis of the primary data—for some sociology studies the metadata themselves could even form the primary data—two additions have recently been made to the app. The first is that it is now possible to fill in metadata outside the theme flow, so that people can review their metadata. They can then correct these if they like, and do not need

to complete all themes. The second is the possibility of sending "push notifications" to the mobile devices, which can in principle be used to nudge participants to fill in particular metadata questions.

4. Prospective research topics

So far we have only described the design and implementation of the *Sprekend Nederland* database, and given some descriptive statistics. In this section we want to indicate some of the potential research topics that can be addressed with this data. Some of these topics have explicit relations to some of the metadata or judgement questions in the app.

4.1 Speech technological research and development

4.1.1 Automatic accent location

In a recent paper (van Leeuwen and Orr 2016) we have shown that *Sprekend Nederland* enables investigation of a new area in speech technology, namely that of *automatic accent location*. Here, the task for the machine is to determine *where someone is from* just from the speech signal. This task description is somewhat ambiguous, as people might have been in several places in their lives, and have picked up accents from any of these. In such case, an automatic accent location system should indicate that there have been multiple influences. This is, of course, a daunting task, not only from the engineering point of view, but also from the user's point of view: when will a system be "good" and how should the outcome be presented? In forensic speaker characterisation, it is not unreasonable to state something like that the speaker in an incriminating recording speaks Dutch with a French accent and is likely to live in or near Amsterdam.

There will be many influencing factors for the rate at which someone acquires their accent from interactions with other members of society: their age, their susceptibility to accents, the typicality of the local accent, the ease of production of accent-characteristic vocalisations, the social status associated with the accent, etc. This data collection caters for some of these factors. In the metadata questions, we ask about the mobility of the participant, and the locations of primary and secondary school, longest residence, and the more subjective questions *"where do you come from?"* and *"where on the map do you put your own accent?"* Also, there are perception questions about intelligibility and accent strength, that others answer about the participant. Information about social status of a particular accent could be found from several judgement questions like *"How much would you like this person as neighbour?"* (colleague / friend / boss) and *"How well is this person suited as TV news presenter?"* (teacher / mayor / DJ / quiz master).

4.1.2 Accent and L1 recognition

A quantised version of accent location is accent recognition, where the location (history) is boiled down to a single class. This is a more traditional line of research (Bahari et al. 2013, Choueiter et al. 2008, Hautamäki et al. 2015), but we can include in the accent classes sociolects and ethnolects (Hanani et al. 2013). Finding such class labels will require some data mining as not all of these labels are metadata items, and perhaps some metatdata items need to be combined with judgments to generate a label. Similarly, this data can be used to study automatic native/non-native detection, and even L1 (native language) recognition.

4.1.3 Automatic Speech Recognition

Sprekend Nederland contains a fair amount of speech data, most of which is read speech. This means that for a large part of the recording the transcription is known, if we assume that most participants produced proper recordings. This means the database can in principle be used to train acoustic models for automatic speech recognition. Research aspects in training acoustic models

fall in the area of automatically verifying transcriptions. In our case the transcriptions come from text prompts, more generally they can come from other sources, such as close captioning, meeting minutes or even automatic speech recognition.

4.2 Linguistic research

4.2.1 Sociolinguistic research

In order to gauge the dynamics in present-day spoken Dutch, the Sprekend Nederland app elicits both production and perception data: participants do not only produce scripted and free, 'spontaneous' speech, they also categorise and evaluate other contributors' speech. The extracted production and perception data are rich in a great number of respects.

First, the speech recordings concern elicited production (notably on 5 phonetic/phonological phenomena known to manifest remarkable variation across the language area, in both the dialects and the standard varieties, sets *a* and b) but also unscripted, 'spontaneous' speech (set *h*). Further metadata comprises the participants' socio-biographical background (origin and year of birth, educational career, language background, etc.). These metadata allow drawing stratified samples of speakers, enabling in-depth variationist analyses along the major geographical, social and ethnic dimensions of the variation in spoken standard Dutch, observed in a range of segmental and supra-segmental phonological, as well as grammatical phenomena. Of course the metadata also enable in 'apparent time' comparison to detect ongoing changes as well as their potential social meaningfulness ('indexicality').

