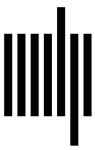
This is a section of doi:10.7551/mitpress/12200.001.0001

The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022



The MIT Press

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122710/c045900_9780262366076.pdf by guest on 27 September 2023

50 Managing Speech Perception Data Sets

Anne Cutler, Mirjam Ernestus, Natasha Warner, and Andrea Weber

1 Speech perception data sets

Sizeable sets of speech perception data can be highly valuable to researchers; consider that the early reports on the identification of American English vowels (Peterson & Barney 1952) and consonants (Miller & Nicely 1955) have racked up citation counts, respectively, of 4329 and 2455 (in June 2021; after more than half a century, the citations are still coming in). Understandably, most experiments on the perception of speech are focused on specific questions (testing models of spoken-language comprehension, comparing processing across structurally different languages, assessing perceptual outcomes for differing listener populations) and accordingly use the minimum data set size necessary for their target statistical power (although see Sedlmeier & Gigerenzer 1989). But other reasons to collect speech perception data can make for large data sets with wide relevance or usefulness.

One reason might be to build the basis for a computational system such as an automatic speech recognizer. A database of human recognition achievement to support such a computational system will need to have broad scope; for example, it should include all phonemes or syllables of the language in question, ideally in every potential phonemic context. This then makes for a data set on which comparative perceptual questions of many other kinds can be tested. Another reason might be to set norms for useful stimulus selection control measures, such as the relative effects of word occurrence frequency in listening, or the effects of vocabulary structure such as word class. Again, once the norms are established, the data sets remain useful for answering a range of further questions. As long as the data sets are easily accessible, they can be put to many uses, similar to the method of "virtual experiments" suggested by Kuperman (2015). Thus, these large data sets on speech perception can be viewed both as "big data" and as "open data" (Borgman 2015).

The following section covers four data sets. The first two indeed came into being in service of a computational model: SHORTLIST-B (Norris & McQueen 2008), a Bayesian model of human spoken-word recognition that draws its probability estimates from human recognition scores for all possible two-phoneme speech sound sequences in all their legal contexts in the language (which in this case is actually the two languages Dutch and English, enabling two language-specific SHORTLIST-B instantiations). These two studies were nicknamed DADDY (the Dutch Auditory Diphone Database) and EDDY (the English Diphone Database), respectively, but given their identical design and guiding principles they are described together in section 2.1. The third study was analogously nicknamed NINNY (Noise-masked Identifications by Native and Non-native listeners) and as described in section 2.2 was aimed mainly at providing a larger-scope data set able to clear up contradictory claims based on smaller-scale studies, many of these incorporating also confounding properties such as variations in lexical familiarity. The fourth study (section 2.3) was motivated by the proven usefulness of an existing data set of responses in a lexical decision study using visually presented stimuli; it was the first set of such lexical decision data using auditory presentation and was accordingly named BALDEY, because it was then the Biggest Auditory Lexical Decision Experiment (Yet). In the following sections, we approach each data set from the questions of "Why?" (what was the goal of the project), "How?" (how was the study done), "What?" (main results), "For whom?" (who the users of the data set are), and "Where?" (data management and storage).

Accessibility of such data is here defined as open access (OA) to the speech stimuli presented and the resulting

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122710/c045900_9780262366076.pdf by guest on 27 September 2023

listener responses. These four data sets all include both stimuli and responses (the latter being response times, or identification choices, or both). Participant-identifying features are anonymized; the issues here for language data are discussed by Warner (2014). Each of the data sets in section 2 has been made available in more than one OA way, and together they cover several options. The choices we have made about how to make these data sets available, and about what information to store in them, relate to the issues of data management and planning discussed by Mattern (chapter 5, this volume). Even when the plan from the beginning is to create a large data set and make it publicly available, the issues Mattern discusses still lead to challenges. All of these data sets are freely licensed for use by others. Authors using these data sets should cite the relevant publication reporting each data set in order to give credit appropriately. Because the purpose of all of these data sets is basic science, we considered this sufficient without requiring other researchers to ask our permission for further use. See Collister (chapter 9, this volume) for discussion of copyright issues for such data sets.

Further questions arise with respect to the use of large OA data sets on speech perception such as these. An obvious initial question is: how do other researchers find such data sets? One method of making data sets known is based in traditional scientific publication: for each of the projects discussed herein, the research team published one or more papers on specific analyses of the data set before making the data publicly available, with each such publication, from the first one on, at least mentioning the availability of the data set for other researchers. The research teams also gave talks at major conferences that served both to present specific results from the data sets and to publicize their availability for further studies to other researchers. Internet searches are a likely method of turning up such data sets when one is seeking them. However, it would be desirable to find additional ways to publicize the existence of such data sets that would continue to reach audiences after conference talks on the projects have concluded.

