Investigating perceptual biases, data reliability, and data discovery in a methodology for collecting speech errors from audio recordings

John Alderete, Monica Davies
Simon Fraser University

Abstract. This work describes a methodology of collecting speech errors from audio recordings and investigates how some of its assumptions affect data quality and composition. Speech errors of all types (phonetic, phonological, lexical, syntactic, etc.) were collected by eight data collectors from audio recordings of unscripted English speech. Analysis of these errors showed that (i) different listeners find different errors in the same audio recordings, but (ii) the frequencies of error patterns are similar across listeners; (iii) errors collected "online" using on the spot observational techniques without recourse to a recording are more likely to be affected by perceptual biases than "offline" errors collected from audio recordings, and (iv) when properly trained, listeners are able to collect gradient phonetic errors that have not traditionally been included in speech error collections. The striking differences between the errors collected online versus offline have potential to contribute new knowledge about the structure of speech error patterns.

Keywords: speech errors, methodology, perceptual bias, data reliability, capture-recapture, phonetics of speech errors

1. Introduction

Many of the foundational assumptions of language production have come from the scientific study of speech errors. Pioneering work in the 1970's and 1980's created significant collections of speech errors that supported detailed descriptions of the types of errors that humans make in natural speech (Dell & Reich, 1981; Fromkin, 1971, 1973; Garrett, 1975; Harley, 1984, 1996; Shattuck-Hufnagel, 1979; Stemberger, 1982/1985). This taxonomy of speech errors, in turn, led to a functional organization of language production processes that continues into present-day research. For example, sound errors like *beef needle* (for *beef noodle*) are a by-product of phonological encoding, or the retrieval of the phonological make-up of already activated words. Such errors are distinct from errors resulting from other production processes, e.g., lexical errors from lemma selection, syntactic errors from positional processing, etc. (see Bock and Levelt (1994) for review). Even models like WEAVER++ that focus on non-erroneous speech draw on speech error evidence to justify model assumptions (Levelt, Roelofs, & Meyer, 1999).

Beyond the discovery of production processes, speech error research has supported detailed investigation of the nature of these processes, including the influence of frequency (Dell, 1990; S. Frisch, 1996; Stemberger, 1991), the interactivity of linguistic levels and production processes (Dell, 1986; Rapp & Goldrick, 2000; Stemberger, 1985; Vigliocco & Hartsuiker,

2002), and the serial ordering of language (Acheson & MacDonald, 2009; Dell, Burger, & Svec, 1997; MacKay, 1970; Shattuck-Hufnagel, 1979), to name just a few themes. Furthermore, by probing the operational units in errors, speech error research has provided evidence for the psychological reality of many linguistic structures commonplace in linguistic analysis (Berg, 1987; Fromkin, 1971, 1973; Kubozono, 1989; Stemberger, 1983).

Despite these contributions, early speech error research has been subject to considerable criticism, largely because of the observational techniques used to collect speech errors. First, data collection is rather labour-intensive. This is because speech errors appear to make a relatively rare appearance, and even the best listeners can only detect about one in three errors in natural speech (see Bock (1996) for review). As a result, large collections like the Stemberger corpus (Stemberger, 1982/1985) or the MIT-Arizona corpus (Garrett 1975, Shattuck-Hufnagel 1979) tend to be multi-year projects that can be hard to justify. Second, speech error collection is error-prone, with opportunities for mistakes at all stages of the collection process. Errors are often missed or misheard, and approximately a quarter of errors collected by trained listeners are false positives because they do not meet the standard definition of a speech error (Cutler, 1982; Ferber, 1991, 1995). Furthermore, once collected, errors can be misclassified and exhibit several types of ambiguity, resulting in many errors being set aside (Cutler, 1988). Third, there is a significant literature documenting a set of perceptual biases in speech error collection that may skew the statistical distribution in large datasets (S. A. Frisch & Wright, 2002; Pérez, Santiago, Palma, & O'Seaghdha, 2007; Pouplier & Hardcastle, 2005). As we clarify below, perceptual biases are likely to reduce the number of errors that are difficult to detect, and furthermore, psychological biases make the documentation of certain errors, including errors of gradient phonetic structure, nearly impossible.

The goal of this article is to describe a methodology for collecting speech errors from natural speech that mitigates many of these problems. This methodology is a variant of Chen's (1999, 2000) approach used to collect speech errors in Mandarin from audio recordings of radio talk-shows. We have adapted some of the key assumptions in this approach to the collection of English errors. We describe our revised methods in detail and explore some of the consequences of the decisions we make on the structure of our data. Our research team includes eight data collectors listening to audio recordings of spontaneous speech, and two analysts verifying and classifying these errors. This approach strikes a balance between two approaches taken in past research: (i) data collection and classification by one or two experts (as in the Stemberger and the MIT-Arizona corpora) and (ii) collection by a very large number of untrained or semi-trained listeners (Dell & Reich, 1981; Pérez et al., 2007). We argue below that our training regime and the use of multiple listeners enables our team to achieve a high degree of quality control, and at the same time, address a set of sampling problems that arises with collections created by one or two individuals. We further argue that the benefits of having access to an audio recording far outweigh any potential drawbacks this approach may have in data quality. We do this by directly comparing "offline" collection from audio recordings with "online" collection techniques that use direct observation. The results of this comparison show that there are many striking differences between errors collected online versus those collected offline. These findings are of relevance to models of language production because they have potential to contribute new knowledge about the structure of speech error patterns.

The picture emerging from this approach is a far more optimistic one than the one arising from the critical assessment of the classic research summarized above. In less than a year, a small team of advanced undergraduate students working part time, guided by a psycholinguist

working full time, can generate a high quality speech error corpus large in size (i.e., greater than 8,000 errors) that can be explored in ways that the classic collections cannot. The ability to explore speech afforded by audio recordings also creates new opportunities for investigating language production that are not possible from direct observation. Audio recordings allow researchers to examine speech rate effects, assess data collection with new metrics, and estimate the frequency of speech errors in natural speech in new ways. Recourse to an audio recording also supports phonetic investigation, including studying gradient sound patterns that lie on a continuum between two sound categories. We summarize some of these new empirical directions, and also give some pilot results on gradient sound errors and estimating the frequency of speech errors in the general population.

2. Background

The goal of most methodologies for collecting speech errors is to produce a sample of speech errors that is representative of how they occur in natural speech. Below we summarize some of the known problems in achieving a representative sample and the best practices used to reduce the impact of these problems.

2.1 Data reliability

Once a researcher is alerted to the existence of speech errors, s/he can usually spot speech errors in everyday speech with relative ease. However, the practice of collecting speech errors systematically and in large quantities is a rather complex rational process that requires much more care. This complexity stems from the standard characterization of a speech error as "an unintended, nonhabitual deviation from a speech plan" (Dell, 1986, p. 284). Speech errors are unintended slips of tongue, and not dialectal or idiolectal variants, which are habitual behaviors. Marginally grammatical forms and errors of ignorance are also arguably habitual, and so they too are excluded (Stemberger 1982/85). A problem posed by this definition, which is widely used in the literature, is that it does not provide clear positive criteria for identifying errors, or a blueprint that collectors can listen for (Ferber, 1995). In practice, however, data collection can be guided by templates of commonly occurring errors, like the inventory of 11 error types given in Bock (2011), or the taxonomies proposed in Dell (1986) and Stemberger (1993).

These templates are tremendously helpful, but as anyone who has engaged in significant error collection will attest, the set of errors included in these templates is rather heterogeneous. Data collectors must listen to words at the sound level, attempting to spot various slips of tongue (anticipations, perseverations, exchanges, shifts), at the same time that they are attending to the phonetic details of the slipped sounds to see if they are accommodated phonetically to their new environment. Data collectors must also pay attention to the message communicated, to confirm the intended words are used, and that word errors of various kinds do not occur (word substitutions, exchanges, blends). Adding to this list, they are also listening for word-internal errors, like affix stranding and morpheme additions and deletions, as well as syntactic anomalies like word shifts, phrasal blends, and morpho-syntactic errors like number attraction. One collection methodology addresses this "many error types" problem by requiring data collectors to exclusively collect a specific type of speech error, namely word-initial sound errors (Dell & Reich, 1981). However, many collection methodologies do not restrict data collection in this way and include all of these error types in their search criteria.

This already difficult task is made considerably more complex by the need to exclude intended and habitual behavior. Habitual behaviors include a variety of phonetic and phonological processes that typify casual speech. For example, [gʊn nuz] for 'good news' does not involve a substitution error, swapping [n] for [d] in the first word, because this kind of phonetic assimilation is routinely encountered in causal speech (Cruttenden, 2014; Shockey, 2003). In addition to mastering these casual speech rules, data collectors must also have an understanding of dialectal variants and the linguistic background of the speakers they are listening to. A third layer of filtering involves attending to individual level variation, or the idiolectal patterns found in all speakers that involve every type of linguistic structure (sound patterns, lexical variation, sentence patterns). Data collectors must also exclude changes of the speech plan, a very common kind of false positive, where the speaker begins an utterance with a particular message, and then switches to another message mid-phrase. Examples like, *I was, we were going to invite Mary,* are not errors (in this case a mis-selection of the subject pronoun), because the speech plan is accurately communicated in both attempts of the evolving message. What makes data collection mentally taxing, therefore, is listeners have a wide range of error types they are listening for, and while they are listening for these different kinds of errors, they must exclude potential errors with several kinds of filters.

It is not a surprise, therefore, that mistakes can happen at all stages of data collection. Given the characterization of speech errors above, many errors are missed by data collectors because the collection process is simply too mentally taxing (see estimates below). The speech signal can also be misheard by the data collector in a "slip of the ear" (Bond, 1999; Michael S Vitevitch, 2002), e.g., spoken*: Because they can answer inferential questions* …; heard: *Because they can answer in French* … (Cutler, 1982), which can lead to either an actual error being missed or a false positive. Furthermore, sound errors can be incorrectly transcribed, which again can lead to false positives or an inaccurate record of the speech event. Finally, errors can be correctly recalled and then documented in such a way that does not actually meet the criteria laid out above, as in false positives due to habitual speech features or changes of the speech plan discussed above. It should be emphasized that all of these problems are more acute with "online" data collection because there is no recourse to the audio recording. An important goal of this article is to document the differences in data reliability between online and offline data collection.

These empirical issues have been documented experimentally on a small scale in Ferber (1991), a study that we build on in our experiment 1. In Ferber's study, four data collectors listened to a 45 minute recording of samples from German radio talk shows and recorded all the errors that they heard. All data collectors had been given an introduction to speech errors. Two of them were linguists who had collected speech errors previously in two separate audio recordings, one was a mathematician, and the last was the author, an experienced psycholinguist. The recording was played without stopping, so this test is comparable to online data collection because the audio could not be repeated. The author then listened again to the same recording offline, stopping and rewinding when necessary. A total of 51 speech errors were detected using both online and offline methods, or an error about every 53 seconds. On average, two thirds of the total of 51 errors were missed by each listener, but there was considerable variation, ranging between missing 51% to 86% of the 51 errors. More troubling is the fact that approximately 50% of the errors were recorded incorrectly, involving transcription errors of the actual sounds and words in the errors. In addition, while all the collectors were trained in speech error collection, one listener found no sound errors, and two listeners found no lexical (word) errors. These

differences in the error types collected by individual collectors raises serious questions about the reliability of using observational techniques to collect speech errors. It also poses a problem for the use of multiple data collectors, since different collectors seem to be hearing different kinds of errors. For this reason, we propose to expand on Ferber's experiment below by probing data collector differences with offline data collection (section 4).

2.2 Perceptual biases and other problems with observational techniques

We have seen some of the ways in which human listeners can make mistakes in speech error collection, given the complexity of the task. A separate line of inquiry examines how constraints on the perceptual systems of human collectors lead to problems in data composition. An important thread in this research concerns the salience of speech errors, arguing that speech errors that involve more salient linguistic structure tend to be over-represented. Thus, errors involving a single sound are harder to hear than those involving larger units, such as a whole word, multiple sounds, or exchanges of two sounds (Cutler, 1982; Dell & Reich, 1981; Tent & Clark, 1980). It also seems to be the case that sound errors are easier to detect word-initially (Cole, 1973), and that speech errors are easier to detect in highly predictable environments, like *... smoke a cikarette (cigarette)* (Cole, Jakimik, & Cooper, 1978), or when they affect the meaning of the larger utterance. Finally, sound errors involving a change of more than one phonological feature are easier to hear than substitutions involving just one feature (Cole, 1973; Marslen-Wilson & Welsh, 1978).