Second, every participant supplied perception data in the form of the regional identification and social evaluation of other participants' input. In view of the fact that social evaluation (as a variation and change determinant) is typically studied in stringently controlled experimental designs, in most cases with limited numbers of subjects, the Sprekend Nederland app returns perceptual data of an unprecedented richness. While experimental designs such as Grondelaers and Steegs (2010) typically elicit evaluations of mild accents produced by male speakers from four regions, the Sprekend Nederland database contains accents of variable strength produced by younger and older, male and female, native and ethnic speakers from every conceivable region (on any level of granularity) in the Netherlands.

The extent to which experimentally controlled and publicly elicited data converge remains an empirical matter, but we have done everything in our power to collect speech data with a high level of intra-speaker, cross-speaker and cross-regional comparability, and to extract judgments which are as indirect and unpremeditated as possible. Evaluations are elicited on six semantic differentials in function of three recurrent evaluation dimensions (prestige, solidarity, dynamism), but also on more indirect measures such as the speaker's appropriateness for a range of job types.

Most importantly, perhaps, the confrontation of richly stratified production and evaluation data enables us to identify language change determinants which are difficult to track in investigations which focus either on production or perception (the absolute majority, in fact). While many studies have assumed a causal link between prestige considerations and language change-—to the extent that we appropriate what we appreciate and admire—the Sprekend Nederland-corpus is one of the first databases which enables us to carry out empirically valid, linguistically as well as extra-linguistically balanced and principled research into the triggers of change.

Owing to its size and stratification, but also the high recording quality of the majority of the speech samples, the Sprekend Nederland-database is technically suited to all sorts of instrumental acoustic analyses which require 'clean' speech. Methodologically, elicited and hence fully comparable speech data can supplement studies based on written stimuli (as in the online questionnaire on which Bennis and Hinskens (2014) is based). 'Clean' speech also is a prerequisite for the selection of experimental samples for speaker evaluation experiments, for which the Sprekend Nederland data turn out to be unexpectedly suited. While Grondelaers and Steegs (2010) and Grondelaers et al. (2011) were based on the high quality standard speech of the teachers in the Spoken Dutch

Corpus 'CGN' (Oostdijk and Broeder 2003)—which contains some regional flavouring but little accent strength variation—-the Sprekend Nederland corpus (even in its present format) features spontaneous speech on preselected, neutral topics which is for the most part cleanly recorded and noisefree, and which has a much wider accent strength range than available in the Teacher part of CGN.

4.2.2 Phonetic research

With the primary speech data and metadata it will be possible to make a comparison of traditional (regional), new (Moroccan, Turkish, Antillean) and non-native (non-Dutch residents) Dutch accents. A by-product of the acoustic model training for speech recognition described above is a forced-alignment of the phone sequences of the transcription to the spoken recording. This would allow researchers to compute acoustic-phonetic parameters over specific phonemes of Dutch, and compare them across the various accents.

In a similar way it is possible to do phonetic research on Dutch accents over generations taken the year of birth of the participant in account.

4.2.3 Language variation research

Two sets of stimuli were chosen to study language variation research. The hand drawn pictures (set *f*, for an example see Figure 1) will elicit different lexical choices from participants, which may depend on regional or sociological background. Such a study would require transcriptions at the word level. In order to obtain these, automatic speech recognition will be useful as a tool, probably in combination with manual transcription of a subsample of the data in order to detect low-frequency lexical alternatives.

It is possible that the data acquisition setting (participants recording utterances that are going to be judged by other participants) has an influence on the accent that they speak, e.g., participants may tend to speak more like what is perceived as "standard Dutch." In order to study the importance of such an effect, the data from set *c*, the loan words, could be used, as these words are more likely to elicit one's own accent.