A separate issue concerns how other researchers who succeed in finding and downloading these data sets then make use of and maintain them. Individual researchers are encouraged to download a copy of the data files for local use. A given researcher is likely to edit the data file to add additional columns during the process of analyzing the data to answer their own questions. At this point, other researchers using these data files need to take responsibility for keeping track of any restructuring they have done and for citing the source of the original data correctly.

2 DADDY, EDDY, NINNY, and BALDEY

2.1 Diphones: Identification of sounds in two-sound sequences over time

2.1.1 The goal (Why?) The goal of the Diphones project (DADDY: Smits et al., 2003; EDDY: Warner, McQueen, & Cutler, 2014) was to provide information about how listeners extract acoustic cues to segments from *all* possible sequences of two sounds of a language over time. Data were collected for two languages: Dutch and English. Many detailed findings were already available on how native listeners perceive specific sequences of sounds, such as /ba, da, ga/, /f/ versus /θ/ before vowels, and so on. However, those studies did not all use the same methods so that the results are not comparable.

The principal stimulus for the project was to provide input data for a probabilistic model of spoken-word recognition, the SHORTLIST-B model of Dutch spoken-word recognition (Norris & McQueen 2008), and a corresponding English model. The Diphones data provide an account of what sounds listeners think they are hearing as the speech signal unfolds. This information can then be fed into the model of what words are consequently considered as candidates for word recognition. For example, if the input is actually "book," the Diphones data show at what probability listeners believe that they are hearing /bʊ/ and then /ʊk/ as the signal progresses. The data also provide information about the probability at any given time with which the listeners may think they have heard something else, such as /bu/, /bʌ/, or /pʊ/. This information probabilistically influences the model's estimation of how likely the listener is to think they are hearing the word *book* as opposed to the words *boot*, *but*, or *put*. For this purpose, the data must include information for all diphones that could occur in the language, even across word boundaries, to allow modeling of recognition of any string of words.

Thus for example the English Diphones stimulus set includes vowel-vowel (VV) diphones such as /oʊaʊ/ (as in "row out") and consonant-consonant (CC) /pʃ/ (as in "upshot") as well as the more commonly studied CV (/ba/ but also the less common /ðoɪ/) and VC (/ab/ but

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122710/c045900_9780262366076.pdf by guest on 27 September 2023

also /ʊv/). All diphones that cannot be ruled out as impossible in the language are included, even if they could only occur across a word or morpheme boundary, as for example /ðv/, which does not occur within any word of English, but could occur in the sequence "loathe vegetables." Thus, phonotactically impossible diphones such as /ɛh/ are the only ones excluded (in English, no syllable can end in a lax vowel like /ɛ/ and /h/ cannot appear in a syllable coda, so this diphone cannot occur even across a word boundary).

2.1.2 The study (How?) The stimuli for each study comprised a list of all the possible two-sound sequences of the language, whether CV, VC, CC, or VV (e.g., /ba, iz, fp, io^{ʊ}/). For diphones containing a vowel, a version with the vowel stressed and a separate version with the vowel unstressed were used, thus there was a stressed and an unstressed /ba/ diphone, and four /io^{ʊ}/ diphones (stressed-stressed, stressed-unstressed, unstressed-stressed, and unstressed-unstressed). Each of these languages has almost 2,300 possible diphones, when stress is counted in this way.

Each diphone (two-sound sequence) was gated at six time points at thirds of the duration of each segment, so that on Gate 1, listeners heard only the first third of the first segment, while on Gate 5 they heard from the beginning of the diphone through two-thirds of the duration of the second sound. Gate end points were placed at thirds of the duration of each segment for most segment types. Thus, for example, Gate 1 of /sa/ would allow listeners to hear from the beginning of the /s/ up until one-third through the /s/; Gate 3 would allow them to hear from the beginning up to the end of the /s/; and Gate 5 would allow them to hear from the beginning of the /s/ up to two-thirds through the /a/. (Only for stops and affricates, the end points of Gate 2 or Gate 5 were set just before the onset of the burst rather than at two-thirds through the duration, so that the burst and aspiration/frication noise always occurred within the same gate.) At the gate end point, the amplitude of the speech was ramped down over the course of five milliseconds while the amplitude of a square wave was simultaneously ramped up and added to the speech wave. The square wave (resembling a computer beep) then continued for a few hundred milliseconds. The amplitude ramp from speech to square wave and presence of the square wave prevents the creation of artifactual cues to some sound. The beep also helps encourage listeners to believe that more sound would follow if they were allowed to hear the entire string. Regardless of gate, the listeners' task was always to identify what two sounds they heard or might have heard the start of, even though at Gate 1 they might be extremely unsure. Full details of methods and stimulus creation appear in Smits et al. (2003) and Warner, McQueen, and Cutler (2014). There were over 12,000 stimuli for each language.