In sound errors, the detection of sound substitutions also seems governed by overall salience of the features that are changed in the substitution, but the salience of these features depends on the listening conditions. In noise, for example, human listeners often misperceive place of articulation, but voicing is far less subject to perceptual problems (Garnes & Bond, 1975; Miller & Nicely, 1955). However, Cole et al. (1978) found that human listeners detected word-initial mispronunciations of place of articulation more frequently than mispronunciations of voicing, and that consonant manner matters: mispronunciations of fricative voicing were detected less frequently than stop voicing. These feature-level asymmetries, as well as the general asymmetry towards salient errors, has the potential to skew the distribution of error types and specific patterns within these types.

Another major problem concerns a bias in many speech error corpora towards discrete sound structure. Though speech is continuous and presents many complex problems in terms of how it is segmented into discrete units, when documenting sound errors, most major collections transcribe speech errors using discrete orthographic or phonetic representations. Research on categorical speech perception shows that human listeners have a natural tendency to perceive continuous sound structure as discrete categories (see Fowler and Magnuson (2012) for review). The combination of discrete transcription systems and the human propensity for categorical speech perception severely curtails the capacity for describing fine-grained phonetic detail. However, various articulography studies have shown that gestures for multiple segments many be produced simultaneously (Pouplier & Hardcastle, 2005), and that speech errors may result in gestures that lie on some gradient between those that would be expected for two different segments (S. A. Frisch, 2007; Stearns, 2006). These errorful gestures may or may not result in audible changes to the acoustic signal, making some of them nearly impossible to document using observational techniques.

Acoustic studies of sound errors have also documented perceptual asymmetries in the detection of errors that can have the effect of skewing substitution errors in collections made

using purely observational techniques (S. A. Frisch & Wright, 2002; Mann, 1980; Marin, Pouplier, & Harrington, 2010). For example, using acoustic measures, Frisch and Wright (2002) found a larger number of $z \rightarrow s$ substitutions than $s \rightarrow z$ in experimentally elicited speech errors, which they attribute to an output bias for frequent segments (*s* has higher frequency than *z*). This asymmetric pattern is the opposite of that found in (Stemberger, 1991), using observational techniques. Thus, different methods for detecting errors (e.g., acoustic vs. observational) may lead to different results.

Finally, a host of sampling problems arise when collecting speech errors. Different data collectors have different rate of collection and frequencies of types of errors they detect (Ferber, 1991). This collector bias is related to the talker bias, or preference for talkers in the immediate environment of the collector who may exhibit different patterns (Dell & Reich, 1981; Pérez et al., 2007). The theoretical bias is also related to the collector bias. Data collectors may be informed by certain theoretical assumptions, and therefore primed to recognize errors that confirm or disconfirm certain patterns related to those assumptions. Finally, speech error collections are subject to distributional biases in that certain error patterns may be more likely because of the opportunities for them in the language under examination are greater than other error patterns. For example, speech errors that result in lexical words are much more likely to be found in monosyllabic words than polysyllabic words because of the richer lexical neighborhoods of monosyllables (Dell & Reich, 1981). Therefore, speech error collections must be assessed with these potential biases in mind.

2.3 Best practices

The list below explains some of the principal ways researchers have addressed the issues above in their research.

- **Documentation best practices**: only document errors when collector has a high degree of confidence, within 30 seconds of the speech act; do not multi-task, i.e., make a conscious effort to collect errors as an observer of speech; see especially Shattuck-Hufnagel (1979) and Stemberger (1982/1985)
- **Interviewing speaker after error production**: if the intended utterance or facts about the context are unclear, some researchers advocate interviewing the speaker to confirm these details; see especially Harley (1984), Harley (1996), and Vousden, Brown, and Harley (2000)
- **Many data collectors**: to reduce the collector and talker biases, and to increase size, some researchers have recruited a large number of data collectors; Dell and Reich (1981) used 200 undergraduate students, Pérez et al. (2007) utilized over 700 undergraduate students
- **Psycho-linguistically naïve**: to address the theoretical bias, some researchers avoid using data collectors that are currently working on a research project related to data collection (Stemberger, 1982/1985)
- **Focus on a specific type of error/error environment**: Dell and Reich (1981) reduced the complexity of speech error collection by instructing data collectors to only collect word errors and sound errors with initial consonant changes.

While it is clear that past research is based in sound methodological decisions, two important problems persist in all projects based in online on the spot observation. The first is

data reliability. Given the rather poor record documented in Ferber (1991), we must be concerned about errors in transcription, acceptance of false positives as errors, and perhaps most importantly, the large number of missed errors. It seems likely that the research champions that built large collections (the Fromkin/UCLA corpus, the Stemberger corpus, the Harley corpus, and MIT-Arizona corpus) achieved a higher degree of accuracy and detection rate than Ferber's data collectors, but without recourse to a record of the speech they documented, it is hard to assess even these carefully prepared datasets. Second, the Toronto corpus (Dell & Reich, 1981) and Pérez (Pérez et al., 2007) corpora aside, most major speech error collections were compiled by one or two individuals. This has some advantages in data quality, because the data is more likely to be internally consistent. However, they are also more subject to collector and talker biases. Given the known perceptual biases, these collectors may have specific tendencies that could have influenced the specific error patterns.

One speech error corpus that addresses many of the data reliability issues is described in Chen (1999, 2000). In this study, the focus was on collecting speech errors in Mandarin from audio recordings of radio programs. In particular, two research assistants were trained to collect speech errors, and, after a data reliability test, the two collectors each listened to 120 programs approximately 40 minutes in length. These listeners collected 1,317 speech errors in total, but this number was reduced to 987 errors after certain false positives were removed by a third data analyst. This approach has two features that help considerably with data reliability. First, the ability to listen and relisten to the recording improves the ability to document hard to hear errors. Second, there was a collection stage involving the two research assistants, and then a later verification stage that removed large numbers of false positives, approximately 25% of the initial submissions. Thus, the existence of an audio recording both supports careful examination of the underlying speech data and verification by an experienced psycholinguist.

It should be noted that Stemberger (1982/1985), and also later Stemberger (1993), discuss the potential using audio recordings to collect speech errors and mentions several advantages that we discuss here in detail. However, Stemberger ultimately opts for online recording because of the additional time need to create the recordings. With the advent of the Internet and ubiquity of social media, these time constraints have been significantly reduced, making offline data collection much more practical.

A recent website for crowd-sourcing speech error collection is also worth mentioning in how it addresses some of the issues raised above. Michael S. Vitevitch et al. (2015) demonstrates how speech error collection and other types of speech facts can be collected and also shared quickly and efficiently with a community of researchers. We believe that this is an excellent tool for both research and education on speech errors. However, as the authors note, it has some of the drawbacks in terms of data quality as the course-based collection methods of e.g., Dell and Reich (1981). Given our focus on data reliability and data discovery, we opt for a model that has a rigorous training component and allows access to audio recording, as in the Chen model.

In the rest of this article, we describe a methodology of collecting English speech errors based in audio recordings from podcast series similar to Chen's approach. One of the important differences is that we associate each error with the individual(s) that found the error. Based on the variation found in Ferber's (1991) experiment, we are interested in asking if data collectors detect substantively different error types. Experiment 1 (section 4) is designed to address this question. We are also interested in determining if there are important effects of the online vs. offline distinction, and section 5 offers the first detailed examination of this factor in speech

error collection. Before we investigate these methodological decisions, we give a detailed description of how speech errors are collected in our methodology.

3. The Simon Fraser University Speech Error Database (SFUSED)

3.1 General methods

Our methodology is characterized by the following decisions and practices, which we elaborate on below in detail.

- **Many data collectors**: to reduce the data collector and talk biases, and also increase productivity, eight data collectors were employed to collect a relatively large number of errors.
- **Training**: to increase data reliability, data collectors go through about twenty five hours of training, including both linguistic training and feedback on error detection sessions.
- **Focus on offline data collection**: also to increase data quality, errors are collected primarily from audio recordings.
- **Allowance for gradient errors**: data collectors use a transcription system that accounts for gradient phonetic patterns that go beyond normal allophonic patterns.
- **Data collection separate from data classification**: data collectors submit speech errors using a spreadsheet template, but they do not directly classify errors; analysts then verify error submissions and then assign a host of variables that classify the error using established standards.

Our approach strikes a balance between employing one or two expert data collectors, as in many of the classic studies discussed above, and a small army of relatively untrained data collectors (Dell & Reich, 1981; Pérez et al., 2007). The many data collectors decision allows us to study individual differences in error detection (since collector identity is part of each record), and contextualize speech error patterns to adjust for any differences. Also, the underlying assumption is that if there are data collector biases, their effect will be reduced to the specific individuals that exhibit it. We report in section 4 these data collector differences, which appear to be quite small.

We have collected speech errors in two ways: (i) online as spectators of natural speech in the daily lives of the data collectors, and (ii) offline as listeners of podcast series available on the Internet. Six data collectors collected 1,041 speech errors over the course of approximately seven months, following the best practices mentioned above. After finding a number of problems with this, we turned to offline data collection, and a different team of six research assistants collected 7,500 errors over a period of approximately 11 months, which was reduced by approximately 20% by removing false positives.

As for the selection of audio recordings, a variety of podcasts series available for free on the Internet were reviewed and screened so that they met the following criteria. First, podcasts were chosen that were conversations of natural unscripted speech, largely free of reading or set routines for the speakers. Furthermore, speech errors were not collected from introductions and advertisements with a set script, and these portions of the recordings were removed from our calculations of recording length. Second, we focused on podcasts with Standard American English used in the U.S. and Canada. That is, most of our speakers were native speakers of some

variety of the Midlands dialect of American English, and all speakers with some other English dialect were carefully noted. Both dialect information and idiolectal features of individual speakers were noted in each podcast recording, and profiles were created for each speaker that summarizes the features of that speaker. The podcasts also differed in genre, including entertainment podcasts like *Go Bayside* and *Battleship Pretension*, technology and gaming podcasts like *The Accidential Tech* and *Rooster Teeth*, and science-based podcasts like *The Astronomy Cast.* Speech errors were collected from on average of 50 hours of speech in each podcast, typically resulting in about one thousand errors per podcast. This enabled the team to acquire a large amount of data from individual talkers, and therefore study individual talker differences in some detail.

In terms of what data collectors are listening for, we follow the standard characterization in the literature of a speech error given above, as an "unintended nonhabitual deviation from the speech plan" (Dell, 1986, p. 284). As explained above, this definition excludes words exhibiting casual speech processes, false starts, and other disfluencies involving a change of the speech plan, and dialectial and idiolectal features. We note that the offline collection method aids considerably in reducing the inclusion of false positives from these features because collectors develop strong intuitions about typical speech patterns of individual talkers, and factor out these traits. For example, one talker was observed to have an intrusive velar before post-alveolars in words like *much* [mʌ^{k}tʃ]. The first few instances of this pattern were initially classified as a speech error because it is aberrant in the larger population. After additional instances were found in other words, e.g., *such* and *average,* an idiolectal pattern was established and noted in the speaker profile of this talker. This note in turn entailed exclusion of these patterns in all future and past submissions. Our experience is that such idiolectal features are extremely common in our corpus and so data collectors need to be trained to find these features and any database of speech errors should have a mechanism for systematically registering them.

The focus of our collection is on speech errors from audio recordings, and data collectors use special tools for searching them. All podcasts are MP3 files, generally of high production quality. These files are opened by the data collector in the speech analysis software Audacity on a personal computer and the speech stream is viewed as an air pressure wave form. Data collectors are instructed to attend to the main thread of the conversation, so that they follow the main topic and the discourse participants involved. However, they are told not to listen for content, but instead focus on what is actually said, and listen specifically for the kinds of speech errors that they have been trained to detect. Data collectors can listen to any interval of speech as many times as possible, and they are also shown how to slow down the speech in Audacity in order to pinpoint specific speech events in fast speech. When a speech error is observed, a number of record field values are assigned (file name, time stamp, date of collection, identity of collector) together with the example itself, showing the position of the error and as much of the speech to give the linguistic context of the error. In addition, data collectors document the inferred intended word, degree of confidence in the intended word, whether the error was corrected or not, and any observations the data collector feels is important to contextualize the example. All examples are input into a spreadsheet template and submitted to analysts for incorporation into the SFUSED database.

