# PiNCeR: a corpus of cued-rate multiple picture naming in Dutch

Joe Rodd[a b], Hans Rutger Bosker[a c], Louis ten Bosch[b a], Mirjam Ernestus[b a], and Antje S. Meyer[a c]

[a] *Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands*
[b] *Radboud University, Centre for Language Studies, Nijmegen, the Netherlands*
[c] *Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, the Netherlands*

*This version: 2019-10-17*

**Abstract**

PiNCeR is a corpus of speech recordings from Dutch speakers who named pictures at different speaking rates. Participants named pre-familiarised '(C)CV.CVC words (e.g., *snavel* ['snaː.vəl] "beak") from line drawings displayed in groups of 8 arranged on a 'clock face'. A cursor moved clockwise from picture to picture to indicate at which of three trained rates (fast, medium and slow) participants were required to name the pictures. Annotation was performed using the POnSS tool (Rodd, Decuyper, & ten Bosch, 2019), where manual and automatic segmentation is combined to yield accurate word onsets and offsets. To detect the onset and offset times of syllables within words, we identified excursions of above-average acoustic instability between the vowel of the initial syllable and the first consonant of the second syllable (Rodd, Bosker, ten Bosch, & Ernestus, 2019). This approach was licensed by careful control of segmental content in the target words to maximise correspondence between acoustics and articulation. The PiNCeR corpus was intended for use in modelling control of speaking rate (Rodd, Bosker, Ernestus, et al., 2019), but may be of interest for other purposes. Trial-level recordings from two related experiments are made available for 25 participants (12 for Experiment 1, 13 for Experiment 2), along with the onset and offset times of the words and the syllables.

## Background

This paper presents the PiNCeR (Picture Naming at Cued Rates) corpus, which was collected to serve as a dataset for the modelling of cognitive control of speech rate (Rodd, Bosker, Ernestus, et al., 2019), and may also be of interest for investigation of phonetic variation as a consequence of speech rate change. The corpus contains productions of experimentally elicited disyllabic Dutch words at three predetermined speaking rates, and temporal annotations of word and syllable onsets and offsets. This paper also documents the procedures used in the preparation of the corpus, notably a distributed annotation system (POnSS) that allowed us to efficiently annotate the corpus (Rodd, Decuyper, & ten Bosch, 2019) and a metric that allowed us to detect syllable onset and offset times from the acoustic signal (Rodd, Bosker, ten Bosch, & Ernestus, 2019).

Two multiple picture naming experiments were conducted, in which the required speaking rate (fast, medium or slow) was indicated with a cueing dot that jumped from picture to picture on a display with 8 pictures. Since the corpus was intended to be used to model cognitive aspects of the preparation of speech, a task that engaged all phases of speech planning before articulation was desirable. Picture naming is the gold standard task for eliciting single word productions, ensuring that all planning phases need to be completed.

In Experiment 1, speakers were explicitly instructed to avoid pausing between words, and instead to adjust their speaking rate by adjusting the duration of the words. In this fashion, we attempted to ensure that we would elicit variation in the way individual words were articulated, rather than variation in the usage of pauses. In Experiment 2, this instruction was not given, to ensure that differences in strategy adopted in the slow speaking rate were the result of speaker-intrinsic processes rather than purely an effect of task.

A Bayesian mixed effects regression analysis was run to characterise the durations of words in the different speaking rates, to assess their compliance with the required rate, and to verify whether the different instructions given to participants in each experiment resulted in different word durations, which would indicate different task strategies.

# Methods

## Experiment 1

The speech was elicited with a multiple picture naming task, forcing speakers to complete all planning phases before articulation of each picture name could begin. Different sets of eight pre-familiarised line drawings were displayed in each trial, in an arrangement reminiscent of a clock face (c.f. Meyer, Wheeldon, Van der Meulen, & Konopka, 2012). A cursor indicated which picture was to be named, moving in a clockwise direction from picture to picture at three predetermined, participant-independent rates: fast, medium, and slow.