Eighteen Dutch and twenty American English native listeners participated in the experiments (each for only their own language). Each listener heard all the stimuli for their language once, visiting the lab for a series of up to thirty sessions, which were each an hour long. All stimuli were randomized. On hearing each stimulus, listeners had to identify both the first and second sound. For a Gate 6 stimulus, the listener might hear both sounds of a stimulus /az/ clearly and be able to identify both sounds correctly. However, at Gate 1 of the same stimulus, the listener might be only somewhat sure what the first segment was and have no idea what the second segment was at all, as Gate 1 ends at one-third through the duration of the first segment, and very little information about the upcoming /z/ spreads into the first third of the preceding /a/. In this case, the listener would have to respond to the second segment by guessing. For each stimulus, listeners saw a computer screen showing buttons for every phoneme of the language on the left half of the screen for the first segment response and the right half of the screen for the second segment response. They used the computer mouse to select what two sounds they heard or might have heard from among the full phoneme set of the language. See Smits et al. (2003) and Warner, McQueen, and Cutler (2014) for additional details. Because each listener gave two judgments (first and second segment) for each of the more than 12,000 stimuli, the total data set across both languages comprises approximately a million judgments.

2.1.3 Main results (What?) The Diphones studies of Dutch and English have elucidated or confirmed many patterns about the timing of speech perception. For example, the results (for Dutch also in Warner et al. 2005) make the difference very clear between segments that strongly change quality over the course of the segment (diphthongs and affricates) and those that remain relatively stable. Listeners seem to perceive whatever sound they hear during the stimulus as a phoneme and do not allow for the possibility that additional

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122710/c045900_9780262366076.pdf by guest on 27 September 2023

acoustic cues to the sound they are currently hearing could still follow. Therefore, if the stimulus ends during the closure of an affricate (e.g., Gate 5 of /at͡ʃ/, which ends just before the burst), listeners typically respond with a stop rather than the affricate. The data for affricates therefore shows very poor perception up until the frication noise, and then a very steep and sudden improvement in perception accuracy when the stimulus includes frication noise of the release. A similar effect for diphthongs at the stimulus that first includes the second quality of the vowel means that patterns of recognition for diphthongs are delayed relative to the patterns for recognition of monophthongs. Diphthong perception accuracy generally lags behind accuracy for monophthongs by one gate (one-third of duration of the segment).

The Diphones studies also allow comparison of speech perception in English and Dutch. One major comparative finding is that unstressed vowels are recognized far more poorly than stressed ones in English, while this effect is small and limited to a few vowels in Dutch. In Warner and Cutler (2017), we argue that this difference comes not from acoustic differences in the unstressed vowel space, but rather from listeners' differing need to distinguish among unstressed vowel qualities. Dutch has more unstressed vowels with full vowel quality (not schwa-like quality), while unstressed English vowels are usually schwa. Hence there is more potential for the quality of an unstressed vowel to aid the listener in determining what word they are hearing in Dutch. Dutch listeners therefore pay more attention to vowel quality even in unstressed vowels than English listeners do.

2.1.4 The users (For whom?) The Diphones data set is primarily of interest for two groups of researchers: those interested in questions about speech perception, and those interested in modeling spoken-word recognition. The data can most straightforwardly be used to answer questions about what information listeners can extract from the acoustic signal at what time point. In addition to our own work, graduate students at other institutions have contacted us about uses they are making of the data for speech perception topics. For modeling of spoken-word recognition, the Diphones data provide input data for SHORTLIST-B, but could also be used for other models. Indeed the Diphone studies are cited by these primary interest groups. However, the papers have also been regularly cited on issues of language acquisition (Altvater-Mackenson, van der Feest, & Fikkert 2014; Law & Edwards 2015; Wagensveld et al. 2013) and more recently also with respect to clinical research (Hajiaghbaba, Marateb, & Kermani 2018) and historical linguistics (Minkova 2016).

2.1.5 The data management (Where?) The Dutch diphones data were initially made available through the Max Planck Institute for Psycholinguistics (Nijmegen) website, and a reference to this site was included in the 2003 publication. None of the authors still work there, however, and the website has been radically upgraded several times, making it difficult to maintain the accessibility. That data set was also deposited in the Alveo Human Communication Science Virtual Lab, a secure Australian repository accessible from Research Data Australia (http://researchdata.ands.org.au/human-communication-science-virtual-laboratory-hcs-vlab). At the time of writing, the Dutch materials remain available at MPI (https://www.mpi.nl/world/dcsp/diphones/index.html). Both the Dutch and English diphones data sets are available through Warner's website and her lab's website (https://nwarner.faculty.arizona.edu/content/7). We plan to deposit the data at an online location that is intended as a long-term archive, such as the University of Arizona library system's archive.