Data collectors use a transcription scheme that accounts for both phonological errors and gradient phonetic patterns. For many errors, orthographic representation of the error word in context is sufficient to account for the relevant observations, and so data collectors are instructed to simply write out error examples using standard spelling if the speech facts do not deviate from

normal pronunciation of these words. Many sound errors need to be transcribed in phonetic notation, however, because it is more accurate and many nonsense error words do not have standard spellings. In this case, data collectors transcribe the relevant words in broad transcription, making sure that the conditioned allophones of the phonemes used in their transcriptions obey the standard rules of English allophones (which is provided in training; see below). When this is not the case, or if a non-English sound is used, a more narrow transcription is used that simply documents all the relevant phonetic facts. Thus, IPA symbols for non-English sounds, like front rounded vowels, are sometimes used, and also phonotactically illicit allophones are transcribed in this way, but both of these patterns are relatively rare.

It is sometimes the case that this system is not able to account for all of the phonetic facts, either because there is a transition from one phoneme to another (other than the accepted diphthongs and affricates of English), or because sounds do not sound like good examples of a particular phoneme. To capture these facts, we employ a set of tools commonly used in the transcription of children's speech (Stoel-Gammon, 2001). In particular, we recognize ambiguous sounds that lay on a continuum between two poles, transitional sounds that go from one pole to another, and intrusive sounds, which are weak sounds short in duration that are clearly audible but do no have the same status as fully articulated consonants or vowels. Table 1 illustrates these three distinct types and explains the transcription conventions we employ (the SFUSED record ID numbers are given as suffixes on "sfused"). The procedure for discovering these gradient error types was simply inductive, built up gradually out of a need to characterize erroneous speech that did not fit the usual patterns of phonological sound errors. We discuss some of our preliminary findings with gradient errors in section 6.3.

Table 1. Gradient sound errors (/ = error word, ^ = sound word)

Ambiguous segments [X|Y]: segments that are neither [X] or [Y] but appear to lay on a continuum between these two poles, and in fact slightly closer to [X] than [Y].

Ex. sfused21: … a whole lot of red photons and a ^few ^blue /ph[u|ʊtɑ]= photons and a ^few green photons and I translate that into a colour.

Transitional segments [X-Y]: segments that transition from [X] to [Y].

Ex. sfused1162: … which is the largest /hip-[f-h]op ^festival in the country I guess.

Intrusive segments [^{X}]: weak segments that are clearly audible but do not have the status of a fully articulated consonant or vowel.

Ex. sfused4742: I'm January ^/[eɪ^{n}tinθ]teenth and it's typically January nineteenth.

3.2 Training

The data collectors were recruited from the undergraduate program at Simon Fraser University and worked as research assistants for at least one semester, though most worked for a year or more. Two research assistants started out as data collectors and then scaffolded into analyst positions involved in data classification, but the majority of the undergraduates did data collection exclusively. All students had taken an introductory course in linguistics and another introduction to phonetics and phonology, so they started with an awareness of the linguistic structures of English. Students also receive both phonetic and psycholinguistic training in speech errors collection.

To brush up on English transcription, research assistants were required to read a standard textbook introduction to phonetic transcription of English, i.e., chapter 2 of Ladefoged (2006), and also assigned a set of drills to practice English transcription. They were then given a seven page document explaining the transcription conventions of the project, which also illustrated the main dialect differences of the speakers they were likely to encounter in the audio recordings, including information about the Northern Cities, Southern, and African American English dialects. After this refresher, they were tested twice on two separate days on their transcription of 20 English words in isolation, and research assistants with 90% accuracy or better were allowed to continue. Research assistants were also given an eight page document describing casual speech processes in English and given illustrations of all of the 29 patterns described in that document.

The rest of the training involved a one hour introduction to speech errors and feedback in a set of listening tests over several days. In particular, research assistants were given a five page document explaining what is and what is not a speech error, given the exclusions discussed above, and also multiple examples of all types of errors based on the taxonomies given in Dell (1986) and Stemberger (1993). After this introduction, the research assistants were given the task of spending one hour outside the lab collecting speech errors as a passive observer of spontaneous speech. The goal of this task is give the data collectors a concrete understanding of the concept of a speech error and its occurrence in everyday speech. They were told to write them down on paper and bring them to the lab for review, though these initial errors were not included in the larger database.

After this introduction, research assistants were given listening tests in which they were asked to identify the speech errors in three 30-40 minute podcasts that had been pre-screened for speech errors. The research assistants were instructed in how to open a sound file in Audacity, navigate the speech signal, and repeat and slow down stretches of speech. As explained above, they were instructed to listen to the speech as much as needed, but listen specifically to what is actually said and not to listen to the content of the podcast as if it was a lecture for a class. They submitted their speech errors using a spreadsheet template, which were then checked by an analyst, generally the first author. In particular, the errors submitted by the data collector trainee were classified into three groups: false positives (do not meet the definition), correct known errors, and new unknown errors (which had not been identified earlier). Also, the number of missed speech errors was calculated (i.e., errors found in the pre-screening but not found by the trainee). From this information, the percentage of missed errors, counts of false positives and new errors were calculated and used to further train the data collector. In particular, the analyst and trainee met and listened to each submitted error, and the analyst explained why the false positives were not errors, and identified missed errors so the collector could learn from these mistakes. Also, the average ‘minutes per error’ (MPE), or the average number of minutes elapsed per error collected, was assessed and used to train the listener. We do not have a set criteria for success for students to continue, because other mechanisms are used to remove false positives and ensure a representative sample. However, the goal of the training is to achieve approximately 75% accuracy in submissions and an MPE of 3 or lower, i.e., be able to detect an error, on average, every three minutes.

3.3 Classification

As explained above, data collectors make speech error submissions in spreadsheet form, which are then batch imported into the SFUSED database. Speech errors are documented as a record in a speech errors data table that contains 46 fields, which are subdivided into six field

types (see appendix for a description of all fields). For example, example fields document the actual speech error and encode facts like if the speech error was corrected, and if a word was aborted mid-word. Record fields document facts about the source of the record, like the researcher who collected the speech error, what podcast it came from, and a time stamp. The data provided by the data collectors is a subset of the example and record fields, and the analyst fills in the rest of the fields from these field types, as well as filling in a host of fields that analyze the the properties of the error. This latter portion, which constitutes the bulk of the classification duties, involve filling in major class fields, word fields, sound fields, and special class fields which apply to only certain classes of errors. As we do not focus on classification in this work, we leave the explanation of the myriad of analytical decisions that goes into classification to the SFUSED manual. The key aspect of our workflow is that there are two parts to documenting errors: initial detection by the data collector, and then data verification and classification by a senior analyst. This separation of work, also assumed in Chen's (1999, 2000) Mandarin study, leads to better data reliability because there is a verification stage. We also believe that it leads to greater internal consistency because classification involves a large number of analytical decisions that are best handled by a small number of individuals focused on just this task.

4. Experiment 1: same recording, many collectors

To reduce the impact of the talker and data collector biases, and investigate them quantitatively, our methodology involves many data collectors. The many collectors assumption is a good one in principle, but it introduces another factor in the database because each data collector may be subject to known perceptual biases to different degrees. As discussed in 2.1, Ferber (1991) found that when collecting speech errors online, listeners differed drastically in counts of the major error types, i.e., lexical word errors versus sublexical sound errors. In experiment 1, therefore, we aim to investigate individual differences in collecting error patterns to determine if the same amount of collector variation is found with offline data collection.

4.1 Methods

In this experiment, nine podcasts of approximately 40 minutes in length were listened to by three data collectors. Two data collectors listened to all nine podcasts, and a pair of data collectors split the same nine recordings because of time constraints. All of the listeners were experienced data collectors, and had at that point collected over 200 speech errors using a combination of online and offline collection methods. The data collectors were instructed to collect errors of all types outlined above. They were also allowed to listen to the recordings as many times as they wished, and could slow the recording to listen for fine-grain phonetic detail. After each data collector had made her/his submission, the speech errors were combined for each podcast, and all three data collectors relistened to all of the errors as a group to confirm that they met the definition of a speech error. False positives were then excluded by majority decision, though the three listeners found consensus on the inclusion or exclusion of an error in almost every case.

The podcasts came from three podcast series: three podcasts from an entertainment podcast, three from a technology and entertainment podcast, and the last three from a science podcast. Each podcast episode was centered on a set of themes and the talkers generally spoke freely on these themes and issues raised from them. There was a balance of male and female

talkers. Removing introductory and closing material and advertisements, the total length of the nine podcasts came to approximately 370 minutes.

4.2 Results and discussion

The three data collectors found 380 speech errors in all nine podcasts, or about an error every 58 seconds. However, 94 speech errors (24.74%) were excluded because, upon relistening, the group determined that it did not meet the definition of a speech error described above. Thus, 286 confirmed errors were detected collectively in all podcasts, which amounted to an error heard every minute and 17 seconds, or an MPE of 1.29. Table 2 breaks down the percentage of correctly detected errors and MPE and by listener (listeners 1 and 2 split the nine podcasts, as explained above). While there are some differences in MPE, it appears that listeners are broadly similar, achieving about 78% accuracy and a mean MPE of 3.22. Another way to probe internal consistency in error detection is to count how often listeners detect the same error. In Table 3, we see that roughly 2/3rds of all errors are heard by just one data collector, and independent detection of the same error by all three listeners is rather rare at approximately 14% of the confirmed errors.

Table 2. Accuracy and Minutes Per Error by data collector.

	Total	False positives	% correct	MPE
Listener 1	50	16	68%	4.85
Listener 2	85	18	78.82%	3.21
Listener 3	177	33	81.36%	2.64
Listener 4	206	32	84.47%	2.18

Table 3. Consistency across confirmed errors

Heard by just one person	193 (67.48%)
Heard by just two people	53 (18.53%)
Heard by all three people	40 (13.99%)
Heard by more than one	93 (32.52%)

From these counts, we can conclude that data collection in general is error prone, because even the data collectors with the highest accuracy produce a large number of false positives. Furthermore, the majority of the speech errors are heard by a single individual, which raises the question of whether individuals are hearing the same types of errors. In other words, it is a fact that listeners detect different speech errors, but do they detect different types of errors? Below in Table 4, we track counts of speech errors by listener (which may overlap because they listened to the same recordings), divided into the following major error type categories: sound errors involving one or more phonological segments, word errors, and other errors involving morphemes or phrases. This three-way breakdown is the most natural way to divide up errors into major types given the small baselines per data collector, and because it allows direct comparison with Ferber's (1991) findings, which used similar error types. As shown in Table 4, the percentages of sound and word errors are broadly similar, though listener 1 did collect a larger percentage of word errors than the other listeners. A chi-square analysis of these frequencies indicates that there is no association between listener and error type ($\chi^2 = 7.84$, $P = 0.2501$). Across all listeners, sound errors are in the majority, and all listeners are also detecting morphological and syntactic errors. This contrasts with Ferber's findings in which some listeners

found no word errors, and one listener found no sound errors, though in Ferber's study used an online methodology.

Table 4. Distribution of major error types, sorted by listener

	Sound	Word	Other	Total
Listener 1	17 (50%)	14 (41.18%)	4 (11.76%)	34
Listener 2	38 (56.72%)	15 (22.39%)	15 (22.39%)	67
Listener 3	89 (61.38%)	40 (27.59%)	16 (11.03%)	145
Listener 4	100 (57.80%)	46 (26.59%)	27 (15.61%)	173

Another way to investigate listener differences is by examining how susceptible they may be to perceptual biases. One way of investigating the salience of speech errors heard is by comparing across listeners the percentage of errors that were corrected by the talker in the utterance. Data collectors are instructed to document whether the error is corrected, and such corrections are often (though not always) a red flag of the occurrence of an error. In Table 5, we see that listeners range from 37.24% to 55.88% in the percentage of errors that are corrected by the speaker. Listeners 1 and 2 seem to be relying a bit more on talker corrections, but these associations are not significant ($\chi^2 = 5.95$, P=0.1141). These two listeners also had higher MPEs than listeners 3 and 4, and therefore lower rates of error detection, which is consistent with the assumption that these listeners are hearing less uncorrected and therefore harder to detect errors.