**Participants** Native Dutch participants ($N = 12$, two males, ten females, $M_{age} = 22$ years) with normal hearing and normal or corrected-to-normal vision were recruited from the participant pool of the Max Planck Institute for Psycholinguistics, with informed consent as approved by the Ethics Committee of the Social Sciences Faculty of Radboud University (Project Code: ECSW2014-1003-196).

**Materials** Twelve disyllabic Dutch concrete nouns with stress on the first syllable were selected as target words for the production experiment. The first syllable was always open, and the second syllable was always closed (C(C)V.CVC, where C = consonant, V = vowel). Vowels were always monophthongs and consonants were never stops. This means that we selected only segments where the articulators do not move during the production of the segment, in contrast to diphthongs or stop consonants, where changing articulatory configuration during the segment is inherent to the segment identity. This is required for the derivation of the onset and offset times of syllables within words. Additionally, words with an ambisyllabic consonant were excluded. This yields words such as *snavel* [ˈsnaː.vəl] "beak", *vriezer* [ˈvriː.zər] "freezer" and *wafel* [ˈwaː.fəl] "waffle". In addition, twelve similar filler words were selected. A full list is provided in the Appendix.

Lists of 70 sets of eight words were pseudo-randomly created from the vocabulary of filler and target words, one set for each of the 70 trials in each rate condition. A different list was used for each participant in each rate condition. Within each set, no word appeared more than once. Within each list, the number of times each word was used was matched as closely as possible (average s.d. in used lists: 0.4769, minimum frequency 26 words per 560, maximum frequency 28 words per 560), as was the frequency with which each word appeared in each of the five "target" positions on the clock face (average s.d. in used lists: 2.107), and the frequency of each pair of words co-occurring in a set (an analogue of transition probability, average s.d. in used lists: 2.294).

For each word, a line drawing was either taken from the Snodgrass and Vanderwart (1980) picture library, or prepared in the same style (see the Appendix).

**Experimental procedure** Participants were tested individually in a sound attenuated booth. Stimulus presentation, eye-tracker synchronisation and audio recording were controlled by Presentation software (Version 16.5; Neurobehavioral Systems, Berkeley, CA, USA). A Sennheiser ME64 directional microphone was used to record the participants' speech at a sampling rate of 48 kHz.

The session began with familiarisation of the pictures and their names, by means of (1) a printed card and (2) naming of the pictures as they were displayed individually on screen, in a pseudo-randomised order with two repetitions of each picture. The experimenter immediately gave the correct name when the participant named a picture incorrectly. After the structure of the experiment was described (three blocks, each at a different rate condition, in a random order), the participant was instructed to *"name the exact picture that the marker indicates"*.

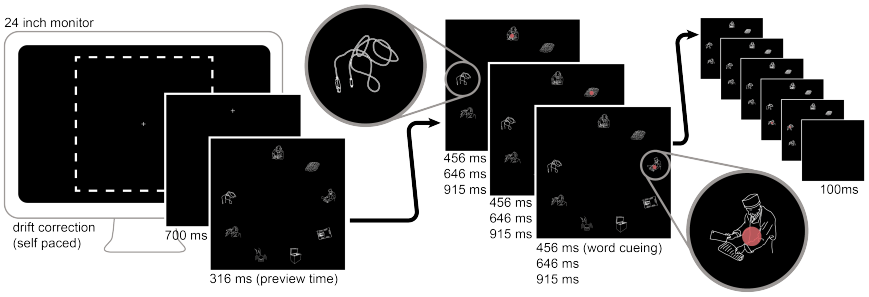



Figure 1: The trial sequence. The trial began with a drift-correction fixation cross (self-paced). A fixation cross was then presented at the location of the first picture for 700 ms, followed by 316 ms preview time. Cueing then began: each word was cued for 456, 646 or 915 ms by overlaying a translucent red dot on the relevant picture. The trial concluded with a blank display for 100 ms.