For both Diphones projects, all stimuli and all responses are available. Researchers can thus calculate any response percentages or confusion matrices they seek, or they can work directly with the raw individual responses. A large zipped file containing all the stimulus sound files (including the appended square wave beep) can be downloaded for each project, enabling acoustic analyses of the stimuli for comparison to responses. Both languages' data sets also have a README file documenting transcription systems, file organization, and such. The Dutch files on the MPI site contain some additional materials, such as premade confusion matrices and the label files used to create the stimuli from the recordings (which contain information about where boundaries were placed).

2.2 NINNY: Noise-masked Identifications by Native and Non-native listeners

2.2.1 The goal (Why?) Adverse listening conditions, for example noisy backgrounds, disrupt listening to non-native speech more strongly than they disrupt listening to native speech (see Garcia Lecumberri, Cooke, & Cutler 2010 for a review). The main goal of the study

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122710/c045900_9780262366076.pdf by guest on 27 September 2023

NINNY (Cutler, Weber, Smits, & Cooper, 2004) was to identify the source of this asymmetry. One obvious possibility was that this disadvantage for non-native listeners was due to greater difficulty in phoneme identification. Where the phoneme categories of a non-native language fail to match those of the native language, phonetic decisions can be influenced by the native repertoire (e.g., Strange 1995), and this influence may become stronger when stimuli are harder to perceive, for example, because they are embedded in noise. In order to render higher-level factors such as lexical frequency or contextual plausibility irrelevant, Cutler et al. tested phoneme identification in VC or CV syllables in noise. Identification responses by American English (native) and Dutch (non-native) listeners to all American English vowels and consonants were collected under three levels of noise masking.

Phonetic identification data of any kind are highly valuable for speech comprehension research. They are valuable because sounds differ in how easily they can be recognized (even in native listening), and the data sets provide identification accuracy and confusion patterns sound by sound. It is also possible to estimate the contribution of phoneme perceptibility to recognition of any spoken word with such data sets. Such large data collections are scarce, however, because collecting the data is time-consuming and laborious. Data for non-native listening are even harder to find, because data for any given language pair might not provide a full set of answers relevant to another language pair.

2.2.2 The study (How?) For the NINNY data set, native speakers of American English and Dutch non-native speakers of English listened to English syllables and identified either the consonant or the vowel. All 645 possible standard CV and VC sequences of American English, excluding those with schwa, were recorded by a female native speaker and centrally embedded in one second of multi-speaker babble noise. The multi-speaker babble was combined with the test syllables at three levels of signal-to-noise ratios (SNRs): zero, eight, and sixteen decibels. These SNRs were chosen on the basis of a pretest to yield difficult, intermediate, and easy English phoneme perception for Dutch non-native listeners. Sixteen native listeners of American English and sixteen non-native Dutch listeners who were highly proficient in English were presented with the syllables in noise, and identified each phoneme of each syllable at each noise level separately (3,870 trials per listener). Testing was spread across eight sessions, each lasting approximately thirty to forty minutes. To guide listeners' responses, illustrative words for all phonemes were shown on a display (e.g., the word *very* for the consonant response /v/), and listeners signaled responses by clicking on the word matching the phoneme they decided was presented. Collected responses comprised correct responses (e.g., a click on *very* when identifying the consonant in the syllable /vi/) as well as errors (e.g., a click on *very* when identifying the consonant in the syllable /bi/). The full identification response set contains 123,840 data points in total: 32 participants, each taking part in eight sessions and contributing 3,870 identification responses.

2.2.3 Main results (What?) With these isolated syllables, all listeners were adversely affected by an increase in noise, and the phoneme identification performance of non-native listeners was overall less accurate than that of native listeners. Crucially, however, the disadvantage for non-native listeners was not proportionally greater at higher noise levels. It was concluded that the frequently reported asymmetry of non-native versus native listening under difficult listening conditions is not due to greater masking and hence greater difficulty of phoneme identification, but rather to non-native listeners' lesser, and less efficient use of, higher-level information (e.g., lexical and statistical information) for recovery from the effect of noise masking.

The combination of native and non-native phonetic identification data is important because it allows us to distinguish the roles of general auditory and language-independent processes from those involving prior knowledge of a given language. Thus a principal theory-driven finding of the NINNY study was that non-native listeners do not need better low-level evidence than native listeners do (i.e., a less noisy environment) to overcome listening difficulties; instead, they could best match native performance by having a larger vocabulary and increased listening experience.