Table 5. Salience measures, all errors

	Errors corrected	Errors uncorrected	Total	Percentage corrected
Listener 1	19	15	34	55.88%
Listener 2	34	33	67	50.75%
Listener 3	54	91	145	37.24%
Listener 4	73	100	173	42.20%

Sound errors can also be probed for salience measures. Speech errors can be distinguished by whether than occur in phonetically salient positions, including stressed syllables and word-initial position. Another way to distinguish sound errors is if they involve aberrant phonetic structure, i.e., one of the three gradient phonetic errors discussed above. Gradient phonetic errors are more difficult to detect because they involve fine-grained phonetic judgments. Table 6 shows that there seems to be broad consistency across data collectors in terms of the salience of sound errors. Roughly 80% of all errors are heard in stressed syllables, and while some listeners heard a few more gradient errors and errors in non-initial position, no data collector stands out as head and shoulders above the others on any single measure.

Table 6. Salience measures, sound errors

	Total	Error in stressed syllable	Error in initial segment	Gradient errors
Listener 1	17	14 (82.35%)	7 (41.18%)	4 (23.53%)
Listener 2	38	29 (76.32%)	13 (34.21%)	8 (21.05%)
Listener 3	89	73 (82.02%)	31 (34.83%)	25 (28.10%)
Listener 4	100	77 (77%)	44 (44%)	25 (25%)

Let us summarize the principal findings of experiment 1. While there is some variation across listeners in terms of the types of errors and what environments they occur in, two more general findings dominate the results. First, regardless of their accuracy or error detection rate, all data collectors produce a large number of false positives: between 16-32% of the errors collected by individual listeners had to be excluded. Any methodology for collecting speech errors from spontaneous speech therefore requires a robust mechanism for checking for and discarding false positives. Second, data collectors detect different specific speech errors. After excluding false positives, 2/3rds of all the errors collected were heard by only one of the three listeners. And yet, upon relistening, the other listeners agreed that the errors that they missed were indeed errors.

Despite these differences among data collectors in the actual errors found, we did find broad consistency across four listeners in terms of the major error types collected and the salience of the collected error. However, other speech error collections may not be characterized by a similar degree of consistency, as Ferber's (1991) findings suggest. We discuss in the final section some of the practical implications of these findings, but it should be noted that a major factor in the variation found across data collectors is likely to be the open-ended nature of the collection task. Data collectors were instructed to relisten as many times as they felt necessary, and so some collectors may have spent more time on certain portions of the recordings than others, resulting in more errors and different errors. Given this freedom to select different portions of the recording and relisten at will, we actually expect some variation.

5. Experiment 2: online vs. offline collection

Our decision to collect speech errors primarily from audio recordings allows for better data verification and data discovery (see section 6). But how does it compare with online data collection so common in earlier speech error collections? Below we probe the effects of this decision by comparing data that we collected online using traditional observational techniques with data collected offline from audio recordings. We expect to find differences in the frequencies of patterns due to (i) the different impact of perceptual biases, (ii) differences due to the source data, and (iii) the ability to verify errors and relisten to the speech facts.

5.1 Methods

Our research team began collecting speech errors in 2015 using traditional observational techniques characteristic of classic speech error studies (e.g., (Shattuck-Hufnagel, 1979; Stemberger, 1982/1985). In particular, six research assistants were given an hour long introduction to speech errors, phonetic training, and instructed in best practices in speech error collection described in sections 2.3 and 3.2. They were then instructed to find set time intervals in their daily lives to collect speech errors, documenting the time, date, speaker information, and as much of the linguistic context of the error as possible. A total of 1,058 errors were collected by the six data collectors in this way.

During this period, a subset of the research assistants also collected speech errors from audio recordings, and two new research assistants were trained to collect speech errors exclusively from podcasts. The benefits of offline collection in terms of data reliability led the entire team to switch to exclusive offline collection, and over 9,000 speech errors to date have been collected in this way. This logistical decision, however, leads to a problem in terms of

comparing online and offline errors because many of the offline errors were collected at points in time when the collectors themselves had far more data collection experience. To balance for this, we will examine a subset of the data submitted from each data collector so that they match in experience level. In particular, a set of 100-215 errors were taken from each collector after he or she had successfully completed the training and submitted their first 30 valid speech errors. This selection procedure resulted in a total of 533 offline errors and 839 online errors, because we had more data collectors trained initially to collect errors online. There is perhaps a small effect of experience for some of the offline data collectors, but many of the statistical effects we discuss below are so strong that we doubt they could be the result of different experience levels.[1]

5.2 Results and discussion

We begin with some baseline data to give a general sense of the pattern frequencies. Breaking down errors by their linguistic level, as done in Table 7, we find broad similarity between the two collection types. The number of sound errors and word errors are comparable, and the only real difference observed is that errors involving individual morphemes are a bit more common in online errors, while phrase errors like phrasal blends and substitutions are a little less common.

Table 7. Error levels, sorted by collection method.

	Offline	Online
Morpheme	18 (3.38%)	51 (6.08%)
Phrase	24 (4.5%)	19 (2.26%)
Sound	315 (59.1%)	506 (60.31%)
Word	176 (33.02%)	263 (31.35%)

Table 8 breaks down the sound errors by type, again showing similar pattern frequencies between the two collection methods. Gradient errors are of course far more common with offline collection, but this is simply due to the fact that they are rather difficult to collect online without recourse to the audio recording. Once gradient errors and shifts, which are too small in number to assess, are removed, there is no significant association between error type and collection method (χ^2= 0.04, P=0.8415).

Table 8. Sound errors, sorted by type and collection method.

	Offline	Online
Addition	55 (17.46%)	72 (14.23%)
Deletion	19 (6.03%)	36 (7.11%)
Gradient	39 (12.38%)	3 (0.59%)
Shift	1 (0.32%)	4 (0.79%)
Substitution	201 (63.81%)	391 (77.27%)

Sound errors can be distinguished by two salience measures, namely the percentage of errors that occur in the stressed syllable, and also the percentage of corrected errors. In these, we again find only small insignificant differences, as shown in Table 9 and Table 10. We might have

[1] We note that a prior study, Boomer & Laver (1968), examines a small number of errors collected both online and offline, but this study was not designed to investigate this difference, nor did it report on it directly.

expected a larger difference in percentage of corrected errors than the 3% difference reported in Table 10, but there is reason to believe that this difference is greater because of differences in reporting. We find in practice that the fact that an error was corrected is an afterthought in error collection that is easy to miss with online errors, but always possible to discern with offline collection. Therefore, we expect this difference to be greater, with online errors having an even higher percentage of corrected errors.

Table 9. Sound errors sorted by stress and collection method.

	Offline	Online
Error in main stressed syllable	240 (76.19%)	370 (73.12%)
Not in main stressed syllable	75 (23.81%)	136 (26.88%)

Table 10. Sound errors, sorted by correction and collection method.

	Offline	Online
Corrected	129 (58.65%)	192 (61.68%)
Not corrected	183 (41.35%)	309 (38.32%)

Two more subtle measures, however, reveal an important difference between the two collection methods. Research has shown that sound errors are subject to a repeated phoneme effect (Dell, 1984; MacKay, 1970; Wickelgren, 1969), or the tendency for the phonetic environment of the intruding sound to be the same in both the source and error word. For example, in "… *they're /plas= passing over the ^plains of the …"* (sfused10), the intruding sound [l] occurs before the phoneme [p] in both the source *plains* and intended *pass*. This effect seems to be stronger in online errors than offline errors, as shown in Table 11 (χ^2= 6.85, *P*=0.0089).

Table 11. Sound errors, repeated phoneme effect sorted by collection method.

	Offline	Online
Repeated phoneme	51 (16.19%)	122 (24.11%)
No repeated phoneme	264 (83.81%)	384 (75.89%)

In terms of perceptual biases, one may conjecture that errors exhibiting the repeated phoneme effect are more salient, perhaps due to priming by the phonetic context in the source word. However, we think a more likely explanation for this difference is that the repeated phoneme effect is affected by speech rate. Our data collectors were instructed in the best practices from prior research, and one important instruction in online collection is to only collect errors that they are absolutely confident in. As a result, online collectors are likely to have collected errors that were produced at a slower rate, because these are naturally easier to detect and document with confidence. Offline collectors, however, have the ability to replay errors as much as possible, and so we conjecture that they will have collected many more errors spoken at faster rates. The fact that the repeated phoneme is stronger in online errors can be seen therefore as a consequence of the general fact that this effect is stronger at slower speech rates (Dell 1986). We have no way to directly contrast speech rates for offline and online errors, because speech rate cannot be measured for the online errors. However, another effect known to be influenced by speech rate seems to corroborate this finding.

Sound errors are known to be subject to a lexical bias, or the tendency for errors to result in actual lexical items, and this effect is also stronger in slower speech (Baars, Motley, & MacKay, 1975; Dell & Reich, 1981; Stemberger, 1984). As shown in Table 12, it seems that online errors have a stronger lexical bias than offline errors. This result needs to be qualified, however, by the fact that there are also many more offline errors in which the error word is clipped by the talker, where the full realization of the intended word is aborted by the talker. In our entire corpus, 20.83% of offline errors are clipped, while only 5.96% of online errors are clipped, perhaps again the result of the need to be certain of the error and the lack of replay with online collection. Though some clipped words may result in actual words, we take the prudent position that we are ignorant of their lexical status (because we do not know how they would be completed). This means that the lexical bias could in fact be greater for offline errors, which would reduce the difference between the two. However, with these clipped words removed, there is a non-significant difference in the strength of the lexical bias that is consistent with the contention that offline errors are on average spoken more quickly than the online errors.

Table 12. Sound errors, lexical bias sorted by collection method.

	Offline	Online
Lexical item	39 (19.40%)	104 (27.23%)
Not a lexical item	104 (51.74%)	245 64.14%)
Clipped word	58 (28.86%)	33 (8.64%)

Another measure of salience in sound errors is the frequency of errors occurring in word-initial position, given research showing that initial position is more salient perceptually (see section 2.2). Interestingly, errors in word-initial position seem to be more common with offline collection than online, though the difference reported in Table 13 is not statistically significant (χ^2=4.22, P=0.1212).

Table 13. Sound errors, word onset effect sorted by collection method.

	Offline	Online
Initial	110 (41.04%)	143 (31.22%)
Medial	123 (45.9%)	255 (55.68%)
Final	35 (13.06%)	60 (13.1%)

Next let's examine some differences stemming from the context and direction of sound errors. Table 14 gives the relative frequencies of contextual and noncontextual errors, where contextual errors are standardly defined as errors that contain a source word with the phonological content of the intruder. Online errors are more likely to be contextual than offline errors (χ^2=23.04, P<0.0001).

Table 14. Sound errors, contextual/noncontextual by collection method.

	Offline	Online
Contextual	192 (60.95%)	389 (76.88%)
Noncontextual	123 (39.05%)	117 (23.12%)

It could be that the phonological content in the source word effectively primes the recognition of an error, and therefore that noncontextual are less salient than contextual errors.

Within the set of contextual errors, there are important differences that stem from the direction of the source sound. In Table 15, we show the relative directions of contextual sound errors. "Anticipations and preservation" errors are simply errors in which the intruding sound can be found in both a prior (perseveration) and following word (anticipation), and "incompletes" are errors that are ambiguous between anticipations and exchanges because there is a break between the error word and the source word downstream. There is a significant association between direction and collection type, where offline errors have much less perseveration and exchange errors (χ^2=4.38, P=0.0364).

Table 15. Sound errors, direction sorted by collection type.