They were instructed to achieve slow speech rates by slowing down their speech, not by producing longer pauses in between words. In this fashion, we attempted to ensure that we would elicit variation in the way individual words were articulated, rather than variation in the usage of pauses. Instructions were presented on screen. Six practice trials at the medium rate then followed, after which the remote eye-tracker (Eyelink 1000 in remote mode; SR Research, Ottawa, ON, Canada) was prepared and calibrated with a standard 9-point calibration procedure. The gaze position measurements, originally collected with future computational simulation work in mind, are not discussed in this article.

A block of seventy trials was presented for each rate condition, followed by a short break. The order of the three rate blocks presented in the experimental session was counterbalanced across participants.

The trial structure is illustrated in Figure 1. Before each trial, the participant performed a self-paced drift-correction procedure for the eye tracking measurements. After successful drift-correction, a fixation cross was presented at the location of the first picture ("12 o'clock") for a duration of 700 ms. Then, the pictures appeared without the cursor, and were presented for 316 ms of "preview time", to allow the participant to prepare for naming.

The pictures were displayed in sets of eight, in a clock-face arrangement with 9 positions. Positions 2 to 6 were occupied by target pictures. The first, seventh and eighth positions were occupied by filler pictures, since these positions were expected to be particularly susceptible to listing intonation. The ninth position (at "10 o'clock") was always left empty to visually reinforce the beginning and end of the sequence of pictures. This arrangement is illustrated in 1. The whole display fitted into an area of 780 x 780 pixels. Each picture was scaled such that it would occupy an area of 90 x 90 pixels.

Once the preview time had elapsed, a cursor was overlaid on each picture in turn for the duration appropriate to the rate condition; fast, medium or slow. The cursor was a translucent red circle with a diameter of 20 pixels, which appeared in the centre of each picture whilst that picture was to be named. The cursor jumped from picture to picture, starting with the topmost and proceeding in a clockwise direction. After all the pictures had been cued, the pictures and the cursor disappeared and a blank screen was presented for 100 ms, after which the drift correction procedure for the next trial started immediately.

The three cueing rates tested were 456 ms/word (2.19 Hz, fast condition), 646 ms/word (1.54 Hz, medium condition) and 915 ms/word (1.09 Hz, slow condition). These rates were derived from the non-cued speaking rates realised by three further participants in a small pre-test (research assistants with a similar background to the participants tested in the main study). This pre-test data was also used to establish an appropriate length of "preview time" to allow the participants to prepare for the naming task.

## Experiment 2

The second elicitation experiment was identical to the first experiment, except participants were given no explicit instruction to avoid pausing. For Experiment 2, 13 further participants were tested (two males, eleven females,