2.2.4 The users (For whom?) The NINNY data set is of interest for researchers who work on either native listening or non-native listening (in comparison to native listening). By November 2019, the 2004 NINNY study had received 280 citations (Google Scholar). The data have been analyzed to compare phoneme confusion patterns to predictions of speech perception models (e.g., Silbert & de Jong 2007) and to guide the selection of sound

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122710/c045900_9780262366076.pdf by guest on 27 September 2023

contrasts for word recognition studies (e.g., Darcy, Daidone, & Kojima 2013; Escudero, Hayes-Harb, & Mitterer 2008; Weber & Cutler 2004). Although most citations are in publications on the central issue of native versus non-native speech perception, the study has also been cited on non-nativeness effects at higher levels of linguistic processing (e.g., Hopp 2010; Van Engen & Bradlow 2007), on the effect of multi-speaker babble as a type of masking noise (e.g., Garcia Lecumberri & Cooke 2006), and for comparisons with clinical data due to age-related hearing deficits (Kumar Kalaiah et al. 2016) or cochlear implant use (Lee & Mendel 2016).

2.2.5 The data management (Where?) The 2004 study again listed the MPI website as a location for accessing the data. All stimuli in WAV format and all individual identification responses were made available there (https://www.mpi.nl/people/cutler-anne/research), and at the time of writing, they are still available. The data set was again also deposited in the Alveo Human Communication Science Virtual Lab. In 2018, both the primary research data (identification responses and audio files) and metadata according to ISO 24622-1 (CMDI) were further archived in the Tübingen CLARIN-D Repository (https://talar.sfb833.uni-tuebingen.de/about/). CLARIN-D (https://www.clarin-d.net/en/) is a research-oriented infrastructure for the Humanities and Social Sciences and covers a wide range of expertise ranging from annotated corpora to psycholinguistic experiments and from speech databases to web-based services for language and speech processing. Archiving in CLARIN-D is sustainable, data are stored in non-proprietary formats, and data sets can easily be found by different search engines (as attested by the awarded Data Seal of Approval; https://www.datasealofapproval.org/en/). With the Virtual Language Observatory (https://vlo.clarin.eu/?4), CLARIN-D also offers a search engine that specializes in finding available metadata for language resources worldwide, including of course resources from the Tübingen CLARIN-D Repository.

2.3 BALDEY: Biggest Auditory Lexical Decision Experiment Yet

2.3.1 The goal (Why?) One of the word recognition researcher's favorite workhorses is the lexical decision task. The literature reports thousands of lexical decision experiments. The studies have taught us about multiple different aspects of written and spoken language processing, for instance about where semantic processing takes place in the brain (e.g., Beeman et al. 1994) or which lexical characteristics affect ease of word recognition (e.g., Connine et al. 1990; Schreuder & Baayen 1995), providing information for language and speech processing models. For every new research question, a new experiment is typically designed, with a small number of target words fulfilling all kinds of constraints related to the research question.

The English Lexicon Project (ELP; Balota et al. 2007), in contrast, contains a huge visual lexical decision experiment, with 40,481 real words and 40,481 pseudowords. Because the data are freely available on the internet, researchers can test hypotheses about visual word processing without conducting new experiments. The frequent citation of the ELP data set in all kinds of analyses proves that this indeed occurs and suggested that it would also be worth conducting a large auditory lexical decision experiment to similarly further the research needs of spoken-word recognition, in particular by including word types for which almost no auditory data were previously available. BALDEY (Ernestus & Cutler, 2015) was designed for this purpose.

2.3.2 The study (How?) The 5,541 BALDEY experimental stimuli consist of 2,780 spoken Dutch real content words and 2,761 pseudowords, the latter differing from real words in just one or two segments. The words represent a large number of categories differing in word class (adjective, noun, or verb), morphological structure (simple or complex, with a restricted set of derivational and inflectional affixes), the position of stress (initial vs. non-initial), and the number of syllables in the stem (one or two). Most of these features have not been systematically varied in prior studies. The stimuli were recorded by a single female speaker and presented to twenty native listeners of standard Dutch (ten male, ten female). Each participant heard all words, distributed over ten experimental sessions, which were an hour long and were held one week apart. The final data set thus contains both accuracy and reaction times of 110,820 responses.