	Offline	Online
Anticipation	54 (27.98%)	119 (30.36%)
Anticipation and Perseveration	53 (27.46%)	52 (13.27%)
Incompletes (broken anticipation)	29 (15.03%)	47 (11.99%)
Perseveration	56 (29.02%)	149 (38.01%)
Exchange	1 (0.52%)	25 (6.38%)

One might be tempted to write off the difference in perseverative errors, given the high number of ambiguous anticipation and perseveration errors in offline collection, which could account for the difference in unambiguous perseverations. However, the difference in exchange errors is striking, and, as shown in Table 16, it is true of all types of exchanges.

Table 16. Exchange errors, by linguistic level.

	Offline	Online
Morphemes		6
Phrases		1
Sounds	1	25
Words	1	15
Totals	2 (0.38% of 533)	47 (5.6% of 839)

The most direct account of this difference is that exchange errors are far more salient than other errors because there are two intruders, and in practice they impede significantly in comprehension (see Stemberger 1982/1985: 22). Because attentional resources are more limited in online collection, these rare but easier to hear errors have a much higher frequency. This interpretation entails that other harder to hear errors have a lower frequency in online collections than they actually occur in natural speech. The large difference we observed between offline and online exchange errors is thus a strong indication that online errors are more subject to perceptual bias, in particular the attention and content biases. We note that the observed 5.6% exchange errors from online collection compares with some prior studies (Boomer & Laver, 1968; Nooteboom, 1969), but greatly undershoots the percentage of exchanges in other studies where exchanges have a much higher percentage (e.g., Pérez et al. (2007), Söderpalm (1979)).

Finally, the specific substitutions observed in sound errors can be contrasted by collection method. Confusion matrices of consonant substitutions constructed separately for online and offline errors have some obvious similarities, and also some striking differences. Table 17 reports the top ten most frequent substitutions of single consonants in both collection methods. Both methods have high counts of certain substitutions, e.g., *s* → *ʃ*, a substitution likely due to the palatal bias ((Shattuck-Hufnagel & Klatt, 1979; Stemberger, 1991), and the intruder exhibits a high degree of phonological similarity with the supplanted intended sound (Broeke & Goldstein, 1980; Cutler, 1980; Fromkin, 1971; Levitt & Healy, 1985). However, other substitutions seem to be dominant in only one collection method. The difference between the occurrence of *r* → *l* and *l* → *r* in online errors, and their absence in the offline top ten, is striking.

Table 17. Top ten sound substitutions, sorted by collection method.

Offline (of 140 total)			Online (of 257 total)		
Substitution	*N*	F change	Substitution	*N*	F change
p → b	7	1	r → l	24	1
s → ʃ	7	1	s → ʃ	14	1
k → t	5	2	l → r	13	1
n → d	4	2	m → n	9	1
k → g	4	1	n → m	6	1
w → r	4	2	ʃ → s	6	1
z → s	4	1	k → t	6	2
m → b	3	2	p → b	5	1
m → n	3	1	b → d	5	2
p → t	3	2	n → l	5	2

Generally speaking, it seems that the confusion matrix for online consonant substitutions is more symmetric and concentrated in certain substitutions than the offline matrix. Thus, 6 of the top ten online substitutions have symmetric counterparts, e.g., *m* → *n, n* → *m*, while the offline top ten have none. The top ten substitutions also account for 36.19% of all observations online, compared to 31.43% in offline errors. Correspondingly, the offline matrix is more diffuse, with 69.32% of all substitutions observed only once, cf. 59.83% for online errors. The difference in symmetry leads to another difference, which is that the offline matrix seems to exhibit stronger output biases in which high frequency sounds replace relatively lower frequency sounds, e.g., *z* → *s, p* → *t*, cf. *s* → *ʃ*, an antifrequency bias (Stemberger, 1991). Whether or not these differences stem from difference in the impact of perceptual biases or frequency biases is difficult to tell from this data. However, it seems clear that the two are different generally, and that offline substitutions are more asymmetrical and diffuse than online errors.

Let's move now to a set of comparisons within word errors. The frequencies of different types of word errors classified in our corpus are shown in Table 18. Errors in stress and intonation are typically very rare, and nearly impossible to document online. If we remove these

errors and also the low frequency additions, we find a significant association between error type and collection method (χ^2=12.99, P=0.0015). Thus, while word substitutions dominate both online and offline errors, there is a higher percentage of substitutions and blends in online errors, offset by a higher number of additions and deletions in offline errors. Blends, because they produce rather odd nonsense words, compare with exchanges in their overall salience. It is not a surprise, then, that we find more than twice as many blends in online errors than offline errors.

Table 18. Word errors, sorted by type and collection method.

	Offline	Online
Additions	7 (3.98%)	3 (1.15%)
Blends	9 (5.11%)	30 (11.45%)
Deletions	17 (9.66%)	8 (3.05%)
Stress/Intonation	3 (1.70%)	0
Substitutions	140 (79.55%)	221 (84.35%)

Word errors can also be classified by the semantic relationship between the intended and error word, as substituted words and blends are often semantically related in some formal sense. There is no established set of semantic relations, but we offer the follow four classes of relatedness that seem most appropriate to our data. Two words can be in the same semantic field (e.g., *cherry, blueberry)*, thematically related in that they often co-occur (e.g., *spider, web*), antonyms, or they can be related in the sense that they are one of a few options from a fixed set (e.g., *dinner, lunch*). This last option is rather common in entertainment podcasts in which the talkers need to refer to characters in TV shows and films, so this perhaps accounts for the high number of mis-selection from fixed set word pairs in offline errors. When these errors are removed, we still find important differences in antonyms and thematically related words in the directions shown below, but they are not significant (χ^2=2.25, P=0.3247).

Table 19. Word errors, sorted by semantic relation and collection method.

	Offline	Online
Same semantic field	21 (28.38%)	33 (37.50%)
Thematically related	11 (14.86%)	25 (28.41%)
Antonyms	7 (9.46%)	6 (6.82%)
Misselection from fixed set	35 (47.30%)	24 (27.27%)

Word errors can also be distinguished by the word class of the intended word, and any changes in word class from the intended to error word. Table 20 shows the word class of the intended word in word substitution errors. It is clear that there is a strong preference for nouns in online errors, where offline errors make up for the difference in noun substitutions with a greater number of substitutions with names, pronouns, and prepositions (χ^2=23.11, P=0.0001, with prepositions removed). Thus, as with the confusions matrices for consonant substitutions, we find that offline word substitutions are more diffuse.

Table 20. Part of speech of intended words in word substitutions.

	Offline	Online
Nouns	27 (27.27%)	101 (51.53%)
Verbs	25 (25.25%)	50 (25.51%)
Adjectives	9 (9.09%)	19 (9.69%)
Names	16 (16.16%)	13 (6.63%)
Pronouns	15 (15.15%)	10 (5.10%)
Prepositions	7 (7.07%)	3 (1.53%)
Totals	99	196

Perhaps nouns are more salient and therefore have a greater overall percentage in online errors, due to the limits on attention. Another measure is how well word substitutions obey the category constraint, or the preference for substitutions that retain the same word class as the intended word (Bock, 1982; Garrett, 1975). There is a greater tendency for intended words to substitute with a word in the same word class in online errors (88.78%) relative to offline errors (84.85%), though this difference is not significant.

To summarize the above findings, there seem to be some differences between the online and offline errors that are artifacts of the collection method or setting. For example, word errors in which the substituted error is a member of a fixed set is much more common in offline errors (see Table 19) simply because in the data sources, entertainment podcasts, the talkers need to refer to characters in movies and TV frequently and they just mix up the names. More importantly, there are differences that seem to relate to the attentional resources intrinsic to the collection method. Online errors exhibit stronger psycholinguistic effects (repeated phoneme effect and lexical bias) that suggest errors are taken from slower more careful speech. They also have a much larger number of errors that require less attention, like exchange errors and word blends. On the other hand, there are some differences that are not predicted by known perceptual biases, and therefore must have another explanation. For example, the asymmetrical consonant substitution patterns are not predicted by perceptual bases, but they may be predicted by output biases for high frequency sounds. At any rate, it is clear that there are significant differences between collections of speech errors collected online versus those collected offline. We believe these differences show that future work documenting offline errors with larger baselines will contribute significantly to the empirical knowledge base of the structure of speech errors.

6. Data discovery

In this section, we investigate some of the new directions that speech error research can take with our methodology. Our approach, inspired by Chen (1999, 2000), involves data collection from audio recordings by many listeners. The existence of an audio recording provides the obvious benefit of allowing another pass at the speech facts to confirm empirical observations. In addition, it gives the researcher a chance to "dig deeper" into the data. For example, citing a finding from Ferreira and Swets (2005) that more errors are found in longer utterances and more complex speech, MacDonald (2016) notes that the lack of context typically recorded in speech errors collections precludes full assessment of this claim. Such requests for more data are not uncommon in the speech error literature, and access to an audio recording allows the researcher to find the additional necessary information. However, the benefits of working with an actual recording go beyond the ability to drill down into the data. Audio recordings also make it

logistically straightforward for multiple listeners to assess the same speech facts, and furthermore, they embed speech to an explicit metric for time. Below we examine some of the new opportunities for data discovery that can be built upon these assumptions.

6.1 Data collection metrics and the frequency of speech errors

As discussed in 3.2, because audio recordings have a specific length of time, we can assess how frequently, on average, a data collector is observing errors. In particular, we use the measure of minutes per error (MPE) to gauge if a data collector is collecting errors are a reasonable rate. After some experimentation, our team has settled on an MPE of 3.0, or a speech error collected every three minutes, or lower. Error submissions with higher MPEs, meaning that more errors have been missed, usually triggers the data collector to relisten to the recording or the database manager to assign the recording to a different data collector in order to achieve a more representative sample.

Our methodology also makes it possible to provide better estimates of speech error frequency in the general population. Estimates of the frequency of speech errors are typically based on counts of attested errors relative to some baseline in a corpus. For example, Ferber (1991)'s team collected 51 speech errors in a 45 minute interval composed of 15 separate samples stitched together. Though the sample is small, and somewhat artificial given the disjointedness of the speech, it yields a MPE of 51/45 or 0.88, which is equal to an error about every 53 seconds. Chen's (1999) corpus of Mandarin speech errors is larger, with 987 errors collected from approximately 4,800 minutes of speech. This sample produces a much larger MPE of 4.86, but Ferber's team had two additional data collectors, and also Chen threw out many errors because they did not meet his stricter definition of a speech error. The London-Lund corpus (Garnham, Shillcock, Brown, Mill, & Cutler, 1981) recorded 191 errors out of approximately 17,000 words. If we take 2.5 words a second as the average speaking rate (Maclay & Osgood, 1959), or 150 words a minute, that converts to an MPE of 5.93, which is still rather high compared to Ferber's findings. The important point is that these estimates are based on actually observed errors, though researchers freely acknowledge that there are errors that have been missed. For example, Garnharm et al. (1981, p 806) note, "There can be no pretence that all slips of the tongue in the corpus have been listed. Thus the estimate of the frequency of speech errors in conversation is a conservative one. A number of factors have prevented a complete listing from being made."

By taking multiple samples from the same recording, a more realistic estimate of the total number of errors can be made by using capture-recapture methods. Capture-recapture methods are commonly used in ecology, for example, to estimate animal populations when it is not practical to attempt a count of all members of the population. Capture-recapture involves multiple samples of the population and marking the individuals found in different samples. Estimates of the total population are then calculated as a function of the proportion of individuals found in all samples (see Chao (2001) for an overview).

The collection of speech errors is parallel in many ways with the kinds of problems investigated with capture-recapture methods. The difficulty in exhaustively counting the number of speech errors makes complete counts impractical. It might seem plausible that a researcher can collect all of the errors in a given recording. After all, they can listen and relisten to every second of speech. However, the facts of experiment 1, as well as Ferber (1991) findings, strongly suggest this is not the case. When the same recording is heard by many listeners, most errors are only heard by one listener. This conclusion is also consistent with our findings in training new data collectors. Three of the 40 minute recordings we used in experiment 1 are used in training

new data collectors, and new trainees regularly find large numbers of new errors in recordings that have already been scoured by three listeners. It is simply impractical to exhaustively count the speech errors in any sizable speech corpus, as attested in the quotation above from Garnharm et al. (1981).