4.3 Sociological research

The data lend themselves also for sociological and sociolinguistic research. We highlight two possibilities. First of all, it possible to examine attitudes towards speakers with particular accents (Giles and Billings 2004). The app includes judgements based on stereotype traits that fall under three standard dimensions of warmth, competence and morality, see (Fiske et al. 2002, Leach et al. 2007). Examples are traits such as friendly (warmth), intelligent (competence), and trustworthy (morality). This allows us to test on a large scale some often assumed ideas such as: Are people from Limburg seen as warm but less competent, and people from the North of the Netherlands as competent but cold? And in general, what stereotypes do Dutch people from a certain region hold of speakers from another region? We can link these evaluations to the background characteristics of the speakers and listeners, and test if, e.g., higher educated or younger speakers are evaluated more positively even when they come from a less liked accent area, as well as whether age, gender, level of education, conservatism, and ethnic identification of the listener have an effect on their stereotypes of others.

Second, participants were asked to indicate their own perceptions of language discrimination, that is, whether they have the impression that others admire their accent and find it interesting or devalue and ridicule it. This makes it possible to find out which people perceive more discrimination based on their language (e.g., from certain regions, social classes, or age categories) but also to check whether these perceptions are based in reality, that is, whether people who perceive discrimination also tend to get more negatively evaluated by other app users.

5. Conclusions

The collaboration between academia and the media leads to an opportunity for large scale data collection. *Sprekend Nederland* is an example of such data collection, and we have found clear evidence that media events in which the data collection effort is mentioned boosts the amounts of participants contributing. For instance, the number of viewers for the second media event of Figure 2 is estimated at about 180 000, including web-views after the broadcast. We can estimate the number of participants that were triggered to participate by this event at roughly 3000, which means about 1.6 % of the viewers takes the action to participate in the research after having seen the television programme.

The voluntary nature of participation makes it also hard to fill all cells in the experimental design. We need a different analysis than is appropriate for an experimental setting in the lab, where there is better control over the completeness of the data being collected. Where normally missing data items are just exceptions that need to be dealt with, in this approach significant fractions of (meta)data may be missing.

For the researchers, the opportunity to gather large amounts of data with relatively little effort is a very attractive one. The interests of the media partner (making television and radio content for a wide variety of people) may be different from the academic partners (exploring new research directions), but there are many common grounds. Both aim at a large participation and a wide coverage of regional and social accents, good recording quality, a scientific basis for data collection, and answers to questions about the attitude towards various accents.

The intention of the NTR is to make the data collection available for research, and some parts of the (meta)data have already been used by undergraduate and master students for thesis work, as well as for the selection of speech utterances for evaluation experiments (see Section 4). For a wider release, we first need to complete the data collection, and process the data w.r.t. privacy and ethical aspects. We are confident that this data collection will lead to many research opportunities in linguistics, sociology and spoken language technology.

References

Bahari, Mohamad Hasan, Rahim Saeidi, Hugo Van hamme, and David A. van Leeuwen (2013), Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech, *Proc. ICASSP*, pp. 7344–7348.

Bennis, H. and F. Hinskens (2014), Goed of fout. niet-standaard inflectie in het hedendaags standaardnederlands, *Nederlandse Taalkunde* **19** (2), pp. 131–184.

Choueiter, Ghinwa, Geoffrey Zweig, and Patrick Nguyen (2008), An empirical study of automatic accent classification, *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, IEEE, pp. 4265–4268.

Demuynck, K., A. Puurula, D. Van Compernolle, and P. Wambacq (2009), The ESAT 2008 system for N-Best Dutch speech recognition benchmark, *Proc. ASRU*, pp. 339–344.

Despres, Julien, Petr Fousek, Jean-Luc Gauvain, Sandrine Gay, Yvan Josse, Lori Lamel, and Abdel Messaoudi (2009), Modeling Northern and Southern varieties of Dutch for STT, *Proc. Interspeech*, ISCA, Brighton, pp. 96–99.

Fiske, S. T., A. J. Cuddy, P. Glick, and J. Xu (2002), A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition, *Journal of Personality and Social Psychology* **82** (6), pp. 878.

Giles, H. and A. C. Billings (2004), Assessing language attitudes: Speaker evaluation studies., *in* Davies, A. and C. Elder, editors, *The handbook of applied linguistics*, Blackwell, pp. 187–209.