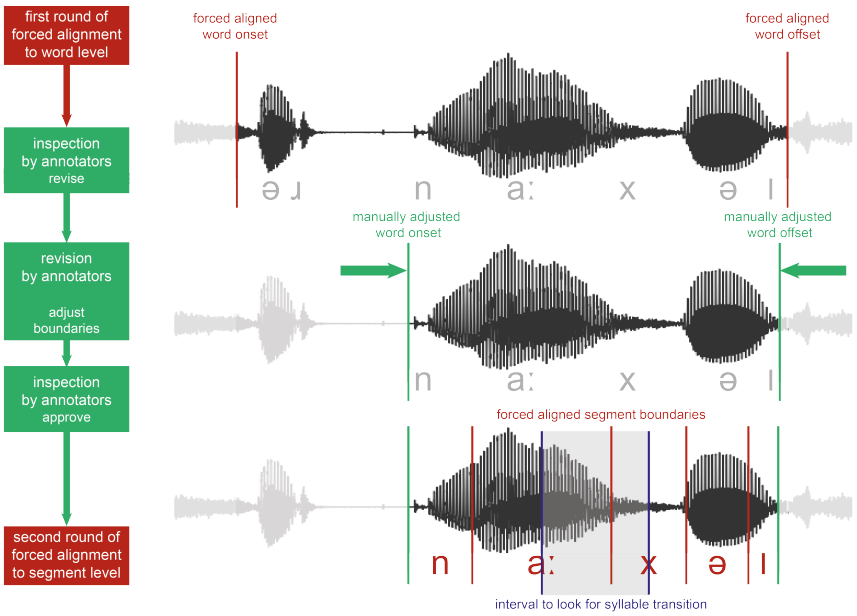



Figure 2: An example of the pipeline for annotating the word *nagel* [ˈnaːxəl] "fingernail". First, an the initial forced alignment run identifies candidate word boundaries. These are inspected by a human annotator. In this case, they are wrong, so the word is marked as needing revision. Later, the same or another annotator adjusts the boundaries. The revised word is checked again, and approved. Then, forced alignment is applied to the single word recording, to identify segment boundaries. An interval is defined, spanning from the centre of the vowel of the first syllable to the centre of the first consonant of the second syllable, as identified by the segment level forced alignment. This interval is used to direct the search for the syllable transition, using the metric developed by Rodd, Bosker, ten Bosch, and Ernestus (2019).

$M_{age}$ = 22 years), recruited from the same pool of native Dutch speakers as the participants tested in Experiment 1, under the same ethics approval.

## Word boundary finding

The extent of the speech data collected (5,250 trials, yielding up to 26,250 target words and 15,750 filler words if no errors were made) precluded fully manual annotation. A fully automatic annotation was also not possible since the nature of the task resulted in many hesitations, omissions and deviations from the canonical productions. Instead, a multi-step acoustic analysis and forced alignment pipeline, POnSS (Pipeline for Online Speech Segmentation), was used to create automatic transcriptions of the speech materials, which were then adjusted as necessary by a panel of ten phonetically trained annotators, including research assistants and the first author. POnSS was developed and validated by Rodd, Decuyper, and ten Bosch (2019). For completeness, we also describe it here. Use of the pipeline results in equivalently reliable transcriptions compared to conventional annotation with Praat software, with greater annotator comfort and greater time efficiency.

The POnSS pipeline is illustrated for an example word in Figure 2. First, the harmonicity (autocorrelation method, default settings) of the trial recordings was analysed using Praat software (Version 6.0.18, Boersma & Weenink, 2015). Each harmonicity peak can be assumed to correspond to one vowel in the recording, allowing the number of disyllabic words produced (i.e. not omitted) to be estimated. We observed that when speakers produced fewer than the full eight words, the words occurring later in the sequence were much more frequently omitted than earlier ones. Based on this observation, the peak counts were used to produce candidate orthographic transcriptions for the forced alignment. If fifteen or sixteen harmonicity peaks were detected (indicating sixteen syllables), all eight words were included in the transcription. If there were fourteen or fifteen, the first seven words were included, and so on. This was done with the aim of achieving better forced alignment results than simply forced aligning against the "script" including all eight words would have done.

From these candidate orthographic transcriptions, forced alignment to the word-level was performed using the MAUS software (Schiel, 2015), which offers good quality forced alignment for Dutch using HTK (Young et al., 2006).

A specially constructed web application using the Django framework (Holovaty & Kaplan-Moss, 2009) was used by the annotators to screen out words that had been poorly aligned or labelled by MAUS and therefore needed revision. Each annotation was presented individually with the waveform and spectrogram of the relevant audio. Annotators could listen to the audio as many times as they wished. For each of the 23,218 annotations produced by MAUS, they decided whether the complete word was isolated, with no material from surrounding words included. If that was not the case (because, for example, some part of the word was missing, or part of the following word was included), they flagged the annotation as requiring further attention. They also had the option to discard annotations containing non-speech or speech errors, 1,095 annotations were discarded for this reason.

The annotations that were flagged by any one of the annotators, but were not outright discarded (81.5% of 5,400 words from the fast rate; 64.5% of 7,864 words from the medium rate; 45.6% of 7,812 words from the slow rate) were subsequently re-trimmed by other annotators from the panel. This was done by dragging word boundaries on a visual display of the waveform and spectrogram.