The availability of an audio recording facilitates making multiple samples that are needed for capture-recapture techniques. However, recent work on capture-recapture (Mao, Huang, & Zhang, 2016/To appear) argues that it is not possible to estimate the population size when the items being counted are heterogeneous in nature, because there can be arbitrarily many hard to find items. Instead, Mao et al. recommends estimating the lower bound and provides a formula for doing so (their (22-23)). As discussed in detail section 2, speech errors are clearly heterogeneous because they occur with different levels of linguistic structure, and they also clearly differ in detection difficulty. As a result, we can estimate the lower bounds of the number of speech errors for a given recording, but not the actual population.

Table 21 below shows the count data from the nine recordings from experiment 1. In particular, it shows the specific number of unique errors found only by the three listeners A, B, and C, as well as the counts of unique errors by all three pairings, e.g., "AB" is the number of errors found only by both A and B, and finally the observations from all three listeners "ABC". n is the total number of actually observed errors, and $\tilde{v}$ is the estimated lower bound using Mao et al.'s formulas, which is equal to $\tilde{m}$ (estimated lower bound of missed errors) + n. These estimates bring us into the time scale of seconds, so we report SPE or seconds per error. While there is some variation in the podcasts, averaging across these nine recordings gives an SPE of 48.49, which is a bit lower than the frequency of attested errors from Ferber's recordings (though recall that this estimate did not calculate missed errors). It should also be noted that this number is conservative. It is a lower bound estimate, so the actual population of errors will likely be larger, and consequently, the average SPE will be smaller. Indeed, the count for the 2,377 second recording (second row from the bottom) has been used in our training regime for new listeners, and after four new listeners have examined this recording, 24 additional errors have been found, bringing the total to 54, which far exceeds the lower bound of 43.39.

Table 21. Count data and estimates from individual recordings.

Seconds	A	B	C	AB	AC	BC	ABC	n	$\tilde{m}$	$\tilde{v}$	SPE
2,100	2	18	3	2	0	3	5	33	16.3	49.3	42.60
1,690	6	5	4	5	0	2	9	31	13.48	44.48	38.00
1,993	2	9	5	1	0	1	5	23	20.08	43.08	46.26
2,385	6	6	5	8	2	1	5	33	11.7	44.70	53.36
4,143	24	9	1	5	1	1	3	44	21.84	65.84	62.93
3,000	9	2	7	3	5	1	2	29	10.63	39.63	75.70
1,800	9	9	3	2	0	1	1	25	29.87	54.87	32.81
2,377	15	2	4	3	2	1	3	30	13.39	43.39	54.78
2,400	18	4	6	1	2	0	7	38	41.93	79.93	30.03

It is possible that the estimate SPE of an error every 48.49 seconds is lower than other estimates because we include gradient errors, and other collections do not recognize this type. However, many of these gradient errors would be counted as regular sound errors in other collections, so we do not think it affects the overall rate to a large degree. The fundamental difference between our estimate and those of prior research is that we use capture-recapture

methods to estimate missed errors. We know from experiment 1, and indeed the acknowledgement by other research teams, that many errors are not counted simply because they have not been found. As a result, we believe that prior research has significantly under-estimated the frequency of speech errors in natural speech. This finding is relevant to the larger field of language production research, because a common thread running through this literature is that speech errors occur very infrequently, and thus, research should focus on normal non-erroneous speech (Levelt et al., 1999).

6.2 Speech rate effects

The effects of speech rate have long been an factor of interest in speech production research. The spreading-activation model of language production proposed in Dell (1986), for example, predicts a simple trade-off between speech and accuracy, with more errors in faster speech. Furthermore, some psycholinguistic effects are known to be stronger at slower rates, including the lexical bias effect and the repeated phoneme effect discussed in section 5. These speech rate effects have been documented in speech errors collected in experimental settings (Dell, 1985; Dell, 1986; Dell, 1995; MacKay, 1971), but they remain to be corroborated in natural speech.

The luxury of having an audio recording at one's disposal makes testing these hypotheses a tractable problem. By adopting an accepted measure of speech rate, either phonemes per time unit (Cucchiarini, Strik, & Boves, 2002) or syllables per time unit (Kormos & Dénes, 2004), relative speech rate can then be assigned to an interval of the recording, a procedure made considerably more efficient by the existence of automatic tools for assessing speech rate (de Jong & Wempe, 2009). Assigning a speech rate measure to speech chunks in turn makes it possible to test speech rate effects in natural corpora. For example, to test the general speech-accuracy trade off, regions of existing recordings that exhibit speech rate differences can be segmented and their rate measured. If speech rate affects incidence of speech errors, we expect faster speech rates to have lower MPEs (=more errors) than regions with slower rates. Moreover, specific psycholinguistic effects can also be tested by assigning speech rate values to smaller intervals. Thus, speech errors can be associated with the speech rate, e.g., syllables per second, of a ten second envelop and then compared in a larger sample. To test the effect of speech rate on the lexical bias, one can bin errors into qualitatively distinct rate types, and then test if the lexical bias is stronger at slower rates. In sum, the ability to situate specific errors in a system for measuring speech rate opens up new doors for empirical investigation.

6.3 Gradient errors

Another opportunity supported by our methodology is exploration of gradient phonetic errors. As discussed in 2.2, critical assessment of speech error collection and analysis has led to a growing interest in the phonetic structure of speech errors (see Pouplier and Hardcastle (2005) and Goldrick and Blumstein (2006) for review). Whereas classic speech error studies focused largely on categorical sound errors, and indeed lacked the tools to describe fine-grained phonetic structure, new research paradigms have emerged that probe the articulatory, acoustic, and perceptual structures of speech errors (S. A. Frisch & Wright, 2002; Goldrick & Chu, 2014; Goldstein, Pouplier, Chena, Saltzman, & Byrd, 2007; Marin et al., 2010; Mowrey & MacKay, 1990; Pouplier & Goldstein, 2005; Slis & Van Lieshout, 2016). The scope of this research is too diverse to engage with here, but many of the results run counter to the research findings based on data collected from observational techniques. These findings include a general assessment that speech errors in articulation are more common than categorical errors (Mowrey & MacKay, 1990), the existence of graded sound categories or blends between two discrete sound categories

(S. A. Frisch & Wright, 2002; Goldrick & Blumstein, 2006), a higher occurrence of phonotactially ill-formed errors (Mowrey & MacKay, 1990), and documentation of asymmetries in the direction of sound errors not observed in categorical errors (S. A. Frisch & Wright, 2002; Pouplier & Goldstein, 2005).

From this research, a kind of consensus view has emerged that recognizes both categorical phonological errors and gradient phonetic errors (S. A. Frisch, 2007; S. A. Frisch & Wright, 2002; Marin et al., 2010; Stearns, 2006). Phonological errors, on the one hand, are pre-articulatory errors that involve higher level production planning. They are essentially a phonological segment that has been mis-selected, resulting in an exemplar of an unintended sound category. Gradient sound errors, on the other hand, involve a mis-selection of, or competition within, an articulatory plan, producing an output sound that falls between two sound categories. Gradient errors can be perceptible or imperceptible, depending on the phonetic structure, and often require deeper phonetic analysis to establish. A standard approach has been to examine the continuous phonetic parameters that distinguish two categories, and define gradient phonetic errors as sounds that are outside the normal range of variation of the intended sound (typically two standard deviations from an average value), and also outside the normal range of another sound category. Gradient errors therefore differ from categorical errors because categorical errors are inside the normal range of an unintended sound category. We acknowledge that there are theoretical perspectives that can capture both categorical and gradient speech errors as we define them here (Goldstein et al., 2007; Smolensky, Goldrick, & Mathis, 2014), but we follow this consensus view that there exists an empirical distinction at least between categorical phonological errors and gradient phonetic errors.

The examination of gradient phonetic errors has in large part been conducted with experimentally elicited speech errors. While some speech error collections acknowledge the existence of phonetic errors (e.g., Stemberger (1993)'s taxonomy recognizes sound blends), the practice of most speech error collection has been to focus on categorical errors, and indeed, transcription practice in the past has tended to require this. Our experience with extensive listening for speech errors from sound recordings, however, is that many errors on closer investigation are indeed gradient in nature and fall between two sound categories. As described in 3.1, we adapt a transcription commonly used in child language (Stoel-Gammon, 2001) that recognizes ambiguous segments and other indeterminate sound categories. In particular, categorical errors involve discrete sound categories, typically an addition, deletion, or substitution of a phoneme of English. Gradient errors, on the other hand, involve all of these error types, but the output may be ambiguous between two poles (A|B), transitional between two poles (A-B), or intrusive. We acknowledge that there will be aberrant speech that cannot be collected from listening to audio recordings because they involve phonetic structure that are imperceptible (Mowrey & MacKay, 1990). However, we believe that by acknowledging the distinction between perceptible phonological and perceptible phonetic structure supported by prior research, will lead to a better understanding of both classes of speech errors.

The results below offer a preliminary look at the structure of phonetic errors collected from audio recordings of natural speech. From a sample of 1,393 offline errors, 839 (60.23%) of which are sound errors, our team has collected 163 gradient errors, or 19.43% of all sound errors. These are shown below in Table 22, sorted by gradient error type (see Table 1) and whether or not the error is contextual. Ambiguous errors are by far the most common, followed by transitional, then intrusive. Interestingly, the percentage of contextual transitional errors is rather close to the percentage of contextual errors in phonological sound errors, which is 60.95% (Table

14), but the percentage of contextual ambiguous errors is much lower at 38.26%. Therefore, while many of the phonetic errors seem to be tied to production planning of nearby segments, at least some ambiguous errors seem to require a different mechanism.

Table 22. Gradient sound errors, sorted by type and contextual/noncontextual.

	Ambiguous	Transitional	Intrusive
Contextual	44 (38.26%)	24 (57.14%)	1 (16.67%)
Noncontextual	71 (61.74%)	18 (42.86%)	5 (83.33%)
Totals	115	42	6

Of the 115 ambiguous errors, 76 (66.09%) are C|C errors between two consonantal poles, and 39 (33.91%) are between two vowel poles. To flesh out these patterns, Table 23 below lists all ambiguous errors with more than three observations. For ambiguous C|C errors, it seems that voicing in stops and nasality between nasals and corresponding voiced stops are the salient dimensions, though the frequencies reported here are too small to make any conclusion on the direction of these changes (e.g, voiced to voiceless and vice versa). In all of the prominent ambiguous sound errors reported below, the difference between the two poles can be described with a single phonological feature, something that is not always true with categorical phonological errors (see Table 17).

Table 23. Ambiguous sound errors, sorted by C/V type

C|C errors		V|V errors	
b|p	6	ɛ|æ	3
b|m	5	i|ɪ	3
ʃ|s	5		
m|b	4		
d|n	3		
g|k	3		
g|ŋ	3		
k|g	3		
p|b	3		

To summarize, our pilot results substantiate some basic conclusions that can be pursued in future work. First, gradient phonetic errors do exist with some frequency in natural speech, substantiating the claim based on experimentally induced errors that speech errors may involve sounds that lie on the continuum between two discrete categories. While our study is limited to just errors that can be perceived by trained listeners, they occur at a relatively high frequency, or roughly one in five sound errors. Second, trained data collectors can, and indeed often feel compelled to, distinguish between phonological and phonetic errors. Gradient errors have been observed by all data collectors (see experiment 1), and so the ability to perceive these is really a matter of sufficient training. Finally, there do seem to be some subtle differences between perceptible phonological errors and perceptible phonetic errors, as shown by high frequency of noncontextual ambiguous errors and the specific shape of these errors reported in Table 23. We believe that further investigation of gradient sound errors in natural speech with larger baselines will be a fruitful line of investigation.

7. General discussion

This article probes a methodology for collecting speech errors from audio recordings, both to determine if it is a viable way of collecting data, and to determine how it compares with traditional observational techniques used in online collection. The results (experiment 1) show that it is valid to collect large number of errors using many data collectors, because different data collectors are broadly consistent in the types of errors they collect, even though they detect different specific errors. Also, speech error collection requires a mechanism to verify speech errors, because even trained and experienced data collectors produce large numbers of false positives (16-32%). Experiment 2 compared online and offline data collection and found a host of differences, including important differences in the strength the speech rate effects and the word onset effect, the overall percentage of contextual errors, the frequencies of specific sound error types and directions, consonant substitution patterns, and word substitution patterns.