2.3.3 Main results (What?) Two initial analyses were presented at a conference (Ernestus & Cutler 2014) and were included in the publication (Ernestus & Cutler 2015) to illustrate how the data set might be exploited. The first analysis concerned the point at which a word

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122710/c045900_9780262366076.pdf by guest on 27 September 2023

has no further neighbors and effectively reaches the criterion for recognition. Listeners' response times were more strongly predicted by the duration of the spoken item than by any property of the word's competitor population, indicating that listeners adopt a rational approach to the task of auditory lexical decision (including the possibility that the input may be a pseudoword), and their responses are driven by these task-based considerations. The second analysis was concerned with how well four different measures of frequency of occurrence (from written corpora, spoken corpora, subtitles, and frequency ratings by listeners) predicted the study outcomes. The results were better predicted by form frequencies in a very large database compiled from film subtitles than by subjective ratings, or by frequencies of forms in written text, or by frequencies in spoken corpora, either of spontaneous or rehearsed speech. The size of the subtitles corpus and its constant objective of naturalness in dialogue were suggested to be the primary underlying drivers of its greater predictive power.

2.3.4 The users (For whom?) The data set is of interest to all researchers working on human or automatic spoken-word recognition. Examples of studies based on BALDEY include Ernestus and Cutler (2014) on spoken-word identification points and which frequency measure best reflects a listener's experience, Brysbaert et al. (2016) on the impact of a word's prevalence on its recognition, and ten Bosch, Boves, and Ernestus (2013), who tested their computational model of spoken-word recognition (Diana) on BALDEY. Although most citations to date have occurred in the context of discussion of large-scale studies, including recently a similar collection of auditory lexical decision data for English (Tucker et al. 2018), the study has also been cited on issues of morphological processing (Goodwin Davies 2018) and of cross-modality consistency (Hasenäcker, Verra, & Schroeder 2018).

2.3.5 The data management (Where?) BALDEY is available as supplemental information for the published article (https://journals.sagepub.com/doi/suppl/10.1080/17470218.2014.984730). In addition, all data are freely available at two other sites: (a) at the first author's site http://www.mirjamernestus.nl/Ernestus/Baldey/index.php and (b) via the Language Archive of the Max Planck Institute for Psycholinguistics (https://archive.mpi.nl).

The OA package contains a list of all stimuli, specifying the phonological and morphological properties of each stimulus, including number of phonemes, number of syllables in the stem, stress location, word class, morphological stem, affixes, and such. The frequency information lists each stimulus' frequencies of occurrences in several databases. The identification point information lists for several definitions of these points their locations in the stimuli. The database also contains the acoustic signals with text grids aligned at phone level. Furthermore, the database lists participant information (e.g., age, gender, handedness, language background), and their accuracy and reaction times for every stimulus. All information is arranged in files that can easily be imported in the statistical package R.

3 Summary

As these four case studies show, there are multiple reasons why large speech perception data sets should be collected, leading in consequence to multiple prompts for them to become OA and multiple ways in which that goal can be realized. We have not exhausted the possibilities. For instance, some journals insist on accepted publications providing OA data sets and contribute to maintaining secure and lasting storage sites for such data. If a speech perception study is accepted by such a journal, then that is another way for speech perception data sets to be sharable via OA, though they may not qualify at all as Big Data.

One way for megastudies in speech perception to become known is similar, in that it involves participation in a special issue or the like devoted to Big Data (as for our BALDEY case). This does not necessarily involve a guarantee of permanent storage, however. The majority of large speech perception data sets are collected for research reasons devised by the collectors, and at the time when we carried out these studies, accessibility was initially dependent on university or personal sites. We are in favor of the establishment and use of more permanent sites, which may best be managed by long-term agreements between multiple universities and professional associations, and, if funding organizations require OA from grant recipients, they should ideally contribute to the permanent maintenance of such sites. Services such as GitHub are a welcome addition to the means of data sharing, for maintaining long-term access to data; the future will probably bring a range of accessible sites, from minimally supervised depositories to professionally curated archives.

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122710/c045900_9780262366076.pdf by guest on 27 September 2023

It is now twenty years since the earliest of these studies (DADDY) was first designed, and in that time not only has researchers' knowledge of best practices in collection and sharing of large behavioral data sets developed considerably, but available technology has multiplied. The chapters in the rest of this volume attest to the explosion of knowledge in this area, plus the associated maintenance and communication options. It is tempting to consider what we might have chosen to do had alternative options been available two decades back. Nonetheless, these four large data sets on speech perception are available and in use, so they apparently meet a need. We therefore look forward to the speech perception community providing many more shared troves of useful data!

Acknowledgments

The order of authorship is alphabetical. The research described in this contribution was financially supported by the Max Planck Society and all authors were previously associated with the Comprehension Group at the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands. Additional funding for part of the work was provided by a Spinoza award to the first author from the Dutch Scientific Research Council (NWO), and by NICHD Grant No. 00323 to Winifred Strange.