### Automatic syllable boundary finding

After annotation, syllable onset and offset times were derived using the automatic metric developed and validated by Rodd, Bosker, ten Bosch, and Ernestus (2019), using an analysis interval spanning from the centre of the vowel of the first syllable to the centre of the first consonant of the second syllable, as identified by the phone-level forced alignment. Syllable planning units tend to overlap, so a method was required to identify the onsets and offsets of syllables where they overlap with neighbouring syllables. The dynamics of the acoustics of speech broadly reflect the dynamics of the articulation that produces it: when the configuration of the articulators is stable, the acoustic signal is also stable. It was therefore possible to identify periods of articulatory stability from the acoustic signal, and periods of transition. We interpreted the period of acoustic transition (the *acoustically evident planning unit overlap*) as coterminous with the period of planning unit overlap, allowing us to identify the onsets and offsets of planning units from the acoustic signal. A similar approach was adopted by Hoang and Wang (2015) to identify phone transitions.

## Confirmatory analysis: word duration

To confirm that participants were indeed performing the task as we expected, that is, primarily modulating speaking rate rather than merely adjusting pause durations, we first examined overall word durations.

A Bayesian mixed effects model was constructed using the brms R package (Bürkner, 2018; R Development Core Team, 2008; Stan Development Team, 2018) to model the log-transformed word duration. We used the log-transformed word duration in order to reduce skewness in the distribution.

Dummy coded fixed effects of cueing rate (categorical predictor, dummy coded with medium rate on the intercept) and experiment (categorical predictor, dummy coded with Experiment 1 on the intercept) were included in the model, along with the interaction of cueing rate by experiment. Random intercepts were included for speaker. Random slopes were included for the log-transformed trial-level residual rate for each speaker-cueing rate combination, grouped by experiment. The trial-level residual rate was calculated as the difference between the realised speaking rate in a trial (the total contiguous speaking time divided by the number of words produced) and the target rate (the duration for which each word was cued with the cursor). For the cueing rate predictor, low-informative priors were set centred at each target speaking rate, with a $\sigma$ of 3.4 log ms (equivalent to 30 linear ms). For the effect of experiment, a normally distributed low-informative prior was set, centred at 0

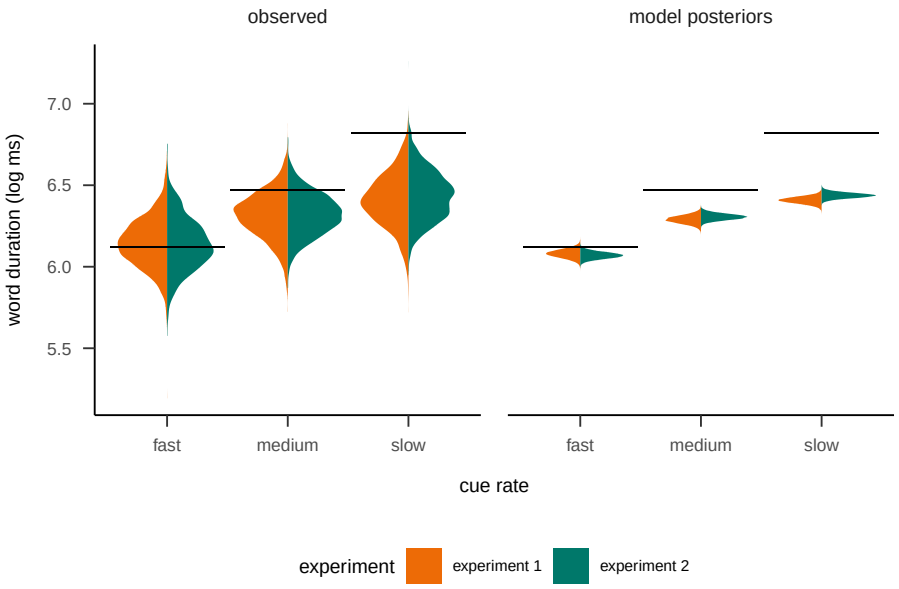



Figure 3: Left panel: word duration (log ms) as a function of cued speaking rate, plotted as violins (the width of the violin shows the distribution of values). The colour indicates which experiment the data come from. Black horizontal lines indicate the target speaking rates. Right panel: the model posterior distributions for the mean, shown as violins.

with a $\sigma$ of 3.4 log ms (equivalent to 30 linear ms, roughly five times the noise prevalent in the annotation task Rodd, Decuyper, & ten Bosch, 2019). The model was well converged (assessed by the Gelman-Rubin diagnostic $\widehat{R}$, effective number of samples and visual inspection of traceplots) after running eight chains of 3,000 warm-up and 3,000 critical iterations.