With these results in mind, it appears that the offline approach with many data collectors has many advantages over an online approach. First, it offers a direct mechanism to verify the data patterns via an audio recording. Second, it appears that the offline methodology results in many more errors that are missed with online collection. One form of support for this claim is that the percentage of errors that require a small amount of attention (e.g., exchanges and word blends) is much higher in online errors, showing that there is a higher percentage of errors that require more attention in offline errors. Another relevant fact concerns the speech rate effects: the lexical bias and repeated phoneme effect are stronger in online errors, suggesting that online errors are collected in slower speech than errors collected offline. This accords with the simple fact that offline errors can be slowed down and relistened to, whereas online collectors only get one shot at detecting the error. Third, given the first two points, it seems logical to conclude that offline error collections better reflect the natural occurrence of speech errors. While the generalizations supported by our offline sample are still preliminary, the differences we observe between offline and online errors strongly suggest that larger collections of offline errors will uncover empirical generalizations that differ from some of the patterns in past research. Fourth, data collectors trained in English phonetics and casual speech processes can document fine-grained phonetic facts of speech errors, facts which also may support new knowledge in phonetic research. However, this point is limited to essentially acoustic facts, and there remains a rich empirical domain of articulatory data that can only be examined with other methods. Finally, the ability to resample the speech and locate errors in specific time intervals supports more realistic estimates of the frequency of speech errors in natural speech and investigation of hypotheses tied to speech rate.

While the advantages of offline collection seem to be considerable, it is appropriate at this time to also consider some of the drawbacks of this approach. One limitation is that because we rely exclusively on audio recordings, our approach does not easily incorporate visual input that may have influenced a talker's errors. Most of the major English speech error collections recognize visual input as a potential source for speech errors. For example, Stemberger (1993) classifies errors with objects and actions external to the speaker that influence the error as "environmental errors". However, there is a sense in which the data we collect is more controlled because our offline errors are usually confined to static studio environments devoid of changing visual information. Occasionally, a talker will pull up a web-browser and comment on something in a webpage, which may provide visual stimuli that influences their speech. However, for the most part, almost all of the data we collect was created in static visual environments. Thus, while

our methodology does not lend itself to studying the impact of visual information, our data also does not seem to be influenced by it to the same degree as speech collected in other settings.

Another limitation of our methodology is that, because the recordings are made by a third party in the past, we cannot question a speaker's intention or thought process after an error is produced. In some methodologies, an effort is made to immediately question the speaker after an error has been made to ascertain the speaker's thoughts, the intended utterance, and any environmental influences (Harley, 1984; Vousden et al., 2000). It is claimed that this practice guards against perceptual biases of the very kinds we are concerned with in this article, so it is important to address this limitation.

While we think this practice is fine in principle, we do wonder how useful it is in the collection of large number of errors. If errors occur with the frequency we think they do, about once every 48.5 seconds, interviewing speakers immediately after they produce an error is likely to lead to rather artificial conversations. Furthermore, in our own experience with online collection, we find that such interruptions of the flow of a conversation are unnecessary because the context and intended utterance are usually self-evident. Moreover, it is not always the case that interviewing speakers leads to additional or more accurate information. As a practical mechanism to address this uncertainty, we allow data collectors to rate errors they collect, and in particular allow them to assign a low confidence value to errors that they do not have a reasonable sense of what the intended word or phrase is. In the offline errors, 11 of the 533 errors of all types (2.06%) were rated with low confidence. Thus, while the concern about the uncertainty of intended utterances is real, it is really only a concern for a small fraction of our errors.

Finally, an interesting difference found in experiment 2 is that sound errors in word-initial position have higher frequency in offline errors than online errors (see Table 13). We expect from prior research that word onsets should be special, because of the word onset asymmetry showing a preference for sound errors in word-initial position (MacKay, 1970; Shattuck-Hufnagel, 1987; Wilshire, 1998). But why would this effect be stronger in offline errors, when it is known that errors are easier to detect in this position (section 2)? One of the important findings from Wilshire's (1998) study is that in experimentally induced errors, the word onset effect is only found in contextual errors with actual words; the asymmetry goes away with nonsense words and in non-contextual errors. It is interesting in this light to return to the word onset data and distinguish contextual and non-contextual errors, as done below in Table 24. Overall, these patterns do not exhibit the magnitude of the word onset effect found in prior studies, which ranges between 50-90% in spontaneous speech error collections. While we find differences in offline and online errors in both error types, the strongest difference is in non-contextual offline errors, which approaches a 50%-50% split between initial and non-initial errors.

Table 24. Sound errors, word onset effect, contextual vs. non-contextual (in column percentages)

	offline		online	
	contextual	non-contextual	contextual	non-contextual
initial	64 (37.4%)	46 (47.4%)	116 (30.1%)	27 (32.5%)
non-initial	107 (62.6%)	51 (52.6%)	259 (69.9%)	56 (67.5%)

These findings are interesting because recent work has suggested that putative word onset effects in German are a result of the way the errors are collected (Marin & Pouplier, 2016). When

collected from audio recordings, sound errors were found to be more common word-medially and there is no preference for word-initial errors. Our methods are comparable to Marin and Pouplier's in that we work primarily from audio recordings, and the lower frequency of word-onset errors in our study, relative to prior studies using online data collection, is consistent with their findings. However, the difference between offline and online errors seems contrary to their larger conclusion that an offline methodology reduces the word-onset effect because our finding is that it is increased in offline errors. The careful listening afforded in offline collection actually results in a higher number of word-initial errors, and more strikingly so with non-contextual errors. When we examine the non-contextual initial sound errors, we find that 18 of the 46 are gradient errors that require a certain amount of attention that is only possible with offline collection. Perhaps this explains how the word onset effect has a stronger impact in non-contextual errors, contrary to Wilshire's (1999) findings, which may have focused on phonological errors only. We also believe that these findings, together with Marin and Pouplier's (2016) results, highlight the need to distinguish specific classes of sound errors when establishing broad generalizations about sound errors and the psycholinguistic effects that affect them.

Appendix

The fields of the speech error data tables of SFUSED are organized by type below, with a sketch of the value lists or typical values of each field.

Example fields [10]

Example: text for speech error with mark-up of error position and source words
Intended word: inferred intended word
Intended word in IPA: phonetic transcription of intended word
Confidence of intended word: confidence rating of inferred intended word
Error word/phrase: text of error word/phrase
Error word in IPA: phonetic transcription of error word/phrase
Confidence in transcription: confidence rating of transcribed error
Corrected error: y/n, is the error is corrected in response?
Error word bounded?: y/n, are the error and source unit in same word?
Clipped word?: y/n, is the error word is clipped?

Record fields [15]

Record identifier: number used to identify error
Researcher found: label for researcher who found the error
Found date: date of error detection
File: file name of podcast
Podcast: label for podcast series
Time stamp: h:mm:ss time stamp on associated file
Online/offline: online or offline data collection
Talker self?: for online errors, is the researcher the talker?
Speaker: for offline errors, label for talker
Sex: sex of talker
Spreadsheet: source spreadsheet for original submission

Personal info?: y/n does the error example include personal information
Record completed?: y/n, is the classification of this record complete?
Confirmed by: label for researcher(s) who did classification and confirmed the record details
Last modified: date stamp for last modification of the record

Major class fields [7]
Level: Sound/Morpheme/Word/Phrase
Type: Substitution/Addition/Deletion/Shift/Blend/Gradient/Stress or Intonation Error/Complex Set of Processes/One of a Kind — See Notes
Direction: Anticipation/Perseveration/Exchange/ Anticipation+Perseveration/Incomplete/Noncontextual
Contextual?: y/n, is source for error taken from the linguistic context?
Form rule violation?: y/n, is there a form rule violation (e.g., subject-verb agreement error)
Alternate level: Sound/Morpheme/Word/Phrase (for ambiguous errors)
Check box for complex processes: all above types

Word fields [8]
Error properties:
Error is right lemma?: y/n, is the error the intended lemma (e.g., sound error)?
Error is a lexical word?: y/n, is the error a lexical word?
Error POS: part of speech of error word/phrase (phrase determined from head of phrase)
Intended properties:
Intended POS: part of speech of intended word/phrase
Intended open/closed: is the intended word open or closed class?
Intended regular/irregular: is the intended word a regular or irregular word (e.g., irregular verb, noun or adjective)
Error-intended semantic relationship: Same semantic field/"Goes with" (thematically related)/Misselection from fixed set/Same meaning (same lexeme)/Same meaning, different POS/Antonyms/Synonyms (but not same lexeme)/Near synonyms/See Notes/Not obviously related
Error-intended morphological relationship: Same Lexeme (differ only in grammatical function)/Same Word Family (shares base)/Shares a morpheme/Not obviously related

Special class fields [6]
Blend type: Failed Lemma Selection/Blend of Sequence (for word and phrasal blends)
Gradient type: Ambiguous/Transitional/Intrusive/See Notes (for gradient phonetic errors)
Prosody type: Word Stress/Phrase-Compound Stress/Sentence Stress/Pitch Accent/Intonation/See Notes (for stress and intonation errors)
Form rules type: Subject-Verb Agreement/Pronoun Agreement/Auxiliary Selection/Pronoun Case/Nonfinite verbs in Embedded Clauses/Determiner-Noun Agreement/See Notes (for morpho-syntactic errors, like agreement errors)
Morphological categories: Derivational Prefix/Derivational Suffix/Inflectional Suffix/Stem/Bound Stem/See Notes (category of morpheme error)

Locality of shifts: Adjacent/Non-adjacent (for shift errors only)

Sound variables [10]

Of the supplanted portion of the intended word:

Supplanted intended: phonological segment supplanted

Syllable role of supplanted intended:
Onset/Nucleus/Coda/Rime/C1ofOnsetCC/C2ofOnsetCC/C1ofCodaCC/C2ofCodaCC/WholeSyllable/C – Appendix/Appendix + Onset/Coda + Appendix/See Notes/NotDiscernable/NotAppl

Word position of supplanted intended:
Initial/Medial/Final/InitialAndFinal/NotDiscernable/NotAppl

Of the intruder sound in error word:

Intruder: phonological segment that intrudes

Syllable role of intruder: see above

Word position of intruder: see above

Of the sound in source word (same as intruder):

Syllable role in source: see above

Word position in source: see above

Of all sound errors:

Identical neighboring segment?: y/n, is there the same segment in both the error and source words that is adjacent to the intruder? (cf. repeated phoneme effect)

Syllable with error has: Main stress/Secondary Stress/No Stress

Acknowledgements

We are grateful to Stefan Frisch, Alexei Kochetov, and Paul Tupper and for audiences at the Vancouver Phonology Group (April 2016) and the Phonetics and Experimental Phonology Lab at New York University (May 2015) for helpful comments and suggestions. We are also indebted to Rebecca Cho, Gloria Fan, Holly Wilbee, Jennifer Williams, and two other research assistants for their tireless work collecting speech errors. This work has been funded in part by a standard SSHRC research grant awarded to the first author. Any errors or omissions that remain are the sole responsibility of the authors.

References

Acheson, D. J., & MacDonald, M. C. (2009). Verbal working memory and language production: Common approaches to the serial ordering of verbal information. *Psychological Bulletin, 135*, 50-68.

Baars, B. J., Motley, M. T., & MacKay, D. G. (1975). Output editing for lexical status from artifically elicited slips of tongue. *Journal of Verbal Learning and Verbal Behavior, 14*, 382-391.

Berg, T. (1987). *A cross-linguistic comparison of slips of the tongue*.

Bock, K. (1982). Toward a cognitive psychology of syntax: information processing contributions to sentence formulation. *Psychology Review, 89*, 1-47.