References

Altvater-Mackenson, N., S. van der Feest, and P. Fikkert. 2014. Asymmetries in early word recognition: The case of stops and fricatives. *Language Learning and Development* 10 (2): 140–178. http://dx.doi.org/10.1080/15475441.2013.808954.

Balota, D. A., M. J. Yap, K. A. Hutchison, M. J. Cortese, B. Kessler, B. Loftis, J. H. Neely, D. L. Nelson, G. B. Simpson, and R. Treiman. 2007. The English Lexicon Project. *Behavior Research Methods* 39 (3): 445–459. http://dx.doi.org/10.3758/BF03193014.

Beeman, M., R. B. Friedman, J. Grafman, E. Perez, S. Diamond, and M. B. Lindsay. 1994. Summation priming and coarse semantic coding in the right hemisphere. *Journal of Cognitive Neuroscience* 6 (1): 26–45. http://dx.doi.org/10.1162/jocn.1994.6.1.26.

Borgman, C. L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge: MIT Press. https://doi.org/10.7551/mitpress/9963.001.0001.

Brysbaert, M., M. Stevens, P. Mandera, and E. Keuleers. 2016. The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance* 42 (3): 441. http://dx.doi.org/10.1037/xhp0000159.

Connine, C. M., J. Mullennix, E. Shernoff, and J. Yelen. 1990. Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16 (6): 1084–1096. http://dx.doi.org/10.1037/0278-7393.16.6.1084.

Cutler, A., A. Weber, R. Smits, and N. Cooper. 2004. Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America* 116 (6): 3668–3678. http://dx.doi.org/10.1121/1.1810292.

Darcy, I., D. Daidone, and C. Kojima. 2013. Asymmetric lexical access and fuzzy lexical representations in second language learners. *Mental Lexicon* 8 (3): 372–420. https://doi.org/10.1075/ml.8.3.06dar.

Ernestus, M., and A. Cutler. 2014. BALDEY: The Biggest Auditory Lexical Decision Experiment Yet. Paper presented at the 9th International Conference on the Mental Lexicon, Niagara, Ontario, September 30–October 2.

Ernestus, M., and A. Cutler. 2015. BALDEY: A database of auditory lexical decisions. *Quarterly Journal of Experimental Psychology* 68 (8): 1469–1488. https://doi.org/10.1080/17470218.2014.984730.

Escudero, P., R. Hayes-Harb, and H. Mitterer. 2008. Novel second-language words and asymmetric lexical access. *Journal of Phonetics* 36 (2): 345–360. http://dx.doi.org/10.1016/j.wocn.2007.11.002.

Garcia Lecumberri, M. L., and M. Cooke. 2006. Effect of masker type on native and non-native consonant perception in noise. *Journal of the Acoustical Society of America* 119:2445–2454. http://dx.doi.org/10.1121/1.2180210.

Garcia Lecumberri, M. L., M. Cooke, and A. Cutler. 2010. Non-native speech perception in adverse conditions: A review. *Speech Communication* 52:864–886. http://dx.doi.org/10.1016/j.specom.2010.08.014.

Goodwin Davies, A. J. 2018. Morphological representations in lexical processing. PhD dissertation, University of Pennsylvania.

Hajiaghbaba, F., H. R. Marateb, and S. Kermani. 2018. The design and validation of a hybrid digital-signal-processing plug-in for traditional cochlear implant speech processors. *Computer Methods and Programs in Biomedicine* 159:103–109. http://dx.doi.org/10.1016/j.cmpb.2018.03.003.

Hasenäcker, J., L. Verra, and S. Schroeder. 2018. Comparing length and frequency effects in children across modalities. *Quarterly Journal of Experimental Psychology* 72 (7): 1682–1691. http://dx.doi.org/10.1177/1747021818805063.

Hopp, H. 2010. Ultimate attainment in L2 inflection: Performance similarities between non-native and native speakers. *Lingua* 120:901–931. http://dx.doi.org/10.1016/j.lingua.2009.06.004.

Kumar Kalaiah, M., D. Thomas, J. S. Bhat, and R. Ranjan. 2016. Perception of consonants in speech-shaped noise among young and middle-aged adults. *Journal of International Advanced*

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122710/c045900_9780262366076.pdf by guest on 27 September 2023

Otology 12 (2): 184–188. http://dx.doi.org/10.5152/iao.2016.2467.

Kuperman, V. 2015. Virtual experiments in megastudies: A case study of language and emotion. *Quarterly Journal of Experimental Psychology* 68 (8): 1693–1710. https://doi.org/10.1080/17470218.2014.989865.

Law, F., and J. R. Edwards. 2015. Effects of vocabulary size on online lexical processing by preschoolers. *Language Learning and Development* 11 (4): 331–355. http://dx.doi.org/10.1080/15475441.2014.961066.