The observed durations of the words produced in each condition are presented in the left panel in Figure 3, measured from the onsets and offsets established by the annotation procedure described in section 2.4. The posterior distributions for the means of each speaking rate in each experiment are shown in the right panel, along with HDI (highest density interval) covering 95% of the posterior.

The results of the Bayesian mixed effects model are summarised in Tables 1 and 2. The model confirmed that, in both experiments, speakers produced shorter words in the fast condition than in the medium condition, and longer words in the slow condition than in the medium with large effects (Cohen's d minimally 0.687, maximally 1.366). Since there were large differences in word duration between each rate condition, we concluded that the participants produced different speaking rates for each cueing condition. However, in all cases, the speakers produced words somewhat shorter than the target rate. This effect is smallest for the fast cueing condition and largest for the slow cueing condition. This arises because the target rates assume continuous production without pauses between words. This suggests that speakers were, even when explicitly asked to try to modulate their word duration, also modulating pause duration to comply with the cued speaking rate.

The model also confirmed that there was no effect of experiment, since all 95% credible intervals included 0, and all effect sizes were small (Cohen's d minimally 0.061, maximally 0.168), and all 95% credible intervals overlapped with a ROPE (region of practical equivalence; Kruschke, 2018) defined to include all effects smaller than 15ms, a reasonable estimate of the degree of noise prevalent in annotation data (Rodd, Decuyper, & ten Bosch, 2019). This means that the word durations measured from speakers instructed to try to avoid pausing between words did not differ from those who did not receive this instruction.

Table 1: Results of the Bayesian mixed effects model for comparisons of realised word duration by cued rate, within experiments.

| experiment | comparison | estimate | CI | Cohen's d |
|---|---|---|---|---|
| experiment 1 | medium → fast | -0.210 | [-0.172, -0.247] | -1.213 |
| experiment 1 | medium → slow | 0.119 | [0.156, 0.081] | 0.687 |
| experiment 2 | medium → fast | -0.236 | [-0.199, -0.275] | -1.366 |

| | | | | |
|---|---|---|---|---|
| experiment 2 | medium → slow | 0.132 | [0.169, 0.093] | 0.761 |

Table 2: Results of the Bayesian mixed effects model for comparisons of realised word duration by experiments, within cued rates.

| cued rate | comparison | estimate | CI | Cohen's d | ROPE percentage |
|---|---|---|---|---|---|
| medium | experiment 1 → experiment 2 | 0.016 | [0.055, -0.022] | 0.093 | 99.49% |
| fast | experiment 1 → experiment 2 | -0.010 | [0.032, -0.054] | -0.061 | 99.36% |
| slow | experiment 1 → experiment 2 | 0.029 | [0.069, -0.01] | 0.168 | 97.6% |

## Data availability

The speech materials for 25 participants, consisting of trial-level recordings,, along with the onset and offset times of the words and the syllables in csv format and as R data format are archived at the Language Archive, and available on request from https://hdl.handle.net/1839/7c210d30-bb55-4cbe-9eeb-baf18570460c.

## References


Boersma, P., & Weenink, D. (2015). Praat: Doing phonetics by computer [computer program]. Version 5.4.15. Amsterdam: University of Amsterdam. Retrieved August 15, 2015, from http://www.fon.hum.uva.nl/praat/

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395–411. Retrieved September 26, 2019, from https://journal.r-project.org/archive/2018/RJ-2018-017/index.html

Hoang, D.-T., & Wang, H.-C. (2015). Blind phone segmentation based on spectral change detection using Legendre polynomial approximation. *The Journal of the Acoustical Society of America*, *137*(2), 797–805.