Godfrey, J. J., E. C. Holliman, and J. McDaniel (1992), Switchboard: telephone speech corpus for research and development, *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 517–520.

Grondelaers, Roeland van Hout, Stefan and Mieke Steegs (2010), Evaluating regional accent variation in standard dutch, *Journal of Language and Social Psychology* **29**, pp. 101–116.

Grondelaers, Stefan, Roeland van Hout, and Dirk Speelman (2011), A perceptual typology of standard language situations in the Low Countries, *in* Kristiansen, Tore and Nikolas Coupland, editors, *Standard Languages and Language Standards in a Changing Europe*, Novus, Oslo, pp. 199–222.

Hanani, Abualsoud, Martin J Russell, and Michael J Carey (2013), Human and computer recognition of regional accents and ethnic groups from british english speech, *Computer Speech & Language* **27** (1), pp. 59–74, Elsevier.

Hautamäki, Ville, Sabato Marco Siniscalchi, Hamid Behravan, Valerio Mario Salerno, and Ivan Kukanov (2015), Boosting universal speech attributes classification with deep neural network for foreign accent characterization, *Sixteenth Annual Conference of the International Speech Communication Association*.

Huijbregts, M., R. Ordelman, L. van der Werff, and F.M.G. Jong (2009), SHoUT, the University of Twente submission to the N-Best 2008 speech recognition evaluation for Dutch, *Proc. Interspeech*, ISCA, pp. 2575–2578.

Kolly, Marie-José, Adrian Leemann, Volker Dellwo, Jean-Philippe Goldman, Ingrid Hove, and Ibrahim Almajai (2014), Voice äpp: A smartphone application for crowdsourcing swiss german dialect data., *Digital Humanities*, Lausanne.

Leach, C. W., N. Ellemers, and M. Barreto (2007), Group virtue: the importance of morality (vs. competence and sociability) in the positive evaluation of in-groups., *Journal of Personality and Social Psychology* **93** (2), pp. 234.

Lee, Kong Aik, Anthony Larcher, Guangsen Wang, Patrick Kenny, Niko Brümmer, David A. van Leeuwen, Hagai Aronowitz, Marcel Kockmann, Carlos Vaquero, Bin Ma, Haizhou Li, Themos Stafylakis, Md. Jahangir Alam, Albert Swart, and Javier Perez (2015), The reddots data collection for speaker recognition, *Proc. Interspeech*, ISCA, Dresden, pp. 2996–3000.

Leemann, Adrian, Camilla Bernardasci, and Francis Nolan (2015a), The effect of speakers' regional varieties on listeners' decision-making, *Proc. Interspeech*, ISCA, Dresden, pp. 1670–1674.

Leemann, Adrian, Marie-José Kolly, Jean-Philippe Goldman, Volker Dellwo, Ingrid Hove, Ibrahim Almajai, Sarah Grimm, Sylvain Robert, and Daniel Wanitsch (2015b), Voice äpp: A mobile app for crowdsourcing swiss german dialect data., *Proc. Interspeech*, ISCA, Dresden, pp. 2804–2808.

Oostdijk, N. H. J. and D. Broeder (2003), The Spoken Dutch Corpus and its exploitation environment, *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03).*, Budapest, Hungary.

Schäfer, Roland (2015), Processing and querying large web corpora with the COW14 architecture, *in* Bański, Piotr, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen, and Andreas Witt, editors, *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)*, pp. 28–34.

van Leeuwen, David A. and Rosemary Orr (2016), The "Sprekend Nederland" project and its application to accent location, *Proc. Odyssey 2016: The Speaker and Language Recognition Workshop*, ISCA, Bilbao, pp. 101–108.

Wieling, Martijn, Clive Upton, and Ann Thompson (2013), Analyzing the bbc voices data: Contemporary english dialect areas and their characteristic lexical variants, *Literary and Linguistic Computing*. http://llc.oxfordjournals.org/content/early/2013/03/04/llc.fqt009.abstract.

Zue, Victor, Stepanie Seneff, and James Glass (1990), Speech database development at MIT: TIMIT and beyond, *Speech Communication* **9**, pp. 351–356.