Lee, S., and L. L. Mendel. 2016. Effect of the number of maxima and stimulation rate on phoneme perception patterns using cochlear implant simulation. *Clinical Archives of Communication Disorders* 1 (1): 87–100. http://dx.doi.org/10.21849/cacd.2016.00066.

Miller, G. A., and P. E. Nicely. 1955. An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America* 27:338–352. http://dx.doi.org/10.1121/1.1907526.

Minkova, D. 2016. From stop-fricative clusters to contour segments in Old English. In *Studies in the History of the English Language VII: Generalizing vs. Particularizing Methodologies in Historical Linguistic Analysis*, ed. D. Chapman, C. Moore, and M. Wilcox, 29–59. Berlin: de Gruyter. http://dx.doi.org/10.1515/9783110494235-003.

Norris, D., and J. M. McQueen. 2008. Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review* 115 (2): 357–395. http://dx.doi.org/10.1037/0033-295X.115.2.357.

Peterson, G. E., and H. L. Barney. 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24 (2): 175–184. http://dx.doi.org/10.1121/1.1906875.

Schreuder, R., and R. H. Baayen. 1995. Modeling morphological processing. In *Morphological Aspects of Language Processing*, ed. L. B. Feldman, 131–157. Hillsdale, NJ: Erlbaum.

Sedlmeier, P., and G. Gigerenzer. 1989. Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin* 105 (2): 309–316. http://dx.doi.org/10.1037/0033-2909.105.2.309.

Silbert, N. H., and K. J. de Jong. 2007. Laryngeal feature structure in 1st and 2nd language speech perception. *Proceedings of the Sixteenth International Congress of Phonetic Sciences* 16:1901–1904.

Smits, R., N. Warner, J. M. McQueen, and A. Cutler. 2003. Unfolding of phonetic information over time: A database of Dutch diphone perception. *Journal of the Acoustical Society of America* 113 (1): 563–574. http://dx.doi.org/10.1121/1.1525287.

Strange, W. 1995. *Speech Perception and Linguistic Experience: Issues in Cross-language Speech Research*. Baltimore, MD: York Press.

ten Bosch, L., L. Boves, and M. Ernestus. 2013. Towards an end-to-end computational model of speech comprehension: Simulating a lexical decision task. In *Proceedings of Interspeech 2013*, 2822–2826.

Tucker, B., D. Brenner, D. K. Danielson, M. C. Kelley, F. Nenadic, and M. Sims. 2018. The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods* 51 (3): 1187–1204. http://dx.doi.org/10.3758/s13428-018-1056-1.

Van Engen, K. J., and A. R. Bradlow. 2007. Sentence recognition in native- and foreign-language multi-talker background noise. *Journal of the Acoustical Society of America* 121 (1): 519–526. http://dx.doi.org/10.1121/1.2400666.

Wagensveld, B., E. Segers, P. van Alphen, and L. Verhoeven. 2013. The role of lexical representations and phonological overlap in rhyme judgments of beginning, intermediate, and advanced readers. *Learning and Individual Differences* 23 (1): 64–71. http://dx.doi.org/10.1016/j.lindif.2012.09.007.

Warner, N. 2014. Sharing of data as it relates to human subject issues and data management plans. *Language and Linguistics Compass* 8 (11): 512–518. http://dx.doi.org/10.1111/lnc3.12107.

Warner, N., and A. Cutler. 2017. Stress effects in vowel perception as a function of language-specific vocabulary patterns. *Phonetica* 74 (2): 81–106. http://dx.doi.org/10.1159/000447428.

Warner, N., J. M. McQueen, and A. Cutler. 2014. Tracking perception of the sounds of English. *Journal of the Acoustical Society of America* 135:2995–3006. http://dx.doi.org/10.1121/1.4870486.

Warner, N., R. Smits, J. M. McQueen, and A. Cutler. 2005. Phonological and frequency effects on timing of speech perception: A database of Dutch diphone perception. *Speech Communication* 46 (1): 53–72. http://dx.doi.org/10.1016/j.specom.2005.01.003.

Weber, A., and A. Cutler. 2004. Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language* 50 (1): 1–25. http://dx.doi.org/10.1016/S0749-596X(03)00105-0.

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122710/c045900_9780262366076.pdf by guest on 27 September 2023

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122710/c045900_9780262366076.pdf by guest on 27 September 2023

MIT Press Direct

© 2021 The Massachusetts Institute of Technology

This work is subject to a Creative Commons CC-BY-NC license. Subject to such license, all rights are reserved.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122710/c045900_9780262366076.pdf by guest on 27 September 2023