Holovaty, A., & Kaplan-Moss, J. (2009). *The definitive guide to django: Web development done right*. Apress.

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. doi:10.1177/2515245918771304

Meyer, A. S., Wheeldon, L., Van der Meulen, F., & Konopka, A. (2012). Effects of speech rate and practice on the allocation of visual attention in multiple object naming. *Frontiers in psychology*, *3*.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

Rodd, J., Bosker, H. R., Ernestus, M., Alday, P. M., Meyer, A. S., & ten Bosch, L. (2019). Control of speaking rate is achieved by switching between qualitatively distinct cognitive 'gaits': Evidence from simulation. *Psychological Review*. In press. doi:10.1037/rev0000172

Rodd, J., Bosker, H. R., ten Bosch, L., & Ernestus, M. (2019). Deriving the onset and offset times of planning units from acoustic and articulatory measurements. *The Journal of the Acoustical Society of America*, *145*(2), EL161–EL167. doi:10.1121/1.5089456

Rodd, J., Decuyper, C., & ten Bosch, L. (2019). Efficient, reliable semi-manual annotation of speech materials with POnSS. Manuscript in preparation.

Schiel, F. (2015). A statistical model for predicting pronunciation. In M. Wolters, J. Livingstone, B. Beattie, J. Stuart-Smith, & J. Scobbie (Eds.), *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow: University of Glasgow.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology: Human learning and memory*, *6*(2), 174.

Stan Development Team. (2018). RStan: The R interface to Stan. R package version 2.18.2. Retrieved from http://mc-stan.org/

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., ... Povey, D., et al. (2006). The HTK book (for HTK version 3.4). *Cambridge University Engineering Department*, *2*(2).

## Appendix: elicitation materials

Table 3: Filler words were included in the first, penultimate and last slots of each trial.

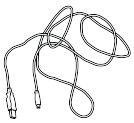
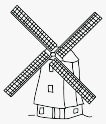
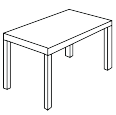
| orthography | phonetic form | meaning | |
|---|---|---|---|
| gieter | ˈxi.tər | watering can |  |
| kabel | ˈkaː.bəl | cable |  |
| lasser | ˈlɑ.sər | welder |  |
| lichaam | ˈlɪx.aːm | body |  |
| molen | ˈmoː.lən | windmill |  |
| monnik | ˈmɔ.nɪk | monk |  |
| spiegel | ˈspi.xəl | mirror |  |
| tafel | ˈta.fəl | table |  |
| trommel | ˈtrɔ.məl | drum |  |

Table 3: Filler words were included in the first, penultimate and last slots of each trial. *(continued)*

| orthography | phonetic form | meaning | |
|---|---|---|---|
| vinger | ˈza.ŋər | finger |  |
| zanger | ˈza.ŋər | singer |  |

Table 4: Target words were included in the second to sixth slot of each trial.

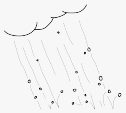
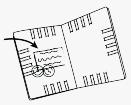
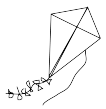
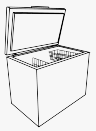
| orthography | phonetic form | meaning | |
|---|---|---|---|
| hagel | ˈhaː.xəl | hail |  |
| hamer | ˈhaː.mər | hammer |  |
| havik | ˈhaː.mər | hawk |  |
| nagel | ˈnaː.xəl | fingernail |  |
| navel | ˈnaː.vəl | navel |  |
| sinus | ˈsi.nʉs | sine wave |  |
| slager | ˈslaː.xər | butcher |  |
| snavel | ˈsnaː.vəl | beak |  |

Table 4: Target words were included in the second to sixth slot of each trial. *(continued)*

| orthography | phonetic form | meaning | |
|---|---|---|---|
| visum | ˈvi.sum | visa |  |
| vlieger | ˈvli.xər | kite |  |
| vriezer | ˈvri.zər | freezer |  |
| wafel | ˈwaːˌfəl | waffle |  |
| zoemer | ˈzu.mər | alarm |  |