Bock, K. (1996). Language production: Methods and methodologies. *Psychonomic Bulletin & Review, 3*, 395-421.
Bock, K. (2011). How much correction of syntactic errors *are* there, anyway? *Language and Linguistic Compass, 5*, 322-335.
Bock, K., & Levelt, W. J. M. (1994). Language production. Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945-984). San Diego: Academic Press.
Bond, Z. S. (1999). *Slips of the ear: Errors in the perception of casual conversation*. San Diego: Academic Press.
Boomer, D. S., & Laver, J. D. M. (1968). Slips of the tongue. *International Journal of Language and Communication Disorders, 3*, 2-12.
Broeke, V. D. M. P. R., & Goldstein, L. (1980). Consonant features in speech errors. In V. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* (pp. 47-65). London: Academic Press.
Chao, A. (2001). An overview of closed capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics, 6*, 158-175.
Chen, J.-Y. (1999). The representation and processing of tone in Mandarin Chinese: Evidence from slips of the tongue. *Applied Psycholinguistics, 20*, 289-301.
Chen, J.-Y. (2000). Syllable errors from naturalistic slips of the tongue in Mandarin Chinese. *Psychologia, 43*, 15-26.
Cole, R. A. (1973). Listening for mispronunciations: A measure of what we hear during speech. *Perception and Psychophysics, 1*, 153-156.
Cole, R. A., Jakimik, J., & Cooper, W. E. (1978). Perceptibility of phonetic features in fluent speech. *Journal of the Acoustical Society of America, 64*, 45-56.
Cruttenden, A. (2014). *Gimson's pronunciation of English (Eighth edition)*. London: Routledge.
Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America, 111*, 2862-2873.
Cutler, A. (1980). Errors of stress and intonation. In V. Fromkin (Ed.), *Errors in linguistic performance: Slips of tongue, ear, pen, and hand* (pp. 67-80). New York: Academic Press.
Cutler, A. (1982). The reliability of speech error data. In A. Cutler (Ed.), *Slips of the tongue and language production* (pp. 7-28). Berlin: Mouton.
Cutler, A. (1988). The perfect speech error. In L. M. Hyman & C. N. Li (Eds.), *Language, speech, and mind: Studies in honour of Victoria A. Fromkin* (pp. 209-233). London: Routledge.
de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*, 385-390.
Dell, G. S. (1984). Representation of serial order in speech: Evidence from the repeated phoneme effect in speech errors. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10*, 222-233.
Dell, G. S. (1985). Positive feedback in hierarchical connectionist models. *Cognitive Science, 9*, 3-23.
Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*, 283-321.

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes, 5*, 313-349.
Dell, G. S. (1995). Speaking and misspeaking. In L. R. Gleitman & M. Liberman (Eds.), *An invitation to cognitive science, Language, Volume 1*. Cambridge, MA: The MIT Press.
Dell, G. S., Burger, L. K., & Svec, W. R. (1997). Language production and serial order: A functional analysis and a model. *Psychological Review, 104*, 123-147.
Dell, G. S., & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior, 20*, 611-629.
Ferber, R. (1991). Slip of the tongue or slip of the ear? On the perception and transcription of naturalistic slips of tongue. *Journal of psycholinguistic research, 20*, 105-122.
Ferber, R. (1995). Reliability and validity of slip-of-the-tongue corpora: A methodological note. Linguistics. *Linguistics, 33*, 1169-1190.
Ferreira, F., & Swets, B. (2005). The production and comprehension of resumptive pronouns in relative clause "Island" contexts. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 263-278). Mahwah, NJ: Erlbaum.
Fowler, C. A., & Magnuson, J. S. (2012). Speech Perception. In M. J. Spivey, K. McRae, & M. F. Joanisse (Eds.), *The Cambridge handbook of psycholinguistics* (pp. 3-25). Cambridge: Cambridge University Press.
Frisch, S. (1996). *Similarity and frequency in phonology.* (Doctoral Doctoral dissertation), Northwestern University.
Frisch, S. A. (2007). Walking the tightrope between cognition and articulation: The state of the art in the phonetics of speech errors. In C. T. Schutze & V. S. Ferreira (Eds.), *The State of the Art in Speech Error Research, MIT Working Papers in Linguistics, Vol. 53* (pp. 155–171). Cambridge, MA: The MIT Press.
Frisch, S. A., & Wright, R. (2002). The phonetics of phonological speech errors: An acoustic analsis of slips of the tongue. *Journal of Phonetics, 30*, 139-162.
Fromkin, V. (1971). The non-anomalous nature of anomalous utterances. *Language, 47*, 27-52.
Fromkin, V. (1973). Speech Errors as Linguistic Evidence. The Hague: Mouton.
Garnes, S., & Bond, Z. S. (1975). Slips of the ear: Errors in perception of casual speech *Proceedings of the 11th regional meeting of the Chicago Linguistics Society* (pp. 214-225).
Garnham, A., Shillcock, R. C., Brown, G. D., Mill, A. I., & Cutler, A. (1981). Slips of the tongue in the London-Lund corpus of spontaneous conversation. *Linguistics, 19*, 805-818.
Garrett, M. (1975). The analysis of sentence production. In G. H. Bower (Ed.), *The psychology of learning and motivation, Advances in research and theory, vol. 9* (pp. 131-177). New York: Academic Press.
Goldrick, M., & Blumstein, S. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes, 21*, 649-683.
Goldrick, M., & Chu, K. (2014). Gradient co-activation and speech error articulation: Comment on Pouplier and Goldstein (2010). *Language, Cognition and Neuroscience, 29*, 452-458.
Goldstein, L., Pouplier, M., Chena, L., Saltzman, E. L., & Byrd, D. (2007). Dynamic action units slip in speech production errors. *Cognition, 103*, 386-412.
Harley, T. A. (1984). A critique of top-down independent level models of speech production: Evidence from non-plan-internal speech errors. *Cognitive Science, 8*, 191-219.

Harley, T. A. (1996). *The role of syllable structure in verbal short-term memory.* (Doctoral dissertation), University of London, England.
Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*(32), 145-164.
Kubozono, H. (1989). The mora and syllable structure in Japanese: Evidence from speech errors. *Language and Speech, 32*, 249-278.
Ladefoged, P. (2006). *A Course in Phonetics*. Boston: Thomson.
Levelt, W. J. M., Roelofs, A., & Meyer, A., S. (1999). A theory of lexical access in speech production. *Behavorial and Brain Sciences, 22*, 1-75.
Levitt, A., & Healy, A. (1985). The roles of phoneme frequency, similarity, and availability in the experimental elicitation of speech errors. *Journal of memory and language, 24*, 717-733.
MacDonald, M. C. (2016). Speak, Act, Remember: The language-production basis of serial order and maintenance in verber memory. *Current Directions in Psychological Science, 25*, 47-53.
MacKay, D. G. (1970). Spoonerisms: The structure of errors in the serial order of speech. *Neuropsychologia, 8*, 323-350.
MacKay, D. G. (1971). Stress pre-entry in motor systems. *American Journal of Psychology, 84*, 35– 51.
Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word, 15*, 19-44.
Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception and Psychophysics, 28*, 407-412.
Mao, C. X., Huang, R., & Zhang, S. (2016/To appear). Petersen estimator, Chapman adjustment, list effects, and heterogeneity. *Biometrics*.
Marin, S., & Pouplier, M. (2016). Spontaneously occurring speech errors in German: BAS corpora analysis. In A. Gilles, V. B. Mititelu, D. Tufis, & I. Vasilescu (Eds.), *Errors by humans and machines in multimedia, multimodal and multilingual data processing*. Bucharest: Romanian Academy Press.
Marin, S., Pouplier, M., & Harrington, J. (2010). Acoustic consequences of articulatory variability during productions of /t/ and /k/ and its implications for speech error research. *The Journal of the Acoustical Society of America, 127*(1), 445-461.
Marslen-Wilson, W., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*, 29-63.
Miller, G. A., & Nicely, P. (1955). An analysis of perceptual confusions among some English Consonants. *Journal of the Acoustical Society of America, 27*, 338-352.
Mowrey, R., & MacKay, I. R. A. (1990). Phonological primitives: Electromyographic speech error evidence. *Journal of the Acoustical Society of America, 88*, 1299-1312.
Nooteboom, S. G. (1969). The tongue slips into patterns. In A. J. van Essen & A. A. van Raad (Eds.), *Leyden studies in linguistics and phonetics* (pp. 114-132). The Hague: Mouton.
Pérez, E., Santiago, J., Palma, A., & O'Seaghdha, P. G. (2007). Perceptual bias in speech error data collection: Insights from Spanich speech errors. *Journal of psycholinguistic research, 36*, 207-235.
Pouplier, M., & Goldstein, L. (2005). Asymmetries in the perception of speech production errors. *Journal of Phonetics, 33*, 47-75.

Pouplier, M., & Hardcastle, W. (2005). A re-evaluation of the nature of speech errors in normal and disordered speech. *Phonetica, 62*, 227-243.
Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review, 107*, 460-.
Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial-ordering mechanism in sentence production. In W. E. Copper & E. C. T. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* (pp. 295-342). Hillsdale, NJ: Erlbaum.
Shattuck-Hufnagel, S. (1987). The role of word onset consonants in speech production planning: New evidence from speech error patterns. In E. Keller & M. Gopnik (Eds.), *Motor and sensory processes of language* (pp. 17-51). Hillsdale, NJ: Erlbaum.
Shattuck-Hufnagel, S., & Klatt, D. H. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior, 18*, 41-55.
Shockey, L. (2003). *Sound patterns of spoken English*. Malden, MA: Blackwell Publishing.
Slis, A., & Van Lieshout, P. H. H. M. (2016). The effect of phonetic context on the dynamics of intrusions and reductions. *Journal of Phonetics, 57*, 1-20.
Smolensky, P., Goldrick, M., & Mathis, D. (2014). Optimization and quantization in gradient symbol systems: A framework for integrating the continuous and the discrete in cognition. *Cognitive Science, 38*, 1107-1138.
Söderpalm, E. (1979). *Speech errors in normal and pathological speech (Travaux de L'Institute de Linguistique de Lund #14)*. Lund: Gleerup.
Stearns, A. M. (2006). *Production and Perception of Articulation Errors.* (MA thesis), University of South Florida.
Stemberger, J. P. (1982/1985). *The lexicon in a model of language production*. New York: Garland.
Stemberger, J. P. (1983). *Speech errors and theoretical phonology: A review*. Bloomington: Indiana University Linguistics Club.
Stemberger, J. P. (1984). *Lexical bias in errors in language production; Interactive components, editors, and perceptual biases. Manuscript, Carnegie-Mellow University*.
Stemberger, J. P. (1985). An interactive activation model of language production. In A. W. Ellis (Ed.), *Progress in psychology of language, Volume one* (pp. 143-186). Hillsdale, NJ: Lawrence Erlbaum Associates.
Stemberger, J. P. (1991). Apparent antifrequency effects in language production: The addition bias and phonological underspecification. *Journal of memory and language, 30*, 161-185.
Stemberger, J. P. (1993). Spontaneous and evoked slips of the tongue. In G. Blanken, J. Dittmann, H. Grimm, J. C. Marshall, & C.-W. Wallesch (Eds.), *Linguistic disorders and pathologies. An international handbook*. Berlin: Walter de Gruyter.
Stoel-Gammon, C. (2001). Transcribing the speech of young children. *Topics in Language Disorders, 21*, 12-21.
Tent, J., & Clark, J. E. (1980). An experimental investigation into the perception of slips of the tongue. *Journal of Phonetics, 8*, 317-325.
Vigliocco, G., & Hartsuiker, R. J. (2002). The interplay of meaning, sound, and syntax in sentence production. *Psychological Bulletin, 128*, 442-472.
Vitevitch, M. S. (2002). Naturalistic and experimental analyses of word frequency and neighborhood density effects in slips of ear. *Language and Speech, 45*, 407-434.

Vitevitch, M. S., Siew, C. S. Q., Castro, N., Goldstein, R., Gharst, J. A., Kumar, J. J., & Boos, E. B. (2015). Speech error and tip of the tongue diary for mobile devices. *Frontiers in psychology, 13, Article 1190*.

Vousden, J. I., Brown, G. D. A., & Harley, T. A. (2000). Serial control of phonology in speech production: A hierarchical model. *Cognitive Psychology, 41*, 101-175.

Wickelgren, W. A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review, 76*, 1-15.

Wilshire, C. E. (1998). Serial order in phonological encoding: an exploration of the 'word onset effect' using laboratory-induced errors. *Cognition, 68*, 143-166.